A Case for Virtualizing the Electric Utility in Cloud Data Centers
Position paper

Cheng Wang, Bhuvan Urgaonkar, George Kesidis, Uday V. Shanbhag, Qian Wang
The Pennsylvania State University *

Abstract

Since energy-related costs make up an increasingly significant component of overall costs for data centers run by cloud providers, it is important that these costs be propagated to their tenants in ways that are fair and promote workload modulation that is aligned with overall cost-efficacy. We argue that there exists a big *gap* in how electric utilities charge data centers for their energy consumption (on the one hand) and the pricing interface exposed by cloud providers to their tenants (on the other). Whereas electric utilities employ complex features such as peak-based, time-varying, or tiered (load-dependent) pricing schemes, cloud providers charge tenants based on IT abstractions. This gap can create shortcomings such as unfairness in how tenants are charged and may also hinder overall cost-effective resource allocation. To overcome these shortcomings, we propose a novel idea of a *virtual electric utility* (VEU) that cloud providers should expose to individual tenants (in addition to their existing IT-based offerings). We discuss initial ideas underlying VEUs and challenges that must be addressed to turn them into a practical idea whose merits can be systematically explored.

1 Introduction and Motivation

The energy consumption of data centers is an important contributor to their overall costs, with large data centers spending millions of dollars per year on their electric utility bills [19, 3]. Figure 1 presents a comparison of these power-related costs for a state-of-the-art large data center based on the amortized monthly costs under Duke electric company's pricing scheme [12]. As shown, a 10MW data center with roughly 20K servers incurs a monthly electric bill of around $730K - comparable to the amortized monthly costs for procuring the IT resources (such as servers, storage, and networking switches/routers).

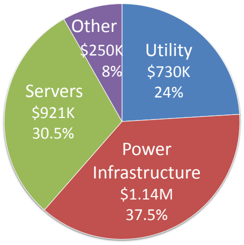


Figure 1: Amortized monthly costs for the physical infrastructure of a 10MW data center. Sources: [4, 18, 28]

In fact, the relative contribution of energy costs to overall costs might grow in the future; this is rooted in the observation that whereas IT equipment generally tends to become cheaper over time, power costs are likely to grow (especially, the cost associated with peak power consumption) [20] [1]. Reducing data center energy costs is, therefore, widely recognized as an important problem by cloud providers.

Since these energy costs are eventually recouped from the consumers (tenants) of cloud platforms, one expects any improvements in these costs to trickle down to the tenants. More importantly, given how significant a contributor energy-related costs are to overall data center costs, one expects that the cloud provider recoups these in a "fair" manner. [2]

Unfortunately, as we argue in this paper, current clouds and their tenants may fail to achieve this intuitively desirable behavior. The root cause behind this shortcoming is the big *gap* that exists today between how electric utility companies charge data centers for their energy usage (on the one hand) and how these costs are passed on to cloud tenants (on the other). In what follows, we first elaborate upon this gap and then upon shortcomings it may create.

Gap between utility and cloud pricing: Real-world electric utility companies employ pricing schemes wherein large consumers like cloud-scale data centers are

[*] This work was supported, in part, by the following: NSF CAREER award 0953541, NSF 1000699, and NSF 1228717.

[1] Predicting the overall trend would likely require more sophisticated arguments since certain sources of energy "wastage" have continuously improved both in IT hardware and others such as cooling. To our knowledge, such analysis is lacking.

[2] We are using the term "fair" somewhat informally here and are not proposing a specific definition/measure. We will discuss an initial idea for formalizing it in Section 2.

charged not just based on their aggregate energy consumption over the billing cycle, but in much more complex ways that may incorporate load-dependent and/or time-varying pricing. There are three main examples of such complexity, and real-world schemes often employ hybrids of these.

First, in *peak-based pricing*, in addition to energy-based charges, the utility bill also has a component based on the peak power drawn over the billing cycle. As a simplified example, Duke electric utility imposes a peak charge of about 12-15 $/kW/month on its larger consumers in addition to an energy charge of 5-10 c/kWh [12]. This amounts to a peak charge that is up to 600 times higher than the energy charge over a typical time window of 30 minutes. Second, in *time-varying pricing* the per-unit electricity price changes over time and the utility bill is a weighted sum of the consumption time-series. A variety of schemes of this kind are found with price variations at different time granularity and different kinds of lookahead/predictability (e.g., "day vs. night" pricing vs. prices changing every few minutes, hour-ahead vs. day-ahead prices, etc.) An important example within this class is that of "coincident peak" pricing [10] wherein the electric utility imposes a high per-unit energy price during periods where many consumers simultaneously impose high loads causing an overall high demand. Finally, in *tiered pricing* some demand thresholds define different per unit prices for electricity; a special case has only one threshold, upon exceeding which a high rate (or additionally a "penalty") is imposed [27].

A typical contemporary cloud provider, on the other hand, charges its tenants based on *virtualized IT abstractions*. Specifically, this pricing interface consists of a variety of options for (usually virtualized) IT resources - machines/instances, storage, networking, etc., (as in IaaS public clouds or many private clouds) or even software abstractions (as in SaaS clouds) with different price vs. performance vs. availability trade-offs. The prices within these interfaces may be exposed explicitly (as is the case with public clouds [1, 2, 30, 16]) or they may be more indirect (as in certain private clouds). Regardless, these pricing schemes, in effect, bundle the energy-related costs incurred by the data center into prices of the IT abstractions that are offered to the tenants. Whereas this way of propagating energy costs onto the tenants may work fine if energy costs were based only on aggregate/raw energy consumed, it fails to do so given the complexities of electric utility pricing discussed above.

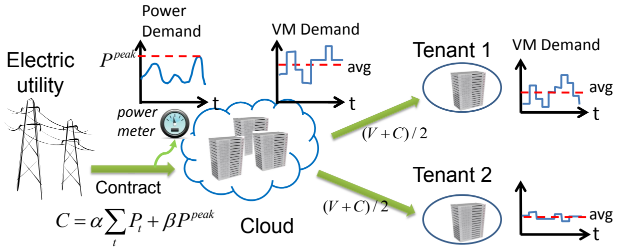


Figure 2: An example illustrating unfairness in how tenants with different demand variances are charged under a peak-based electricity pricing scheme. (V: VM cost charged by the cloud.)

Shortcomings this gap may cause: Whereas the IT-centric pricing interface exposed by cloud providers has the merit of being simple for a tenant to understand and use, this simplification may come at a heavy price. To keep the discussion concrete, we consider a cloud provider with a single data center whose electric utility employs a peak-based pricing scheme with its energy charge and peak power charge parameters denoted by α and β, respectively. We will use this peak-based pricing in our discussions throughout the rest of the paper as well. Our arguments, however, easily extend for other kinds of electric utility pricing schemes. The gap we identified above can result in shortcomings related to the following aspects:

Fairness: Consider the example shown in Figure 2, wherein two tenants procure resources from an IaaS cloud in the form of virtual machines (VMs). Ignoring complexities such as the diversity in the types of VMs offered by most cloud providers [1, 2, 30, 16], "reserved" or "spot-priced" VMs, etc., each tenant employs a profit optimizing VM allocation algorithm (e.g., several recent papers present such algorithms [42, 43, 45]) based on its predicted workload. Let us consider a thought experiment wherein tenant 1 and tenant 2 have the same total VM demand (and hence create the same *raw* energy demand under the simplified assumption that energy consumption is in proportion to number of VMs used) over a given period of time. However, tenant 1 has a higher demand variation than tenant 2. Suppose that the cloud is charged C\$ by the electric utility and that the other (non-energy such as IT costs) costs that the cloud needs to recoup from the tenants amount to V. The cloud would charge both these VMs equally ($\frac{V+C}{2}$). This example is consistent with existing cloud pricing. However, this is obviously unfair to tenant 2 - tenant 1, who has the higher consumption variation, is contributing more to the *peak* power consumption (and hence to the electric utility bill) of the cloud than tenant 2. Furthermore, this unfairness will only become worse as the relative share of energy-related costs becomes higher. Additionally, workloads can exhibit great diversity in the energy consumption of the VMs, making matters even more complex.

Cost-efficacy: Due to the gap described above, any demand-response (DR) carried out by the data center for optimizing its energy-related costs may end up being done without effective participation of tenants since they are not aware of the electricity pricing schemes imposed on data centers and thus cannot react to the electricity price changes by the utility via modulating their own demand locally (as proposed in [38]). The tenants cost objectives are not aligned with those of the data center, again because of the gap in pricing schemes. For example, as in Figure 2, tenant 1 will focus on minimizing its own VM usage regardless of the aggregate peak power consumption experienced by the data center. Consequently, the data center's DR, which might place a high emphasis on reducing the aggregate peak power, may suppress its power demand in ways or at times that are not the best choices for tenant 1.

Sustainability: Finally, the larger energy consumption and higher peak power may be incurred because of this gap, with negative implications for the sustainability of both the data center and the grid. By removing the gap between cloud pricing and electric utility pricing, the data center and the tenants can collaborate and modulate their demands together to achieve better sustainability.

Our Proposal: Virtualize the Electric Utility: To bridge this gap between cloud and electric utility pricing, we propose to enhance the interface exposed to the tenant to also include (in addition to the virtualized IT resources it already contains) a *virtual electric utility* (VEU). Our focus is on "big"/long-lasting tenants (e.g., a large e-commerce site hosted on a public cloud that has its own clients that are effectively "fleeting"/short-lived tenants of the cloud) of the cloud. The cloud provider would create the VEU for its tenants with the dual goals of (i) providing enough information to tenants about the actual energy-related costs and their own contribution to it, and (ii) letting the tenants play a part in controlling their own energy-related costs via incorporating information offered by the VEU into their own decision-making. Evaluating the efficacy of this idea would require us to investigate two main areas. The first area, which we discuss in Section 2, is concerned with how a cloud provider might construct an effective VEU abstraction and involves issues such as: what kind of pricing design would achieve fairness and other desirable behaviors, how should the VEU design be informed by the cloud provider's understanding of the tenants' workloads and modulation behavior, how should a VEU meter be designed to help the tenants understand their individual VEU and make proper decisions, and what role should the tenants have in negotiating the pricing scheme to be used by their VEU? The second area, which we explore in Section 3, is concerned with tenant operation in a data center that exposes the VEU interface.

Figure 3 demonstrates the key ideas in a cloud exposing the VEU interface to its tenants. Different from Figure 2 wherein the tenants are only exposed an IT-based interface (possibly combined with energy charge implicitly), here the tenants are offered a combination of IT-based and VEU-based interfaces and apply appropriate control knobs to optimize total costs. With the help of VEUs, the DR done by the cloud along with the participation of tenants may improve the cost-efficacy of both the cloud and the tenants. This is the central hypothesis underlying our work.

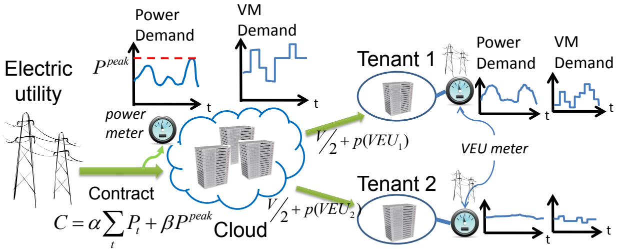


Figure 3: An overview of the key ideas in a cloud exposing VEUs to its tenants. ($p(VEU_i)$: VEU cost associated with tenant i.)

2 Designing Effective VEUs

In this section, we discuss two main issues related to a cloud provider's design of VEUs for its tenants. First, what kind of pricing design for VEUs might allow the provider to harmonize its charging of tenants with the electric utility costs it incurs? Second, what kind of systems infrastructure and tools might the provider need to effectively implement VEUs?

A Preliminary VEU Pricing Design: We present a simple game-based[3] VEU pricing mechanism design which improves the "unfairness" described in Section 1. Let κ_i and s_i be the expected energy consumption and the standard deviation of consumption of tenant i over a given interval of time, respectively. Denote as $\underline{\kappa}$ and $\underline{s}$ vectors of κ_i and s_i. Define $S^2 = \sum_i s_i^2$ as the variance of the aggregate energy consumption of the data center. Consider per-unit pricing policies of the form

$$p_i(\underline{\kappa}) \quad = \quad \alpha + \beta + \beta 2 S \kappa_i^{-1} g(\kappa_i, s_i) / \sum_j g(\kappa_j, s_j)$$

wherein $g(\kappa, s) = (\frac{\kappa}{\kappa_{max}})^{\gamma(s - s_{\max})}$, $\infty > s_{\max} > \max_i \max_{\kappa_i} s_i(\kappa_i)$, $\infty > \kappa_{\max} > \max_i \kappa_i$ and $\gamma > 0$.

In this game, tenant i will optimize net utility

$$\max_{\kappa_i} v_i(\underline{\kappa}) \quad = \quad u_i(\kappa_i) - p_i(\underline{\kappa}) \kappa_i, \qquad (1)$$

where the utility u_i is increasing, concave and bounded. For example, $u(\kappa) = u_{\max}(1 - (1 + \kappa/a)^{-1})$ where $a > 0$.

As a special case illustrating the merits of the proposed pricing design, consider two identical tenants (as in Figure 3) $i \in \{1, 2\}$ except fixed $s_1 > s_2$. Solving the first order necessary conditions gives (for $\forall i$)

$$\frac{u'(\kappa_i^*) - \alpha - \beta}{(s_i - s_{\max}) \kappa_2^*} \quad = \quad \frac{2 \beta S \gamma \prod_j g(\kappa_j^*, s_j)(\kappa_j^*)^{-1}}{(\sum_j g(\kappa_j^*, s_j))^2}$$

[3] For simplicity we omit the proof for existence and uniqueness of Nash equilibrium.

To see the impact of this proposed pricing design on the "unfairness" pointed our earlier, consider the case where $s_1 \approx s_2$ (but $s_1 > s_2$), and that the parameter a is taken sufficiently small so that $\kappa u'(\kappa)$ and $\kappa(u'(\kappa) - \alpha - \beta)$ are decreasing in κ. In this case, $\kappa_1^* < \kappa_2^*$ when $s_1 > s_2$ and $s_1 \approx s_2$, which achieves "fairness" since tenant 1 who contributes more to the aggregate peak (because of higher s_1) will be charged more when the game converges to the Nash equilibrium and therefore has to modulate its own demand to a lower level (κ_1)[4].

Note that here p_i is increasing w.r.t. s_i at the Nash equilibrium, which is desirable since the tenant with higher demand variance will experience a higher price.

Challenges and Ideas: We identify the following areas of work related to VEU pricing design:

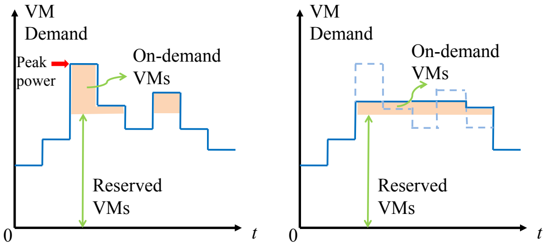
- In addition to our desire that the VEU pricing design help alleviate the unfairness problem that we identified in Section 1, we may also want it to offer certain other appealing characteristics. One example is that the price per unit demand is decreasing in consumption (i.e., $p_i(\underline{\kappa}, \underline{s})$ is decreasing in κ_i), which may be valuable in capturing volume discounts (similar to those offered for IT resources to consumers with large demands - either explicitly by cloud providers [1] or realized implicitly by such consumers via using cheaper "reserved" instances [2]).
- A key simplification inherent in the above model is that it deals with decision-making during a single time interval. One key direction along which the VEU pricing design must be enhanced concerns capturing the tenants' ability to modulate their workloads over multiple time slots. One useful idea would be to extend our game-based formulation to accommodate tenant workload modulation as developed in Section 3.
- A fundamental question concerns the "structure" of the VEU pricing scheme: Instead of the scalar VEU price explored above, should VEU prices resemble those exposed by the (actual) electric utility? For example, might a data center subjected to peak pricing design its VEUs to also expose peak pricing to the tenants? If this were to be the case, might the data center then offer choices for energy demand charge and peak power charge that tenants could select from? For our running example of two tenants with identical average demands but different variances, it is intuitively clear that the tenant with a lower variance would be willing to pay a higher charge for peak power if that allowed it to choose a lower energy demand charge.

Systems Software and Tools: The key enabling facility a cloud provider would need to actually implement VEUs (spanning various choices represented by the discussion above), is similar to what has been called "energy accounting" in the literature [5, 21]. The data center may want to provide its tenants with facilities (a.k.a. VEU meter) that can be used to infer their local power consumption and estimate the individual VEU, which can be done based on inference techniques that can translate easily to measure numbers like utilization into accurate power consumption employing well-designed models. One particular challenge is attributing the energy consumed by shared components and indirectly exercised components. We will build upon related work as well as our own prior experience on vPath for this [33, 34].

3 Tenant Operation with a VEU

Next, we consider the following question: how might tenants operate with this newer interface?

Novel Resource Allocation Problems: With a VEU interface exposed by the data center, tenants will face fundamentally novel resource allocation problems compared to those addressed in the current literature because they will optimize costs that are summations of IT costs and VEU bills. Figure 4 shows a simple example of how the control problems with VEU become different from and more challenging than existing ones in the literature.



(a) (b)

Figure 4: A simple example of how the control problems become different from existing ones.)

Suppose one tenant wants to procure VMs from a cloud provider which exposes the options of both on-demand and reservation-based VM pricing schemes (e.g., Amazon EC2 [1]) to tenants. As in Figure 4(a), the **typical decision making (without VEU)** involves finding a threshold based on the cost-benefit trade-offs between on-demand and reservation prices: any VM demand below the threshold can be fulfilled by reserved VM; the rest will be obtained as on-demand instances.

However, if the same tenant is charged by the cloud based on both VEU and IT resource (VMs in this example) prices, it might find completely different control decisions to be cost-optimal. Suppose that the VEU interface is designed to be the same as the peak-based electricity pricing. For example, besides the VM costs, the tenant might also be charged for its peak power and total energy consumption. Assume that the tenant's power consumption is proportional to the number of VMs be-

[4] Here we choose to modulate power demand via load-shedding; other control knobs based on load-shifting will be considered in our future work.

ing used. If this tenant is running batch workload that can tolerate delay to some extent, it might want to postpone part of its VM procurement during the peak time and smooth the power peak, as shown in Figure 4(b).

Whereas consumers solving such problems will likely enable better DR for the data center (assuming appropriate pricing design and VEU implementation), they will require **additional complexity** at the tenant: (i) solving more complex stochastic optimization problems and (ii) having to reason about predicting/converting resource allocations into power consumption.

Implementation Considerations and Enabling Systems Software: We believe that (i) can be addressed using approaches similar to those we have been pursuing. In our prior work [38], given the computational complexity of optimizing power cost with various existing control knobs all together, we propose to modulate the power demand via abstract knobs: demand dropping and demand delaying, based on the observation that many of the real knobs are either dropping demand or delaying demand or both. This abstraction can be combined together with the existing IT resource-based control techniques for the tenants to make better decisions under the VEU interface. For example, as in Figure 4(b), a tenant running delay tolerant MapReduce workload could employ demand delaying; whereas a tenant running delay sensitive web search engine might exploit demand dropping to shave the peak power (e.g., partial execution [44]).

Regarding (ii), we can leverage variety of related work. One such example is "vPower" [39], a software system which allows applications to specify their power needs and dynamically monitors power usage. Other related work that might help in converting resource allocation into power consumption can be found in Section 4. However, these works cannot provide sufficient VEU-related information to tenants to help them make cost-effective decisions as discussed in Section 4.

4 Related Work and Other Discussion

Reducing Energy-related Costs: A large body of work exists on reducing data center energy costs and we may view it as being of two types. The "first line of attack" on these costs has been based on techniques to reduce the (raw) energy consumption of the data centers by improving the power proportionality [35] of hardware/software or reducing sources of energy wastage in IT or non-IT infrastructure (thereby improving the data center's "PUE" [6]). Some salient examples of such work include the use of IT equipment capacity modulation/shutdown (e.g., CPU [14], memory and disks [41, 37], entire servers [13, 22], software re-design [7], improvements in the design and operation of the cooling system [23], and combinations of these options [31]. Since these approaches are based on reducing energy consumption, their positive impact on sustainability is straightforward to appreciate.

A second (and more indirect/subtle) line of work involves employing more general DR within data centers in response to pricing complexities discussed in Section 1 (e.g., coincident peaks [25], peak pricing [38], time-varying prices [36], etc.) and also to avail of cost benefits offered by such "supply-side" options as local generation (including renewables) [11, 32, 15] or such offerings from the grid as ancillary services [9], carbon credits [32], etc. This body of work has explored the use of a variety of demand-side control knobs such as DVFS [8], geographical load balancing [24, 29], partial execution [44], energy storage within the data center [40, 17], etc.

We consider VEUs to be complimentary to this entire body of work: *VEUs can serve as a mechanism for propagating the energy/cost benefits offered by the above techniques to individual tenants fairly.* Furthermore, as described in Section 3, many optimization and control techniques developed in these works can be serve as starting points for a tenant using a VEU for devising its own cost-aware resource allocation.

Alternate Approaches based on Work that "Virtualizes" Power: One line of work in the literature worth comparing our VEU-based approach against is based on some recent proposals for "virtualizing power" in the data center. This idea of virtualizing power in the data center is not new. Briefly, these proposals argue for treating energy as a first-class resource (at par with IT resources such as the CPU or entire servers, etc.) which can be directly/explicitly controlled by individual applications. Data centers based on these proposals track/estimate the power consumed by individual applications, make this information available to them via software calls, and even let applications control their energy usage. Salient examples of such proposals are "VirtualPower" [26], "energy containers" [5], and "palloc" [39].

Given this line of work, one might ask if replacing the current IT-based interface/pricing offered to tenants with a completely (virtual) energy-based interface would be a good idea. The key distinction between our work and these ideas is that *we propose to virtualize not just power but the electric utility itself.* By allowing a tenant to interact with (what it sees as) its *own* electric utility, we allow it to negotiate its own energy pricing with the cloud and carry out subsequent resource allocation that is aware of the energy-related costs it will incur (as dictated by this negotiated pricing scheme). In fact, these techniques are complimentary to VEUs because our implementation can benefit from their ideas related to application-level energy consumption monitoring and accounting. However, they do not focus on bridging the gap between electric utility pricing and cloud pricing, and the fairness and efficiency related problems resulting from it.

References

[1] Amazon EC2 Pricing, 2014. http://aws.amazon.com/ec2/pricing/#on-demand.

[2] Windows Azure Pricing, 2014. https://www.windowsazure.com/en-us/pricing/overview/.

[3] Barroso, L. A., and Hölzle, U. The datacenter as a computer: An introduction to the design of warehouse-scale machines. *Synthesis lectures on computer architecture 4*, 1 (2009), 1–108.

[4] Barroso, L. A., and Hölzle, U. *The Datacenter as a Computer: An Introduction to the Design of Warehouse-Scale Machines*. Synthesis Lectures on Computer Architecture. Morgan & Claypool Publishers, 2009.

[5] Bellosa, F. The Benefits of Event-Driven Energy Accounting in Power-sensitive Systems. In *Proceedings of the 9th Workshop on ACM SIGOPS European Workshop: Beyond the PC: New Challenges for the Operating System* (New York, NY, USA, 2000), EW 9, ACM, pp. 37–42.

[6] Bemis, P., and Marshall, L. Improving data center pue through airflow management. *Applied Math Modeling Inc., Concord, NH, USA* (2010).

[7] Chen, Y., Alspaugh, S., Borthakur, D., and Katz, R. H. Energy efficiency for large-scale mapreduce workloads with significant interactive analysis. In *EuroSys* (2012), P. Felber, F. Bellosa, and H. Bos, Eds., ACM, pp. 43–56.

[8] Chen, Y., Das, A., Qin, W., Sivasubramaniam, A., Wang, Q., and Gautam, N. Managing server energy and operational costs in hosting centers. In *Proceedings of the 2005 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems* (New York, NY, USA, 2005), SIGMETRICS '05, ACM, pp. 303–314.

[9] Chiu, D., Stewart, C., and McManus, B. Electric grid balancing through lowcost workload migration. *SIGMETRICS Perform. Eval. Rev. 40*, 3 (Jan. 2012), 48–52.

[10] Fort Collins Coincident Peak, 2013. http://www.fcgov.com/utilities/business/rates/electric/coincident-peak.

[11] Deng, N., Stewart, C., Gmach, D., Arlitt, M., and Kelley, J. Adaptive green hosting. In *Proceedings of the 9th International Conference on Autonomic Computing* (New York, NY, USA, 2012), ICAC '12, ACM, pp. 135–144.

[12] Duke utility bill tariff, 2012. http://www.considerthecarolinas.com/pdfs/scscheduleopt.pdf.

[13] Gandhi, A., Chen, Y., Gmach, D., Arlitt, M., and Marwah, M. Minimizing data center sla violations and power consumption via hybrid resource provisioning. In *Proceedings of the 2011 International Green Computing Conference and Workshops* (Washington, DC, USA, 2011), IGCC '11, IEEE Computer Society, pp. 1–8.

[14] Gandhi, A., Harchol-Balter, M., and Kozuch, M. A. Are sleep states effective in data centers? In *IGCC* (2012), pp. 1–10.

[15] Goiri, I., Katsak, W. A., Le, K., Nguyen, T. D., and Bianchini, R. Parasol and greenswitch: managing datacenters powered by renewable energy. In *ASPLOS* (2013), pp. 51–64.

[16] Google Cloud Pricing, 2014. https://cloud.google.com/products/cloud-storage/.

[17] Govindan, S., Wang, D., Sivasubramaniam, A., and Urgaonkar, B. Leveraging stored energy for handling power emergencies in aggressively provisioned datacenters. In *ACM ASPLOS* (Mar 2012), pp. 75–86.

[18] Hamilton, J. Internet-scale service infrastructure efficiency. *SIGARCH Comput. Archit. News 37*, 3 (June 2009), 232–232.

[19] James Hamilton's Blog, 2014. http://perspectives.mvdirona.com/.

[20] INTERNATIONAL ENERGY OUTLOOK 2013, July 2013. http://www.eia.gov/forecasts/ieo/world.cfm.

[21] Krishnan, B., Amur, H., Gavrilovska, A., and Schwan, K. VM Power Metering: Feasibility and Challenges. *SIGMETRICS Performance Evaluation Review 38*, 3 (2010), 56–60.

[22] Lin, M., Wierman, A., Andrew, L., and Thereska, E. Dynamic right-sizing for power-proportional data centers. In *INFOCOM, 2011 Proceedings IEEE* (April 2011), pp. 1098–1106.

[23] Liu, Z., Chen, Y., Bash, C., Wierman, A., Gmach, D., Wang, Z., Marwah, M., and Hyser, C. Renewable and cooling aware workload management for sustainable data centers. In *Proceedings of the 12th ACM SIGMETRICS/PERFORMANCE Joint International Conference on Measurement and Modeling of Computer Systems* (New York, NY, USA, 2012), SIGMETRICS '12, ACM, pp. 175–186.

[24] Liu, Z., Lin, M., Wierman, A., Low, S. H., and Andrew, L. L. Greening geographical load balancing. In *Proceedings of the ACM SIGMETRICS joint international conference on Measurement and modeling of computer systems* (New York, NY, US-A, 2011), SIGMETRICS '11, ACM, pp. 233–244.

[25] Liu, Z., Wierman, A., Chen, Y., and Razon, B. Data center demand response: Avoiding the coincident peak via workload shifting and local generation. In *ACM SIGMETRICS* (Jun 2013).

[26] Nathuji, R., and Schwan, K. Virtualpower: Coordinated power management in virtualized enterprise systems. In *Proceedings of Twenty-first ACM SIGOPS Symposium on Operating Systems Principles* (New York, NY, USA, 2007), SOSP '07, ACM, pp. 265–278.

[27] Ontario Energy Board: Tired Pricing, 2013. http://www.ontarioenergyboard.ca/OEB/Consumers/Electricity/Electricity+Prices#tiered.

[28] Patterson, M. K., Costello, D., Grimm, P., and Loeffler, M. Data center tco; a comparison of high-density and low-density spaces. *Thermal Challenges in Next Generation Electronic Systems (THERMES 2007)* (2007).

[29] Qureshi, A., Weber, R., Balakrishnan, H., Guttag, J. V., and Maggs, B. M. Cutting the electric bill for internet-scale systems. In *SIGCOMM* (2009), pp. 123–134.

[30] Rackspace Public Cloud Pricing, 2014. http://www.rackspace.com/cloud/public-pricing/.

[31] Raghavendra, R., Ranganathan, P., Talwar, V., Wang, Z., and Zhu, X. No "power" struggles: Coordinated multi-level power management for the data center. In *Proceedings of the 13th International Conference on Architectural Support for Programming Languages and Operating Systems* (New York, NY, USA, 2008), ASPLOS XIII, ACM, pp. 48–59.

[32] Ren, C., Wang, D., Urgaonkar, B., and Sivasubramaniam, A. Carbon-aware energy capacity planning for datacenters. In *MASCOTS* (2012), IEEE Computer Society, pp. 391–400.

[33] Tak, B. C., Kwon, Y., and Urgaonkar, B. Towards An Effective and General Resource Accounting and Control Framework in Consolidated IT Platforms. In *Proceedings of the Seventh Workshop on Large-Scale Distributed Systems and Middleware, (LADIS 2013), colocated with ACM SOSP* (Nov. 2013).

[34] Tak, B. C., Tang, C., Zhang, C., Govindan, S., Urgaonkar, B., and Chang, R. N. vPath: Precise Discovery of Request Processing Paths from Black-box Observations of

Thread and Network Activities. In *Proceedings of the 2009 Conference on USENIX Annual Technical Conference* (Berkeley, CA, USA, 2009), USENIX'09, USENIX Association, pp. 19–19.

[35] Thereska, E., Donnelly, A., and Narayanan, D. Sierra: practical power-proportionality for data center storage. In *EuroSys* (2011), pp. 169–182.

[36] Urgaonkar, R., Urgaonkar, B., Neely, M. J., and Sivasubramaniam, A. Optimal power cost management using stored energy in data centers. In *ACM SIGMETRICS* (Jun 2011), pp. 221–232.

[37] Verma, A., Koller, R., Useche, L., and Rangaswami, R. Srcmap: Energy proportional storage using dynamic consolidation. In *Proceedings of the 8th USENIX Conference on File and Storage Technologies* (Berkeley, CA, USA, 2010), FAST'10, USENIX Association, pp. 20–20.

[38] Wang, C., Urgaonkar, B., Wang, Q., and Kesidis, G. A hierarchical demand response framework for data center power cost optimization under real-world electricity pricing. In *Proceedings of the IEEE International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems (MASCOTS 14)* (2014), IEEE Computer Society.

[39] Wang, D., Ren, C., and Sivasubramaniam, A. Virtualizing power distribution in datacenters. In *Proceedings of the 40th Annual International Symposium on Computer Architecture* (New York, NY, USA, 2013), ISCA '13, ACM, pp. 595–606.

[40] Wang, D., Ren, C., Sivasubramaniam, A., Urgaonkar, B., and Fathy, H. K. Energy storage in datacenters: What, where and how much? In *ACM SIGMETRICS* (Jun 2012).

[41] Wang, J., Yao, X., and Zhu, H. Exploiting in-memory and on-disk redundancy to conserve energy in storage systems. *Computers, IEEE Transactions on 57*, 6 (June 2008), 733–747.

[42] Wang, W., Li, B., and Liang, B. To reserve or not to reserve: Optimal online multi-instance acquisition in iaas clouds. In *ICAC* (San Jose, California, USA, 2013), pp. 13–22.

[43] Wang, W., Niu, D., Li, B., and Liang, B. Dynamic cloud resource reservation via cloud brokerage. In *ICDCS* (2013), pp. 400–409.

[44] Xu, H., and Li, B. Reducing Electricity Demand Charge for Data Centers with Partial Execution. *ArXiv e-prints* (July 2013).

[45] Zafer, M., Song, Y., and Lee, K.-W. Optimal bids for spot vms in a cloud for deadline constrained jobs. In *Proceedings of the 2012 IEEE Fifth International Conference on Cloud Computing* (Washington, DC, USA, 2012), CLOUD '12, IEEE Computer Society, pp. 75–82.