The Electrical Grid and Supercomputer Centers: An Investigative Analysis of Emerging Opportunities and Challenges

Fname Lname
Institute Name
City State Zip
Country
Name@email.adr

Fname Lname
Institute Name
City State Zip
Country
Name@email.adr

Fname Lname
Institute Name
City State Zip
Country
Name@email.adr

Fname Lname
Institute Name
City State Zip
Country
Name@email.adr

Fname Lname
Institute Name
City State Zip
Country
Name@email.adr

Fname Lname
Institute Name
City State Zip
Country
Name@email.adr

Some of the largest supercomputer centers in the United States are developing new relationships with their electricity service providers. These relationships, similar to other commercial and industrial facilities, are driven by mutual interest to save energy costs and improve electrical grid reliability. Supercomputer centers are concerned about price, quality, environmental impact and availability. Electricity service providers are concerned about the supercomputer center‘s impact on the electrical grid reliability in terms of energy consumption, peak power and fluctuations in power. Supercomputer center power demand can be greater than 20MW or more—the theoretical peak power requirements are greater than 45MW—and re-occurring intra-hour variability can exceed 8MW. Consequently, the electricity service providers for some supercomuter centers are asking for hourly forecasts of power demand, a day in advance. This paper evaluates today‘s relationships, potential partnerships and possible integration between supercomputer centers and their electricity service providers. The paper uses feedback from a questionnaire submitted to supercomputer centers on the Top 500 List in the United States to list opportunities for overcoming the challenges to HPC-Grid integration.

Categories and Subject Descriptors

C.4 [**Performance of Systems**]: Measurement Techniques—*power, energy*

General Terms

Management, Measurement, Performance

SC14 '14 New Orleans, Louisiana USA

Keywords

Electricity Service Provider, demand response

1. INTRODUCTION AND BACKGROUND

Supercomputing centers(SCs) with petascale [1] systems for high-performance computing (HPC) realize the large impact they are putting on their electricity service providers (ESPs) with peak power demands of 20MW and instantaneous power fluctuations of 8MW. Today, with the impetus towards Exascale [2], the need arises to aid electrical grid reliability with even larger peak power demands and instantaneous power fluctuations.

Electrical grid reliability is extremely important when optimally managing and linking electric supply with demand. Changes in electrical usage by end-users from normal consumption patterns can affect the electric supply infrastructure. To moderate such effects, it is important to understand how end users like SCs with high power demand and instantaneous fluctuations affect power-supply reliability. Consequently, ESPs are now seeking hourly forecasts of power demand from SCs, a day in advance.

The large power demand of SCs also make them pivotal in aiding ESPs to maintain reliable supply to other end-users. For example, when the triple digit temperature hits the valley, the domestic household power consumption surges. Lawrence Livermore National Laboratory (LLNL) becomes a friendly neighbor and lowers its power usage, thus decreasing its load on the grid and helping out its ESP. This is just one among many other instances where SCs have shown to be resourceful in aiding their ESPs.

In the past, large smelters and manufacturing industries extended great influence on the electrical grid; but today, becasue of their increasing power demands, supercomputing centers are the key players. While supercomputing centers

[1]Petascale computing refers to computing systems capable of at least 10^{15} floating point instructions per second(FLOPS).

[2]Exascale computing refers to computing systems capable of at least 10^{18}FLOPS.

often focus on energy efficiency to lower costs, it is important to understand if this focus is also a key interest for ESPs. Likewise, it is important for the SC to understand the significance of electrical grid reliability—a priority for ESP. A mutual understanding of concerns between the SC and ESP can produce a symbiotic relation that goes beyond the current producer-consumer paradigm, paving the way for possible integration of HPC with the electrical grid.

The Energy Efficient HPC Working Group (EE HPC WG) has been investigating opportunities for large supercomputing sites to integrate more closely with their electricity service providers. The objectives of this investigation are to understand the willingness of SCs to cooperate with their ESP, their expectations from their ESP, and the feasible measures that the SC could employ to help their ESP. To achieve our objectives we deployed a questionnaire to Top100 List class supercomputer centers in the United States, the results of which are documented in subsequent sections.

This work leverages prior work on data center and grid integration opportunities done by Lawrence Berkeley National Laboratory's (LBNL) Demand Response Research Center **(cite LBNL DR datacenter integration paper)** that describes challenges and opportunities in which data centers and electricity service providers may interact and how this integration can advance new market opportunities[3]. This integration model describes programs that are used by some of the electricity service providers to encourage particular responses by their customers and methods used to balance the electric grid supply and demand. Perhaps one of the most straightforward ways that SCs can begin the process of participating in demand response is by using this existing software infrastructure to manage their electricity requirements in a tightly coupled manner with their ESPs, facilitating both energy efficiency and grid reliability.

Section 2.2 of this paper describes in greater detail the model for integrating supercomputer centers and the electrical grid. Section 3.4 is a review of prior work on HPC center strategies that might be deployed for managing electricity and power. In order to further understand today's relationships, potential partnerships and possible integration between HPC centers, their electricity providers and the grid, a questionnaire was deployed whose respondents were Top100 List class supercomputer centers in the United States. Section 4 of this paper describes the results of that questionnaire. Section 5 of the paper describes opportunities, solutions and barriers. Section 6 describes conclusions and next steps. Finally, the last section recognizes additional authors.

2. ELECTRICITY SERVICE PROVIDERS AND HPC-GRID INTEGRATION

The EE HPC WG Team took as their starting point a model developed by LBNL's Demand Response Research Center that describes strategies that data centers might employ for utility programs to manage their electricity and power requirements to lower costs and benefit from utility incentives. The EE HPCWG Team adopted this model with slight tweaks to reflect the supercomputing environment focus (versus the data center as described by LBNL's Demand Response Research Center).

[3]LBNL Data Center Grid Integration Activities: http://drrc.lbl.gov/projects/dc

This paper distinguishes a supercomputer center as having unique characteristics that distinguish it from a datacenter. As opposed to datacenters, supercomputer centers have very high system utilization and are not likely to use virtualization as a strategy. Also, supercomputer applications are generally not easily portable between geographic locations for a variety of reasons: security, data- locality, system tuning. Finally, supercomputing centers are more energy efficient than the average data center. According to prior studies, the average data center has a power usage effectiveness (PUE) of between 1.91 and 2.9, whereas the highest PUE reported by the supercomputer center respondents in this study is 1.53.

2.1 Electricity Provider Programs and Methods

One key goal of the electricity provider is to provide efficient and reliable generation, transmission and distribution of electricity. Methods and programs employed by the electricity providers and their consumers are key to managing and balancing the supply and demand of electricity. While the methods describe how electricity providers manage supply, the programs describe the activities that the electricity providers can offer to their consumers to balance demand with supply.

Although critical to the eletricity service providers, *methods* are generally not visible to the consumer of the electricity because they operate within the generation or transmission stations. These methods are the major means by which supply and demand of electricity are managed.

Electricity provider *programs* encourage customer responses to target both energy efficiency and real-time (day-ahead or day-of) management of demand for electricity. An example of an electricity provider program that encourages energy efficiency would be to provide home consumers a financial incentive for replacing single pane with double pane windows. On the other hand, an example that illustrates programs that help with real-time demand management would be to provide a financial incentive for reducing load during high demand periods (such as hot summer afternoons when air conditioners are heavily utilized).

The following is a list and brief definitions of key methods and programs.

2.1.1 Methods

- Regulation (Up or Down): Methods used to maintain that portion of electricity generation reserves that are needed to balance generation and demand at all times. Raising supply is up regulation and lowering supply is down regulation. There are many types of reserves (e.g., operating, ancillary services), distinguished by who manage them and what they are used for.

- Transmission Congestion: Methods used to resolve congestion that occurs when there is not enough transmission capability to support all requests for transmission services. Transmission system operators must re-dispatch generation or, in the limit, deny some of these requests to prevent transmission lines from becoming overloaded.

- Distribution Congestion: Methods used to resolve congestion that occurs when the distribution control sys-

tem is overloaded. It generally results in deliveries that are held up or delayed.

- Frequency response: Methods used to keep grid frequency constant and in-balance. Generators are typically used for frequency response, but any appliance that operates to a duty cycle (such as air conditioners and heat pumps) could be used to provide a constant and reliable grid balancing service by timing their duty cycles in response to system load.
- Grid Scale Storage: Methods used to store electricity on a large scale. Pumped-storage hydroelectricity is the largest-capacity form of grid energy storage.
- Renewables: Methods used to manage the variable uncertain generation nature of many renewable resources.

2.1.2 Programs

- Energy Efficiency: Programs used to reduce overall electricity consumption.
- Peak Shedding: Programs used to reduce load during peak times, where the reduced load is not used at a later time.
- Peak Shifting: Programs where the load during peak times is moved to, typically, non-peak hours.
- Dynamic Pricing: Time varying pricing programs used to increase, shed or shift electricity consumption. There are two types of pricing, peak and real-time. Peak pricing is pre-scheduled; however, the consumer does not know if a certain day will be a peak or a non-peak day until day-ahead or day-of. Real-time pricing is not pre-scheduled; prices can be set day-ahead or day-of.

These methods and programs have historically not been relevant to supercomputing centers; however, the following example illustrates their potential relevance. The generation capacity requirements and response timescales vary across the country for electricity providers and operators. For example, the New England independent system operator (ISO-NE) uses a method of regulation and reserves that relies heavily on a day-ahead market program. This provides an opportunity for demand side resources- like supercomputer centers with renewable energy sources- to participate in the market supplying the ISO-NE with electricity. It also makes the NE-ISO particularly sensitive to major fluctuations in electricity demand, which, as discussed further in the questionnaire section, is an emerging characteristic of the largest supercomputer centers. [4]

2.2 Supercomputing Centers and HPC-Grid Integration

As described in Section 2.1, electricity providers offer programs like peak shifting that may request a change in timing and/or magnitude of demand by supercomputing centers. In response, supercomputing centers may employ one or more of a number of strategies to control their electricity demand.

Although these strategies can be used to temporarily modify loads in response to a request from an electric service provider, some of strategies could eventually be used at all times to improve energy efficiency if the HPC sees no operational issues. It is the former that is of primary interest to this investigation - what HPC systems can do in response to a grid request that they cannot do all the time? Two examples may help to clarify this distinction. Temporary load migration is an example of a strategy that is well suited to responding to an electric service provider request, but is not likely to improve energy efficiency (lowering aggregate energy use). Fine grained power management at all times, on the other hand, is more likely to be used for improving energy efficiency, unless the strategy is specifically used in response to a service provider's request. Below is a list of strategies:

- Fine grained power management refers to the ability to control HPC system power and energy with tools that are high resolution control and can target specific low level sub-systems. A typical example is voltage and frequency scaling of the Central Processing Unit (CPU).
- Coarse grained power management also refers to the ability to control HPC system power and energy, but contrasts with fine grained power management in that the resolution is low and it is generally done at a more aggregated level. A typical example is power capping.
- Load migration refers to temporarily shifting computing loads from an HPC system in one site to a system in another location that has stable power supply. This strategy can also be used in response to change in electricity prices.
- Job scheduling refers to the ability to control HPC system power by understanding the power profile of applications and queuing the applications based on those profiles.
- Back-up scheduling refers to deferring data storage processes to off-peak periods.
- Shutdown refers to a graceful shutdown of idle HPC equipment. It usually applies when there is redundancy.
- Lighting control allows for data center lights to be shutdown completely.
- Thermal management is widening temperature set-point ranges and humidity levels for short periods.

3. PRIOR WORK

We now describe prior work done in Power and Thermal Management, Job Scheduling, and Load Migration in the HPC and data center communities. Software and hardware techniques to save energy and power have been studied extensively. However, most of this work does not take into consideration power management (especially cooling systems and IT equipment) in response to a request from an electrical service provider [20].

[4] http://drrc.lbl.gov/sites/drrc.lbl.gov/files/LBNL-5958E.pdf

3.1 Power Management

DVFS and power capping are two popular ways to manage node power. Prior work in the HPC domain looked at analytical models to understand energy consumption [39, 19, 27] and at trading execution time for lower power/energy [2, 22]. Several DVFS algorithms have also been proposed, such as CPUMiser [19] and Jitter [25]. Varma et al, 2003 [44] demonstrated system-level DVFS techniques. They monitored CPU utilization at regular intervals and performed dynamic scaling based on their estimate of utilization for the next interval. Springer et al. [40] analyzed HPC applications under an energy bound. Rountree et al. used linear programming to find near-optimal energy savings without degrading performance [36] and implemented a runtime system based on this scheme [35].

There also has been work in the real-time systems community to solve the DVFS scheduling problem using mixed integer linear programming on a single processor[24, 37, 41, 42]. Other real-time approaches looked at saving energy [31, 29, 30, 49, 47].

In addition, there has been active research in the domain of virtual machines. Von Laszewski et al. [45] presented an efficient scheduling algorithm to allocate virtual machines in a DVFS-enabled cluster by dynamically scaling the supplied voltages. Dhiman et al. designed vGreen [8], which is a system for energy efficient computing in virtualized environments. They linked online workload characteristics to dynamic VM scheduling decisions and achieved better performance, energy efficiency and power balance in the system. Curtis-Maury et. al [4, 6, 5] introduced Dynamic Concurrency Throttling, which is a technique to dynamically optimize for power and performance by varying the number of active threads in parallel codes.

Chip power measurement and capping techniques were initially introduced with the Running Average Power Limit (RAPL) interface on Intel Sandy Bridge processors [23, 7]. In the HPC domain, Rountree et al. [34] proposed RAPL as an alternative to DVFS and analyzed application performance under hardware-enforced power bounds. They also established that variation in power directly translates to variation in application performance under a power bound. Patki et al. [33] used power capping techniques to demonstrate how hardware overprovisioning can improve HPC application performance under a global power bound significantly. Overprovisioning was also explored in the data center community [15].

3.2 Thermal Management

Thermal and cooling metrics are becoming important in HPC resource management. Runtime cooling strategies are mostly job-placement-centric. These techniques either aim to place incoming computationally intensive jobs in a thermal-aware manner on servers with lower temperatures or attempt to migrate or load-balance jobs from high-temperature servers to servers with lower temperatures.

Kaushik et. al [26] proposed T^*, a system that is aware of server thermal profiles and reliability as well as data semantics (computation job rates, job sizes, etc). This system saves cooling energy costs by using thermal-aware job placements without trading off performance. This paper assumes that the given grid is a constant. However, it is expected that future grid infrastructures will evolve due to grid integration solutions [21]. This may make thermal management more challenging. Aikema et. al [1] explored the potential for HPC centers to adapt to dynamic electrical prices, to variation in carbon intensity within an electrical grid, and to availability of local renewables. Their simulations demonstrated that 10- 50 % of electricty costs could potentially be saved. They also concluded that adapting to the variation in the electrical grid carbon intensity was difficult, and that adapting to local renewables could result in significantly higher cost savings. Sarood et. al [38] designed a runtime system that does temperature-aware load balancing in data centers using DVFS and task migration. They also discussed how hotspots could be avoided in data centers, and showed cooling costs can be reduced by up to 48% with temperature-aware load balancing.

3.3 Job Scheduling

The problem of scheduling jobs has been extensively studied. Most resource managers implement the First Come First Serve (FCFS) policy as a simple but fair strategy for scheduling jobs. However, FCFS suffers from low system utilization. A common optimization is backfilling [28, 32, 14]. Backfilling improves system utilization by executing jobs with small resource requests out of order on idle nodes.

Power-aware resource management without degrading utilization has been proposed as a DR strategy to reduce electricity costs [46, 48]. The novelty of the proposed job scheduling mechanism is its ability to take the variation in electricity price into consideration as a means to make better decisions about job start times. Experiments on an IBM Blue Gene/P and a cluster system as well as a case study on Argonne's 48-rack IBM Blue Gene/Q system have demonstrated the effectiveness of this scheduling approach. Preliminary results show a 23% reduction in the cost of electricity for HPC systems.

Fan et al. [13] discussed power-aware job scheduling in the data center domain. They discussed a power monitoring system that could use power capping (based on a power estimation method such as RAPL or direct power sensing) and a power throttling mechanism. Such as system works well when is a set of jobs with loose service level guarantees or low priority that can be forced to reduce consumption when the datacenter is approaching the power cap value. Etinski et al. [10, 9, 11, 12] explored scheduling under a power budget in supercomputing and analyzed bounded slowdown of jobs. In their series of papers, they introduced three policies. Their first policy is based looks at current system utilization and uses DVFS during job launch time to meet a power bound. Their second policy meets a bounded slowdown condition without exceeding a job-level power budget. Their third policy improves upon the former by analyzing job wait times and adding a reservation condition.

There are many use cases in a grid computing environment that require QoS guarantees in terms of guaranteed response time, including time-critical tasks that must meet a deadline. Foster et. al [17, 16] proposed *advance reservations* to achieve time guarantees. Advance reservation is a guarantee for the availability of a certain amount of resources to users and applications at specific times in the future. The advance reservation feature requires scheduling systems to support reservation capabilities in addition to backfilling-based batch scheduling. Modern resource management systems such as Sun Grid Engine, PBS, OpenPBS, Torque, SLURM, Maui, and Moab support advance reser-

vation capabilities.

3.4 Load Migration

Chiu et. al [3] discussed a electrical grid balancing problem that was experienced in the Pacific Northwest. In order to match electricity supply and balance the electrical grid, they proposed low-cost geographic load migration. They also suggested that a symbiotic relationship between datacenters and electrical grid operators that leads to mutual cost benefits could work well. Ganti et al. [18] looked at two applied cases for distributed data centers. The results show that load migration is possible in both homogenous and heterogeneous systems. Their migration strategies were based on a manual process and can benefit from automation.

4. QUESTIONNAIRE

We used a questionnaire to understand the current experiences of a supercomputer center's interaction with their electricity providers. We restricted the analysis to sites in the United States because the results of the survey and practices of demand response is highly correlated and driven by energy policies in the country. [43].

Nineteen Top100 List sized sites in the United States were targeted for the questionnaire. Eleven sites responded (Oak Ridge National Laboratory, Lawrence Livermore National Laboratory, Argonne National Laboratory, Los Alamos National Laboratory, LBNL, Wright Patterson Air Force Base, National Oceanic Atmospheric Administration (NOAA), National Center for Supercomputing Applications, San Diego Supercomputing Center (SDSC), Purdue University and Intel Corporation). The questionnaire was sent to a sample that was not randomly selected. It was sent to those sites where it was relatively easy to identify an individual based on membership within the EE HPC WG. The sample is more representative of Top50 sized sites (1 Top50 sized site was not in the sample and 60% (9/15) of the sample responded). Only 4 additional sites were sampled from the Top51-Top100 List and, of those, 2 responded (Intel and National Oceanic and Atmospheric Administration).

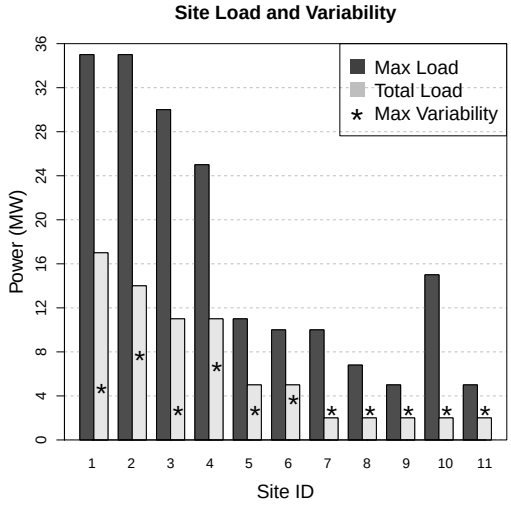

Figure 1: Site Load and Variability

The total power load as well as the intra-hour fluctuation of these sites varied significantly (Figure 1). Total power load includes all computing systems plus ancillary systems such as power delivery and cooling components. There were four sites with total power load greater than 10MW, two sites with ˜5MW total power load and five sites with less than 2MW total power load. For those with total power load greater than 10MW, the intra-hour fluctuation varied from less than 3MW to 8MW. One of ˜5MW sites said that they experienced 4MW variability. We chose less than 3MW intra-hour variability as the bottom of the scale because we assumed that the electricity providers would not be affected by 3MW (or less) fluctuations. The rest of the sites all reported less than 3MW intra-hour fluctuation. Most of the intra-hour variability was due to preventative maintenance (again, the power variation includes both computing and ancillary systems).

We asked if the supercomputer centers had talked to their electricity providers about programs and methods used to balance the grid supply and demand of electricity. About half of them have had some discussion, but it has mostly been limited to programs (e.g., peak shed, dynamic pricing) and not methods (e.g., regulation, frequency response, congestion).

Table 1: Caption Number 1

Discussions with Energy Providers	% Yes
Demand-side programs	
Shedding load during peak demand	54
Responding to pricing incentive programs	45
Shifting load during peak demand	36
Supply-side programs	
Enabling use of renewables	36
Congestion, Regulation, Frequency Response	18
Contributing to electrical grid storage	10

Approximately half of the respondents are not currently interested in shedding load during peak demand. LANL reports that the "technical feasibility" and "business case has yet to be developed." There is slightly more interest in shifting than shedding load. SDSC reports that "Automatic load shedding is being explored/deployed today" for the entire campus, not just the supercomputing center.

Responding to pricing incentive programs is also not considered currently interesting to approximately half of the repondents, although the reasons for this low interest may be organizational. Several open-ended comments revealed that pricing is fixed and/or done by another organization at the site level and outside of their immediate control.

Eighty percent of the respondents have not had discussions with their electricity providers about congestion, regulation and frequency response. Los Alamos National Laboratory (LANL) is one of two who have had discussions and who commented that they are "learning about the process" and that it is "outside of [their] visibility or control".

There were many more respondents who have had discussions with their electricity providers about enabling the use of renewables; 36% have already had discussions and more than half are interested in further and/or future discussions. SDSC already has a site-wide program; "the campus has a large fuel cell (2.5+ MW) and works with the utility with renewables." Other responses suggest that the interest is at

the site level and not unique to the supercomputer center.

An open-ended question was posed as to whether or not there was information either requested of the supercomputer sites by their providers or, conversely, requested of the providers by the sites. In both cases, well over 75% of the respondents answered no. Lawrence Livermore National Laboratory (LLNL) and LANL were the exceptions. LLNL is "responding to requests for additional data on an hourly, weekly and monthly basis." They are also working to develop an automated capability to share data with their electricity providers, which would provide automated additional detailed forecasting and ultimately real time data." LANL has also been requested to provide "power projections, hour by hour, for at least a day in advance" and, perhaps as a consequence, would like to have more information on "sensitivity of power distribution grid to rapid transients (random daily step changes of 10 MW up or down within a single AC cycle)."

Given the low levels of current engagement between the electricity providers and the supercomputer centers, it is not surprising that none of the supercomputer centers are currently using any power management strategies to respond to grid requests by their energy service providers. SDSC's *supercomputer center* is not an exception, but they did respond that their entire "campus is leveraging parallel electrical distribution to trigger diesel generators and other back-up resources to respond to grid and non-grid requests."

It was suggested by ORNL that some of the power management strategies are of questionable business value even for energy efficiency, let alone grid integration. For example, ORNL comments that "these assets have very clear depreciation schedules, and the modest cost savings in terms of electricity consumption due to some of these methods may not (or frequently will not) outweigh the capital investment cost in the computer. I.e. If a site spent $100M for a computer that will remain in production for 60 months, then the apparent benefit of power capping, etc can easily be outweighed by lost productivity of the consumable resource.

Similarly, another comment by ORNL suggested that the rapid deployment of hardware features, like P-states, may outpace the need for strategies like power aware job scheduling.

We tried to evaluate if power management strategies will be considered relevant and effective for grid integration at some point in the future. Two questions were asked: is there interest in using the strategies and what impact did they think that the strategies would have? When combining interest and impact, the results showed that power capping, shutdown, and job scheduling were both potentially interesting and of high impact.

Load migration, back-up scheduling, fine-grained power management and thermal management were of medium interest and impact. Lighting control and back-up resources were of low interest and impact.

Temperature control and lighting management are utilized as strategies, but considered medium to low interest and impact for responding to requests from electricity service providers because they are used as strategies for energy efficiency rather than grid response. The infrastructure energy efficiency of the responding supercomputer sites is high, as reflected in their reported Power Usage Effectiveness (PUE) (Figure 2). Two sites reported a PUE below 1.25, the majority were between 1.25 and 1.5 and the highest was 1.53.

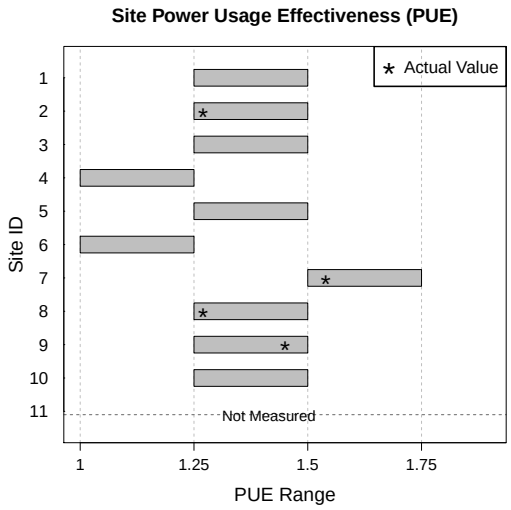

Figure 2: Site Power Usage Effectiveness

Approximately half of the respondents said that they used temperature control and lighting management as strategies, but not for grid requests. Temperature control and lighting management are well documented and understood strategies for improving energy efficiency, so it is not surprising that sites with PUEs below 1.5 are using them. NOAA comments that their "lights automatically shut off 24x7 when there is no motion in the data center." There is a value in lighting control for energy efficiency purposes, as demonstrated by its having been fully implemented. NOAA also comments that the impact of further lighting control "is so small compared to the HPC demand load that" they would "be surprised if the utility is interested."

Distinguishing interest from impact sheds further insight; some strategies are considered high impact, but not interesting enough to consider deployment. Facility shutdown is rated as having a high impact, but only considered interesting by 36% of the respondents. NOAA commented that, "We've had too many instability and equipment failures to utilize this as a strategy." This divide is even more apparent with load migration. It is rated as having a high impact by 36% of the respondents, but only interesting to 10% .

5. OPPORTUNITIES/SOLUTIONS AND BARRIERS

The responses to the questionnaire presented in Section 4 represent a variety of desires and experience regarding interactions between supercomputer centers and energy service providers. For example, the responses from the two centers with the largest power draws, Lawrence Livermore National Laboratory (LLNL) and Oak Ridge National Laboratory (ORNL), diverge in several areas. This divergence is perhaps primarily due to characteristics of their respective energy service providers. In contrast, San Diego Supercomputer Center (SDSC) stands out as a leader in integrating with their energy service provider on a site-wide level. To that end, the responses from SDSC may exemplify some of the opportunities available to other supercomputer centers

Table 2: HPC Strategies Responding to Electricity Provider Requests

HPC strategies for responding to Electricity Provider requests (listed from highest to lowest interest + impact)	% Interested	% High Impact	% Medium Impact
Coarse grained power management	64	46	27
Facility shutdown	36	64	10
Job scheduling	36	27	18
Load migration	10	36	18
Re-scheduling back-ups	45	0	10
Fine-grained power management	27	0	36
Temperature control beyond ASHRAE limits	27	0	18
Turn off lighting	18	0	0
Use back-up resources (e.g., generators)	0	10	27

that are willing to pursue this degree of integration.

The responses to the questionnaire also suggest that some energy service providers are requesting that their supercomputer center customers develop capabilities for informing the provider of expected periods of exceptional power consumption and for responding to requests from the provider to consume less power for specified periods of time. Upon initial consideration, this idea might seem to run counter to the primary mission objective of most supercomputer centers of delivering as many uninterrupted computational cycles as possible to their users. In some extreme cases, supercomputer centers may not have a choice in the matter as the size and energy requirements of supercomputers increase; indeed, some energy service providers may *require* large centers to develop a demand-response capability. However, a direct business case may exist to encourage supercomputer centers to develop this negotiation capability on their own. For example, if energy service providers were to offer electricity at a significantly reduced rate on the condition that the supercomputer center customer develop demand-response capabilities, the long-term cost savings to the center could make undertaking such a project worthwhile.

Perhaps one of the most straightforward ways that supercomputer centers can begin the process of developing a demand-response capability is by enhancing existing system software used for managing computing resources within the center. Indeed, the questionnaire responses from Section 4 as well as the literature review presented in Section 3.4 both strongly support the idea that the greatest opportunities for supercomputer centers to develop integration capabilities are related to system software. Specifically, and presented in approximate order of decreasing interest and expected impact to the questionnaire respondents, system software in this context consists of coarse-grained power management in the form of power capping, job scheduling, load migration, rescheduling backups, and fine-grained power management.

Coarse-grained power capping may be one of the most straightforward methods of power management. In the simplest form, this technique may entail human intervention to adjust computing resources so they operate at a reduced capacity or to entirely shut down some of the computing capacity of a supercomputer center. By attenuating resources, the supercomputer center manager can ensure that power consumption stays below some defined level. This defined level may be a pre-arranged power cap negotiated between the supercomputer center and the energy service provider and maintained on an ongoing basis, or, perhaps more likely, a power draw level that is requested by the energy service provider to handle unanticipated loads somewhere else in the energy service provider's system. One important thing to notice in using this approach to power management is that the savings in power may not need to come entirely from attenuating computing resources. Rather, reducing power consumption in computing resources is likely to result in a corresponding reduction in thermal load within the supercomputer center. This reduction in thermal load may allow a significant degree of power saving to occur by requiring less power to be spent on air conditioning.

The coarse-grained power capping technique described above assumes that the supercomputer center environment has some amount of instrumentation and metering that allows for the collection of power telemetry data. This telemetry is necessary for the supercomputer center facility manager in order to understand how the power supplied by the energy service provider is distrubted to resources within the center. Further, this telemetry is likely important to automated solutions for power management, such as the job scheduling techniques described below. In light of the fact that many system integrators such as Cray and IBM are now delivering supercomputer systems that include telemetry capabilities, the assumption that this information is available seems acceptable. According to the responses to the questionnaire presented in the previous Section, supercomputer center facility managers perceive this accounting data as distinct from per-user or per-job accounting data that is typically collected and indicate that this data should be retained for electricity provisioning planning purposes.

Techniques that involve job scheduling may offer more automated approaches to power management. Due to the unique role that the job scheduler and resource manager play within a supercomputer center, these techniques may involve adjusting either the workflow of jobs within the center or characteristics of the computational resources within the center.

On one hand, the job scheduler has knowledge of and control over the upcoming workflow within the supercomputer center simply by examining and manipulating the job queue. One easily-accessible technique is for a human operator to use capabilities such as advanced reservations to reserve prearranged blocks of time in which jobs with high power loads will run. These blocks of time could be negotiated with the energy service provider on an ongoing basis or could be in

response to on-demand requests made by the energy service provider. Even more automatic techniques are possible if the job scheduler is given enough information about the workflow to make intelligent decisions about job scheduling. For example, jobs may be submitted with various metadata that enable the job scheduler to understand characteristics of each job such as *priority*, the relative importance of a job compared to other jobs, and *urgency*, the rate at which the value of a job decreases as time elapses. These characteristics are not only important to a job scheduler for ensuring efficient utilization of a supercomputer center's resources under traditional circumstances, but they are also a vital piece of successfully implementing a demand-response capability for at least two reasons. First, they provide a set of metrics by which the supercomputer center can estimate the cost in terms of the "lost opportunity" of responding to an energy service provider's request to run with attenuated resources. Second, they allow the supercomputer center to prioritize jobs in the queued workflow in order to understand how to best utilize computational resources. This capability is important under normal circumstances, but becomes even more essential in a demand-response scenario.

On the other hand, the job scheduler has knowledge of and control over the computational resources within the supercomputer center, giving the job scheduler several mechanisms for implementing a demand-response capability. Most of these mechanisms could be considered fine-grained power control mechanisms because they mostly tune low-level settings on the nodes and processors within the supercomputer center. For example, the job scheduler knows which nodes within a supercomputer are occupied with running jobs or are expected to become occupied in the near future. To that end, the job scheduler can use its control over the resource management process to place idle nodes into a sleep state in which they draw significantly reduced power. This strategy is especially effective in supercomputer environments containing at least some resources that are used at irregular intervals, allowing opportunities to utilize sleep states effectively during periods when the resources are idle. In environments where all computing resources are heavily utilized most of the time, more sophisticated strategies that require the job scheduler to rely on knowledge about each batch job may be necessary. Such knowledge might come from the type of metadata described in the previous paragraph or from a database that is maintained based on previous runs of jobs submitted by each user. For example, if the job scheduler knows that a given job contains mostly I/O operations, or consists of discrete phases where I/O occurs, the job scheduler might choose to adjust the Performance State (P-state) for each processor running the job in a way that reduces the job's overall power consumption. The P-state mechanism is a way of scalably adjusting a CPU's frequency and voltage operating points which in turn causes the processor to consume less power directly and to produce less thermal load indirectly. In cases where a processor is executing a processing-intense task, gating the processor's P-state often has a noticeable impact on the overall task performance; however, in cases where a processor is executing a mostly I/O-bound task, gating the processor's P-state typically does not make a noticeable impact on the overall task performance due to the fact that the processor spends a great deal of time blocked waiting for I/O operations to complete.

Even more interesting scenarios are possible in cases where the job scheduler combines its knowledge of the upcoming queued workflow with its knowledge and control over the computational resources within the supercomputer center. These scenarios are most appropriate when the supercomputer scenario contains a pervasively heterogeneous mix of computational resources. For example, many contemporary supercomputer centers contain several different types of compute nodes with various types of processors and accelerator cards. In some circumstances, the job scheduler may be able to choose which resource to use for running a given job among several candidate resources. The trade-off here is not only in terms of the time necessary to complete the job (i.e., different resources could potentially complete the job in very different amounts of time) but also in terms of the energy consumed in completing the job (i.e., different resources could potentially consume very different amounts of energy in completing the job). Further, other resources such as memory access patterns, disk access patterns, and network use affect the energy signature of a job and may be observed by the scheduler. By maintaining a database of job-to-resource mappings that record the time and energy taken for each job, the scheduler can, over time, improve its ability to decide which jobs have the highest affinity to each type of resource. Using this knowledge to optimize a supercomputer center's workflow in terms of job throughput or energy consumption is admittedly complex, but the potential rewards are likely to be compelling both to the day-to-day operation of the center and to demand-response capabilities.

Opportunities may also exist for supercomputer centers to cooperate with each other in scenarios in which computational loads are migrated from one site to another where energy costs are less expensive. This scenario is challenging for both technical and business reasons. Technical challenges include issues such as user authentication and authorization (i.e., a user may be authorized to use resources at one site but not at another site) and data movement (i.e., it may be infeasible to migrate large datasets from one site to another site). To some extent, some of these technical challenges may be mitigated by the use of advanced reservation capabilities in the scheduling systems at each site, allowing resources to be simultaneously reserved while large datasets are properly staged. Business challenges include the notion that a supercomputer center currently has little incentive to migrate jobs to another "competing" center. Indeed, the questionnaire results reflect low interest in load migration strategies. It seems likely that in order to be a feasible scenario, the structure of payment and rewards to a supercomputer center to cooperate with other centers would need to be structured differently than they are currently.

In a very broad sense, demand-response techniques such as job scheduling, power capping, and load migration can be considered to be coarse-grained approaches because they involve considering "big picture" views of the workload and computational resources in a supercomputer center. According to the questionnaire results presented in the previous Section, facilities managers view these approaches as the most likely candidates for creating effective demand-response capabilities.

Finally, this Section has focused heavily on the opportunities available to supercomputer centers that come from developing demand-response capabilities. This notion is pri-

marily due to the fact that the questionnaire presented in Section 4 was distributed to high-performance computing centers in the United States, not to energy service providers. That said, opportunities do exist for energy service providers that develop demand-response capabilities. At one level, the negotiation process itself requires integration in terms of the communication and messaging protocols that are necessary. To that end, opportunities exist for adapting and extending existing standards currently used within the industry, thus creating new use cases and capabilities for energy service providers. At a higher level, energy service providers will most likely need to improve their ability to determine in near real time the important places within the electrical grid where demands exceed supply. Determining this is likely to be a complex optimization problem. While this Section focuses on solving these problems to the end of developing a demand-response strategy in conjunction with supercomputer centers, these capabilities are likely applicable to a broad range of customers.

6. CONCLUSIONS AND NEXT STEPS

This paper explores the possibility of a new relationship between electricity service providers and supercomputer centers with increased communication and engagement from both parties

Because supercomputer centers have an increasingly large and fluctuating demand for power, they challenge their providers to supply a reliable source of electricity. Electricity service providers are interested in partnering with customers, like supercomputer centers, to create a more dynamic and resilient grid by obtaining predictable demand forecasts and engaging in programs like demand response.

We focused our attention on the largest supercomputer centers in the United States. The two supercomputer centers with the largest electricity demand, ORNL and LLNL, have very different experiences. ORNL's experience is that its electricity demand and fluctuations are not significant factors for their electricity service provider. LLNL's experience is opposite to that of ORNL. Because of large swings in power usage, the LLNL supercomputer center was approached by their electricity service provider with a request for daily predictable demand forecasts. That request began an ongoing relationship.

The LANL supercomputer center's experience is similar to that of LLNL. SDSC has an even tighter relationship with their electricity service provider, but this relationship involves the entire campus and not just the supercomputer center.

As previous research with datacenters has shown [**need a ref here**], supercomputer centers can serve as resources to the grid. To enable this, automation technologies and data communication standards, which can link the supercomputer centers with the electric grid and on-site power management strategies for grid services will play a key role to ease adoption and lower the participation costs. Power capping, shutdown, and job scheduling are identified as the most interesting management strategies with the highest leverage for responding to requests from electricity service providers.

Nonetheless, the business case for the grid integration of supercomputer centers remains to be demonstrated. Supercomputer centers have concerns that deploying these strategies might have an adverse impact on their primary mission. One of the key enablers for supercomputing centers to participate in electricity markets (e.g., demand response, electricity prices) is having markets that value their participation. In other areas like commercial buildings and select industrial facilities, benefits to both electricity service providers and customers are well documented. However, as the electrical grid and new dynamic loads such as supercomputer centers evolve, the markets need mechanisms to identify and provide value of participation (e.g., cost, energy, carbon).

We are planning to pursue several areas in our future work.

We are planning a similar survey for Europe to explore if there is a more compelling business case in other geographies. We expect that the business value of such grid integration to be enhanced where the price of electricity is expensive, varies dynamically, or where there is strong reliance on expensive back-up generation (e.g., India),

We plan on following-up with the ESPs that support these US-based supercomputer centers. We note that this work's focus was from the perspective of the supercomputer center, and we are interested in hearing from the ESPs about what makes a customer more or less interesting or challenging with respect to grid integration.

With increasing variable renewable generation and price-based DR programs, the intra-hour fluctuations and demand forecasting is becoming increasingly important. Understanding the timescales of supercomputer center's load response will have different value to the electric grid programs. What are the trends in inter-hour fluctuation patterns? Is this a new behavior, an interim one, or one that is likely to get worse?

7. ADDITIONAL AUTHORS

Fname Lname, Affiliation
Fname Lname, Affiliation
Fname Lname, Affiliation
Fname Lname, Affiliation
Fname Lname, Affiliation
Fname Lname, Affiliation

8. REFERENCES

[1] D. Aikema and R. Simmonds. Electrical cost savings and clean energy usage potential for HPC workloads. In *2011 IEEE International Symposium on Sustainable Systems and Technology (ISSST)*, pages 1–6, 2011.

[2] K. W. Cameron, X. Feng, and R. Ge. Performance-constrained distributed DVS scheduling for scientific applications on power-aware clusters. In *Supercomputing*, Seattle, Washington, Nov. 2005.

[3] D. Chiu, C. Stewart, and B. McManus. Electric grid balancing through lowcost workload migration. *SIGMETRICS Perform. Eval. Rev.*, 40(3):48âĂŞ52, Jan. 2012.

[4] M. Curtis-Maury, F. Blagojevic, C. D. Antonopoulos, and D. S. Nikolopoulos. Prediction-based power-performance adaptation of multithreaded scientific codes. *IEEE Trans. Parallel Distrib. Syst.*, 19(10):1396–1410, Oct. 2008.

[5] M. Curtis-Maury, J. Dzierwa, C. D. Antonopoulos, and D. S. Nikolopoulos. Online power-performance adaptation of multithreaded programs using hardware

event-based prediction. In *International Conference on Supercomputing*, New York, NY, USA, 2006. ACM.

[6] M. Curtis-Maury, A. Shah, F. Blagojevic, D. S. Nikolopoulos, B. R. de Supinski, and M. Schulz. Prediction models for multi-dimensional power-performance optimization on many cores. In *International Conference on Parallel Architectures and Compilation techniques*, New York, NY, USA, 2008. ACM.

[7] H. David, E. Gorbatov, U. R. Hanebutte, R. Khanna, and C. Le. RAPL: Memory Power Estimation and Capping. In *Proceedings of the 16th ACM/IEEE international symposium on Low power electronics and design*, ISLPED '10, pages 189–194, New York, NY, USA, 2010. ACM.

[8] G. Dhiman, G. Marchetti, and T. Rosing. vGreen: a system for energy efficient computing in virtualized environments. In *Proceedings of the 14th ACM/IEEE international symposium on Low power electronics and design*, ISLPED '09, page 243âĂŞ248, New York, NY, USA, 2009. ACM.

[9] M. Etinski, J. Corbalan, J. Labarta, and M. Valero. Optimizing Job Performance Under a Given Power Constraint in HPC Centers. In *Green Computing Conference*, pages 257–267, 2010.

[10] M. Etinski, J. Corbalan, J. Labarta, and M. Valero. Utilization driven power-aware parallel job scheduling. *Computer Science - R&D*, 25(3-4):207–216, 2010.

[11] M. Etinski, J. Corbalan, J. Labarta, and M. Valero. Linear Programming Based Parallel Job Scheduling for Power Constrained Systems. In *International Conference on High Performance Computing and Simulation*, pages 72–80, 2011.

[12] M. Etinski, J. Corbalan, J. Labarta, and M. Valero. Parallel job scheduling for power constrained hpc systems. *Parallel Computing*, 38(12):615–630, Dec. 2012.

[13] X. Fan, W.-D. Weber, and L. A. Barroso. Power provisioning for a warehouse-sized computer. In *The 34th ACM International Symposium on Computer Architecture*, 2007.

[14] D. G. Feitelson, U. Schwiegelshohn, and L. Rudolph. Parallel job scheduling - a status report. In *In Lecture Notes in Computer Science*, page 1âĂŞ16. Springer-Verlag, 2004.

[15] M. E. Femal and V. W. Freeh. Safe overprovisioning: using power limits to increase aggregate throughput. In *International Conference on Power-Aware Computer Systems*, Dec 2005.

[16] I. Foster. The anatomy of the grid: enabling scalable virtual organizations. In *First IEEE/ACM International Symposium on Cluster Computing and the Grid, 2001. Proceedings*, pages 6–7, 2001.

[17] I. Foster, C. Kesselman, C. Lee, B. Lindell, K. Nahrstedt, and A. Roy. A distributed resource management architecture that supports advance reservations and co-allocation. In *1999 Seventh International Workshop on Quality of Service, 1999. IWQoS '99*, pages 27–36, 1999.

[18] V. Ganti and G. Ghatikar. Smart Grid as a Driver for Energy-Intensive Industries: A Data Center Case Study. In *Grid-Interop 2012*, Dec. 2012.

[19] R. Ge, X. Feng, W. Feng, and K. W. Cameron. CPU Miser: A performance-directed, run-time system for power-aware clusters. In *International Conference on Parallel Processing*, Xi'An, China, 2007.

[20] G. Ghatikar, V. Ganti, N. Matson, and M. A. Piette. Demand Response Opportunities and Enabling Technologies for Data Centers: Findings From Field Studies. In *"PG&E/SDG&E/CEC/LBNL"*, 2012.

[21] M. M. He, E. M. Reutzel, X. Jiang, Y. H. Katz, S. R. S, and D. E. Culler. An architecture for local energy generation, distribution, and sharing. In *IEEE Energy2030 Conference Proceedings*, Atlanta, Georgia, USA, Nov. 2008.

[22] C.-H. Hsu and W.-C. Feng. A power-aware run-time system for high-performance computing. In *Supercomputing*, Nov. 2005.

[23] Intel. Intel-64 and IA-32 Architectures Software Developer's Manual, Volumes 3A and 3B: System Programming Guide. 2011.

[24] T. Ishihara and H. Yasuura. Voltage scheduling problem for dynamically variable voltage processors. In *International Symposium on Low power Electronics and Design*, pages 197–202, 1998.

[25] N. Kappiah, V. W. Freeh, D. K. Lowenthal, and F. Pan. Exploiting slack time in power-aware, high-performance programs. In *Supercomputing*, Nov. 2005.

[26] R. T. Kaushik and K. Nahrstedt. T*: a data-centric cooling energy costs reduction approach for big data analytics cloud. SC '12, page 52:1âĂŞ52:11, Los Alamitos, CA, USA, 2012. IEEE Computer Society Press.

[27] J. Li and J. F. Martìnez. Dynamic power-performance adaptation of parallel computation on chip multiprocessors. In *12th International Symposium on High-Performance Computer Architecture*, Austin, Texas, Feb. 2006.

[28] D. A. Lifka. The ANL/IBM SP scheduling system. In *In Job Scheduling Strategies for Parallel Processing*, page 295âĂŞ303. Springer-Verlag, 1995.

[29] B. Mochocki, X. S. Hu, and G. Quan. A realistic variable voltage scheduling model for real-time applications. In *Proceedings of the 2002 IEEE/ACM International Conference on Computer-Aided Design*, 2002.

[30] B. Mochocki, X. S. Hu, and G. Quan. Practical on-line DVS scheduling for fixed-priority real-time systems. In *11th IEEE Real Time and Embedded Technology and Applications Symposium*, 2005.

[31] M. A. Moncusí, A. Arenas, and J. Labarta. Energy aware EDF scheduling in distributed hard real time systems. In *Real-Time Systems Symposium*, December 2003.

[32] A. W. Mu'alem and D. G. Feitelson. Utilization, predictability, workloads, and user runtime estimates in scheduling the IBM SP2 with backfilling. *IEEE Trans. Parallel Distrib. Syst.*, 12(6):529âĂŞ543, June 2001.

[33] T. Patki, D. K. Lowenthal, B. Rountree, M. Schulz, and B. R. de Supinski. Exploring Hardware Overprovisioning in Power-constrained, High Performance Computing. In *International Conference*

on Supercomputing, pages 173–182, 2013.

[34] B. Rountree, D. H. Ahn, B. R. de Supinski, D. K. Lowenthal, and M. Schulz. Beyond DVFS: A First Look at Performance under a Hardware-Enforced Power Bound. In *IPDPS Workshops*, pages 947–953. IEEE Computer Society, 2012.

[35] B. Rountree, D. Lowenthal, B. de Supinski, M. Schulz, V. Freeh, and T. Bletch. Adagio: Making DVS Practical for Complex HPC Applications. In *International Conference on Supercomputing*, June 2009.

[36] B. Rountree, D. K. Lowenthal, S. Funk, V. W. Freeh, B. de Supinski, and M. Schulz. Bounding energy consumption in large-scale MPI programs. In *Supercomputing*, Nov. 2007.

[37] H. Saputra, M. Kandemir, N. Vijaykrishnan, M. Irwin, J. Hu, C.-H. Hsu, and U. Kremer. Energy-conscious compilation based on voltage scaling. In *Joint Conference on Languages, Compilers and Tools for Embedded Systems*, 2002.

[38] O. Sarood and L. V. Kalé. A 'cool' load balancer for parallel applications. In *Proceedings of the 2011 ACM/IEEE conference on Supercomputing*, Seattle, WA, November 2011.

[39] R. Springer, D. K. Lowenthal, B. Rountree, and V. W. Freeh. Minimizing execution time in MPI programs on an energy-constrained, power-scalable cluster. In *ACM Symposium on Principles and Practice of Parallel Programming*, Mar. 2006.

[40] R. C. Springer IV, D. K. Lowenthal, B. Rountree, and V. W. Freeh. Minimizing execution time in MPI programs on an energy-constrained, power-scalable cluster. In *ACM Symposium on Principles and Practice of Parallel Programming*, Mar. 2006.

[41] V. Swaminathan and K. Chakrabarty. Real-time task scheduling for energy-aware embedded systems. In *IEEE Real-Time Systems Symposium*, Nov. 2000.

[42] V. Swaminathan and K. Chakrabarty. Investigating the effect of voltage-switching on low-energy task scheduling in hard real-time systems. In *Asia South Pacific Design Automation Conference*, Jan. 2001.

[43] J. Torriti, M. G. Hassan, and M. Leach. Demand response experience in europe: Policies, programmes and implementation. *Energy*, 35(4):1575–1583, Apr. 2010.

[44] A. Varma, B. Ganesh, M. Sen, S. R. Choudhury, L. Srinivasan, and B. Jacob. A control-theoretic approach to dynamic voltage scheduling. In *Proceedings of the 2003 international conference on Compilers, architecture and synthesis for embedded systems*, CASES '03, page 255âĂŞ266, New York, NY, USA, 2003. ACM.

[45] G. von Laszewski, L. Wang, A. Younge, and X. He. Power-aware scheduling of virtual machines in DVFS-enabled clusters. In *IEEE International Conference on Cluster Computing and Workshops, 2009. CLUSTER '09*, pages 1–10, 2009.

[46] X. Yang, Z. Zhou, S. Wallace, Z. Lan, W. Tang, S. Coghlan, and M. E. Papka. Integrating dynamic pricing of electricity into energy aware scheduling for HPC systems. In *Proceedings of SC13: International Conference for High Performance Computing, Networking, Storage and Analysis*, SC '13, page 60:1âĂŞ60:11, New York, NY, USA, 2013. ACM.

[47] Y. Zhang, X. S. Hu, and D. Z. Chen. Task scheduling voltage selection for energy minimization. In *Proceedings of the 39th annual Design Automation Conference*, 2002.

[48] Z. Zhou, Z. Lan, W. Tang, and N. Desai. Reducing energy costs for IBM blue Gene/P via power-aware job scheduling.

[49] D. Zhu, R. Melhem, and B. R. Childers. Scheduling with dynamic voltage/speed adjustment using slack reclamation in multi-processor real-time systems. *IEEE Transactions on Parallel and Distributed Systems*, 2003.