Saving on Data Center Energy Bills with EDEALS: Electricity Demand-response Easy Adjusted Load Shifting

Will McFadden
University of Chicago

Anita Nikolich
Morgridge Institute

Ray Parpart
University of Chicago

Dr. Birali Runesha
University of Chicago

Abstract

Energy demand response presents a highly cost-effective means to improve the sustainability of data centers whenever there is flexibility in task scheduling. This paper presents an empirical study in the area of data center demand response, with the goal of cost savings on electricity bills for small to medium size data centers drawing 1-5 MegaWatts. Using the SLURM resource manager, we demonstrate a methodology for energy aware load shifting by flexibly reducing compute cycles at times of peak energy demand. Simply reducing the pool of available servers during a brief period of peak energy demand results in tangible cost savings through reduced power consumption of the server cluster, with minimal performance degradation to users. We have developed a data processing pipeline, EDEALS, to determine the potential cost savings of partial data center shutdown to enable demand-response load shifting. As our baseline, we measured the power draw and job scheduling delay of a small-scale test cluster by varying available resources. We then model a production cluster's performance in response to a realistic energy constraint imposed by a utility provider. For our data center, we quantify a potential annual cost savings on electricity of 7% while only causing 16 hours of total increased wait time (0.1%) throughout the entire year.

1 Introduction

Data centers in the US consume an estimated 91 billion kilowatt-hours yearly, equivalent to the annual output of 34 large coal-fired power plants.[2] These same estimates show that only 6-12% of the electricity is used for powering servers while the rest is used to keep machines idling, wasting resources and money in the process. Data center electricity is not inexpensive, costing American businesses $13 billion annually in electricity bills.[2] Because cost is a strong motivating factor for businesses and universities, we consider data center energy efficiency in the context of cost savings for data center operations.

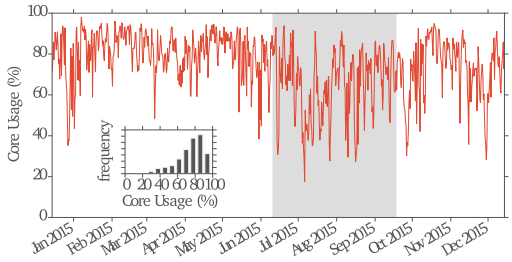

Figure 1: Average core usage for a 244 node shared HPC partition in the Midway cluster. Note that the period of peak energy demand (gray background) coincides with a period of reduced core usage. Insert shows usage statistics histogram.

Demand response (DR) programs provide incentives to induce dynamic management of customers' electricity load in response to power supply conditions, for example, reducing power consumption in response to a request from the utility.[11] Whereas the benefits of demand response programs had previously been focused solely on price reduction[10], the value of demand response towards sustainability and carbon footprint reduction is becoming more apparent with the introduction of the 2015 Environmental Protection Agency (EPA) Clean Power Plan[5]. In an estimate prepared for the Advanced Energy Management Alliance, DR programs could provide as much as a 1% reduction in greenhouse gas emissions through direct and indirect mechanisms[8].

In one of the simplest DR scenarios, many energy providers have Voluntary Load Response (VLR) programs, which encourage commercial consumers to reduce power demands during peak periods, such as partic-

ularly hot summer days, in exchange for electricity supply rebates. We are interested in exploring more active ways for university data centers to participate in VLR programs while minimally impacting user experience.

In many university (typically categorized as Tier 1) data centers, a significant portion of the data center is dedicated to high performance research computing (HPC). While these tasks often take longer periods of time to complete, they are less time sensitive and more flexible than systems which support core business functions such as the university's email. We wish to use the flexibility in scheduling of these jobs to reduce energy consumption of university data centers during periods of peak demand by shifting the load to off-peak periods. While university HPC is our primary focus, more generally, any computing tasks that can be queued will be amenable to this kind of load shifting.

As shown in the example core usage data of Figure 1, although the typical average usage during the school year is a fairly standard 80%, the averaged workload can fall to 65% of full capacity in the hottest summer months from June to September. These months also present the period of greatest electricity demand due largely to increased need for cooling or air conditioning. This presents a valuable opportunity to potentially curtail electricity use in demand response scenarios by shifting loads off the peak periods of energy price. Toward this aim, this paper is our presentation of EDEALS, a data processing system to estimate the economic savings, feasibility, and any potential user impact from cluster shutdown during periods of increased energy demand.

1.1 Alternative Demand Response Options in Data Centers

Although we focus on load shifting for our study, we wish to point out prior work on alternative strategies that may be of relevance for demand response.

Facility changes A study by Lawrence Berkeley National Laboratory (LBNL) found that 5% of the data center load can typically be shed in 5 minutes and 10% of the load can be shed in 15 minutes without changes to how the IT workload is handled, i.e., via temperature adjustment and other building management approaches[4]. Most data centers have local power due to a backup generator, which could also be used to absorb some load during peak time [6]. More recently, methods of energy storage have been proposed[7] in which UPS batteries are re-purposed for provisioning during periods of peak demand in addition to their primary purpose of backup power. However, these methods all entail manual intervention, with close monitoring and control.

Power capping is a strategy by which to run data center equipment within a set of constraints which assume the electricity draw for the data center as a whole cannot grow any larger. Some examples of this include and turning off or constraining CPU/GPU power consumption to values below the CPU Thermal Design Power (TDP) value, which requires less voltage. Many equipment manufacturers - including IBM, Intel, and AMD - have implemented power capping technology that can be monitored at the processor level and applied at the rack level. One approach to power capping is Dynamic Voltage/Frequency Scaling (DVFS). However, as noted by Roundtree[9], no machine in the Top 500 list of supercomputers makes use of DVFS to save power or energy since the performance impact and the amount of power and energy saved was highly application dependent. Power capping doesn't necessarily equate to energy efficiency nor cost savings.

Schedulers Zhou et al[13] present a method for power-aware scheduling by using a combination of a scheduling window and 0-1 knapsack model, which shows promise. However, since SLURM is the workload manager installed on our HPC cluster and test machine, this paper focuses solely on SLURM. Bodas et al[1] demonstrate an integration of power capping into a power-aware scheduler, with the overall goal of maintaining average system power within a budget. Their work demonstrates that SLURM's auto mode can be used to maximize available power.

Server overprovisioning By overprovisioning the amount of servers, one can reduce the load and temperature on each server by utilizing only a subset of active servers, with the rest in standby mode, in order to reduce the idle power. Ahmad et al.[3] proposed to reduce the sum of the cooling power and idle power by trading off idle power and cooling power for each other. Similarly, Liu et al.[12] propose geographical load balancing for massive, distributed, Internet-scale systems, in which route to areas where green energy is available. If one can redesign a data center such that the active servers are geographically distributed throughout it, this is a sound approach. However, high end compute clusters are often in a concentrated physical location within the data center. Since HPC compute clusters are not distributed, the latter approach is not feasible.

2 Problem Statement

Can load shifting of high performance computing tasks save universities money in energy demand response scenarios? To explore the relative costs of implementing load shifting in response to surges in energy demand, we have expressed the problem by modeling total dollar cost. We wish to use this framework to explore the optimization of cost in the presence of various data center

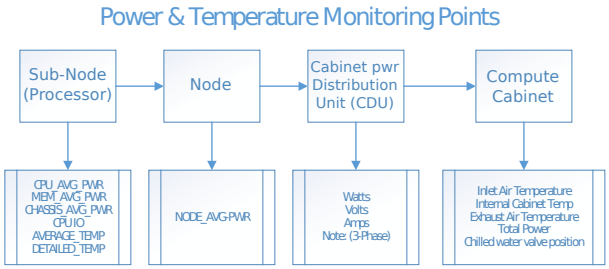

Figure 2: Energy measurement systems in EDEALS.

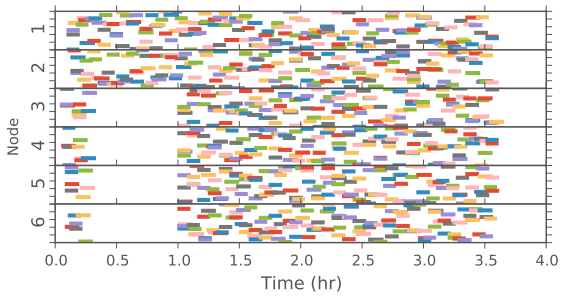

Figure 3: Diagram of job scheduling during a four node temporary shutdown experiment. Each colored rectangle displays the execution of a single LAMPPS test job running for approximately 5 minutes.

usage statistics and price fluctuation schemes.

2.1 Modeling Energy Costs

We generate a model total cost function composed of a fixed cost for purchasing and maintaining nodes plus a variable cost dependent on data center power usage and energy prices. We wish to minimize the cost function

$$C = p_n T n_{max} + \int_0^T dt \cdot p(t) \left(n(t)u(t) + u_w \frac{\Delta n(t)}{\Delta t} \right)$$

where $p(t)$ is the price of power at time t, $0 < n(t) < n_{max}$ is the number of running nodes, $u(t)$ is the average node power usage, u_w is the wasted power from turning on a node, p_n is the amortized lifetime cost of purchasing a node, and n_{max} is the total number of nodes in the cluster.

Based on our cluster usage statistics, we approximate that compute cycles are roughly interchangeable and that the main determiner of power usage is simply the CPU utilization of the node. In this case, node power usage takes the form

$$u(t) = u_0 + u_v \cdot r(t)$$

where $0 < r(t) < 1$ is the fraction of CPU usage, u_0 is the cost of an idling node and u_v is the variable cost for doing r work on a machine.

We wish to minimize the cost function C subject to the constraint that the sum of the submitted CPU cycles, S, are all completed after a period T.

$$\int_0^T dt \cdot n(t)r(t) = S$$

2.2 Response to a Temporary Price Spike

In particular, we wish to use this framework to determine how to run our data center in the situation where every T days, we see a "price spike" from p_0 to p_s, lasting time period t_s. This condition is highly similar to those facility managers face when Utilities provide impose usage tariffs during peak energy demand periods.

In this situation, the number of running machines will change stepwise between a high number of running machines, $n_H = n_{max}$, and a low number of running machines, n_L, and a high and low CPU utilization $r_H = 1$, r_L, with a corresponding u_H and u_L as defined above. The high usage will occur during the normal energy costs, and the low usage will occur during the price spike. Therefore we can rewrite our cost function as

$$\begin{aligned} C = {} & p_n n_H T + p_0(u_0 + u_v) n_H (T - t_s) \\ & + p_s(u_0 + u_v r_L) n_L t_s + p_0 u_w (n_H - n_L) \end{aligned} \quad (1)$$

with the constraint

$$n_H(T - t_s) + n_L r_L t_s = S \quad (2)$$

Inserting the constraint into our cost function to replace r_L yields

$$\begin{aligned} C = {} & p_s u_v S + n_L \cdot (p_s u_0 t_s - p_0 u_w t_w) \\ & + n_H \cdot (p_n T + p_0 u_w t_w - (\Delta p u_v - p_0 u_0)(T - t_s)) \end{aligned} \quad (3)$$

where we have introduced the price difference, $\Delta p = p_s - p_0$.

We can analyze the change in costs as a function of n_L and n_H to determine the optimal cluster setup for known variables, t_s, p_s, p_n, p_0, u_0, u_v, and u_w.

From this analysis, whenever $p_s u_0 t_s < p_0 u_w t_w$, the cost of powering off nodes exceeds the cost of running those nodes idle so $n_L = n_H$ and $r_L = S/n_H t_s - (T - t_s)/t_s$. Otherwise powering off nodes saves money so the nodes that remain on run at full capacity $r_L = 1$, and n_L is minimized subject to constraints giving $n_L = S/t_s - n_H(T - t_s)/t_s$.

If we can freely choose n_H to optimize cost, then whenever $(\Delta p u_v - p_0 u_0)(T - t_s) >$

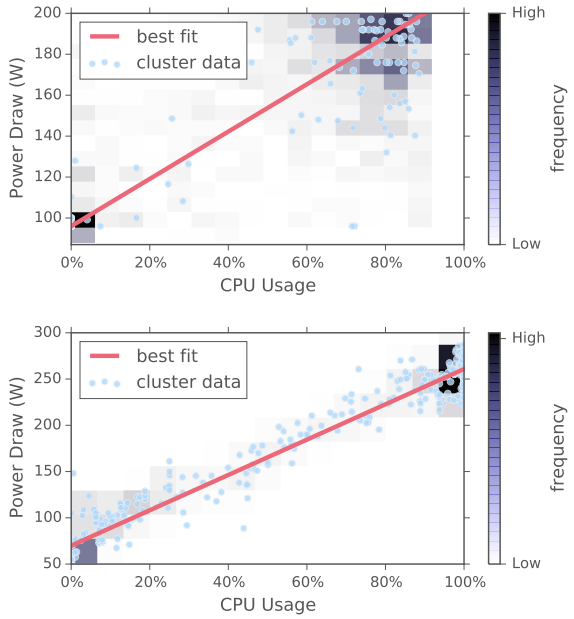

Figure 4: Power data for test cluster (top) and production cluster (bottom) nodes in presence of variable usage. The slope and intercept of the line are used to determine u_v and u_0 respectively.

$p_n T + min(p_0 u_w t_w, p_s u_0 t_s)$, we would increase n_H (i.e. buy more machines) until all the work is done during the normal energy period. Therefore $n_H = S/(T - t_s)$ and either n_L or r_L is 0. Otherwise, the cost of new machines is more than any cost savings achieved from exploiting the price difference, and we would simply ignore the price spike (i.e. set $n_H = n_L = S/T$ and $r_L = 1$).

3 EDEALS: Electricity Demand-response Easy Adjusted Load Shifting

For a data center manager to use the above model to determine their cost savings, they must collect and analyze usage and power data on their system. We have built a cluster data processing pipeline, EDEALS, to assess the magnitude of potential savings available when varying available computer resources. We combine SLURM job scheduling, node level IMM power and usage metrics, and cabinet level CDU measurements to determine the optimum magnitude of demand response cluster shutdowns.

Here we describe our data center instrumentation, so that we ensure accurate measurements of performance of the workload management system and HPC cluster alone without the influence of extraneous components.

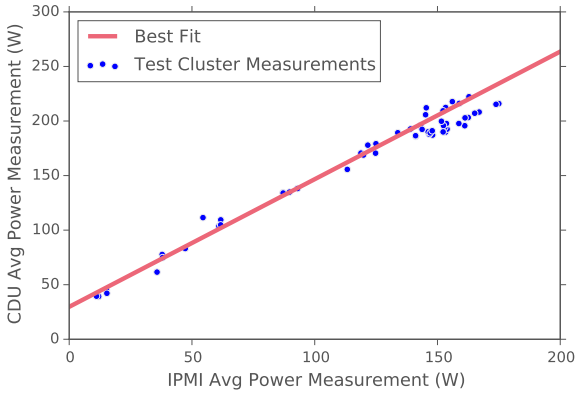

Figure 5: Comparison between node level IPMI measurements and rack level CDU measurements. Best fit shows the model relationship used to convert IPMI data to estimated total power draw.

Since our focus is the HPC cluster and SLURM workload manager, we need to ensure those components alone affect the reduced data center utility bill. As depicted in Figure 2, we take measurements at the core, node, rack, and cabinet level. These data points are combined to detect power losses at each step and to determine the correlation between the power measurements at the machine level and the true power draw at the facility level.

Combining this data with electricity pricing statistics from utility managers allows system administrators to determine when and by how much to reduce their power usage to save money. We have built a set of scripts particular to our system to implement machine level power down in response to predicted energy peaks. At the end of the peak energy period the machines automatically reboot and are added back to SLURM's available server pool. Currently, these power cycling scripts are manually executed by system administrators after evaluation of the likelihood of near-term energy demand peaks. However, as more data centers begin to implement smart metering, it will become possible to automate load shifting in response to real-time energy pricing indicators. We look forward to continuing this as future work.

4 Small-Scale Evaluation of EDEALS

To test our load shifting scheme, we launched a series of small-scale experiments on a 6 node test cluster using SLURM batch management system to schedule jobs. We wished to compare the energy savings and job wait times during both a full and partial cluster shutdown in response to an energy price spike.

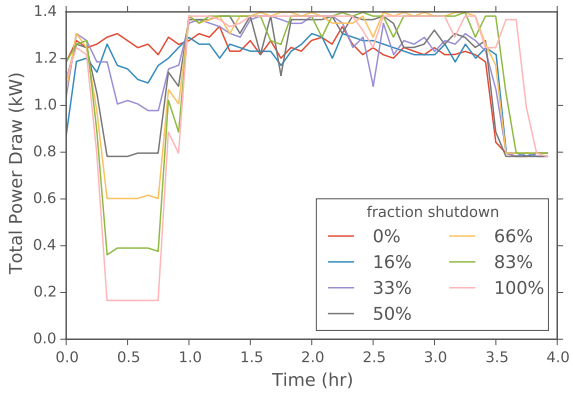

Figure 6: Total power consumed during experiments where variable numbers of machines were shut down during simulated peak pricing.

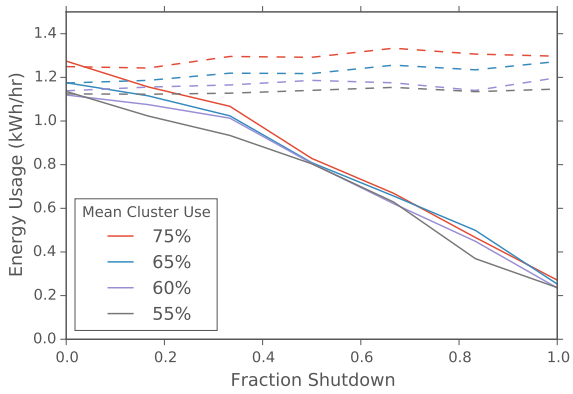

Figure 7: Energy usage of test cluster during partial shutdown experiments. Solid lines indicate power usage during the shutdown, while dashed lines indicate power usage during full operation.

4.1 Experimental Setup

We measured the total energy use over a 3.5 hour window, of which the first 30 minutes comprised a partial cluster shutdown, followed by a 15 minute power-up routine. We explored the impact of shutting down between one and all six nodes during the 30 minute window. The shutdown was carried out by fully powering off nodes. We compared this to the energy usage without the partial shutdown.

Identical sample jobs were submitted to the cluster via the SLURM scheduler at a constant rate to set the average cluster usage inn the range from 55% to 75% capacity. We used custom state control commands to set the power states of individual machines in the test cluster. The SLURM scheduler automatically shifted queued jobs to run on the available machines, as illustrated in the example job schedule of Figure 3 for a four node shutdown experiment. We used our EDEALS data analysis pipeline to measure the changes in energy usage and job wait time in the queue.

4.2 Evaluation of Model Parameters

Importantly, EDEALS allowed us to determine appropriate power parameters, u_0, u_v, and u_w for both our test rack as well as a larger partition of the University of Chicago's Midway production cluster. Figure 4 shows the measured relationship between CPU utilization and energy usage as determined from the machine level IPMI metric data.

To account for losses not measured at the IPMI level, we compare the sum IPMI power usage to the rack level power monitoring, as displayed in Figure 5. This comparison revealed a correction factor of 1.16 between the IPMI measurement and the total rack level energy draw as well as a phantom power draw of 30 W even when machines were in the off state. Although these effects may seem small we find it very important to add these corrections factors when evaluating true cost savings. Using this corrected model, we were able to predict power consumption at the CDU level via CPU utilization under variable scheduler loads.

4.3 Relative Energy Savings and Max Wait Times

Our test cluster provided us with an important baseline in determining the effectiveness of a partial shutdown in reducing energy usage. As shown in Figure 6, the total power draw from the test cluster was reduced dramatically during the shutdown period, and then returned to its baseline level.

These experiments were repeated with different job submission rates such that the average CPU usage varied from 55% to 75%. As shown in Figure 7, the partial shutdowns reduced the total energy usage as measured at the CDU level. Not surprisingly, the power usage during cluster shutdown for all usage levels converged to roughly the same value at the point where all remaining operational machines reached full capacity. Interestingly, the energy savings did not appear to be perfectly directly proportional to the fraction shut down. In particular, there was residual energy use associated with our machine's low power state even when the cluster was entirely shut down.

We also measured the difference between job submission and start time, as depicted in Figure 8. As one

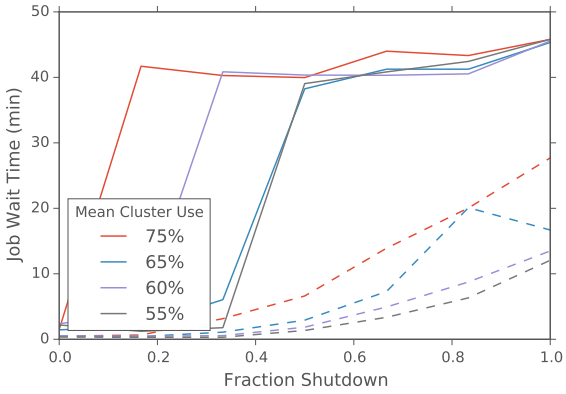

Figure 8: Maximum (solid) and mean (dashed) job wait times during partial shutdown experiments.

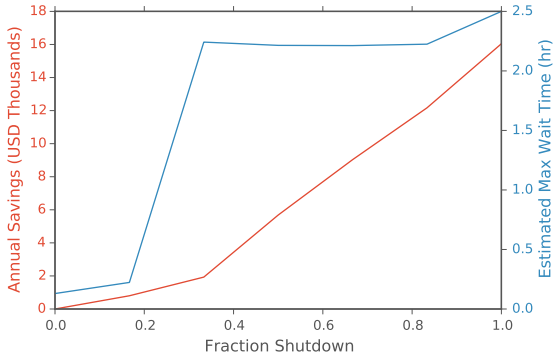

Figure 9: Estimated savings from partial cluster shutdowns.

T	t_s	p_0	p_s
360 hr	2 hr	\$0.03/kWh	\$6.5/kWh

Table 1: Model pricing parameters estimated for medium scale HPC data center.

would expect, both mean and max wait times increased as the shutdown fraction grew and the effect was more pronounced when the cluster usage was higher. However, we were pleasantly surprised to find that max wait times topped out at 45 minutes, which was the duration of the entire cluster down period. In addition, the job runtime was only increased by a maximum of 2%. These indicate that SLURM does not add much additional overhead, and therefore, the worst-case user wait times would not exceed the total period the cluster was shut down.

5 Conclusion: Implication for An Operational HPC Datacenter

In many data centers, the variable cost to supply electricity to a facility can be decomposed into both a nominal cost per kilowatt-hour and a procurement cost from the supplier. Some suppliers impose a substantial procurement tariff based on electricity usage during the five, two hour long periods of highest demand in a year. In this scenario, the savings of load shedding can be orders of magnitude higher than the nominal price per kilowatt-hour. Using historic data of electricity supply costs, we estimate that curtailing 1MW, 8 times per year, will lead to an annual electricity savings of approximately \$100K. For our facility, this corresponds to a total cost savings of roughly 7% annually. Approximating that the 8 curtailment periods are spread out over the 4 month period from June to September, we arrive at the system parameters listed in Table 1.

Combining this pricing data and the power usage measurements from our test cluster, we can extrapolate the yearly savings we expect from demand-response load shifting for our production HPC cluster. Assuming that we can reduce our energy use by a similar fraction across our 1000 node production cluster, we expect our price savings to be proportional to the energy reduction during the shutdown period. Here, we have corrected for differences in average power draw (i.e u_0 and u_v) between our test cluster hardware and our production cluster hardware to improve our estimate. We display this information in Figure 9, as a function of the fraction of the cluster that we would be willing to shut down. If we apply the same shutdown procedure to our 1000 node cluster for a total of 16 hours, we would generate a cost savings of approximately \$18K. More significantly, once this methodology is extended to more HPC clusters in the data center we could save \$100K annually, or 7% of our data center's energy consumption. In terms of the job scheduling impact, the wait time statistics from our test cluster shows that the worst-case impact on user wait-times will not be significantly larger than the curtailment period itself.

Based on our analysis, we believe that many university data centers could enact this type of load shifting curtailment strategy. Moreover, it's highly probable that any mixed use data center with a significant HPC or HTC workload can potentially reduce its annual electricity operating budget by several percent while having only minuscule impact on user experience. We hope that the use of our EDEALS system can help other medium scale data centers evaluate whether load shifting can be an economically viable avenue to increasing their overall sustainability.

Acknowledgments

Special thanks to Brandon Strader for all his work getting our test cluster set up as well as for his useful input on machine pricing information. We also wish to thank Matt Beach for his invaluable knowledge on energy pricing mechanisms and university power plans. And we'd like to thank Dr. Edwin Munro for his support while working on this project.

Availability

On our Github, we have provided all data collection scripts, analysis routines, and experimental setups, as well as detailed calculations and optimization methods for other price models.

`https://github.com/rcc-uchicago/edeals`

References

[1] BODAS, D., SONG, J., RAJAPPA, M., AND HOFFMAN, A. Simple power-aware scheduler to limit power consumption by hpc system within a budget. In *Proceedings of the 2Nd International Workshop on Energy Efficient Supercomputing* (Piscataway, NJ, USA, 2014), E2SC '14, IEEE Press, pp. 21–30.

[2] DELFORGE, P., AND WHITNEY, J. Data center efficiency assessment: Scaing up energy efficiency across the data center industry: Evaluating key drivers and barriers. Tech. rep., National Resources Defense Council, August 2014.

[3] F.AHMAD, AND VJAYKUMAR., T. Join optimization of idle and cooling power in data centers while maintaining response time. In *ASPLOS'10: Proceedings of the 15th conference on Architectural support for programming languages and operating systems* (2010), pp. 243–256.

[4] GHATIKAR, G., GANTI, V., MATSON, N., AND PIETTE, M. A. Demand response opportunities and enabling technologies for data centers: Findings from field studies.

[5] HOGAN, W. W. Electricity markets and the clean power plan. In *Harvard Project on Climate Agreements*. Belfer Center for Science and Internattional Affairs, Harvard Kennedy School, 2015.

[6] LIU, Z., WIERMAN, A., CHEN, Y., RAZON, B., AND CHEN, N. Data center demand response: Avoiding the coincident peak via workload shifting and local generation. *SIGMETRICS Perform. Eval. Rev. 41*, 1 (June 2013), 341–342.

[7] NARAYANAN, L., WANG, D., MAMUN, A.-A., SIVASUBRAMANIAM, A., AND FATHY, H. K. Should we dual-purpose energy storage in datacenters for power backup and demand response? In *6th Workshop on Power-Aware Computing and Systems (HotPower 14)* (Broomfield, CO, 2014), USENIX Association.

[8] NAVIGANT CONSULTING. Carbon dioxide reductions from demand response. Tech. rep., Advanced Energy Management Alliance, 2014.

[9] ROUNTREE, B., AHN, D. H., DE SUPINSKI, B. R., LOWENTHAL, D. K., AND SCHULZ, M. Beyond dvfs: A first look at performance under a hardware-enforced power bound. *2013 IEEE International Symposium on Parallel & Distributed Processing, Workshops and Phd Forum 0* (2012), 947–953.

[10] US DEPARTMENT OF ENERGY. Benefits of demand response in electricity markets and recommendations for achieving them. a report to the united states congress pursuant to section 1252 of the energy policy act of 2005, February 2005.

[11] WIERMAN, A., LIU, Z., LIU, I., AND MOHSENIAN-RAD, H. Opportunities and challenges for data center demand response. In *Green Computing Conference (IGCC), 2014 International* (Nov 2014), pp. 1–10.

[12] Z. LIU, M. LIN, A. W. S. L. L. A. Greening geographical load balancing. In *SIGMETRICS '11 Proceedings f the ACM SIGMETRICS joint international conference on measurement and modeling of computer systems* (2011).

[13] ZHOU, Z., LAN, Z., TANG, W., AND DESAI, N. Reducing energy costs for ibm blue gene/p via power-aware job scheduling. In *Job Scheduling Strategies for Parallel Processing*, N. Desai and W. Cirne, Eds., vol. 8429 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2014, pp. 96–115.