Title Page

Final Technical Report

Project Title: Recovery Act – Power Minimization for Networked Datacenters

DOE Award Number: DE-EE0002890

Project Period: 04:2010 – 06:2011

Principal Investigator(s) –author(s):
Steven H. Low, 626-395-6767, slow@caltech.edu
A. Kevin Tang 607-255-4803, atang@ece.cornell.edu

Recipient organization:
California Institute of Technology
1200 E. California Boulevard, Pasadena, CA 91125

Sub-recipient organization:
Cornell University
Tower Road, Ithaca, NY, 14853

Other project team member organizations : N/A

Date of Report : September 28, 2011

Acknowledgment, Disclaimer and Proprietary Data Notice – DOCUMENT AVAILABILITY

Acknowledgment: “This report is based upon work supported by the U. S. Department of Energy under Award No. DE-EE 0002890“.

Disclaimer: “Any findings, opinions, and conclusions or recommendations expressed in this report are those of the author(s) and do not necessarily reflect the views of the Department of Energy”

Document Availability: Reports are available free via the U.S. Department of Energy (DOE) Information Bridge Website: http://www.osti.gov/bridge

Reports are available to DOE employees, DOE contractors, Energy Technology Data Exchange (ETDE) representatives, and Informational Nuclear Information System (INIS) representatives from the following source:

Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831
Tel: (865) 576-8401
FAX: (865) 576-5728
E-mail: reports@osti.gov
Website: http://www.osti.gov/contract.html

Table of Contents

List of Acronyms

GLB: geographical load balancing

ICT: information and communication technologies

Lists of Figures

0. Executive summary

Our project is motivated by the recent surge of interest to reduce energy consumption of datacenter networks. Data centers now pay more for electricity than servers. They paid $4.5B for and consume 1.5% of total U.S. electricity in 2006, more than the nation's TVs. Moreover, the level of consumption is growing exponentially at an annual rate of 15% [EPA2007]. Despite these alarming statistics, there is a tremendous opportunity to drastically reduce energy consumption through better power management practices and technologies. In this project, we will develop novel power optimization technologies for networked datacenters.

Our *objective* is to develop a mathematical model to optimize energy consumption at multiple levels in networked data centers, and develop algorithms to optimize not only individual servers, but also coordinate the energy consumption of clusters of servers within a data center and across geographically distributed data centers to minimize the overall energy cost and consumption of brown energy of an enterprise.

In this project, we have focused on two aspects, speed scaling and load balancing. Here, speed scaling refers to the dynamic adjustment of processing capacity to save energy. At the level of abstraction of our models, it may represent the dynamic adjustment of the frequency or voltage level of a processor, or the activation or deactivation of servers in a datacenter. Load balancing refers to the routing of job arrivals to one of the available processing stations. Our models may refer to load balancing among multiple servers in a datacenter, or among multiple geographically distributed datacenters.

We have explored the following questions:

1. *Speed scaling*: how to optimally balance energy consumption and response time through speed scaling under process sharing scheduling.
2. *Speed scaling* + *load balancing*: how to optimize energy consumption and response time by load balancing across multiple datacenters and dynamically adjusting the capacities (#active servers) of these datacenters? Can geographical load balancing reduce the use of energy from fossil fuels (so called brown energy)?
3. *Speed scaling* + *load balancing* + *admission control*: how to minimize a rich set of various forms of cost by jointly optimizing service rate control, load balancing, and admission control?

We have formulated a variety of optimization models, and have obtained a variety of qualitative results on the structural properties, robustness, and scalability of the optimal policies. We have also systematically derived from these models decentralized algorithms to optimize energy efficiency, analyzed their optimality and stability properties. Finally, we have conducted preliminary numerical simulations to illustrate the behavior of these algorithms.

We draw the following *conclusions* from our results:

- There is a substantial opportunity to minimize both the amount and the cost of electricity consumption in a network of datacenters, by exploiting the fact that traffic load, electricity cost, and availability of renewable generation *fluctuate over time and across geographical locations*. Judiciously matching these stochastic processes can *optimize the tradeoff* between brown energy consumption, electricity cost, and response time.
- Given the stochastic nature of these three processes, *real-time dynamic feedback* should form the core of any optimization strategy. The key is to develop decentralized algorithms that can be implemented at different parts of the network as simple, local algorithms that coordinate through asynchronous message passing.

- Our research suggests that simple scalable decentralized algorithms to optimize energy consumption at each server (speed scaling, scheduling, admission control), within a datacenter (sleep mode, rate control, admission control), and across multiple datacenters (routing) are possible. We have proposed a few of such algorithms, analyzed their optimality and stability properties, and evaluated their performance through numerical simulations.
- This set of results can form the core of a program that takes some of the algorithms developed in this project to the next level towards eventual commercialization; see Section 5 for more detail on the commercialization potential.

1. Introduction

Our project is motivated by the recent surge of interest to reduce energy consumption of datacenter networks. Data centers now pay more for electricity than servers. They paid $4.5B for and consume 1.5% of total U.S. electricity in 2006, more than the nation's TVs. Moreover, the level of consumption is growing exponentially at an annual rate of 15% [EPA2007]. Despite these alarming statistics, there is a tremendous opportunity to drastically reduce energy consumption through better power management practices and technologies. In this project, we will develop novel power optimization technologies for networked datacenters.

Our *objective* is to develop a mathematical model to optimize energy consumption at multiple levels in networked data centers, and develop algorithms to optimize not only individual servers, but also coordinate the energy consumption of clusters of servers within a data center and across geographically distributed data centers to minimize the overall energy cost and consumption of brown energy of an enterprise.

2. Background

Motivation and goal: Our project is motivated by the recent surge of interest to reduce energy consumption of datacenter networks. A large datacenter houses hundreds of thousands of servers and consumes tens of megawatts of electricity [Katz2009]. [Qureshi2009] shows with a back-of-the-envelope calculation that such consumption could incur an annual electricity cost upward of tens of millions of dollars and thus even a fractional reduction could yield significant savings. Indeed, Google just released in early September 2011 the electricity consumption of its datacenters around the world: last year, they consumed 260 MW, or 2.3 TWh (http://www.sfgate.com/cgi-bin/article.cgi?f=/c/a/2011/09/08/BU711L1T91.DTL). If the average cost of electricity for Goolge was $60/MWh, then electricity cost for Google was $137M last year.

Many approaches to reduce the electricity consumption of datacenters are being developed and deployed. They include improving the cooling system, the power distribution system, the architectural design of new datacenters that reduce or eliminate the need for chillers, deploying more efficient computing and IT hardware, increasing utilization of servers through virtualization, etc. [Google2011, Beloglazov2010, Unsal2003, Irani2005, Kaxiras2008]. We focus in this project on improving electricity efficiency by optimizing software and deployment at various levels. For instance, at the chip level, our work is relevant to the adjustment of the power consumption according to utilization level via dynamic voltage/frequency scaling or speed scaling. At the datacenter level, we dynamically turn on and off servers or network elements. At the network level, we have energy-aware routing that exploits the difference in electricity prices at different locations and at different times and load balance to minimize energy and delay.

The **objective** of our project is to develop a mathematical model for energy optimization at multiple levels in networked data centers, and develop abstract algorithms to optimize not only individual servers, but also coordinate the energy consumption of clusters of servers within a data center and across geographically distributed data centers to minimize the overall energy cost of an enterprise. In particular, we focus on two important aspects, speed scaling and load balancing, and their joint optimization.

Prior work: Here, speed scaling refers to the dynamic adjustment of processing capacity to save energy. At the level of abstraction of our models, it may represent the dynamic adjustment of the frequency or voltage level of a processor, or the activation or deactivation of servers in a datacenter. Load balancing refers to the routing of job arrivals to one of the available processing stations. Our models may refer to load balancing among multiple servers in a datacenter, or among multiple geographically distributed datacenters. We now elaborate on each.

2.1 Speed scaling

There are many previous analytic studies of speed scaling designs. Beginning with [Yao1995], the focus has been on either (i) minimizing the total energy used in order to complete arriving jobs by their deadlines, e.g., [Yao1995, Pruhs2004, Pruhs2008a], or (ii) minimizing the average response time of jobs, i.e., the time between their arrival and their completion of service, given a set energy/heat budget, e.g., [Bunde2009, Pruhs2008b, Zhang2007]. Many settings have neither job completion deadlines nor fixed energy budgets. In these cases, the goal is to optimize a tradeoff between energy consumption and mean response time. In particular, the performance metric can be written as $\mathrm{E}[T] + \mathrm{E}[E]/\beta'$, where T is the response time of a job, E is the energy expended on that job, E represents the expectation operator, and β' controls the relative cost of delay. This metric is a practical choice when both delay and energy incur a financial cost. However, other metrics may be more suitable in other contexts, and qualitatively different conclusions may be drawn in those cases.

This performance metric has attracted attention recently, e.g., [Albers2006, Bansal2009, Bansal2007b, 14]. The related analytic work falls into two categories: worst-case analyses and stochastic analyses. The former provides specific, simple speed scaling designs guaranteed to be within a constant factor of the optimal performance regardless of the workload, e.g., [Albers2006, Bansal2009, Bansal2007b]. In contrast, stochastic results have focused on service rate control in the M/M/1 model under First Come First Served (FCFS) scheduling, which can be solved numerically using dynamic programming. Unfortunately, the structural insight obtained from stochastic models has been limited.

Other studies consider fundamental limits of the worst-case performance of optimal speed scaling algorithms, typically coupled with optimal schedulers such as shortest remaining processing time first (SRPT) [Albers2006, Bansal2007b, Bansal2009, Andrew2010]. However, SRPT is rarely implemented in practice because it requires knowledge of the size of a job before the job completes, which may be impractical.

In our project, we take an analytic approach, but investigate more practical problems, in two ways. First, we study speed scaling designs coupled with a practical scheduler, processor sharing (PS), which serves all jobs currently in the system at equal rates. PS is a tractable model for schedulers currently used in operating systems and many other applications. Second, we study not only the worst-case performance but also the expected performance in a stochastic setting, which is often a better guide for system design. Comparison of the worst-case and stochastic results shows structural similarities, which allows designers to design for optimal expected performance while retaining worst-case guarantees. We expect our results will be of interest to researchers who focus on performance issues in networking systems. See Section 3 for more details.

2.2 Joint load balancing and speed scaling

An important approach to electricity efficiency is through optimization at various levels. At the chip level, one can adjust the power consumption according to utilization level via dynamic voltage/frequency scaling or speed scaling [Fan2007, Andrews2010b]. At the datacenter level, one can dynamically turn on and off servers or network elements [Andrews2010a]. At the network level, one can employ energy-aware routing that exploits the price differentials of electricity prices at different locations and at different times [Qureshi2009,Qureshi2010], and load balance to minimize energy and delay [Rao2010a, Rao2010b]. Hence, dynamic geographical load balancing can balance the revenue lost due to increased delay against the electricity costs at each location. Indeed, many papers have illustrated the potential of geographical load balancing to provide significant cost savings for data centers, e.g., [Pakbaznia2009, Qureshi2009, Rao2010b, Stanojevic2010, Wendell2010, Lin2011] and the references therein.

3. Results and Discussion

In this project, we have explored the following questions:

1. *Speed scaling*: how to optimally balance energy consumption and response time through speed scaling under process sharing scheduling?
2. *Speed scaling* + *load balancing*: how to optimize energy consumption and response time by load balancing across multiple datacenters and dynamically adjusting the capacities (#active servers) of these datacenters? Can geographical load balancing reduce the use of brown energy?
3. *Speed scaling* + *load balancing* + *admission control*: how to minimize a rich set of various forms of cost by jointly optimizing service rate control, load balancing, and admission control?

We have used a variety of models, and have obtained a variety of results from structural properties of the system under energy optimization to distributed optimization algorithms to performance evaluation of these algorithms. In the following, we will summarize the main models and results.

3.1 Power aware speed scaling with process sharing [Wierman2011]

Key Results: There are three main results in this part.

We provide upper and lower bounds on the optimal performance of an M/GI/1 PS system with variable speeds, and upper and lower bounds on the optimal speeds.[1] Surprisingly, these bounds show that, when the arrival process is Poisson[2] of a known rate, dynamic speed scaling improves performance only marginally compared to a simple scheme where the server uses a static speed when busy and speed 0 when idle -- at most a factor of ~2 for typical parameters and often less. Counter-intuitively, these bounds also show that the power-optimized response time remains bounded as the load grows.

We use these bounds to prove that this optimal speed scalar has a finite competitive ratio in the worst case, when the workload need not be Poisson. *This demonstrates that the main benefit of dynamic speed scaling is improved robustness.*

We provide a tighter upper bound on the best competitive ratio achievable by a practical scheduler, showing that a competitive ratio of 10 is achievable for typical parameters, i.e., even an optimal schedule that is not practically implementable cannot achieve more than 10 times

[1] An M/GI/1 PS system is a mathematical model of a processing system (e.g., a computer, a checkout counter) that receives jobs (e.g., web search request, a customer checking out), queues them up, and processes them. The jobs arrive at random times (M), each requires a random amount of time to serve (GI), and all the jobs waiting in the system shares the processor fairly (PS).

[2] A Poisson process describes how job randomly arrive at the processing station. The inter-arrival times between every two consecutive job arrivals are statistically independent with exponential distribution.

the performance of a practical scheduler. We also provide the first competitive analysis of a practical system in which the scheduler does not require knowledge of the power function and can thus be decoupled from the speed scaling mechanism.

Performance: *Static vs. Dynamic schemes*

Figure 1 compares the optimal dynamic speeds with the optimal static speeds. A static scheme always uses a constant processing speed as long as there are one or more jobs in the system; a dynamic scheme can choose a variety of processing speeds depending on the number of jobs in the system. Note that the bounds on the dynamic speeds are quite tight, especially when the number of jobs in the system, n, is large. Although the speed of the optimal scheme differs significantly from that of gated static, the actual costs are very similar, as predicted. This is shown in Fig.2. The bounds on the optimal speed are also very tight, both for large and small energy aware load. Part (a) shows that the lower bound is loosest for intermediate where the weights given to power and response time are comparable. Part (b) shows that the gated static (i.e., the upper bound) has very close to the optimal cost.

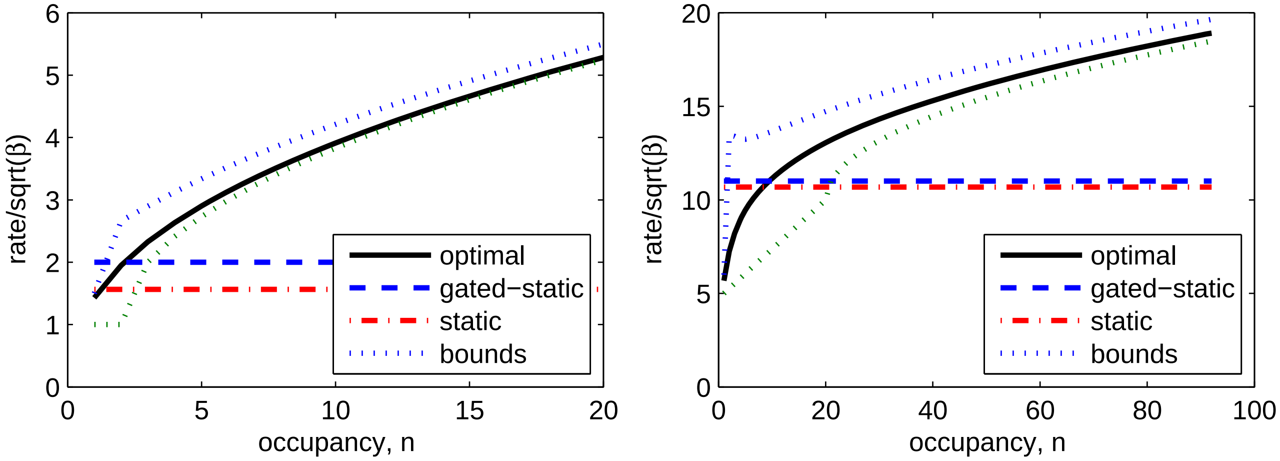

Figure 1. Rate vs n, for different energy-aware-load

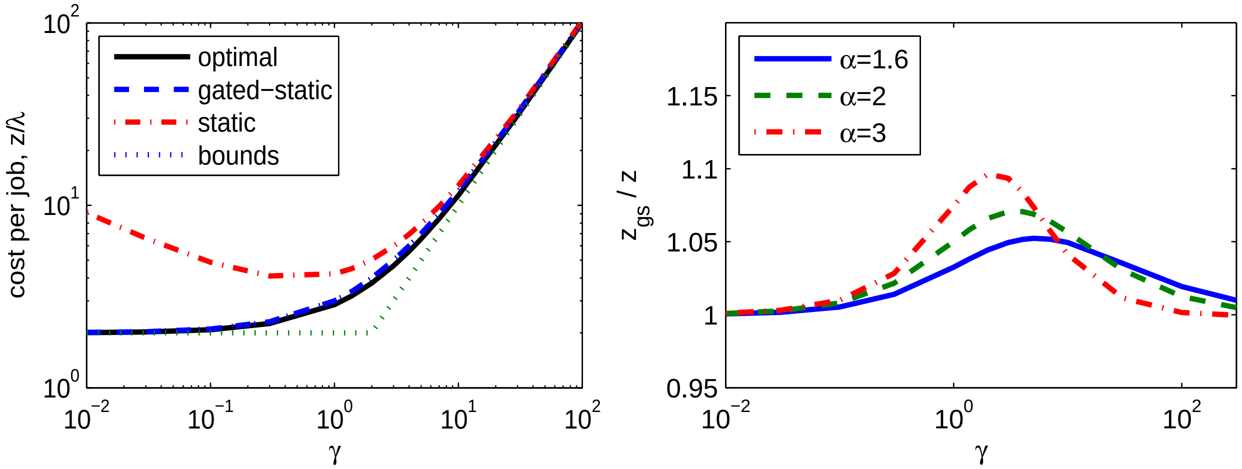

Figure 2: Cost z vs. energy-aware-load

Robust power-aware design

We focus first on robustness with respect to the load, The optimal speeds are sensitive to the average load, but in reality this parameter must be estimated, and will be time-varying. It is easy to see the problems mis-estimation of the average load causes for static speed designs. If the load is not known, then the selected speed must be satisfactory for all possible anticipated loads.

Optimal dynamic scaling is not immune to mis-estimation of the average load. However, because the speed adapts to the queue length, dynamic scaling is more robust. The solid line in Fig 3. shows this improvement. This robustness is improved further by the speed scaling scheme, which we term ``linear'', that scales the server speed in proportion to the queue length, The dotted line in Fig 3. shows that linear scaling provides significantly better robustness than the optimal dynamic scheme. The (significant) price that linear scaling pays is that it requires very high processing speed when the occupancy is high, which may not be supported by the hardware.

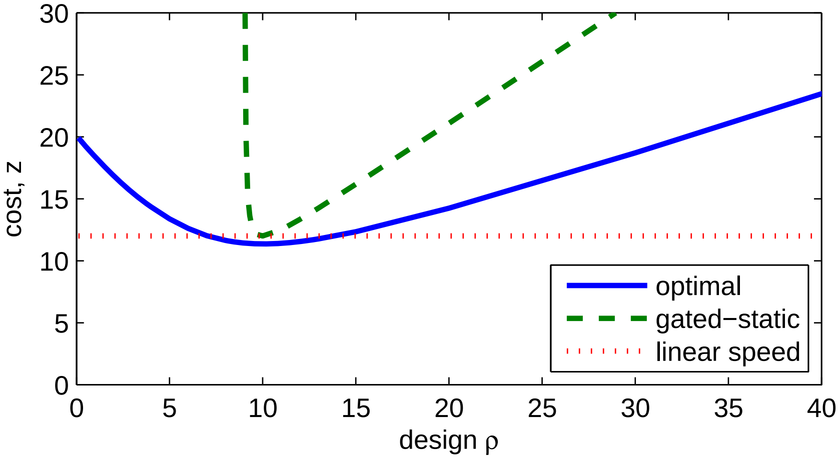

Figure 3: Cost at load=10, when speeds are designed for “design load**”**

3.2 Geographical load-balancing [Chen2011, Lin2011, Liu2011, Liu2012]

Motivation: Our work differs from most previous work on geographical load balancing (GLB) in two ways. First, most previous work proposes heuristic algorithms that are evaluated by simulations only. We start by formulating mathematically the GLB problem and then derive distributed GLB algorithms systematically as solutions to a constrained optimization. We not only evaluate our algorithms using simulations, but also prove rigorously its optimality and stability properties.

Second, we not only attempt to minimize electricity cost, but also explore its potential to reduce brown energy consumption by exploiting the availability of renewable generation. In particular, geographical load balancing aims to reduce energy costs, but this can come at the expense of increased total energy usage: by routing to a data center farther from the request source to use cheaper energy, the data center may need to complete the job faster, and so use more service capacity, and thus energy, than if the request was served closer to the source. In other words, paradoxically, minimizing the *cost* of energy may increase the *amount* of energy used.

In contrast to this negative consequence, geographical load balancing also provides an opportunity for environmental benefit as the penetration of green, renewable energy sources increases. An enormous challenge facing the electric grid is the large-scale integration of intermittent, unpredictable renewable sources such as wind and solar. A key technique for mitigating the uncertainty of renewable sources is demand-response, where elastic electricity load is dynamically adjusted to match fluctuating supply [EPA2007]. Geographical load balancing in an Internet-scale system offers an attractive way to help balance supply and demand at different regions by routing traffic to "follow the renewables", providing demand-response without service interruption or curtailment. Since data centers represent a significant and growing fraction of total electricity consumption, and the IT infrastructure is already in place, geographical load balancing has the potential to provide an inexpensive approach for enabling large scale, global demand-response and real-time power regulation.

The key to realizing the environmental benefits above is for data centers to move from the fixed price contracts that are now typical toward some degree of dynamic pricing, with lower prices when green energy is available. The demand response markets currently in place provide a natural way for this transition to occur, and there is already evidence of some data centers participating in such markets [EPA2007].

Our *contribution* is twofold. (1) We develop distributed algorithms for geographical load balancing with provable guarantees. Unlike most previous work, we jointly optimize datacenter capacity (turning servers in or out of sleep mode) in addition to routing of requests. More importantly, we rigorously prove the optimality and stability of our algorithms. (2) We use the proposed algorithms to explore the feasibility and consequences of using GLB as demand response to minimize the use of brown energy.

Key results: We formulate mathematically the GLB problem as a constrained optimization:

$$\begin{aligned} \min_{\lambda_{ij}, m_i} \quad & \sum_i g_i\left(m_i, \lambda_i\right) + \sum_i \sum_j \lambda_{ij}\left(f_i\left(m_i, \lambda_i\right) + d_{ij}\right) \\ \text{s. t.} \quad & \sum_i \lambda_{ij} = L_j \\ & 0 \le m_i \le M_i \end{aligned}$$

Here, control variables are the routing of request j to datacenter i representing by λ_{ij}, and the number m_i of active servers in datacenter *i*. The first term in the objective function is the total energy cost $g_i\left(m_i, \lambda_i\right)$ as a function of the number of active servers and the amount of job requests routed to datacenter *i*. The second term in the objective function is the total response time. The first constraint says that all job requests are routed to one of the datacenters for processing and the second constraint says that the number of active servers cannot exceed the total number of available servers in a datacenter.

Our *key results* are [Liu2011, Liu2012, Lin2011]:

- Qualitative properties of optimal solution: for example, though there can be multiple optimal solutions, the arrival rate to each server under any optimal policy is unique. Therefore, at optimality, all active servers should have the same utilization (and therefore the same queueing delay). Optimal routing of requests is very sparse, making optimal strategy scalable.

- Distributed algorithms: We have derived two distributed algorithms to solve the GLB problem, one based on Gauss-Seidel iteration and the other, a distributed gradient algorithm. We have proved that both algorithms are stable, i.e., starting from any initial state, the algorithms converge to an optimal policy.
- Performance evaluation: We have evaluated the performance of the algorithms using realistic data for traffic and electricity prices (see below for more detail).
- Greening GLB: We show that if electricity is dynamically priced in proportion to the instantaneous fraction of the total energy that is brown, then geographical load balancing provides significant reductions in brown energy use. However, the benefits depend strongly on the degree to which systems accept dynamic energy pricing and the form of pricing used.

Performance: Using 48 hours of Hotmail traces (a large Internet mail provider, part of Microsoft) starting on August 4, 2008 for traffic load and 14 emulated Google datacenters spread across the US, we conducted numerical experiments to evaluate the performance of our GLB algorithms relative to two other algorithms. The first is the min-load algorithm that routes a new request to a datacenter that is least loaded at the time. This is the most popular load balancing method and will serve as the baseline for performance evaluation. The second is the min-delay algorithm that routes a new request to a datacenter that will yield the least (expected) response time (processing + queueing + propagation delay). We compare these three algorithms in terms of electricity cost, delay, and their weighted sum (objective function of our GLB problem). The results are expressed in terms improvement over the baseline (min-load algorithm), and shown in the following plots where the x-axis is time in hour over two days and the y-axis is the %-improvement of various quantities over baseline (those quantities achieved by the min-load algorithm).

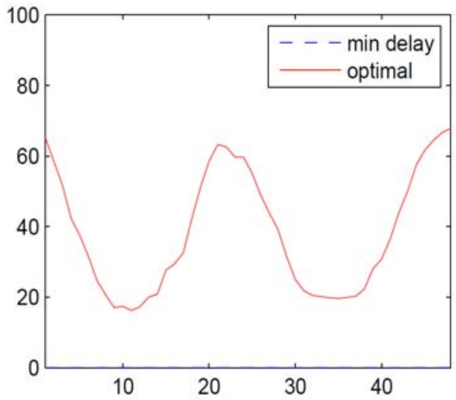

Figure 4a. The %-improvement in electricity cost over baseline (min-load algorithm).

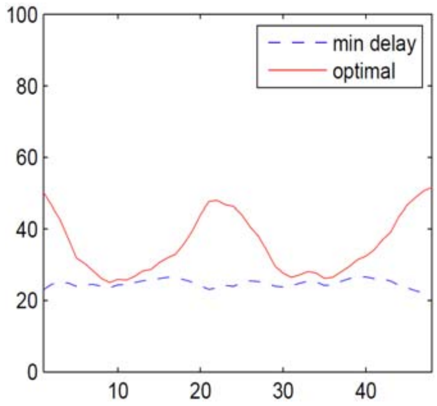

Figure 4b. The %-improvement in weighted sum of cost and delay over baseline.

Figure 4a shows the %-improvement in electricity cost over baseline (min-load algorithm) of the optimal GLB algorithm (red line) and of the min-delay algorithm (blue dash line, which is at 0% for all times) vs time over the 48-hour period. The improvement varies over time because the load and electricity cost varies over time. As the plot shows, the optimal algorithm incurs 25%-65% lower electricity cost relative to the baseline while the min-delay algorithm incurs the same electricity cost (0% improvement). Figure 4b shows the corresponding %-improvement in weighted sum of delay and electricity cost. It shows that the optimal algorithm incurs 25%-50% lower overall cost relative to the baseline while the min-delay algorithms incurs 25%

improvement over the baseline. Therefore the optimal algorithm performs significantly better both in terms of overall performance metric (objective of GLB optimization) and in terms of electricity cost. Does it pay for this improvement in terms of much higher delay? No, according to Figure 4c: the optimal algorithm incurs 30% less delay than the baseline, though it incurs slightly higher delay than the min-delay algorithm.

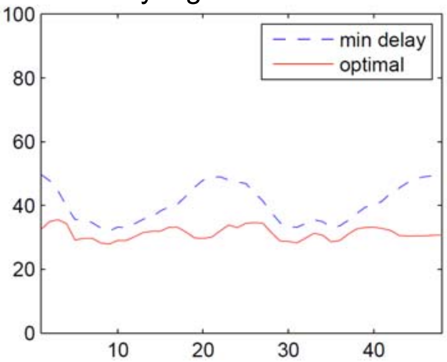

Figure 4c. The %-improvement in delay over baseline.

Thus optimal GLB algorithm strikes a good balance between energy consumption and response time.

For the local renewable case, we perform a trace-based study to evaluate three issues: the impact of geographical load balancing, the role of storage, and the optimal mix of renewables. Our results highlight that geographical load balancing can significantly reduce the required capacity of renewable energy by using the energy more efficiently with "follow the renewables" routing. Further, our results show that small-scale storage can be useful, especially in combination with geographical load balancing, and that an optimal mix of renewables includes significantly more wind than photovoltaic solar.

We have also studied another model to explore the interaction of speed scaling and load balancing [Chen2011]. Here we consider a network setting and characterize the equilibrium resulting from such interaction. We characterize the degree of inefficiency at the load-balancing-speed-scaling equilibrium, in terms of both 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, if the objective is only to minimize delay, we can always decouple the design of load balancing from that of speed scaling without incurring much inefficiency. However, if the objective is to minimize a weighted sum of delay and energy consumption, then decoupling the design of load balancing and speed scaling will not incur large efficiency loss only when the system heterogeneity is small.

3.3 Admission control, load balancing and speed scaling of two parallel queues [Lim2011a, Lim2011b]

Motivation: This part is motivated by the recent surge of interest to reduce power consumption in datacenter networks. A large datacenter houses tens of thousands of servers and can consume electricity at the order of tens of megawatts. To reduce electricity cost, the hardware approach is to replace the existing infrastructure with more power efficient components. A better cooling system, more efficient servers and power distribution, and improved architectural design of the datacenters are some of the approaches that have been implemented by datacenter owners. On the other hand, better efficiency can also be achieved by optimizing software and deployment at various levels. At the chip level, we have dynamic voltage/frequency scaling (DVFS) or speed scaling. At the machine level, virtualization is a widely deployed method to run multiple computer systems on a single set of computer hardware. At the datacenter level, one proposed method is to power down inactive network elements. At the network level, we have power-demand routing which exploits the price differentials of electricity prices for different geographical regions, and load balancing with a constrained average delay.

From a modeling point of view, our work focuses on the network level by considering a joint minimization of energy cost, delay cost and routing cost. We also allow the service rate to be changed dynamically according to the number of demands in the system. From an analytical perspective, our work is a generalization of the optimal service rate control of a queue to an open network of parallel queues with routing costs and no feedback or cascade topologies. In this framework, the costs in consideration are known in the literature as cost of effort and holding cost (or congestion cost), which corresponds to energy cost and delay cost for the datacenter networks, respectively.

Key Results: We aim to find a stationary policy that minimizes, over an infinite horizon, the long-run average cost rate. Using a dynamic programming formulation, we show that the optimal routing policy is acyclic and bipartite. We prove that the relative cost function is monotonically non-decreasing in queue size while for the case of 2 servers, the optimal service policy is non-decreasing in queue size and the optimal routing policy is a threshold policy. We show how upper and lower bounds of the optimal average cost rate can be efficiently calculated numerically. In particular, based on the monotonicity property, we develop an approximate dynamic programming procedure to efficiently compute a good upper bound.

For two-server case, we performed extensive simulation to verify our prediction. Here are a few representative figures.

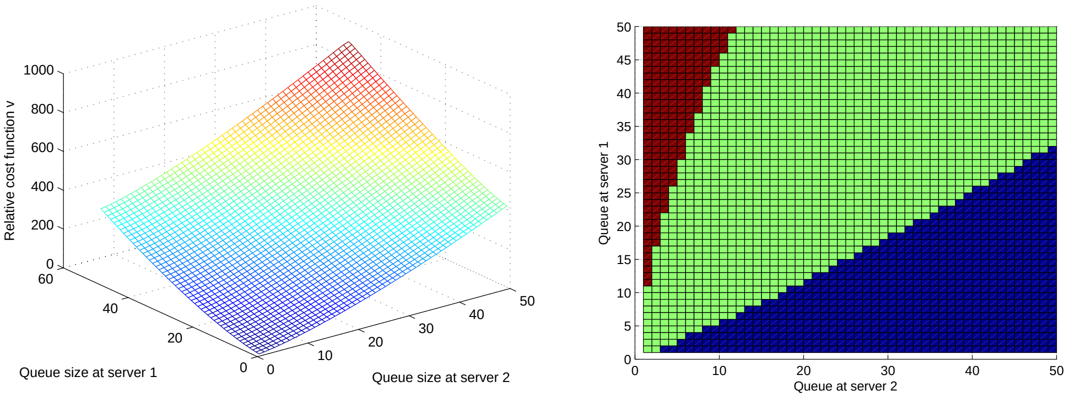

Figure 5: Relative cost function (left) and Routing policy (right)

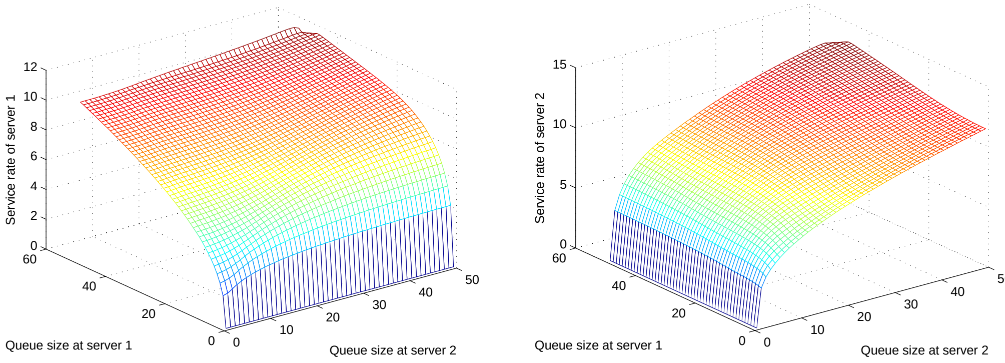

Figure 6: Service rate policy of servers 1 (left) and 2 (right)

These figures verify all our predicted structural properties: the relative cost function is monotonically nondecreasing, the optimal routing policy is a threshold policy and also the optimal service rate is monotonically nondecreasing if the queue size of other servers are held fixed.

4. Benefits Assessment

The focus of the project is to develop mathematical models for energy optimization of ICT, systematically derive optimization algorithms based on the model, understand structural properties of these algorithms, and evaluate their performance through simulations. The emphasis has been on developing a conceptual framework and designing abstract algorithms, and less about large-scale and comprehensive simulation of realistic systems. In particularly, the numerical simulations conducted in our project illustrate the behavior of our algorithms and serve as a sanity check. Accurate estimates of potential benefits will require a more careful, comprehensive, and large-scale simulation of a real network of datacenters than the numerical simulations done in the project.

Nonetheless, based on the results in Section 3.2, we'll expect that an optimal geographical load balancing algorithm together with dynamic adjustment of datacenter capacity can potentially save significant amount of brown energy, possibly 30% or higher, over the most popular algorithm that route requests to minimum-loaded datacenters.

5. Commercialization

The focus of the project is to develop a conceptual framework for energy optimization of ICT, systematically derive distributed algorithms from these mathematical models, and evaluate these algorithms using simulations.

The results of this project can form the core of a commercialization strategy, but much more effort and resources are needed to take these abstract algorithms to a stage that is deployable, including:

1. Comprehensive evaluation and design refinement.
   - Conduct more comprehensive, realistic and large-scale simulation of networks of datacenters to evaluate the benefit and behavior of these algorithms, paying particular attention to corner cases that are usually ignored in academic research.

- Improve algorithms based on simulation results, and repeat the cycle until satisfactory.

2. Prototyping and testing.
   - This will include developing proxy software to route job requests, probably in DNS servers, build software to collect in real time electricity prices or surrogates for energy costs at different datacenters, communication protocol and software between datacenters and proxy servers, optimization software at the proxy and the datacenters, software to turn on and off servers at a datacenter and monitor their utilization.
   - Run experiments to evaluate the performance of the prototype.
   - Improve prototype/algorithm based on experimental results.

3. Commercialization.
   - Productize the prototype, making it robust in real-world network environment and testing in scenarios that cover as many corner cases as possible.
   - Identify potential markets and customer list.
   - Formulate business model and go-to-market strategy.
   - Build founding team, raise funds, and start a company.
   - Alternatively, identify potential partners and explore technology licensing to these larger partners for commercialization.

6. Accomplishments

Deliverables: The original proposal has three main tasks. Task 1 is to formulate appropriate mathematical models that capture the essential features of energy optimization for ICT in datacenters and the expected outcome is at least one such model. Task 2 is to develop efficient algorithms to solve some of the optimization problems formulated in Task 1.0 and the expected outcome is at least one algorithm. Special attention will be paid to the efficiency, simplicity, decentralized structure, and the implementability of these algorithms. Task 3 is to analyze the efficiency and optimality of these algorithms and the expected outcome includes publications reporting models, algorithms, and their performance evaluation. Special attention will be paid to understanding the equilibrium and dynamic properties of these algorithms, the convergence to optimality or bounds on suboptimality of these algorithms.

We have met or exceeded these project objectives. In particular, we have developed multiple models to study different aspects of energy optimization for ICT in datacenters. We have developed multiple abstract algorithms to solve these optimization problems. We have conducted mathematical analysis to understand their structural properties and numerical simulations to evaluate their performance.

Publications: Our results are reported in the following publications:

[Chen2011] Lijun Chen, Na Li, Steven H. Low. *On the Interaction between Load Balancing and Speed Scaling.* Proceedings of ITA Workshop, February 2011

[Lim2011a] C. Lim and A. Tang, *Dynamic Speed Scaling and Load Balancing of Interconnected Queues*, Proceedings of ITA Workshop, February 2011

[Lim2011b] C. Lim and A. Tang, *Dynamic Control of Two Parallel Queues: Admission, Speed Scaling and Load Balancing*, submitted to Computer Networks, June 2011.

[Lin2011] Minghong Lin, Adam Wierman, Lachlan L. H. Andrew and Eno Thereska. *Dynamic right-sizing for power-proportional data centers.* IEEE INFOCOM, April 2011 (Best Paper Award)

[Liu2011] Zhenhua Liu, Minghong Lin, Adam Wierman, Steven Low, Lachlan L. H. Andrew. *Greening Geographical Load Balancing*. ACM Sigmetrics, Jun 7–11, 2011, San Jose, CA.

[Liu2012] Zhenhua Liu, Minghong Lin, Adam Wierman, Steven Low, Lachlan L. H. Andrew. *Geographical Load Balancing with Renewables*. ACM SIGMETRICS Performance Evaluation Review (PER), March 2012. (Best Student Paper Award at GreenMetrics Conference, June 7, 2011 San Jose, CA)

[Wierman2011] A. Wierman, L. L. H. Andrew, A. Tang, *Power-Aware Speed Scaling in Processor Sharing Systems: Optimality and Robustness*, To be submitted for publication, 2011

Award and thesis: The following paper won the Best Student Paper Award at the GreenMetrics Conference in 2011:

[Liu2012] Zhenhua Liu, Minghong Lin, Adam Wierman, Steven Low, Lachlan L. H. Andrew. *Geographical Load Balancing with Renewables*. ACM SIGMETRICS Performance Evaluation Review (PER), March 2012. (Best Student Paper Award at GreenMetrics Conference, June 7, 2011 San Jose, CA)

The following paper won the Best Paper Award at Infocom 2011:

[Lin2011] Minghong Lin, Adam Wierman, Lachlan L. H. Andrew and Eno Thereska. *Dynamic right-sizing for power-proportional data centers.* IEEE INFOCOM, April 2011

The results in [Liu2011] and [Liu2012] are the core in Zhenhua Liu's MS Thesis at Caltech in 2011. The results in [Lin2011] will form a part of the PhD thesis of Minghong Lin at Caltech, expected in 2012/2013.

The results in [Lim2011a] and [Lim2011b] will form a part of the PhD thesis of Chiunlin Lim at Cornell, expected in 2013.

7. Conclusions

We draw the following conclusion from our project:

- There is a substantial opportunity to minimize both the amount and the cost of electricity consumption in a network of datacenters, by exploiting the fact that traffic load, electricity cost, and availability of renewable generation *fluctuate over time and across geographical locations*. Judiciously matching these stochastic processes can *optimize the tradeoff* between brown energy consumption, electricity cost, and response time.
- Given the stochastic nature of these three processes, *real-time dynamic feedback* should form the core of any optimization strategy. The key is to develop decentralized algorithms that can be implemented at different parts of the network as simple, local algorithms that coordinate through asynchronous message passing.
- Our research suggests that simple scalable decentralized algorithms to optimize energy consumption at each server (speed scaling, scheduling, admission control), within a datacenter (sleep mode, rate control, admission control), and across multiple datacenters (routing) are possible. We have proposed a few of such algorithms, analyzed their optimality and stability properties, and evaluated their performance through numerical simulations.

- This set of results can form the core of a program that takes some of the algorithms developed in this project to the next level towards eventual commercialization; see Section 5 for more detail.

8. Recommendations

As explained above, there is substantial opportunity to minimize both the amount and the cost of electricity consumption in a network of datacenters, by exploiting the fact that traffic load, electricity cost, and availability of renewable generation *fluctuate over time and across geographical locations*. The results from this project can form the core of a program that takes some of the algorithms developed to the next level towards eventual commercialization; see Section 5 for more detail.

9. References

[Albers2006] S. Albers, H. Fujiwara, Energy-Efficient Algorithms for Flow Time Minimization, in: Lecture Notes in Computer Science (STACS), vol. 3884, 621–633, 2006.

[Andrew2010] L. L. Andrew, M. Lin, and A. Wierman. Optimality, fairness, and robustness in speed scaling designs. In Proceedings of ACM Sigmetrics, 2010.

[Andrews2010a] Andrews, M., Anta, A., Zhang, L., Zhao, W., 14-19 2010. Routing and scheduling for energy and delay minimization in the powerdown model. In: INFOCOM, 2010 Proceedings IEEE. pp. 1 -5.

[Andrews2010b] Andrews, M., Anta, A., Zhang, L., Zhao, W., 14-19 2010. Routing for energy minimization in the speed scaling model. In: INFOCOM, 2010 Proceedings IEEE. pp. 1-9.

[Bansal2007a] N. Bansal, T. Kimbrel, and K. Pruhs. Speed scaling to manage energy and temperature. J. ACM, 54(1):1–39, 2007.

[Bansal2007b] N. Bansal, K. Pruhs, and C. Stein. Speed scaling for weighted flow times. In Proc. ACM-SIAM SODA, 2007.

[Bansal2009] N. Bansal, H.-L. Chan, and K. Pruhs. Speed scaling with an arbitrary power function. In Proc. ACM-SIAM SODA, 2009.

[Beloglazov2010] A. Beloglazov, R. Buyya, Y. C. Lee, and A. Zomaya. A taxonomy and survey of energy-efficient data centers and cloud computing systems, Technical Report, 2010.

[Bunde2009] D.P.Bunde, Power-aware Scheduling for Makespan and Flow, J.Scheduling, 12(5):489–500, 2009

[Chen2011] Lijun Chen, Na Li, Steven H. Low. *On the Interaction between Load Balancing and Speed Scaling.* Proceedings of ITA Workshop, February 2011

[EPA2007] Environmental Protection Agency. Server and data center energy efficiency, Final Report to U.S. Congress, 2007.

[Fan2007] Fan, X., Weber, W.-D., Barroso, L. A., 2007. Power provisioning for a warehouse-sized computer. In: ISCA '07: Proceedings of the 34th annual international symposium on Computer architecture. New York, NY, USA, pp. 13-23.

[Google2011] Google, Retrieved March 18, 2011. Efficient computing: Data centers. URL http://www.google.com/corporate/green/datacenters/

[Herbert2007] S. Herbert and D. Marculescu. Analysis of dynamic voltage/frequency scaling in chip-multiprocessors. In Proc. ISLPED, 2007.

[Irani2005] S. Irani and K. R. Pruhs. Algorithmic problems in power management. SIGACT News, 36(2):63–76, 2005.

[Kattakayam1996] T. A. Kattakayam, S. Khan, and K. Srinivasan. Diurnal and environmental characterization of solar photovoltaic panels using a PC-AT add on plug in card. Solar Energy Materials and Solar Cells, 44(1):25–36, Oct 1996.

[Kaxiras2008] S. Kaxiras and M. Martonosi. Computer Architecture Techniques for Power-Efficiency. Morgan & Claypool, 2008.

[Katz2009] Katz, R. H., February 2009. Tech titans building boom. IEEE Spectrum.

[Lim2011a] C. Lim and A. Tang, *Dynamic Speed Scaling and Load Balancing of Interconnected Queues*, Proceedings of ITA Workshop, February 2011

[Lim2011b] C. Lim and A. Tang, *Dynamic Control of Two Parallel Queues: Admission, Speed Scaling and Load Balancing*, submitted to Computer Networks, June 2011.

[Lin2011] Minghong Lin, Adam Wierman, Lachlan L. H. Andrew and Eno Thereska. *Dynamic right-sizing for power-proportional data centers.* IEEE INFOCOM, April 2011 (Best Paper Award)

[Liu2011] Zhenhua Liu, Minghong Lin, Adam Wierman, Steven Low, Lachlan L. H. Andrew. *Greening Geographical Load Balancing*. ACM Sigmetrics, Jun 7–11, 2011, San Jose, CA.

[Liu2012] Zhenhua Liu, Minghong Lin, Adam Wierman, Steven Low, Lachlan L. H. Andrew. *Geographical Load Balancing with Renewables*. ACM SIGMETRICS Performance Evaluation Review (PER), March 2012. (Best Student Paper Award at GreenMetrics Conference, June 7, 2011 San Jose, CA)

[Pakbaznia2009] E. Pakbaznia and M. Pedram. Minimizing data center cooling and server power costs. In Proc. ISLPED, 2009.

[Pruhs2004] K.Pruhs, P.Uthaisombut, G.Woeginger, Getting the Best Response for Your Erg, in Scandinavian Worksh. Alg. Theory, 2004.

[Pruhs2008a] K.Pruhs, P.Uthaisombut, G.Woeginger, Getting the Best Response for Your Erg, ACM Trans. Algorithms, 4(3), 2008

[Pruhs2008b] K. Pruhs, R. van Stee, P. Uthaisombut, Speed scaling of tasks with precedence constraints, Theory of Computing Systems 43 (1): 67–80, 2008

[Qureshi2010] Qureshi, A., 2010. Power-demand routing in massive geo-distributed sys tems. Ph.D. thesis, Massachusetts Institute of Technology.

[Qureshi2009] Qureshi, A., Weber, R., Balakrishnan, H., Guttag, J., Maggs, B., 2009. Cutting the electric bill for internet-scale systems. In: Proc. ACM SIG COMM. New York, NY, USA, pp. 123-134.

[Rao2010a] Rao, L., Liu, X., Ilic, M., Liu, J., 2010. Mec-idc: joint load balancing and power control for distributed internet data centers. In: ICCPS '10: Proceedings of the 1st ACM/IEEE International Conference on Cyber Physical Systems. ACM, New York, NY, USA, pp. 188-197.

[Rao2010b] Rao, L., Liu, X., Xie, L., Liu, W., 14-19 2010. Minimizing electricity cost: Optimization of distributed internet data centers in a multi-electricity- market environment. In: INFOCOM, 2010 Proceedings IEEE. pp. 1-9.

[Stanojevic2010] R. Stanojevic and R. Shorten. Distributed dynamic speed scaling. In Proc.

IEEE INFOCOM, 2010.

[Unsal2003] O. S. Unsal and I. Koren. System-level power-aware design techniques in real-time systems. Proc. IEEE, 91(7):1055–1069, 2003.

[Wendell2010] P. Wendell, J. W. Jiang, M. J. Freedman, and J. Rexford. Donar: decentralized server selection for cloud services. In Proc. ACM Sigcomm, pages 231–242, 2010.

[Wierman2009] A. Wierman, L. L. H. Andrew, and A. Tang. Power-aware speed scaling in processor sharing systems. In Proceedings of IEEE Infocom, 2009.

[Wierman2011] A. Wierman, L. L. H. Andrew, A. Tang, *Power-Aware Speed Scaling in Processor Sharing Systems: Optimality and Robustness*, To be submitted for publication, 2011

[Yao1995] F. Yao, A. Demers, S. Shenker, A Scheduling Model for Reduced CPU Energy, in: Proc. IEEE Symp. Foundations of Computer Science (FOCS), 374–382, 1995.

[Yuan2005] 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.

[Zhang2007] S. Zhang, K. S. Catha, Approximation Algorithm for the Temperature-aware Scheduling Problem, in: Proc. IEEE Int. Conf. Comp. Aided Design, 281–288, 2007.

[Zhu2005] Y. Zhu and F. Mueller. Feedback edf scheduling of real-time tasks exploiting dynamic voltage scaling. Real Time Systems, 31:33–63, 2005.