On the Interaction between Load Balancing and Speed Scaling

Lijun Chen, *Member, IEEE*, and Na Li, *Member, IEEE*

***Abstract*—Speed scaling has been widely adopted in computer and communication systems, in particular, to reduce energy consumption. An important question is how speed scaling interacts with other resource allocation mechanisms such as scheduling and routing, etc. In this paper, we study the interaction of speed scaling with load balancing. We characterize the equilibrium resulting from the load balancing and speed scaling interaction, and introduce two optimal load balancing designs, in terms of traditional performance metric and cost-aware (in particular, energy-aware) performance metric respectively. Especially, we characterize the load-balancing-speed-scaling equilibrium with respect to the optimal load balancing schemes in processor sharing systems. Our results show that the degree of inefficiency at the equilibrium is mostly bounded by the heterogeneity of the system, but independent of the number of servers. These results provide insights in understanding the interaction of load balancing with speed scaling and guiding new designs.**

***Index Terms*—Load balancing, Speed scaling, Energy efficiency, Efficiency loss, Data centers.**

I. Introduction

The energy consumption rate of computer and communication systems has been increasing exponentially. Computer and communication systems must make a fundamental tradeoff between performance and energy usage; see, e.g., [1], [2]. The addition of energy to standard performance metrics such as delay, throughput and loss fundamentally changes the problem space of some of resource allocation designs. Not only are new mechanisms needed to optimize energy usage, existing algorithms and protocols must be re-examined as a formerly optimal algorithm may now perform poorly with respect to a new energy-aware metric. Energy management decisions must be decomposed and coordinated spatially as well as temporally, and yet global optimality must be achieved through local algorithms that are implementable in a distributed manner. In this paper we study load balancing and its interaction with speed scaling.

Energy-aware speed scaling – to adapt the speed of the system so as to balance energy and performance metrics – is a widely-adopted power management technique; see, e.g., [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13]. Previous works on speed scaling usually focus on a single server and study its interaction with scheduling; see, e.g, [14], [8], [9], [10], [13]. Here we consider a network setting and study the interaction of speed scaling with load balancing, to provide insights into such issues as: i) How does the system perform under speed scaling in terms of traditional performance metrics as well as energy-aware metrics? ii) How to design energy-aware optimal load balancing and can we decouple the design of load balancing from that of speed scaling? iii) How does the sophistication of speed scaling impact the design and performance of load balancing? We focus on gated-static speed scaling in processor sharing systems, and our results provide useful insights into the first two questions.

Specifically, we characterize the equilibrium resulting from the load balancing and speed scaling interaction, and introduce two optimal load balancing design problems, in terms of traditional performance metric and cost-aware (in particular, energy-aware) performance metric respectively. We study in detail the load-balancing-speed-scaling (LBSS) equilibrium and the optimal load balancing designs in processor sharing systems with gated-static speed scaling, and propose distributed load balancing algorithms to achieve the corresponding equilibrium and optima. Especially, we characterize the degree of inefficiency at the load-balancing-speed-scaling equilibrium, in terms of delay as well as energy-aware metric. We show that the degree of inefficiency is mostly bounded by the heterogeneity of the system, but independent of the number of servers in the system. Our results suggest that, as in many applications a low-order polynomial provides a good approximation to power function, we can decouple the design of load balancing from that of speed scaling without incurring much inefficiency in delay. In terms of power-aware performance metric, our results suggest that, as long as the heterogeneity in the system is small, we can decouple the design of load balancing from that of speed scaling without incurring much efficiency loss; but when the heterogeneity in the system is large, we have to do energy-aware load balancing if the energy consumption is a main concern.

To summarize, we make the following main contributions in this paper:

1) We formulate three different models to study the interaction of load balancing and speed scaling: energy-oblivious load balancing where the dispatcher minimizes the delay experienced by a job, delay-optimal load balancing where the dispatcher minimizes the overall delay incurred at the servers, and energy-aware optimal load balancing where the dispatcher minimizes the overall energy-aware performance metric at the servers.
2) We characterize the equilibrium and optimum of the

Manuscript received March 20, 2015; revised July 10, 2015; accepted September 8, 2015.

L. Chen is with Computer Science and Telecommunications, University of Colorado, Boulder, CO 80309, USA (email: lijun.chen@colorado.edu). N. Li is with Electrical Engineering, Harvard University, Cambridge, MA 02138, USA (email: nali@seas.harvard.edu).

Preliminary result of this paper has been presented at the INFORMS Annual Meeting, Austin, Texas, 2010 and the ITA Workshop, Lo Jolla, California, 2011.

above three load balancing designs in terms of the set of active servers, and propose distributed algorithms for achieving the corresponding equilibrium or optimum. Our algorithms have low implementation complexity, and require only information that can be estimated or measured locally at the dispatcher and the servers.

3) We characterize the efficiency loss at the LBSS equilibrium (i.e., the energy-oblivious load balancing that dominates the current practice) in delay and energy-aware performance metric, and show that the degree of inefficiency is mostly bounded by the heterogeneity of the system but independent of the number of servers. These results provide insights in understanding the interaction of load balancing with speed scaling and guiding new designs.
4) We provide numerical examples to demonstrate the convergence of the proposed distributed algorithms and verify the bounds on the efficiency loss.

The paper is organized as follows. The next section briefly discusses some related work. Section III describes the system model. Section IV gives a brief characterization of the load-balancing-speed-scaling equilibrium, and introduces two optimal load balancing design problems. Section V studies in detail the load-balancing-speed-scaling interaction in processor sharing systems with gated-static speed scaling. Section VI provides numerical examples to complement the theoretical analysis, and Section VII concludes with some discussion on further research.

Notation: Major notation used in this paper is summarized in Table I.

TABLE I
NOTATIONS

N	Set of servers
λ	Job arrival rate at the dispatcher
λ_i, s_i	Job arrival rate and service rate at server $i \in N$
$\mathcal{F}_i$	Performance metric at server $i \in N$
T_i	Delay at server $i \in N$
$c_i(s_i)$	Operating cost at speed s_i of server $i \in N$
$P_i(s_i)$	Power function of server $i \in N$
α_i	Order of polynomial power function of server $i \in N$
$\mathcal{M}_i$	Cost-aware performance metric at server $i \in N$
^{e}	Superscript denoting LBSS equilibrium
*	Superscript denoting delay-optimal LB
$^+$	Superscript denoting energy-aware optimal LB
C^e	Social cost in delay at LBSS equilibrium
C^o	Optimal cost in delay
D^e	Energy-aware social cost at LBSS equilibrium
D^o	Optimal energy-ware cost

II. Related work

Power management techniques have been increasingly adopted in designs from single-device level such as chips to network level such as data centers. It has spurred a new branch of research in its own right. In particular, starting with Yao et al [15], there is extensive research on analytical study of speed scaling; see, e.g., [16], [17], [18], [19], [20], [21], [14], [22], [8], [23], [24], [9], [10], [11], [12], [13]. Bansal et al [8] show that a speed scaling policy (SRPT, $P^{-1}(n+1)$) is 3-competitive for regular power functions in the worst-case analysis. This result has been tighten and extended to PS scheduling as well as to stochastic analysis by Andrew et al [10]. Especially, Andrew et al [10] provide a comprehensive study of speed scaling and its interaction with scheduling, and show a fundamental tradeoff between optimality, fairness and robustness in speed scaling designs. Stanojevic and Shorten [11] study distributed speed scaling to minimize energy consumption subject to performance constraints. Son and Krishnamachari [12] study speed-scaling-aware load balancing for cellular networks, and their model has structural similarity to ours for the energy-aware optimal load balancing (see Section V-C). However, their model includes delay and energy consumption in the networking components, in addition to those in the computing/processing components. Their optimum characterization focuses on user association, while ours focuses on the set of active servers. Their iterative algorithm is based heuristically on a variant of optimum characterization (i.e., user association at the optimum), while ours is based on the gradient method.

Related work also includes [25], [26] that show that the degree of inefficiency in delay for load balancing in processor sharing systems with fixed server speeds scales with the number of servers in the system. This result has been extended to the processor sharing system with multi-class load [27], and to other scheduling policies such as SRPT [28]. In contrast to these results, we show that the degree of inefficiency in delay for load balancing in processor sharing systems with speed scaling is bounded by the heterogeneity of the system, but independent of the number of servers.

III. System model

Consider a system with a set N of servers[1] and a Poisson arrival process of rate $\lambda > 0$; see Figure 1. We assume that job size is i.i.d., and without loss of generality, has a mean of 1. Associated with each server i is a service rate (or speed) s_i. There is a load balancing dispatcher that probabilistically routes the arrivals to the severs according to certain "traditional" performance metric $\mathcal{F}_i$ that end users are concerned with, so that $\mathcal{F}_i$ at each server i is the same and minimal. The metric $\mathcal{F}_i$ can be, for instance, the mean response time $E[T_i]$ at the server, the summation of $E[T_i]$ and propagation delay τ_i, and the blocking probability p_i, etc.

It follows that the resulting arrival process to server i is Poisson with rate λ_i. We assume that server i's performance curve $\mathcal{F}_i = f_i(s_i, \lambda_i)$ (or its analytical approximation) is continuously differentiable, increasing in the arrival rate λ_i, and decreasing in the service rate s_i with $f_i(\infty, \lambda) = 0$. This is a rather general assumption. In order to ensure stability, we must have $\lambda_i < s_i$ for all $i \in N$. We can thus assume that $f_i(s_i, \lambda_i) = \infty$ when $\lambda_i \geq s_i$.

Besides performance metric $\mathcal{F}_i$ that is perceived by end users, each server i incurs certain cost $c_i(s_i)$ per unit time when it runs at a speed of s_i. The cost can be, for instance,

[1]Here a server can be a single server, or represent a cluster of collocated servers in, e.g., a micro-datacenter.

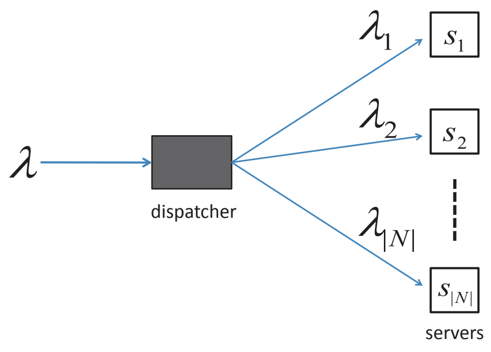

Fig. 1. A pictorial diagram of the system model.

the power expended at the server, or any other types of service costs. Given an incoming rate of λ_i, let $g_i(s_i,\lambda_i)=E[c_i(s_i)]$, the average cost. The average cost depends on the speed as well as the scheduling policy at the server. The cost function $g_i(s_i,\lambda_i)$ (or its analytical approximation) is assumed to be continuously differentiable, increasing in s_i, and non-decreasing in λ_i. Given the arrival rate λ_i and scheduling policy, each server i will choose a speed s_i to minimize a "cost-aware" performance metric $\mathcal{M}_i$:

$$\mathcal{M}_i = g_i(s_i,\lambda_i) + \beta_i\lambda_i f_i(s_i,\lambda_i), \tag{1}$$

where $\beta_i > 0$ is used to characterize the relative weight of internal cost and traditional performance metric.

By the above model, we have actually assumed some kind of static speed scaling, i.e., choose a single speed s_i for a given arrival rate λ_i. With more complicated notation, we can also model dynamic scaling, i.e., adapt speed to different states such as the number of jobs in the server.

Speed scaling can be broadly defined as any behavior of adapting speed to load, and can be due to various reasons, corresponding to different choices of cost function $g_i(s_i,\lambda_i)$. In this paper, we will mostly focus on energy-aware speed scaling as a concrete system to study the interaction between load balancing and speed scaling, and consider the following performance metric:

$$\mathcal{M}_i = E[P_i(s_i)] + \beta_i\lambda_i E[T_i], \tag{2}$$

where $P_i(s_i)$ is the power expended when server i runs at speed s_i. The modeling of the power function $P_i(s_i)$ is an active research topic, and measurements have shown it can take on different forms depending on the system involved. In many applications a low-order polynomial form

$$P_i(s_i) = k_i s_i^{\alpha_i}, \quad k_i > 0, \quad \alpha_i > 1 \tag{3}$$

provides a good approximation. For example, for dynamic power in CMOS P_i is often assumed to be cubic in previous works; see, e.g., [2]. We will focus on polynomial power function (3) in this paper, as in many previous works on speed scaling.

IV. Load-balancing-speed-scaling interaction

In this section, we characterize the equilibrium resulting from the interaction between load balancing and speed scaling for the general model described in Section III. We then introduce two optimal load balancing problems, $\mathcal{F}$-optimal load balancing and cost-aware optimal load balancing, under speed scaling. We intend to characterize the equilibrium with respect to those two optimal load balancing problems, as well as proposing distributed load balancing algorithms to achieve the corresponding equilibrium and optima.

Given the server speeds $(s_i)_{i\in N}$ and denote the set of servers used at load balancing by N_b, i.e., $i\in N_b$ iff $\lambda_i>0$. At load balancing, the $\mathcal{F}_i$ value at any server $i\in N_b$ is thus the same, and not larger than the $\mathcal{F}_j$ value a job would experience if routed to any unused server $j\in N/N_b$. This can be written mathematically as

$$\begin{aligned} f_i(s_i,\lambda_i) &\le f_j(s_j,\lambda_j),\ \forall j\in N,\ \forall i\in N_b, &(4)\\ \sum_{i\in N}\lambda_i &= \lambda, &(5)\end{aligned}$$

where $(\lambda_i)_{i\in N}$ is the arrival rates at the servers at load balancing. Denote the $\mathcal{F}_i$ value at server $i\in N_b$ at load balancing by γ. The load balancing condition (4)-(5) can be equivalently written as: there exists a $\gamma>0$, such that

$$\begin{aligned} (f_i(s_i,\lambda_i)-\gamma)(\bar\lambda_i-\lambda_i) &\ge 0,\ \forall\ \bar\lambda_i\ge 0,\ \forall i\in N, &(6)\\ \sum_{i\in N}\lambda_i &= \lambda. &(7)\end{aligned}$$

To see this equivalence, note that equations (6)-(7) imply that γ must equal the $\mathcal{F}_i$ value at server $i\in N_b$ at load balancing.

Assume that speed scaling problem $\min_{s_i>\lambda_i}\mathcal{M}_i$ has a unique solution $s_i(\lambda_i)$. Under the aforementioned assumptions on f_i and g_i, speed scaling $s_i(\lambda_i)$ satisfies:

$$\frac{\partial g_i(s_i,\lambda_i)}{\partial s_i} + \beta_i\lambda_i\frac{\partial f_i(s_i,\lambda_i)}{\partial s_i} = 0. \tag{8}$$

Notice that the dynamic speed range of a server is usually finite, i.e., $s_i\le r_i$ for some $r_i>0$. For simplicity, we do not consider such a constraint in this paper. Such a constraint does not change the general structure of our model since it does not change the convexity and the distributed decomposition structure of the model. But it will affect the characterization of efficiency loss in Section V. However, to remove the speed range constraint is reasonable for two reasons. First, one key aspect of this paper is to study the impact of speed scaling, but the speed range constraint (that is tight relative to the job arrival rate) will limit the capability of or even "disable" a server's speed scaling. Second, the computing capacity is usually not a constraint; and actually a main motivation for speed scaling is to scale down "idle" server capacity in order to save energy.

Definition 1. *The load-balancing-speed-scaling (LBSS) equilibrium is defined as a triple* $\{(\lambda_i^e)_{i\in N},(s_i^e)_{i\in N},\gamma^e\}$ *that satisfies the variational inequalities (6), (7) and (8).*

The performance of the system under load balancing and speed scaling is determined by the LBSS equilibrium. At the LBSS equilibrium $\{(\lambda_i^e)_{i\in N}, (s_i)_{i\in N}^e, \gamma^e\}$, $s_i^e = s_i(\lambda_i^e)$ and

$$(f_i(s_i(\lambda_i^e), \lambda_i^e) - \gamma^e)(\bar{\lambda}_i - \lambda_i^e) \quad \geq \quad 0, \ \forall\ \bar{\lambda}_i \geq 0, \tag{9}$$

$$\sum_{i\in N} \lambda_i^e \quad = \quad \lambda. \tag{10}$$

The following result is straightforward [29].

Theorem 2. *The LBSS equilibrium satisfies the local optimality condition for the following optimization problem:*

$$\min_{\lambda_i \geq 0} \qquad \sum_i \int f_i(s_i(\lambda_i), \lambda_i) d\lambda_i \tag{11}$$

$$s.t. \qquad \sum_i \lambda_i = \lambda, \tag{12}$$

and $-\gamma^e$ *is the corresponding optimal dual variable.*

Proof. Note that LBSS equilibrium condition (9)-(10) is a variational inequality characterization of optimality condition for optimization problem (11)-(12) and its dual [29]. □

An optimization problem characterization of the equilibrium is usually very useful. It captures the global structure of the problem, and often we can easily tell from the optimization problem if there exists an equilibrium, the multiplicity of the equilibria, as well as derive distributed or efficient algorithm to the equilibrium.

When there is no speed scaling, i.e., s_i is fixed, we recover the optimization problem characterization of usual load balancing. Under this situation, problem (11)-(12) is strictly convex as $f_i(s_i, \lambda_i)$ is an increasing function of λ_i, and the equilibrium is unique. In general, there may be no or multiple LBSS equilibria, depending on properties of the performance curve $f_i(s_i(\lambda_i), \lambda_i)$ under speed scaling. For example, consider performance metric (2) with power function (3) in a processor sharing system with gated static speed scaling (see the next section). Speed scaling $s_i(\lambda_i)$ satisfies

$$\frac{\beta_i}{(s_i - \lambda_i)^2} = k_i(\alpha_i - 1)s_i^{\alpha_i - 2}.$$

When $\alpha_i < 2$, $f_i(s_i(\lambda_i), \lambda_i)$ is decreasing. So, problem (11)-(12) becomes a problem of minimizing a concave objective function, which is usually a hard computing problem and may admit multiple solutions.

In the above load balancing model, the dispatcher routes the arrivals according to "traditional" performance metric $\mathcal{F}_i$ but does not consider the internal cost g_i of the server. We call this model *cost-oblivious* load balancing (e.g., energy-oblivious in the case of energy-aware speed scaling). It can also be seen as a selfish routing game where each job chooses a server with minimal $\mathcal{F}_i$ value; see, e.g., [30]. So, the LBSS equilibrium might not be socially optimal, in terms of metric $\mathcal{F}_i$ as well as energy-aware metric $\mathcal{M}_i$. As we mentioned before, speed scaling brings additional dimension such as energy into the design objective. It is of significant value to study its interaction with the existing algorithms and protocols; for example, if it is optimal with respect to traditional performance metric $\mathcal{F}_i$ as well as a new one $\mathcal{M}_i$, how to design distributed optimal algorithms in terms of new performance metric, and if we can decouple speed scaling from other resource allocation mechanisms. In order to study these questions for load balancing, we consider two new load balancing models, as follows.

$\mathcal{F}$-optimal load balancing: The dispatcher routes arrivals so as to achieve social optimum in terms of traditional performance metric $\mathcal{F}_i$:

$$\min_{\lambda_i \geq 0} \qquad \sum_i \lambda_i f_i(s_i(\lambda_i), \lambda_i) \tag{13}$$

$$\text{s.t.} \qquad \sum_i \lambda_i = \lambda. \tag{14}$$

When $\mathcal{F}_i = E[T_i]$, we call it delay-optimal load balancing.

Cost-aware optimal load balancing: The dispatcher routes arrivals so as to achieve social optimum in terms of cost-aware performance metric $\mathcal{M}_i$:

$$\min_{\lambda_i \geq 0} \qquad \sum_i g_i(s_i(\lambda_i), \lambda_i) + \beta_i \lambda_i f_i(s_i(\lambda_i), \lambda_i) \tag{15}$$

$$\text{s.t.} \qquad \sum_i \lambda_i = \lambda. \tag{16}$$

We call it energy-aware optimal load balancing in the case of energy-aware speed scaling.

The end users as a whole care about problem (13)-(14) and the servers/end users as a whole care about problem (15)-(16). We intend to characterize the LBSS equilibrium with respect to them, as well as proposing distributed algorithms to achieve the corresponding equilibrium or optima. Again, the general problems (13)-(14) and (15)-(16) may be highly nontrivial, depending on the performance curve f_i under speed scaling. In the remainder of this paper, we will focus on load balancing with energy-aware speed scaling in processor sharing systems with performance metric (2) with power function (3), as a concrete system to study the interaction between load balancing and speed scaling. We will leave the general problem to future work.

V. Load-Balancing-Speed-Scaling Interaction in Processor Sharing Systems

In this section, we consider energy-aware speed scaling in processor sharing (PS) systems with performance metric (2) and power function (3). While general speed scaling policies can be taken at a server, we focus on *gated-static speed scaling*, in which the server has a zero speed when there is no job and otherwise runs at a constant speed that balances the response time and the energy usage; see, e.g, [9], [10]. Gated-static speed scaling is the simplest *nontrivial* speed scaling. It requires minimal hardware to support. For example, a CMOS chip may set a constant clock speed but AND it with the gating signal to set the speed to 0 when there is no job; see, e.g., [10]. The gated static speed scaling captures some essence of dynamic speed scaling while admits more tractable analysis.

As mentioned in Section IV, when $\alpha_i < 2$, the problem under gated-static speed scaling may become hard problem of minimizing a concave objective function. We thus focus on the system with $\alpha_i \geq 2$, in order to obtain a clean characterization to gain insights. Power functions with $\alpha_i \geq 2$ is also practically important, as in the server with a power function with $\alpha_i \geq 2$ the energy cost is usually the driving force in deciding on server speed while in the server with a power function with $\alpha_i < 2$ the traditional performance metric is the driving force. Besides, the results obtained for gated-static speed scaling with $\alpha_i \geq 2$ are expected to carry over to static provisioning with $\alpha_i \geq 1$, in which the server runs at a constant static speed that is chosen based on workload to balance the response time and the energy usage. Static provisioning is the simplest form of speed scaling, and is a model often used in energy-aware capacity provisioning in data centers.

A. Energy-oblivious load balancing

Under PS scheduling, the mean response time at server $i \in N$ takes the form:

$$f_i(s_i, \lambda_i) = \frac{1}{s_i - \lambda_i}. \tag{17}$$

Under gated static speed scaling, the energy cost is only incurred during the time when the server is busy. Note that the fraction of the time when the server is busy is λ_i/s_i. So, the server decides on speed s_i by solving the following optimization problem:

$$\min_{s_i > \lambda_i} \quad \beta_i \frac{\lambda_i}{s_i - \lambda_i} + \frac{\lambda_i}{s_i} P_i(s_i). \tag{18}$$

Thus, the speed scaling $s_i(\lambda_i)$ satisfies

$$-\frac{\bar{\beta}_i}{(s_i - \lambda_i)^2} + s_i^{\alpha_i - 2} = 0, \tag{19}$$

where $\bar{\beta}_i = \frac{\beta_i}{k_i(\alpha_i - 1)}$. By equation (19), we have

$$s_i'(\lambda_i) = \frac{2 s_i(\lambda_i)}{\alpha_i s_i(\lambda_i) - (\alpha_i - 2)\lambda_i} > 0, \tag{20}$$

$$s_i''(\lambda_i) = \frac{(2\alpha_i - 4)(s_i(\lambda_i) - \lambda_i s_i'(\lambda_i))}{(\alpha_i s_i(\lambda_i) - (\alpha_i - 2)\lambda_i)^2} \geq 0, \tag{21}$$

where the second inequality follows from the fact that $s_i'(\lambda_i) \leq 1$, and moreover, $s_i'(\lambda_i) = 1$ and $s_i''(\lambda_i) = 0$ if and only if $\alpha_i = 2$. Hence, speed scaling $s_i(\lambda_i)$ is a strictly increasing, convex function of λ_i. Further,

$$f_i(s_i(\lambda_i), \lambda_i) = \frac{1}{s_i(\lambda_i) - \lambda_i} = \sqrt{\frac{(s_i(\lambda_i))^{\alpha_i - 2}}{\bar{\beta}_i}} \tag{22}$$

is also a strictly increasing function of λ_i.

Corollary 3. *There exists a unique LBSS equilibrium for processor sharing systems with gated-static speed scaling.*

Proof. By Theorem 2, the LBSS equilibrium satisfies the optimality conditions for optimization problem:

$$\min_{\lambda_i} \quad \sum_i \int \frac{1}{s_i(\lambda_i) - \lambda_i} d\lambda_i \tag{23}$$

$$\sum_i \lambda_i = \lambda. \tag{24}$$

Since $\frac{1}{s_i(\lambda_i) - \lambda_i}$ is strictly increasing in λ_i, the above optimization problem is strictly convex. The existence and uniqueness of LBSS equilibrium follows from the fact that problem (23)-(24) has a unique optimum [29]. □

Now, let us characterize the equilibrium. For each server i, define the "base" service rate $s_i^0 = s_i(0^+) = \bar{\beta}_i^{\frac{1}{\alpha_i}}$.[2] Without loss of generality, we assume that $s_1^0 \geq s_2^0 \geq \cdots \geq s_{|N|}^0$. For later convenience, we also assume that $s_{|N|+1}^0$=0.

Theorem 4. *The set of servers that are used at the LBSS equilibrium is $N_e = \{1, 2, \cdots, n\}$, with a unique n that satisfies*

$$\sum_{i=1}^{n} (\tilde{f}_i)^{-1}(\frac{1}{s_n^0}) < \lambda \leq \sum_{i=1}^{n} (\tilde{f}_i)^{-1}(\frac{1}{s_{n+1}^0}), \tag{25}$$

where

$$\tilde{f}_i(\lambda_i) = \frac{1}{s_i(\lambda_i) - \lambda_i}. \tag{26}$$

Proof. By equilibrium condition (9), we have $\frac{1}{s_i^0} < \gamma^e$ if $i \in N_e$ and $\frac{1}{s_i^0} \geq \gamma^e$ otherwise. Further,

$$\lambda_i^e = s_i^e - \frac{1}{\gamma^e} > 0, \text{ if } \frac{1}{s_i^0} < \gamma^e \tag{27}$$

$$\lambda_i^e = 0, \text{ if } \frac{1}{s_i^0} \geq \gamma^e. \tag{28}$$

Since s_i^0 is decreasing in i, N_e takes the form of $\{1, 2, \cdots, n\}$.

Note that $\frac{1}{s_n^0} < \gamma^e \leq \frac{1}{s_{n+1}^0}$, and $\tilde{f}_i(\lambda_i)$ is an increasing function. So,

$$\sum_{i=1}^{n} (\tilde{f}_i)^{-1}(\frac{1}{s_n^0}) < \sum_{i=1}^{n} (\tilde{f}_i)^{-1}(\gamma^e) \leq \sum_{i=1}^{n} (\tilde{f}_i)^{-1}(\frac{1}{s_{n+1}^0}),$$

i.e.,

$$\sum_{i=1}^{n} (\tilde{f}_i)^{-1}(\frac{1}{s_n^0}) < \sum_{i=1}^{n} \lambda_i^e = \lambda \leq \sum_{i=1}^{n} (\tilde{f}_i)^{-1}(\frac{1}{s_{n+1}^0}).$$

The uniqueness of n follows from the fact that the LBSS equilibrium is unique. □

We see that the LBSS equilibrium has a water-filling structure. If we see load balancing as a selfish routing problem [30], the arrivals will aggressively occupy fast servers with low delay first.

1) Distributed load balancing algorithm: The (convex) optimization problem characterization of the LBSS equilibrium also suggests a distributed algorithm to achieve the equilibrium.

At k-th iteration:

- Each server i estimates the arrival rate λ_i, and adjusts its speed s_i, according to

$$s_i(k) = s_i(\lambda_i(k)). \tag{29}$$

- The dispatcher measures delay $t_i(k) = \frac{1}{s_i(k) - \lambda_i(k)}$ experienced at each server i. Denote by $E[t(k)]$ the minimal

[2]For a function $f(x) : \mathcal{R} \mapsto \mathcal{R}$, $f(a^+)$ denotes the right hand limit $\lim_{x \to a^+} f(x)$.

$\bar{t}(k)$ at step k such that $\bar{t}(k) = \frac{1}{|\bar{N}(k)|}\sum_{i\in\bar{N}(k)} t_i(k)$ with $\bar{N}(k) := \{i|\lambda_i(k) > 0 \text{ or } t_i(k) \le \bar{t}(k), i \in N\}$.[3] The dispatcher adjusts λ_i to each server i, according to

$$\lambda_i(k+1) = [\lambda_i(k) - \varepsilon(t_i(k) - E[t(k)])]^+, \tag{30}$$

where ε is a positive stepsize, and '+' denotes the projection onto $\mathcal{R}^+$, the set of nonnegative real numbers.

When ε is small enough, the above algorithm converges. Let $\delta_i(k) = \lambda_i(k+1) - \lambda_i(k)$. It is straightforward to verify that

$$\sum_i \delta_i(k) = 0, \tag{31}$$

$$\sum_i \delta_i(k) t_i(k) \le 0. \tag{32}$$

We see that $\sum_i \delta_i(k)t_i(k) = 0$ only if $\delta_i(k) = 0$, which requires $t_i = \bar{t}$, or, $\lambda_i = 0$ and $t_i > \bar{t}$.

The above algorithm actually follows the negative gradient direction of $\sum_i \int \frac{1}{s_i(\lambda_i)-\lambda_i} d\lambda_i$ subject to $\lambda_i = \lambda$ [29]. Any algorithms that follow a properly-chosen negative gradient direction would work, and (30) picks a specific gradient direction that will facilitate the convergence analysis. We skip the convergence proof for brevity.

The above distributed algorithm, as well as the other two proposed in Section V.B.1) and Section V.C.1), has low implementation complexity. All the information required in the algorithm can be estimated or measured locally at the dispatcher and individual servers. Such algorithms are highly desirable in a network setting that may involve a large number of servers.

B. Delay-optimal load balancing

In this subsection, we study the delay-optimal load balancing design:

$$\min_{\lambda_i \ge 0} \quad \sum_i \frac{\lambda_i}{s_i(\lambda_i) - \lambda_i} \tag{33}$$

$$\text{s.t.} \quad \sum_i \lambda_i = \lambda, \tag{34}$$

and characterize the LBSS equilibrium with respect to it.

By equation (19),

$$\frac{\lambda_i}{s_i(\lambda_i) - \lambda_i} = \sqrt{\frac{1}{\beta_i} s_i^{\frac{\alpha_i}{2}}} - 1, \tag{35}$$

which is strictly increasing and convex in s_i. Notice that $s_i(\lambda_i)$ is increasing and convex. It follows that $\frac{\lambda_i}{s_i(\lambda_i)-\lambda_i}$ is a strictly convex function of λ_i.[4] So, problem (33)-(34) is strictly convex, and has a unique optimum. Denote the optimum by $(\lambda_i^*)_{i\in N}$. There exists a unique $\gamma^* > 0$, such that the optimality condition can be written as [29]

$$\left(\frac{s_i(\lambda_i^*) - \lambda_i^* s_i'(\lambda_i^*)}{(s_i(\lambda_i^*) - \lambda_i^*)^2} - \gamma^*\right)(\bar{\lambda}_i - \lambda_i^*) \ge 0, \ \forall\ \bar{\lambda}_i \ge 0, \tag{36}$$

$$\sum_{i\in N} \lambda_i^* = \lambda. \tag{37}$$

Theorem 5. *The set of servers that are used at the optimum is $N_o = \{1, 2, \cdots, n^*\}$, with a unique n^* that satisfies*

$$\sum_{i=1}^{n^*} (\hat{f}_i)^{-1}\left(\frac{1}{s^0_{n^*}}\right) < \lambda \le \sum_{i=1}^{n^*} (\hat{f}_i)^{-1}\left(\frac{1}{s^0_{n^*+1}}\right)\}, \tag{38}$$

where

$$\hat{f}_i(\lambda_i) = \frac{s_i(\lambda_i) - \lambda_i s_i'(\lambda_i)}{(s_i(\lambda_i) - \lambda_i)^2}. \tag{39}$$

Moreover, $\gamma^* \ge \gamma^e$ and $n^* \ge n$.

Proof. Note that $\hat{f}_i(\lambda_i)$ is an increasing function of λ_i, and $\hat{f}_i(0) = \frac{1}{s^0_i}$. The first part of the theorem follows the same proof as in Theorem 4.

For the second part of the theorem. Note that $s_i'(\lambda_i) \le 1$ by equation (20). Thus, $\hat{f}_i(\lambda_i) \ge \tilde{f}_i(\lambda_i)$. If $\gamma^* < \gamma^e$, then $n^* \le n$ and

$$\sum_{i=1}^{n^*} (\hat{f}_i)^{-1}(\gamma^*) < \sum_{i=1}^{n^*} (\tilde{f}_i)^{-1}(\gamma^*) \le \sum_{i=1}^{n^*} (\tilde{f}_i)^{-1}(\gamma^e) \le \lambda.$$

This contradicts $\sum_{i=1}^{n^*} (\hat{f}_i)^{-1}(\gamma^*) = \sum_{i=1}^{n^*} \lambda_i^* = \lambda$. So, $\gamma^* \ge \gamma^e$, and $n^* \ge n$ follows. □

Notice that γ^e has the interpretation as the delay at the energy-oblivious load balancing, but γ^* does not have such an interpretation as delay. So, $\gamma^* \ge \gamma^e$ does not imply a larger delay at the delay-optimal load balancing. In fact, in the delay-optimal load balancing different servers may have different delays and the whole system has the best overall delay performance.

1) Distributed load balancing algorithm: The delay-optimal load balancing is a convex problem. We can apply similar distributed algorithm to algorithm (29)-(30), to guide the optimal load balancing design.

At k-th iteration:

- Each server i estimates the arrival rate λ_i, and adjusts its speed s_i, according to

$$s_i(k) = s_i(\lambda_i(k)). \tag{40}$$

- The dispatcher measures delay $t_i(k) = \frac{1}{s_i(k)-\lambda_i(k)}$ experienced at each server i, and estimates $\hat{f}_i$, according to

$$\hat{f}_i(k) = \hat{f}_i(\lambda_i(k)) = \frac{\alpha_i\lambda_i(k)(t_i(k))^2 + \alpha_i t_i(k)}{2\lambda_i(k)t_i(k) + \alpha_i}. \tag{41}$$

Denote by $E[\hat{f}(k)]$ the minimal $\bar{\hat{f}}(k)$ at step k such that $\bar{\hat{f}}(k) = \frac{1}{|\bar{N}(k)|}\sum_{i\in\bar{N}(k)} \hat{f}_i(k)$ with $\bar{N}(k) := \{i|\lambda_i(k) > 0 \text{ or } \hat{f}_i(k) \le \bar{\hat{f}}(k), i \in N\}$. The dispatcher adjusts λ_i to each server i, according to

$$\lambda_i(k+1) = [\lambda_i(k) - \varepsilon(\hat{f}_i(k) - E[\hat{f}(k)])]^+, \tag{42}$$

[3]$\bar{t}$ and $\bar{N}$ can be determined in a recursive way as follows. In the beginning, let $\bar{N} = N$ and calculate $\bar{t} = \frac{1}{|\bar{N}|}\sum_{i\in\bar{N}} t_i(k)$, and then exclude from $\bar{N}$ those servers i such that $\lambda_i = 0$ and $t_i > \bar{t}$. Repeat the same procedure with the new sets $\bar{N}$, and when it stops we get $E[t]$.

[4]Note that, when $\alpha_i = 2$, $\frac{\lambda_i}{s_i(\lambda_i)-\lambda_i}$ is not strictly convex but linear in λ_i. But this would not change the uniqueness of the optimum.

where ε is a positive stepsize, and '+' denotes the projection onto $\mathcal{R}^+$, the set of nonnegative real numbers.

Notice that the delay-optimal load balancing algorithm (40)-(42) is more complicated than the simple, energy-oblivious load balancing algorithm (29)-(30). It requires to estimate $\hat{f}_i$. In addition, it requires the dispatcher to know the servers' power function characteristic parameters α_i and k_i.

2) Efficiency loss in delay at the LBSS equilibrium: Define the social cost in delay:

$$C = \sum_i \frac{\lambda_i}{s_i(\lambda_i) - \lambda_i}, \tag{43}$$

we now characterize the inefficiency in delay at the LBSS equilibrium.

Lemma 6. *Let $\alpha = \max_i \alpha_i$. Then,*

$$\gamma^e \le \gamma^* \le \frac{\alpha}{2}\gamma^e. \tag{44}$$

Proof. The first inequality has been proved in Theorem 5. It remains to prove the second one.

By equation (35), $\hat{f}_i$ can be written as

$$\hat{f}_i(\lambda_i) = \frac{\alpha_i}{2}\sqrt{\frac{s_i^{\alpha_i-2}}{\bar{\beta}_i}}\, s_i'. \tag{45}$$

Note that $s_i'(\lambda_i)$ is increasing. Thus, $s'(\lambda_i^e) \ge \frac{2}{\alpha_i}$ by equation (20). Combining with $s'(\lambda_i^e) \le 1$, we get

$$\tilde{f}_i(\lambda_i^e) \le \hat{f}_i(\lambda_i^e) \le \frac{\alpha_i}{2}\tilde{f}_i(\lambda_i^e) \le \frac{\alpha}{2}\tilde{f}_i(\lambda_i^e).$$

If $\gamma^* > \frac{\alpha}{2}\gamma^e$, then

$$(\hat{f}_i)^{-1}(\gamma^*) \ge (\tilde{f}_i)^{-1}(\frac{2}{\alpha_i}\gamma^*) > (\tilde{f}_i)^{-1}(\gamma^e).$$

Thus,

$$\sum_{i=1}^{n^*}(\hat{f}_i)^{-1}(\gamma^*) > \sum_{i=1}^{n}(\tilde{f}_i)^{-1}(\gamma^e) = \lambda.$$

This contradicts the fact that $\sum_{i=1}^{n^*}(\hat{f}_i)^{-1}(\gamma^*) = \lambda$ (also notice that $n^* \ge n$). So, $\gamma^* \le \frac{\alpha}{2}\gamma^e$. □

Theorem 7. *Denote the social cost in delay at the LBSS equilibrium by C^e and the optimal cost by C^o. Then,*

$$\frac{C^e}{C^o} \le \frac{\alpha}{2}. \tag{46}$$

Proof. The social cost at the LBSS equilibrium is

$$C^e = \lambda\gamma^e. \tag{47}$$

When $\lambda_i^* > 0$, by equations (22), (45) and (44), we have

$$\frac{1}{s_i(\lambda_i^*) - \lambda_i^*} = \sqrt{\frac{s_i^{\alpha_i-2}}{\bar{\beta}_i}} = \frac{2\gamma^*}{\alpha_i s_i'} \ge \frac{2\gamma^*}{\alpha_i} \ge \frac{2\gamma^e}{\alpha}. \tag{48}$$

So,

$$C^o = \sum_i \frac{\lambda_i^*}{s_i(\lambda_i^*) - \lambda_i^*} \ge \frac{2\gamma^e}{\alpha}\sum_i \lambda_i^* = \frac{2\lambda\gamma^e}{\alpha}. \tag{49}$$

Thus,

$$\frac{C^e}{C^o} \le \frac{\alpha}{2}. \tag{50}$$

□

We see that the degree of inefficiency in delay at the LBSS equilibrium depends only on the order α_i of the power functions. For example, if $\alpha_i = 2$, the LBSS equilibrium achieves the social optimum. As α is a constant independent of the number $|N|$ of the servers in the system, this result is very different from the efficiency loss of the usual load balancing (with fixed server speeds), which scales with $|N|$, see, e.g., [25]. Also, note that $\frac{\alpha}{2}$ can be seen as a measure of heterogeneity in power functions. We can thus say that the degree of inefficiency at the LBSS equilibrium is bounded by the heterogeneity of the system. As the power function can usually be well approximated as a low-order polynomial function, the above result suggests "benign" interaction between energy-oblivious load balancing and power-aware speed scaling, in terms of delay. As the energy-oblivious load balancing is already employed in practice and simple to implement, we may not need to change it as it does not incur a large penalty in delay.

C. Energy-aware optimal load balancing

In this subsection, we study energy-aware optimal load balancing design:

$$\min_{\lambda_i, s_i} \quad \sum_i \beta_i \frac{\lambda_i}{s_i - \lambda_i} + \frac{\lambda_i P_i(s_i)}{s_i} \tag{51}$$

$$\text{s.t.} \quad \sum_i \lambda_i = \lambda, \tag{52}$$

and characterize the LBSS equilibrium with respect to it.

By speed scaling (i.e., solving for s_i first), the above problem reduces to:

$$\min_{\lambda_i} \quad \sum_i h_i(\lambda_i) \tag{53}$$

$$\text{s.t.} \quad \sum_i \lambda_i = \lambda, \tag{54}$$

where

$$h_i(\lambda_i) = \beta_i \frac{\lambda_i}{s_i(\lambda_i) - \lambda_i} + \frac{\lambda_i P_i(s_i(\lambda_i))}{s_i(\lambda_i)}. \tag{55}$$

Note that

$$\begin{aligned} h_i'(\lambda_i) &= \frac{\beta_i s_i(\lambda_i)}{(s_i(\lambda_i) - \lambda_i)^2} + k_i(s_i(\lambda_i))^{\alpha_i - 1} \\ &= \frac{\alpha_i\beta_i}{\alpha_i - 1}\frac{s_i(\lambda_i)}{(s_i(\lambda_i) - \lambda_i)^2}, \qquad (56) \\ h_i''(\lambda_i) &= \frac{\alpha_i\beta_i}{\alpha_i - 1}\frac{2s_i(\lambda_i) - (s_i(\lambda_i) + \lambda_i)s_i'(\lambda_i)}{(s_i(\lambda_i) - \lambda_i)^3}. \qquad (57) \end{aligned}$$

We see that $h_i' > 0$ and $h_i'' > 0$, and thus $h_i(\lambda_i)$ is strictly increasing and convex. So, problem (53)-(54) is a strictly convex problem, and has a unique optimum. Denote the optimum by $(\lambda_i^+)_{i\in N}$. There exists a unique $\gamma^+ > 0$, such that the optimality condition can be written as [29]

$$(h_i'(\lambda_i^+) - \gamma^+)(\bar{\lambda}_i - \lambda_i^+) \geq 0, \; \forall \; \bar{\lambda}_i \geq 0, \tag{58}$$

$$\sum_{i\in N} \lambda_i^+ = \lambda. \tag{59}$$

Note that $h_i'(\lambda_i)$ is strictly increasing, and

$$h_i'(\lambda_i) \geq \frac{\alpha_i\beta_i}{\alpha_i - 1}\hat{f}_i(\lambda_i) \geq \frac{\alpha_i\beta_i}{\alpha_i - 1}\tilde{f}_i(\lambda_i). \tag{60}$$

Let $d_i^0 = \frac{1}{h_i'(0)} = \frac{\alpha_i - 1}{\alpha_i\beta_i}s_i^0$. We can define a permutation $\pi : \{1, 2, \cdots, |N|\} \mapsto \{1, 2, \cdots, |N|\}$, such that d_i^0 is in decreasing order under π. We have the following characterization of the optimum.

Theorem 8. *The set of servers that are used at the optimum is $N_s = \{\pi^{-1}(1), \pi^{-1}(2), \cdots, \pi^{-1}(m)\}$, with a unique m that satisfies*

$$\sum_{i=1}^{m}(h'_{\pi^{-1}(i)})^{-1}(\frac{1}{d^0_{\pi^{-1}(m)}}) < \lambda \leq \sum_{i=1}^{m}(h'_{\pi^{-1}(i)})^{-1}(\frac{1}{d^0_{\pi^{-1}(m+1)}}).$$

Proof. It follows the same proof as in Theorem 4. We skip it for brevity. □

We see that the energy-aware optimal load balancing has a similar water-filling effect, and the arrivals will occupy servers with low marginal cost in energy-aware metric first. As a result, the jobs will be consolidated into a subset of servers that have low energy-aware cost.

1) Distributed load balancing algorithm: The energy-aware optimal load balancing is a convex problem. Again, we can apply similar distributed algorithm to algorithm (29)-(30), to guide the optimal load balancing design.

At k-th iteration:

- Each server i estimates the arrival rate λ_i, and adjusts its speed s_i, according to

$$s_i(k) = s_i(\lambda_i(k)). \tag{61}$$

- The dispatcher measures delay $t_i(k) = \frac{1}{s_i(k) - \lambda_i(k)}$ experienced at each server i, and estimates h_i', according to

$$h_i'(k) = h_i'(\lambda_i(k)) = \frac{\alpha_i\beta_i}{\alpha_i - 1}(\lambda_i(k)(t_i(k))^2 + t_i(k)). \tag{62}$$

  Denote by $E[h'(k)]$ the minimal $\bar{h}'(k)$ at step k such that $\bar{h}'(k) = \frac{1}{|\bar{N}(k)|}\sum_{i\in\bar{N}(k)} h_i'(k)$ with $\bar{N}(k) := \{i|\lambda_i(k) > 0 \text{ or } h_i'(k) \leq \bar{h}'(k), i \in N\}$. The dispatcher adjusts λ_i to each server i, according to

$$\lambda_i(k+1) = [\lambda_i(k) - \varepsilon(h_i'(k) - E[h'(k)])]^+. \tag{63}$$

  where ε is a positive stepsize, and '+' denotes the projection onto $\mathcal{R}^+$, the set of nonnegative real numbers.

Again, the energy-aware optimal load balancing algorithm (61)-(63) is more complicated than the energy-oblivious load balancing algorithm (29)-(30). In addition to the servers' power function characteristic parameters, the dispatcher requires to know their weights β_i.

2) Efficiency loss in energy-aware performance metric at the LBSS equilibrium: Define the social cost in energy-aware performance metric $\mathcal{M}_i$:

$$D = \sum_i \beta_i \frac{\lambda_i}{s_i - \lambda_i} + \frac{\lambda_i P_i(s_i)}{s_i} = \sum_i h_i(\lambda_i). \tag{64}$$

We now characterize the inefficiency in energy-aware performance metric at the LBSS equilibrium.

It is complicated to characterize the efficiency loss for the system with arbitrary power functions and loads. Here we give a partial characterization, focusing on the case with power functions of the same order, i.e., $P_i(s_i) = k_i s_i^\alpha$ for all servers, and in heavy traffic, i.e., $\lambda \gg 1$. We leave a complete characterization of the efficiency loss to future work.

The case with power functions of the same order models a system that employs similar servers but with different scaling factors and weights. Heavy traffic regime is of significant interest, as the inefficiency of load-balancing-speed-scaling interaction is intuitively worst under heavy traffic.

Theorem 9. *Assume that $\alpha = 2$. Denote the energy-aware social cost at the LBSS equilibrium by D^e and the optimal cost by D^o. Under the aforementioned conditions, we have*

$$\frac{D^e}{D^o} \leq \frac{\max_i k_i}{\min_i k_i}|N|. \tag{65}$$

Proof. When $\alpha = 2$, at the LBSS equilibrium $(\lambda_i^e)_{i\in N}$, the arrivals will be routed to the server i^* that has the maximal $\bar{\beta}_i$ value.[5] Under heavy traffic, the energy-aware social cost at the LBSS equilibrium is

$$D^e \approx k_{i^*}\lambda^2 \leq \max_i k_i \lambda^2.$$

At the social optimum $(\lambda_i^+)_{i\in N}$, $\lambda_i^+ \approx \frac{1/k_i}{\sum_j 1/k_j}\lambda$. The optimal social cost is

$$D^0 \approx \sum_i k_i(\lambda_i^+)^2 \approx \frac{\lambda^2}{\sum_i 1/k_i} \geq \frac{\min_i k_i}{|N|}\lambda^2.$$

Thus,

$$\frac{D^e}{D^o} \leq \frac{\max_i k_i}{\min_i k_i}|N|. \tag{66}$$

□

We see that when $\alpha = 2$, the degree of inefficiency at the LBSS equilibrium scales with the number of servers in the system. This happens because the energy-oblivious load balancing uses only the server with the largest base rate, which incurs a huge energy cost at this server, while the energy-aware optimal load balancing will spread load across all servers, which leads to much smaller energy cost at the servers. This suggests that we should do energy-aware load balancing if the energy consumption is a main concern.

[5]There may exist multiple servers that have the maximal $\bar{\beta}_i$ value. But it is reasonable to expect that the number of such servers is bounded by a constant that does not scale with the total number of the servers in the system. For simplicity of presentation, we assume that there is only one server that has the maximal $\bar{\beta}_i$ value. This only brings in a constant factor to the bound on efficiency loss, if there are multiple such servers.

Lemma 10. *Assume $\alpha > 2$. Define $\zeta_i = \alpha k_i \bar{\beta}_i^{\frac{\alpha-1}{\alpha-2}}$ for each server i,. Then,*

$$\min_i \zeta_i (\gamma^e)^{\frac{2\alpha-2}{\alpha-2}} \leq \gamma^+ \leq \max_i \zeta_i (\gamma^e)^{\frac{2\alpha-2}{\alpha-2}}. \quad (67)$$

Proof. By equation (56), h_i' can be written as

$$h_i'(\lambda_i) = \alpha k_i (s_i(\lambda_i))^{\alpha-1} = \zeta_i (\tilde{f}_i(\lambda_i))^{\frac{2\alpha-2}{\alpha-2}}. \quad (68)$$

If $\gamma^+ < \min_i \zeta_i (\gamma^e)^{\frac{2\alpha-2}{\alpha-2}}$, then

$$(h_i')^{-1}(\gamma^+) < (\tilde{f}_i)^{-1}(\gamma^e).$$

Thus,

$$\sum_i (h_i')^{-1}(\gamma^+) < \sum_i (\tilde{f}_i)^{-1}(\gamma^e) = \lambda.$$

This contradicts the fact that $\sum_i (h_i')^{-1}(\gamma^+) = \lambda$. So, $\gamma^+ \geq \min_i \zeta_i (\gamma^e)^{\frac{2\alpha-2}{\alpha-2}}$. The second inequality can be proved similarly. □

Theorem 11. *Assume $\alpha > 2$. Denote the energy-aware social cost at the LBSS equilibrium by D^e and the optimal cost by D^o. Under the aforementioned conditions, we have*

$$\frac{D^e}{D^o} \leq (\frac{\max_i \zeta_i}{\min_i \zeta_i})^{\frac{\alpha}{\alpha-1}}. \quad (69)$$

Proof. Under heavy traffic, $\lambda_i \gg 1$. By Lemma 1 in [9], we have the following approximation for speed scaling under heavy traffic:

$$s_i(\lambda_i) \approx \lambda_i + \sqrt{\frac{\bar{\beta}_i}{\lambda_i^{\alpha-2}}} \approx \lambda_i.$$

Thus,

$$\beta_i \frac{\lambda_i}{s_i - \lambda_i} + \frac{\lambda_i P_i(s_i)}{s_i} \approx \beta_i \frac{\lambda_i^{\frac{\alpha}{2}}}{\sqrt{\bar{\beta}_i}} + k_i \lambda_i^\alpha \approx k_i \lambda_i^\alpha.$$

Note that, at the LBSS equilibrium $(\lambda_i^e)_{i\in N}$,

$$\gamma^e = \sqrt{\frac{(s_i^e)^{\alpha-2}}{\bar{\beta}_i}} \approx \sqrt{\frac{(\lambda_i^e)^{\alpha-2}}{\bar{\beta}_i}}.$$

The energy-aware social cost at the LBSS equilibrium is

$$D^e \approx \sum_i k_i (\bar{\beta}_i \gamma^{e2})^{\frac{\alpha}{\alpha-2}} \leq \sum_i k_i \bar{\beta}_i^{\frac{\alpha}{\alpha-2}} (\frac{\gamma^+}{\min_j \zeta_j})^{\frac{\alpha}{\alpha-1}}, \quad (70)$$

where the inequality follows from (67).

At the social optimum $(\lambda_i^+)_{i\in N}$,

$$\gamma^+ = \alpha k_i s_i^{\alpha-1} \approx \alpha k_i (\lambda_i^+)^{\alpha-1}. \quad (71)$$

The optimal social cost is

$$D^o \approx \sum_i k_i (\frac{\gamma^+}{k_i \alpha})^{\frac{\alpha}{\alpha-1}}. \quad (72)$$

Thus,

$$\begin{aligned} \frac{D^e}{D^o} &\leq \frac{\sum_i k_i \bar{\beta}_i^{\frac{\alpha}{\alpha-2}} (\frac{\gamma^+}{\min_j \zeta_j})^{\frac{\alpha}{\alpha-1}}}{\sum_i k_i (\frac{\gamma^+}{k_i \alpha})^{\frac{\alpha}{\alpha-1}}} \\ &\leq \max_i \frac{k_i \bar{\beta}_i^{\frac{\alpha}{\alpha-2}} (\frac{\gamma^+}{\min_j \zeta_j})^{\frac{\alpha}{\alpha-1}}}{k_i (\frac{\gamma^+}{k_i \alpha})^{\frac{\alpha}{\alpha-1}}} \\ &= \max_i (\frac{\zeta_i}{\min_j \zeta_j})^{\frac{\alpha}{\alpha-1}} \\ &\leq (\frac{\max_i \zeta_i}{\min_j \zeta_j})^{\frac{\alpha}{\alpha-1}}. \end{aligned} \quad (73)$$

□

We see that when $\alpha > 2$, the degree of inefficiency at the LBSS equilibrium depends only on the degree of heterogeneity $\frac{\max_i \zeta_i}{\min_j \zeta_j}$ in the system but not the number of servers $|N|$. If the degree of heterogeneity in the system is small, energy-oblivious load balancing interacts benignly with speed scaling, in terms of the energy-aware cost. In this situation, we may not need complicated energy-aware load balancing, i.e., we can decouple the design of load balancing from that of speed scaling. Otherwise, we must do energy-aware optimal load balancing if energy consumption is a main concern.

VI. Numerical Examples

In this section, we provide numerical examples to complement the analysis in the previous sections. We first show the convergence of the three distributed algorithms proposed in section V and the allocation of the arrival rates and service rates of the servers as well, and then verify the bounds on efficiency loss described in Theorem 7, Theorem 9, and Theorem 11.

A. Distributed algorithms

We consider a system with 10 servers with speed scaling. Half of the servers have a power function of the form $P_i(s_i) = k_i s_i^{\frac{5}{2}}$ and the other half have a power function of the form $P_i(s_i) = k_i s_i^3$. The total load is normalized to be $\lambda = 10$, and the values for parameter k_i and β_i used to obtain numerical results are randomly drawn from $[1, 10]$ and $[5, 15]$, respectively. Figures 2, 3 and 4 show the evolution of the arrival rate and service rate with stepsize $\varepsilon = 0.2$ for the energy-oblivious load balancing, the delay-optimal load balancing and the energy-aware optimal load balancing, respectively. We see that the arrival rates and service rates approach the corresponding equilibrium or optimum quickly. The numerical results confirm previous analysis and intuitions. As we go from the energy-oblivious load balancing to the delay-optimal load balancing, the load is spread more across the servers, which is driven by minimizing the social cost in delay. We also see that the changes in the arrival rate and service rate are not severe, which intuitively confirms Theorem 7 that gives a small bound on efficiency loss at the LBSS equilibrium. As we move to the energy-aware optimal load balancing, the load becomes more evenly distributed.

This is driven by minimizing the energy-aware social cost, and an uneven load distribution will lead to uneven service rate distribution, which may result in large cost in energy at the server(s) with large speed. We also see large changes in the arrival rate and service rate. This implies a large degree of inefficiency at the LBSS equilibrium, which intuitively confirms Theorem 11 even though it is a characterization for the system with power functions of the same order.

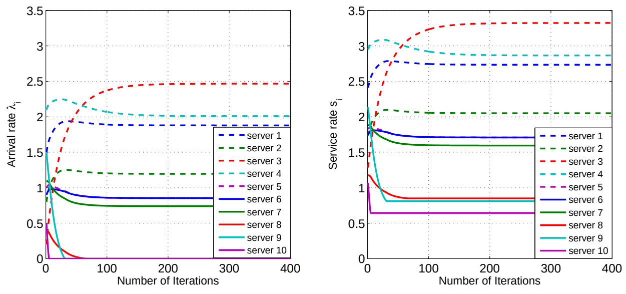

Fig. 2. The arrival rate and service rate evolution of the energy-oblivious load balancing.

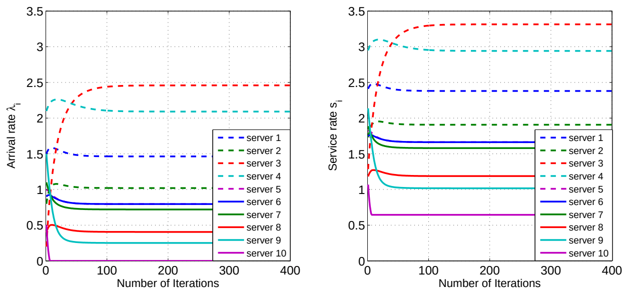

Fig. 3. The arrival rate and service rate evolution of the delay-optimal load balancing.

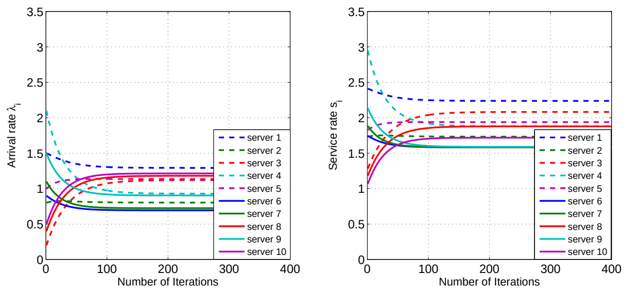

Fig. 4. The arrival rate and service rate evolution of the energy-aware optimal load balancing.

In order to study the impact of different choices of the stepsize on the convergence of the algorithms, we have run simulations with different stepsizes. We found that the smaller the stepsize, the slower the convergence, and the larger the stepsize, the faster the convergence but the system may only approach to within a certain neighborhood of the equilibrium, which is a general characteristic of any gradient based method. In practice, the dispatcher can first choose large stepsizes to ensure fast convergence, and subsequently reduce the stepsizes once the price starts oscillating around some mean value.

B. *Efficiency loss at the LBSS equilibrium*

We now consider the same system but with different numbers of servers and with a normalized load of $\lambda = 10|N|$ that scales with the number of servers, corresponding to a heavy traffic scenario. Also, the values for parameter k_i and β_i used to obtain numerical results are instead randomly drawn from $[6, 9]$ and $[0.4, 0.7]$, respectively. The key consideration in choosing these ranges is to incorporate enough heterogeneity in the system.[6]

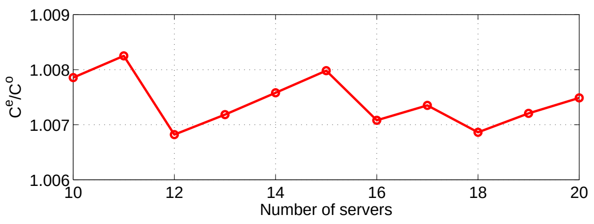

Fig. 5. Efficiency loss in delay.

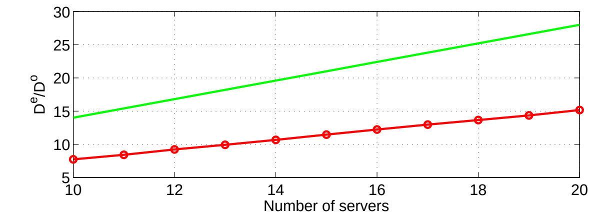

Fig. 6. Efficiency loss in energy-aware performance metric for a system with $\alpha = 2$. The line with dots shows the result of numerical experiments, while the line without dots shows the worst-case bound given in Theorem 9.

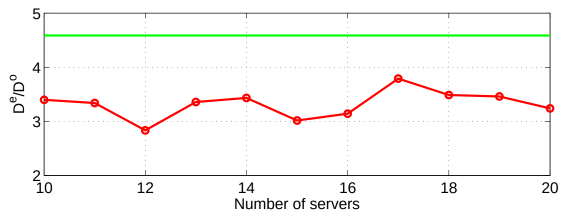

Fig. 7. Efficiency loss in energy-aware performance metric for a system with $\alpha = 3$. The line with dots shows the result of numerical experiments, while the line without dots shows the worst-case bound given in Theorem 11.

We simulate the system with different numbers $|N| = 10, 11, \ldots, 20$ of servers. Figure 5 shows the efficiency loss in delay at the LBSS equilibrium. We see that it is consistent with the (worst-case) bound of 1.5 given by Theorem 7, and the degree of inefficiency does not scale with the number of servers in the system. Figure 6 shows the efficiency loss in energy-aware performance metric at the LBSS equilibrium for a system with $\alpha = 2$. We see that it is consistent with

[6]Notice that in Subsection VI-A on convergence, the system has been chosen to have more heterogeneity in order to contrast clearly the spread of loads in the three load balancing designs.

Theorem 9, and the degree of inefficiency scales linearly with the number of servers in the system. Figure 7 shows the efficiency loss in energy-aware performance metric at the LBSS equilibrium for a system with $\alpha = 3$. We see that it is consistent with Theorem 11, and the degree of inefficiency does not scale with the number of servers but depends only on the degree of heterogeneity of the system.

VII. Conclusion

We have studied the interaction between load balancing and speed scaling. We characterize the equilibrium resulting from the load balancing and speed scaling interaction, and introduce two optimal load balancing designs, in terms of traditional performance metric and cost-aware (in particular, energy-aware) performance metric respectively. We study in detail the load-balancing-speed-scaling equilibrium and the optimal load balancing designs in processor sharing systems with gated-static speed scaling, and propose distributed load balancing algorithms to achieve the corresponding equilibrium and optima. Especially, we characterize the degree of inefficiency at the load-balancing-speed-scaling equilibrium in terms of delay as well as energy-aware metric, and show that the degree of inefficiency is mostly bounded by the heterogeneity of the system, but independent of the number of the servers. These results provide insights in understanding the interaction of load balancing with speed scaling and guiding new designs.

Further research stemming out of this paper includes the following. We are characterizing the efficiency loss in energy-aware metric at the load-balancing-speed-scaling equilibrium for the system with power functions of different polynomial orders. We are also studying the load balancing and speed scaling interaction in the processor sharing system with general power functions (e.g., nonconvex, discontinuous, with possibly a discrete set of allowable speeds), as well as in the system with other scheduling policies such as Shortest Remaining Processing Time (SRPT). We will further study other speed scaling policies and their impact on the design and performance of load balancing. Finally, we will go beyond energy-aware speed scaling, and study other types of speed scaling behaviors and their interaction with load balancing in, e.g., date centers or call centers.

References

[1] O. S. Unsal and I. Koren. System-level power-aware deisgn techniques in real-time systems. *Proc. IEEE*, 97(3):1055–1069, 2003.

[2] S. Kaxiras and M. Martonosi. *Computer Architecture Techniques for Power-Efficiency*. Morgan and Claypool, 2008.

[3] S. Irani and K. R. Pruhs. Algorithmic problems in power management. *SIGACT News*, 36(2):63–76, 2005.

[4] L. Yuan and G. Qu. Analysis of energy reduction on dynamic voltage scaling-enabled systems. *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.*, 24(12):1827–1837, 2005.

[5] Y. Zhu and F. Mueller. Feedback edf scheduling of real-time tasks exploiting dynamic voltage scaling. *Real Time Systems*, 31:33–63, 2005.

[6] N. Bansal, T. Kimbrel, and K. Pruhs. Speed scaling to manage energy and temperature. *J. ACM*, 54(1):1–39, 2007.

[7] S. Herbert and D. Marculescu. Analysis of dynamic voltage/frequency scaling in chip-multiprocessors. In *Proc. ISLPED*, 2007.

[8] N. Bansal, H.-L. Chan, and K. Pruhs. Speed scaling with an arbitrary power function. In *Proc. ACM-SIAM SODA*, 2009.

[9] A. Wierman, L. L. H. Andrew, and A. Tang. Power-aware speed scaling in processor sharing systems. In *Proceedings of IEEE Infocom*, 2009.

[10] L. L. Andrew, M. Lin, and A. Wierman. Optimality, fairness, and robustness in speed scaling designs. In *Proceedings of ACM Sigmetrics*, 2010.

[11] R. Stanojevic and R. Shorten. Distributed dynamic speed scaling. In *INFOCOM, 2010 Proceedings IEEE*, pages 1–5, March 2010.

[12] Kyuho Son and B. Krishnamachari. Speedbalance: Speed-scaling-aware optimal load balancing for green cellular networks. In *INFOCOM, 2012 Proceedings IEEE*, pages 2816–2820, March 2012.

[13] Maryam Elahi, Carey Williamson, and Philipp Woelfel. Decoupled speed scaling: Analysis and evaluation. *Performance Evaluation*, 73(0):3 – 17, 2014.

[14] N. Bansal, K. Pruhs, and C. Stein. Speed scaling for weighted flow times. In *Proc. ACM-SIAM SODA*, 2007.

[15] F. Yao, A. Demers, and S. Shenker. A scheduling model for reduced cpu energy. In *Proceedings of IEEE Symposium on Foundations of Computer Science (FOCS)*, 1995.

[16] J. M. George and J. M. Harrison. Dynamic control of a queue with adjustable service rate. *Operations Research*, 49(5):720–731, 2001.

[17] K. Pruhs, P. Uthaisombut, and G. Woeginger. Getting the best response for your erg. In *Scandinavian Worksh. Alg. Theory*, 2004.

[18] J. R. Bradley. Optimal control of a dual service rate m/m/1 production-inventory model. *European Journal of Operations Research*, 161(3):812–837, 2005.

[19] S. Albers and H. Fujiwara. Energy-efficient algorithms for flow time minimization. *Lecture Notes in Computer Science*, 3884:621–633, 2006.

[20] D. P. Bunde. Power-aware scheduling for makespan and flow. In *Proc. ACM Symp. Parallel Alg. and Arch*, 2006.

[21] S. Zhang and K. S. Catha. Approximation algorithm for the temperature-aware scheduling problem. In *Proceedings of IEEE Conference on Computer Aided Design*, 2007.

[22] N. Bansal, H.-L. Chan, T.-W. Lam, and L.-K. Lee. Scheduling for speed bounded processors. In *Int. Colloq. Automata, Languages and Programming*, 2008.

[23] N. Bansal, H.-L. Chan, K. Pruhs, and D. Katz. Improved bounds for speed scaling in devices obeying the cube-root rule. In *Int. Colloq. Automata, Languages and Programming*, 2009.

[24] T.-W. Lam, L.-K. Lee, I. K. K. To, and P. W. H. Wong. Speed scaling functions for flow time scheduling based on active job count. In *Proc. Euro. Symp. Alg.*, 2009.

[25] M. Haviv and T. Roughgarden. The price of anarchy in an exponential multi-server. *Operations Research Letters*, 35:421–426, 2007.

[26] T. Wu and D. Starobinski. On the price of anarchy in unbounded delay networks. In *Proc. of Game Theory for Comm. and Networks*, 2006.

[27] E. Altman, U. Ayesta, and B. J. Prabhu. Optimal load balancing in processor sharing systems. In *Proceedings of GameComm*, 2008.

[28] H. Chen, J. Marden, and A. Wierman. On the impact of heterogeneity and back-end scheduling in load balancing designs. In *Proceedings of IEEE Infocom*, 2009.

[29] D. P. Bertsekas and J. N. Tsitsiklis. *Parallel and Distributed Computation*. Prentice Hall, 1989.

[30] N. Nissan, T. Roughgarden, E. Tardos, and V. V. Vazirani. *Algorithmic game theory*. Cambridge University Press, 2007.

Lijun Chen (M'05) is an Assistant Professor of Computer Science and Telecommunications at University of Colorado at Boulder. He received a Ph.D. in Control and Dynamical Systems from California Institute of Technology in 2007. He was a co-recipient of the Best Paper Award at the IEEE International Conference on Mobile Ad-hoc and Sensor Systems (MASS) in 2007. His current research interests include optimization and control of networked systems, distributed optimization and control, convex relaxation and parsimonious solutions, and game theory and its engineering application.

Na Li (M'09) is an Assistant Professor in the School of Engineering and Applied Sciences in Harvard University. She received her B.S. degree in Mathematics and Applied Mathematics from Zhejiang University in China and PhD degree in Control and Dynamical systems from California Institute of Technology in 2013. She was a Postdoctoral Associate of the Laboratory for Information and Decision Systems at Massachusetts Institute of Technology. She entered the Best Student Paper Award fnalist in the 2011 IEEE Conference on Decision and Control. Her research lies in the design, analysis, optimization and control of distributed network systems, with particular applications to power networks and systems biology/physiology.