Electric Grid Balancing through Low-Cost Workload Migration

David Chiu
Washington State University
Vancouver, WA 98686
david.chiu@wsu.edu

Christopher Stewart
The Ohio State University
Columbus, OH 43210
cstewart@cse.ohio-state.edu

Bart McManus
Bonneville Power Administration
Vancouver, WA 98666
bamcmanus@bpa.gov

ABSTRACT

Energy production must continuously match demand on the electric grid. A deficiency can lead to service disruptions, and a surplus can place tremendous stress on grid components, potentially causing major blackouts. To manage this balance, grid operators must increase or lower power generation, with only a few minutes to react. The grid balancing problem has also impeded the pace of integrating bountiful renewable resources (e.g., wind), whose generation is intermittent. An emerging plan to mitigate this problem is *demand response*, i.e., for grid operators to alter the electricity usage behavior of the masses through real-time price signals. But due to prohibitively high infrastructure costs and societal-scale adoption, tangible demand response mechanisms have so far been elusive.

We believe that altering the usage patterns of a multitude of data centers can be a tangible, albeit initial, step towards affecting demand response. Growing in both density and size, today's data center designs are shaped by the increasing awareness of energy costs and carbon footprint. We posit that shifting computational workloads (and thus, demand) across geographic regions to match electricity supply may help balance the grid. In this paper we will first present a real grid balancing problem experienced in the Pacific Northwest. We then propose a symbiotic relationship between data centers and grid operators by showing that mutual cost benefits can be accessible. Finally, we argue for a *low cost* workload migration mechanism, and pose overarching challenges in designing this framework.

Keywords

demand response, smart grid, data centers

1. INTRODUCTION

Power grid operators face a challenging balancing act. Because electricity cannot be stored cheaply, it must be generated and consumed at roughly the same time. On one hand, grid operators must orchestrate a set of power sources to produce enough electricity to satisfy their customers' load. On the other hand, if too much electricity is generated, the excess must be disposed via inefficient mechanisms, e.g., circuit breakers. A poor decision could cause power outages, leading to significant economic and societal impact.

Copyright is held by author/owner(s).

Increased green energy penetration further complicates grid balancing. The push to integrate large amounts of renewable energy has led to increased development of wind fleets and solar photovoltaic farms. Although carbon free, wind and solar resources can be variable and unpredictable, which adds hardship to the balancing operation. At the same time, emerging dynamic technologies like electric cars, lithium-ion batteries, and energy-saving appliances will likely increase electricity demand and disrupt load predictability. This further opposes the unreliability of green energy. In this variable demand and supply environment, grid operators must therefore track and respond to customer load in even greater detail and faster response time.

An approach to reconcile variable generation and demand is two-fold: (1) Diversify renewable resources and (2) Alter the load. To understand this vision, consider this scenario: Morning is settling in on Eastern U.S. cities, and their load is increasing as expected. In response, grid operators must ramp up generation through traditional, albeit controllable, methods (e.g., burning coal/gas). Still deep into the night, distant regions to the West may be experiencing great amounts wind with little demand. Without cheap storage, power authorities may be forced to waste an abundance of wind energy. Ideally in this situation, the Western-based utilities could instead transfer their excess wind energy at a diminished cost to the East. Likewise, when wind generation is lowered, or when the Western U.S. increases demand, the costs could rise accordingly. Such dynamic pricing signals serve as a mechanism to influence load.

While this *demand response* mechanism is a key goal for the Smart Grid [10], its implementation has been stifled by high costs of restructuring the current transmission network:

- The electric grid was designed to confine major outages to a region, specifically, the Eastern U.S., Western U.S., and Texas Interconnects. Links between these interconnects are therefore distant and weak, resulting in transmission loss and prohibiting the integration of geographically diverse resources.
- To truly realize a demand-response mechanism with real-time pricing (RTP) will take a nontrivial amount of time. Societal-scale appliance upgrades and acceptance of changes on electricity usage behavior are prerequisite.

We argue, however, that our current infrastructure can potentially affect demand response with minimally invasive change. Our position is that electricity demand can be geographically distributed via large-scale data center workloads, circumventing the above two challenges. We can affect demand with cost incentives to migrate massive amounts of computation across geographically dispersed data centers.

Myriad data centers support a substantial portion of the Internet's infrastructure and operations. Due to their growing number and density, the Department of Energy has recommended a 10% power reduction in today's data center operations [19]. In response to this call, recent research has focused almost unilaterally on minimizing the data center's operational costs. Energy-aware advancements, including low-power multicore architectures, virtualization, and cooling techniques now permeate modern data center design. Data centers have also begun purchasing green energy credits to further reduce their overall carbon footprint [11, 13].

Studies as recent as 2011 have shown that the roughly 6000 data centers in the U.S. consume anywhere from 80 to 120 billion kilowatt-hours (kWh) per year, i.e., 2.2% to 3.5% of all electricity use [31, 17, 18]. Because data centers can command great amounts of energy, we propose shifting power load geographically by migrating computation across data center locations. We believe that a massive, but low-cost, compute migration will not only lower costs for both data centers and electric utilities, but also increase the rate of variable green energy integration.

The remainder of this paper is organized as follows. Section 2 describes a real-world grid balancing problem in the Pacific Northwest. We present the motivating cost benefits in Section 3 and explore initial technical challenges in Section 4. Related works will be discussed in Section 5, and we will conclude in Section 6.

2. MOTIVATING PROBLEM AT BPA

The Bonneville Power Administration (BPA) is a federal power marketing agency, regulating a diverse set of power sources (hydro, wind, thermal/nuclear) in the Pacific Northwest [7]. Just this year (March 2012) BPA announced it had integrated close to 5000 MW of wind power from privately owned wind farms, collectively known as a wind fleet, hoping to market their energy to the public. The fleet's output is predicted on an hourly schedule, but due to imperfect prediction models, BPA must be ready to balance the error margin, forcing incremental and decremental reserves to be running on standby, which explains the inherently higher costs of wind energy [1].

Within the hourly schedule, BPA can ramp up an *incremental reserve* to reconcile a wind deficiency. To compensate for a surplus, operators must call on a *decremental reserve* (e.g., its control over 30+ hydroelectric dams) to offset unexpected increases in wind production [2]. Decrementing is done by spilling water over the dam instead of passing it through the turbines. While spilling has been an effective strategy, it is far from optimal in terms of costs. Here, wind energy would be offsetting an already green resource (hydro), but a deep environmental concern also looms: The spills can hurt the salmon fish population and consequently, the broader ecosystem [2].

With increasing numbers of wind farms requesting for BPA interconnection, this problem's urgency has reached a boiling point, and BPA's Wind Integration Team must seek tangible solutions to this problem.

3. COST ANALYSIS

We argue that myriad data centers, which are distributed and significant contributors of demand, can work in concert with local electric utilities for mutual cost and environmental benefits. In this section, we model the utility and data center costs in a workload migration environment. We emphasize that our preliminary models are simplistic, but can nonetheless communicate the cost benefits at a high level.

3.1 Modeling Utility Cost

We will first show that the workload migration framework would be economically beneficial to electric utilities. Let us define the Regular Operation at an electric utility U as follows. Assuming demand at time t can be interpreted as,

$$D_t = w_t + \overline{w_t} \tag{1}$$

where w_t and $\overline{w_t}$ denote wind and non-wind energy respectively, generated at time t. Suppose that R is the price charged to consumers for 1 MWh of energy, and C_w and $C_{\overline{w}}$ represent the cost to produce 1 MWh of wind and non-wind energy, respectively. Utility U's profit can be expressed as the revenue minus cost of operation,

$$\begin{aligned} P_U(t) = &\ Rev_U(t) - Cost_U(t) \qquad (2)\\ = &\ R \times (w_t + \overline{w_t}) - (C_w \times w_t + C_{\overline{w}} \times \overline{w_t} + \Omega(\Delta_t)) \end{aligned}$$

where $\Omega(\Delta_t)$ denotes the overhead cost to run incremental/decremental reserves on standby to balance $|\Delta_t|$ MWh instantaneously. For simplicity, we will assume that Ω monotonically increases over $|\Delta_t|$.

Now suppose at time $t+1$, wind energy production w_{t+1} is displaced by Δ_w MWh, and demand also changes by Δ_D MWh. We express the *true* displacement at $t+1$ as,

$$\begin{aligned} \Delta_{t+1} = &\ \Delta_D - \Delta_w \\ = &\ (D_{t+1} - D_t) - (w_{t+1} - w_t) \qquad (3) \end{aligned}$$

As wind energy ramps up or down at $t+1$, we may have to displace Δ_D MWh of change in demand. The utility's profit at $t+1$ is defined as,

$$\begin{aligned} P_U(t+1) = &\ Rev_U(t+1) - Cost_U(t+1) \\ = &\ R \times (w_t + \overline{w_t} + \Delta_D) - Cost_U(t+1) \qquad (4) \end{aligned}$$

where

$$\begin{aligned} &Cost_U(t+1) = \qquad (5)\\ &\qquad C_w \times w_{t+1} + C_{\overline{w}} \times \overline{w}_{t+1} + \Omega(\Delta_{t+1}) \end{aligned}$$

The utility's profit $P_U(t+1)$ is constrained by the overhead cost of $\Omega(\Delta_{t+1})$. Specifically, when $\Delta_{t+1} \neq 0$, then an imbalance exists, and we must either ramp up or lower the reserve power by $|\Delta_{t+1}|$ MWh. Notice that in this model, we also affix the same overhead costs for a decrease in generation. The intuition is derived from a waste of otherwise profitable resources. As mentioned in Section 2, the BPA can spill water over its dams instead of turbines to lower generation. The waste of Δ_{t+1} MWh of hydro in this case is also viewed as an overhead cost of $\Omega(\Delta_{t+1})$ in our model.

In general, the electric utility cost optimization problem is simply to minimize $\sum_{i=0}^{t+1} \Omega(\Delta_i)$. Because Ω is an increasing function over Δ_t and $\Delta_t \to 0 \Rightarrow \Omega(\Delta_t) \to 0$, our approach seeks to eliminate Δ_t by providing a conceptually new approach for a grid operator to balance energy on the grid: Migrating some resident data center load to a non-regional compute location and reducing electricity demand. Conversely, compute load can be brought near the utility when there is a surplus of wind energy.

If we let Δ' denote the amount of energy migratable as workloads across data centers, then electric utilities have two ways to curtail costs:

- When $\Delta_{t+1} < 0$ (i.e., wind production decreases and/or demand increases): We reduce local demand by Δ' to circumvent the overhead cost to generate the extra Δ_{t+1} incremental reserve energy.
- When $\Delta_{t+1} > 0$ (i.e., wind production increases and/or demand decreases): We can add to local demand by Δ' to reduce the overhead of wasting decremental reserve energy.

The utility's operational cost at $t+1$ with migration ($UMig$) can now be altered as follows,

$$Cost_{UMig}(t+1) = \quad (6)$$
$$C_w \times w_{t+1} + C_{\overline{w}} \times (\overline{w_t} - (\Delta_{t+1} - \Delta')) + \Omega(\Delta_{t+1} - \Delta')$$

3.2 Mutual Cost Benefits for Data Centers

Data centers, on the other hand, share the same goal of reducing total operational costs. To entice data centers to buy-in on the demand-response mechanism, electric utilities must use dynamic price signals. Suppose we have a data center DC consisting of a set of n geographically disperse locations $DC = \{u_1, u_2, \ldots, u_n\}$. Furthermore, each data center location $u \in DC$ is associated with its own demand $d_u(t)$ and cost R_u for 1 MWh energy from its local utility. The total cost of operation for DC at time t can be defined as,

$$Cost_{DC}(t) = \sum_{u \in DC} R_u \times d_u(t) \quad (7)$$

Suppose the electric utility nearby data center u experiences a power deficiency and signals for an ϵ increase in pricing per MWh. The rise in costs may encourage a transfer of Δ' MWh from u to its remaining data center sites, $DC - \{u\}$ at $t+1$.* The aggregate cost of DC at $t+1$ would be,

$$Cost_{DC}(t+1) = (R_u + \epsilon) \times (d_u(t) - \Delta') + \left[\sum_{v \in DC - \{u\}} R_v \times (d_v(t) + \Delta'_v) + \underset{u \to v}{C}(Work(\Delta'_v)) \right] \quad (8)$$

where Δ'_v denotes the fraction of Δ' that a remote data center v must offset, and $Work(x)$ is the amount of computational work equivalent to x MWh. The migration cost,

$$\underset{u \to v}{C}(Work(\Delta'_u)) = \quad (9)$$
$$C_s(u) + C_{net}(Work(\Delta'_v)) + C_s(v)$$

Clearly, for migration to be economically sensible on either end of the utility and the data center, the following conditions must be satisfied:

$$Cost_{UMig}(t+1) < Cost_U(t+1) \quad \texttt{(C-1)}$$
$$Cost_{DC}(t+1) < Cost_{DC}(t) \quad \texttt{(C-2)}$$

*Due to space limitations, we only show one direction of migration, but the reverse would be similar.

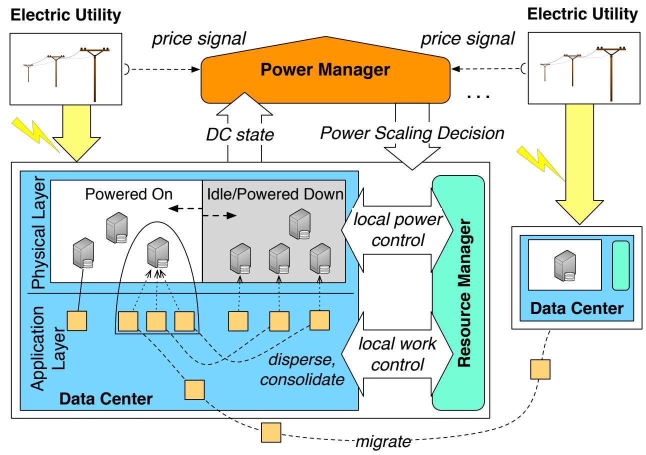

Figure 1: Demand Response Data Center Management Framework

For Condition `C-1` to hold, we observe that $Cost_{UMig}(t+1)$ can be minimized when the utility displaces Δ' approaching true displacement Δ_{t+1}. This would not only require timely and accurate wind prediction models, but also, the utility must make sensible pricing adjustments to lower the risk of a loss in revenue [24]. Condition `C-2` requires that the staging costs (C_s) and data transfer costs (C_{net}) between the migrating data centers, both functions of time, be trivial. Moreover, the change in price, ϵ, issued by the utility must also be significant to encourage the migration process. The function to distribute work Δ_v among the destinations, and the relationship between $Work$ and energy would also affect Condition `C-2`.

These models provide an analytical basis for our intuition that we need systems where the data transfer cost is *extremely low*, while the amount of work migrated is high.

4. LOW COST WORKLOAD MIGRATION

We envision a Demand-Response Data Center Management Framework, shown in Figure 1. We will present a high level description of the framework's major functionalities. The *Physical Layer* of our data center refers to the infrastructure, e.g., machines, lighting, cooling, etc. We could measure energy demand at both coarse-grained (building phase) and fine-grained (rack or machine PDU) levels. The *Application Layer* for each data center refers to the computational components. The unit of migratable computational $Work$ can be defined at this level. The type of work of interest should be both spreadable and malleable, i.e., easily consolidated, dispersed, and/or scaled. We will revisit the applications later in this section.

Each data center will associate an *Electric Utility* which dynamically signals its unit-energy price. The *Power Manager* is responsible for minimizing cost from the global system perspective by handing down power scaling decisions (i.e., Δ') to participating data centers. It monitors the utilities' dynamic price signals and the power state of each data center.

The *Resource Manager* in each data center must decide on how to increase or decrease power demand. For instance, to increase usage, a data center could turn on a set of machines and redistribute its work. Conversely, decreasing usage might be done by consolidating existing work onto a smaller set of physical machines or delay scheduling batch jobs. When consolidation is infeasible, then the data center should determine a candidate set of work to migrate.

For our approach to be viable from both stakeholders, together with BPA grid operator personnel, we have identified following key questions:

Can data centers really affect large scale electricity loads? We analyzed BPA's reserve deployment data over a 7 day period. Incremental reserves were deployed 808 times, while decremental reserves deployed 1883 times, more than doubling the amount of incremental deployment. However, the power displaced in all deployments are highly skewed towards the lower end. For example, 80% of balancing occurs for roughly $\leq$ 200 MW. Let us compare this with Koomey's 2011 report [18] which Google to have 220 MW average power consumption. It suggests that a small number of participating data centers could reduce power reserve deployment. We concede that further investigation is needed to support this speculation.

What is the relationship between energy and computational $Work$? Recent works have shown that power usage and computation share a non-trivial relationship [15, 4, 30, 25, 28]. Further understanding this relationship and the particular applications that influence power is imperative for our proposed paradigm. Clearly, the work distribution function for the participating data centers would be closely related.

How do we measure the success of the workload migration paradigm? We will leverage BPA's live and historical data pertaining to power generation (hydro, wind, thermal) and consumer demand. The data shows instances of when balancing occurred, and for how long. Below, we consider several data center applications, and how the answers to these questions vary with each.

How quickly must the workload transfer take place? From our interviews with BPA personnel, in the worst case, their facilities can wait up to approximately 10 minutes before tapping the power reserves. Our migration approach must consider fast transfers of work. We consider the following pervasive data center applications.

Map-Reduce: By itself, a single Map-Reduce instance appears to be a poor candidate for grid balancing. Having to pause and restart computation elsewhere requires transferring massive amounts of state, that is, $Work(\Delta')$ can be high. However, emerging distributed Map-Reduce environments that can *collectively process* subsets of data shared on a remote host (e.g., a data cloud) can be considered [8]. Because Map-Reduce is idempotent, scaling computational nodes up/down on various data center locations would have no effect on final results. With proper coordination, this can potentially drive $Work(\Delta')$ to be small.

Simulation: Large-scale simulation is characterized by low input with high computational requirements leading to significant electricity demand. Combined with the ability to pause and restart execution with relatively low amounts of saved state, e.g., small $Work(\Delta')$ transferred, makes simulation an excellent candidate for electric grid balancing. We would, however, expect a complicated staging time, as the simulation software must already be resident at all locations. This may mean having to boot up many virtual machine images at the receiving data center locations before restarting the simulation, which can take time.

Elastic Cloud/Web Services: Cloud computing has exploded in popularity. At peak loads, cloud services represent a significant electrical demand and can be very effective for grid balancing because Web-based protocols are stateless and thus can be shifted easily via DNS [20, 24]. Another way cloud providers could shift load is through their IaaS framework. For instance, live migration [16] or VM allocation in certain geographical regions can be controlled. However, cloud workload demand is tantamount to that of electric demand: customer-driven and variable. During low periods in the end-users' request rates, the electrical demand will decrease, making migration less effective.

With BPA collaboration, we will deploy the migration framework over large-scale clusters at Washington State University and the Ohio State University as our initial distributed data center testbed. We will evaluate our approach using BPA's historic generation and load data sets on these three types of applications to simulate demand-response.

5. RELATED WORKS

The emergence of energy-aware and energy-efficient *"green" clouds* [6, 22, 5, 9, 29, 23] has generated much excitement in systems research circles. Particularly, these refer to cloud infrastructures which strive to be energy-aware by leveraging the use of VMs and, particularly, their migration, while meeting performance requirements.

Nathuji and Schwan proposed *VirtualPower* to flexibly manage power in a virtualized data center [25]. They showed that a combination of soft scaling, hard scaling, and VM consolidation can be leveraged to provide viable power management in a data center. Liu, *et al.* proposed the GreenCloud architecture, which combines online monitoring of physical resources with a technique for finding power-saving VM placements [22]. They modeled VM migration time, energy consumed, and physical server load as a function of cost. Le, *et al.* consider VM placement in a cloud environment for high performance applications [21]. The authors propose policies for VM migration across multiple data centers in reaction to power pricing. More recently, Goiri, *et al.* proposed *GreenSlot* [12], a solar power-sensitive scheduling algorithm for data center workloads. The authors argue that "the ideal design for green datacenters connects them to both solar/wind energy sources and the grid (as a backup)." Stewart and Shen considered data center architectures that are equipped with onsite intermittent renewable sources [29]. HP Labs has also recently revealed the Net-Zero Energy Data Center, which can shape its demand based on supply [26].

Rao, *et al.* sought to minimize overall costs for multiple data centers located in disparate energy marketing regions [27]. Akoush, *et al.'s Free Lunch* architecture for cloud data centers shares several aspects of our goals [3]. The authors argue for either pausing VM executions or migrating VMs between sites based on local and remote energy availability. The Canadian-based Greenstar Network provides similar efforts toward developing in green load-following carbon protocol [14]. Liu, *et al.*'s geographical load balancing matches very closely to our goals [24]. The authors assume a general Internet service-request workload for data centers located in various geographical regions. They proposed distributed algorithms for minimizing aggregated costs by solving for an optimal number of active servers per data center and a load balancing policy (request routing). Liu, *et al.* studied workload and cooling management for a data center colocated with a PV microgrid [23]. They evaluated their optimization models over interactive (web requests) and batch jobs, and showed their scheduling algorithm can reduce non-renewable energy costs.

In contrast, we argue for an integration of utilities and data centers. We model the utility's overhead cost grid balancing, and our work distribution framework considers mass migration of several classes of data center workloads. Our workload migration model must meet specific time and energy displacement constraints, set externally by grid operators.

6. CONCLUSION

Due to their variable production, green energy presents a tremendous challenge for system operators when matching load and supply. In this paper we posit that data centers, which are becoming increasingly dense, can be used to influence electricity load. We argue that *increasing* the data centers' power consumption can further help reduce environmental impact. We have presented initial models that show the mutual cost benefits for both utilities and data center operations. We will work directly with the Wind Integration Team at the Bonneville Power Administration (BPA) to solve real-world power management problems in the Pacific Northwest.

7. REFERENCES

[1] *Northwest Wind Integration Action Plan.* Wind Integration Forum document. Northwest Power and Conservation Council, 2007.

[2] Columbia river high-water operations. Technical report, Bonneville Power Administration, 2010.

[3] S. Akoush, R. Sohan, A. Rice, A. W. Moore, and A. Hopper. Free lunch: exploiting renewable energy for computing. In *Proceedings of the 13th USENIX Workshop on Hot Topics in Operating Systems (HotOS 11)*. USENIX, 2011.

[4] L. A. Barroso and U. Hölzle. The case for energy-proportional computing. *Computer*, 40:33–37, December 2007.

[5] A. Beloglazov and R. Buyya. Energy efficient resource management in virtualized cloud data centers. In *CCGrid'10*, pages 826–831, Los Alamitos, CA, USA, 2010. IEEE Computer Society.

[6] A. Berl, E. Gelenbe, M. D. Girolamo, G. Giuliani, H. D. Meer, M. Q. Dang, and K. Pentikousis. Energy-efficient cloud computing. *The Computer Journal*, 53(7):1045–1051, 2009.

[7] Bonneville Power Administration, http://www.bpa.gov.

[8] M. Cardosa, *et al.* Exploring mapreduce efficiency with highly-distributed data. In *MapReduce'11*, pages 27–34, New York, NY, USA, 2011. ACM.

[9] N. Deng, C. Stewart, D. Gmach, and M. Arlitt. Policy and mechanism for carbon-aware cloud applications. In *Proceedings of Network Operations and Management Symposium (NOMS'12)*. IEEE, 2012.

[10] Department of Energy. The Smart Grid: An Introduction. 2008. http://www.oe.energy.gov/SmartGridIntroduction.htm.

[11] Greenhouse data, http://www.greenhousedata.com.

[12] I. Goiri, R. Beauchea, K. Le, T. D. Nguyen, M. E. Haque, J. Guitart, J. Torres, and R. Bianchini. Greenslot: scheduling energy consumption in green datacenters. In *Proceedings of SC'11*, New York, NY, USA, 2011. ACM.

[13] Google green, http://www.google.com/green.

[14] Greenstar Network, http://www.greenstarnetwork.com.

[15] A. E. Husain Bohra and V. Chaudhary. *VMeter: Power modelling for virtualized clouds*, pages 1–8. Ieee, 2010.

[16] C. C. Keir, C. Clark, K. Fraser, S. H, J. G. Hansen, E. Jul, C. Limpach, I. Pratt, and A. Warfield. Live migration of virtual machines. In *In Proceedings of the 2nd ACM/USENIX Symposium on Networked Systems Design and Implementation (NSDI)*, pages 273–286, 2005.

[17] J. G. Koomey. Worldwide electricity used in data centers. *Environmental Research Letters*, 3(3):034008 (8pp), 2008.

[18] J. G. Koomey. Growth in data center electricity use 2005 to 2010. *Analytics Press*, 2011.

[19] P. Kurp. Green computing. *Communications of the ACM*, 51(10):11–13, 2008.

[20] K. Le, R. Bianchini, M. Martonosi, and T. Nguyen. Cost- and Energy-Aware Load Distribution across Data Centers. In *Proceedings of the Workshop on Power Aware Computing and Systems (HotPower'09)*. USENIX, 2009.

[21] K. Le, R. Bianchini, J. Zhang, Y. Jaluria, J. Meng, and T. D. Nguyen. Reducing electricity cost through virtual machine placement in high performance computing clouds. In *Proceedings of SC'11*, pages 22:1–22:12, New York, NY, USA, 2011. ACM.

[22] L. Liu, H. Wang, X. Liu, X. Jin, W. B. He, Q. B. Wang, and Y. Chen. Greencloud: a new architecture for green data center. In *Proceedings of the 6th International Conference on Autonomic Computing and Communications (ICAC'09)*, pages 29–38, New York, NY, USA, 2009. ACM.

[23] Z. Liu, Y. Chen, C. Bash, A. Wierman, D. Gmach, Z. Wang, M. Marwah, and C. Hyser. Renewable and cooling aware workload management for sustainable data centers. In *SIGMETRICS'12*, pages 175–186. ACM, 2012.

[24] Z. Liu, M. Lin, A. Wierman, S. H. Low, and L. L. Andrew. Greening geographical load balancing. In *SIGMETRICS'11*, pages 233–244. ACM, 2011.

[25] R. Nathuji and K. Schwan. VirtualPower: coordinated power management in virtualized enterprise systems. In *Proceedings of the 21st ACM SIGOPS Symposium on Operating Systems Principles*, volume 41 of *SOSP '07*, pages 265–278, New York, NY, USA, Oct. 2007. ACM.

[26] HP Unveils Architecture for First Net Zero Energy Data Center, http://www.hp.com/hpinfo/newsroom/press/2012/120530c.html.

[27] L. Rao, X. Liu, L. Xie, and W. Liu. Minimizing electricity cost: Optimization of distributed internet data centers in a multi-electricity-market environment. In *Proceedings of the 29th IEEE International Conference on Computer Communications (INFOCOM'10)*, pages 1145–1153, 2010.

[28] S. Srikantaiah, A. Kansal, and F. Zhao. Energy aware consolidation for cloud computing. In *Proceedings of the Workshop on Power Aware Computing and Systems (HotPower'08)*. USENIX, 2008.

[29] C. Stewart and K. Shen. Some joules are more precious than others: Managing renewable energy in the datacenter. In *Proceedings of the Workshop on Power Aware Computing and Systems (HotPower'09)*. USENIX, 2009.

[30] A. S. Szalay, G. C. Bell, H. H. Huang, A. Terzis, and A. White. Low-power amdahl-balanced blades for data intensive computing. *SIGOPS Oper. Syst. Rev.*, 44:71–75, March 2010.

[31] U.S. Environmental Protection Agency. EPA Report to Congress on Server and Data Center Energy Efficiency. Technical report, U.S. Environmental Protection Agency, 2007.