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 partnerships, are driven by mutual interest to reduce energy costs and improve electrical grid reliability. Supercomputer centers are concerned about quality, environmental impact and availability, while electricity service providers are concerned about impacts on electrical grid reliability, particularly in terms of energy consumption, peak power and power fluctuations. Supercomputer centers' power demand can be 20MW or more—the theoretical peak power requirements are greater than 45MW—and recurring intra-hour variability can exceed 8MW. 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 Top100 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

Keywords

Electricity Service Provider, demand response

1. INTRODUCTION

Supercomputing centers (SCs) with petascale [1] systems for high-performance computing (HPC) have an outsized impact on their electricity service providers (ESPs), with peak power demands in excess of 20MW and instantaneous power fluctuations of up to 8MW. As the HPC community moves towards exascale computing,[2] we anticipate a growing number of facilities will be reaching or exceeding these service levels, with significant potential effect on electrical grid reliability. To moderate this risk, in this paper we seek to understand how these anticipated usage patterns can be integrated safely into the power grid with minimal cost and disruption.

Being a "good citizen" on the grid has several historical precedents. For example, electrically-intensive industries such as aluminium smelters have received preferential pricing in return for predictable loads and flexibility in reducing power during periods of high consumption. SCs are already adopting these strategies. For example, Lawrence Livermore National Laboratory (LLNL) reduces its power usage when temperatures exceed 100 degrees and residential power usage in the area surges, Other centers are exploring the benefits of predicting hour-ahead and day-ahead use in concert with their providers. 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) investigates opportunities for large supercomputing sites to integrate more closely with their electricity service providers. We seek to understand the willingness of SCs to cooperate with their ESPs, their expectations from their ESPs, and the feasible measures that SCs could employ to help their ESPs. To achieve our objectives we developed a questionnaire and

[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.

distributed it to the Top100 supercomputer centers in the United States.

This work leverages prior work on datacenter and grid integration opportunities done by Lawrence Berkeley National Laboratory's (LBNL) Demand Response Research Center **(cite LBNL DR datacenter integration paper)** that describes the challenges and opportunities for datacenters and electricity service providers to 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 electrical 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.

The paper is organized as follows. Sections 2 and 3 of this paper describe in greater detail the model for integrating supercomputer centers and the electrical grid. Section **??** reviews prior work on HPC center strategies. Section 4 provides the results of the questionnaire. Section 5 discusses the several opportunities, solutions and barriers that have been highlighted by the survey results. We offer our conclusions in Section 6 along with our plan for future work. Additional authors are listed in Section 7 and section Appendix summarizes the survey questions.

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 datacenters 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 datacenter as described by LBNL's Demand Response Research Center).

For purposes of this paper, we define supercomputer centers as distinct from a datacenters as having significantly higher system utilization and thus little or no virtualization. Additionally, supercomputer applications are distinguished by their lack of geographical portability due to security concerns, data size and machine-specific optimization. We also note that supercomputing centers tend to be more energy-efficient than datacenters. In our survey, no SC exceeded a power usage effectiveness (PUE) of 1.53, while the average data center falls between 1.91 and 2.9 (with 1.0 being the ideal). [**?**]

2.1 Electricity Provider Programs and Methods

An electricity provider seeks 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 manages 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 system 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.

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

- 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. The two types of pricing are 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.

Although these methods and programs have historically not been relevant to supercomputer centers, 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 ISO-NE 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]

This paper assumes that the given grid is a constant. However, it is expected that future grid infrastructures will evolve with smart-grid capabilities.

3. SUPERCOMPUTING CENTERS AND HPC-GRID INTEGRATION

In November 2004, the Blue Gene/L system at Lawrence Livermore National Laboratory became the fastest computer in the Top 500, [**?**], displacing the NEC Earth Simulator, the previous champion. This change marked the transition from supercomputing gains based on ever-higher-performance components to systems comprised of far larger numbers of slow but energy-efficient components. However, total system power consumption continued to rise, and we are now poised to begin a second transition to "power-limited computing". The new model has been exemplified by the US Department of Energy issuing guidance that the first DOE exascale machine should not exceed 20MW; effectively a $1000x$ performance improvement with only a $3x$ increase in power.

However, the problem is not as simple as provisioning 20MW. Ultimately, SCs optimize for performance per dollar, not performance per Watt, and flexibility in power consumption can be expected to result in lower overall prices. Use of green technologies such as wind and solar may also lead to cheaper but less predictable sources of power. To adapt to this new landscape, SCs may employ one or more strategies to control their electricity demand.

- **Node level.** Controlling power ultimately requires control of individual components. Historically, this control has been accomplished through Dynamic Voltage/Frequency Scaling (DVFS), which allows the processor to use a lower voltage at the cost of a slower clock frequency. Newer technologies such as Intel's Running Average Power Limit leverage DVFS to guarantee that at user-specified processor power bound will, on average, not be exceeded over the duration of a short time window. DVFS can also be found on accelerator cards such as nVidia's Kepler GPGPU. Other efforts reduce DRAM power by batching reads and writes, thus allowing the memory to spend more time in a lower-power state. Several processor configuration options have indirect but significant effects on power consumption. For example, the choice of the number of cores to use, whether or not to enable hyperthreading, and the use of "turbo" modes will change the power/performance curve.

- **Job level.** Each of the node-level controls requires a tradeoff between power and performance. SCs resources are typically oversubscribed, so degrading performance to save power and energy ultimately results in less science getting done. However, at the job level, load imbalance provides opportunities to slow nodes that are off of the critical path of execution without slowing the overall job execution time. Traditionally, load rebalancing strategies have focused on moving bytes around the job allocation. With power control, we can now rebalance power as well as work.

- **System level.** While most SCs use time and space partitioning (where a node only runs a single job at a time), there are still shared resources that must be managed across jobs. Periodic checkpointing saves sufficient job state to a filesystem shared across jobs so that a job may be restarted from a recent point in case a fault occurs. Because these checkpoints involve much more data motion than normal execution, power spikes can be observed at the node level (particularly DRAM), network, and filesystem. These checkpoints may need to be coordinated across large jobs to prevent unnecessary performance degredation.

- **Scheduler level.** Up through the system level, power control is evaluated using the execution time of individual jobs. The scheduler optimizes for overall throughput rather than individual job performance. At this point, scheduling is a two-dimensional problem: jobs request a certain number of nodes for a certain duration. As power-limited computing becomes more common, schedulers will add power bounds to this mix: a job will be allowed nodes, time, and a certain number of watts (the responsibility for not exceeding the job power bound rests with the system software, not the user or application). The scheduler not only determines when jobs in the queue begin execution, but also what happens when a job exits the system. Depending on the priorities of already-running jobs and the priorities of jobs in the queue, the best solution in terms of throughput may be to idle the recently-freed nodes and redistribute the freed power to running jobs.

- **Site level.** At the level of the machine room (or multiple machine rooms), decisions must be made as to how much power should be allocated for cooling versus computation, which requires understanding how temperature interacts with performance. A higher intake air temperature uses less cooling power but results in

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

higher static processor power and may limit opportunities for ”turbo” mode in processors where it is available. As cooling power varies with outside air temperature, a single machine room temperature setpoint may not be the optimal solution in terms of overall performance.

3.1 Prior Work

We now describe prior work done in Power and Thermal Management, Job Scheduling Load Migration and Dynamic Pricing in the HPC and datacenter communities. Software and hardware techniques to save energy and power have been studied extensively.

3.1.1 Power Management

Dynamic Voltage Frequency Scaling (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 (Springer et al [39], Ge et al [18], Li and Martinez [26]) and at trading execution time for lower power/energy [2, 21]. Several DVFS algorithms have also been proposed, such as CPUMiser [18] and Jitter [24]. 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[23, 37, 41, 42]. Other real-time approaches looked at saving energy [30, 28, 29, 48, 46].

Chip power measurement and capping techniques were initially introduced with the Running Average Power Limit (RAPL) interface on Intel Sandy Bridge processors [22, 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 [14].

Techniques for fine-grained power management have also been proposed. 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. (Add Sridutt’s references on fine grained power management here.)

Power Usage Effectiveness (PUE) is a metric that reveals how much energy is expended on cooling costs and other non-compute operations in a facility [32]. It is the ratio of the energy supplied to the energy used in useful computations in a datacenter or a HPC facility. A PUE of 1.0 is ideal, but studies have shown that on average, a large datacenter has a PUE of 2.9, which indicates that datacenters are fairly energy inefficient.

3.1.2 Thermal Management

Thermal management is a key driver for improving energy efficiency of data centers as well as supercomputer centers. There are many strategies for thermal management that can improve energy efficiency, such as free cooling and proper airflow. This paper discusses two thermal management strategies that have an opportunity for grid integration. The first strategy is controlling the inlet temperature to the computing equipment, raising it as high as possible without causing reliability induced hardware failures. The second strategy is using thermally aware job scheduling.

In 2011, the American Society of Heating, Refrigeration and Air Conditioning (ASHRAE) data center Technical Committee TC9.9 published guidelines that âĂIJexpanded the environmental range for data centersâĂİ and supercomputer centers. The environmental range includes factors such as temperature, humidity and dew point and allowable rate of change. This expansion allows for maintaining high reliability while achieving gains in energy efficiency. These guidelines continue to be updated and the range continues to expand as the industry collects more historical data showing trade-offs between reliability and environmental factors.

It is implicit in the ASHRAE guidelines that a supercomputer center might be able to increase temperature as a response to a request from an energy service provider. The guideline defines both âĂIJrecommendedâĂİ and âĂIJallowableâĂİ environmental ranges. It also specifies a maximum rate of change, which is most stringent for tape drives. For supercomputer centers, the difference between the maximum recommended and allowable dry bulb temperature is a minimum of 9 degrees F. The rate of change for tape drives is 9 degrees F per hour (36 degrees F for solid state computing systems). Therefore, assuming that supercomputer centers normally operate within the recommended range and that they are willing to operate on occasion in the allowable range (or beyond), it is theoretically possible to stay within ASHRAE thermal guidelines and use temperature excursion as a grid-integration strategy.

ASHRAE has also published a guideline on liquid cooling environmental ranges (reference). At this point, however, the guidelines do not document rate of change for liquid temperature. Although it isnâĂŹt explored in this paper, it may be possible to use increases in liquid cooling temperature as a grid-integration strategy as well.

(Ghatikar et al[19]) has done field studies on using thermal management as a grid-integration strategy. They demonstrate increasing âĂIJfacility HVAC temperature set points in order to decrease HVAC power demandâĂİ in two different field locations. There was only a small electricity demand decrease demonstrated.

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 [25] 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.

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.

3.1.3 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 [27, 31, 13]. Backfilling improves system utilization by executing jobs with small resource requests out of order on idle nodes.

Fan et al. [12] 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. [9, 8, 10, 11] 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 [16, 15] 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 reservation capabilities.

3.1.4 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. [17] 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.

3.1.5 Dynamic Pricing

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 electricity 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.

Power-aware resource management without degrading utilization has been proposed as a DR strategy to reduce electricity costs [45, 47]. The novelty of the proposed job scheduling mechanism is its ability to take the variation in electricity price (dynamic pricing) 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.

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).

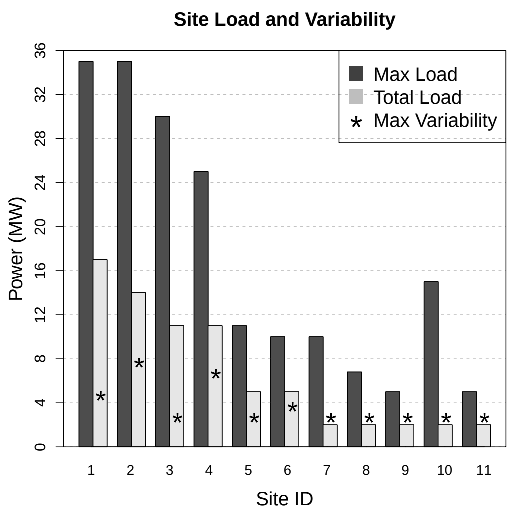

Figure 1: Site Load and Variability

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).

For every respondent, the theoretical peak energy or maximum load is approximately twice the total energy, which is indicative of expected future growth in power and energy requirements for supercomputer centers. Some of the design parameters that may affect theoretical peak limits are the customer switchgear, transformer and chiller water capacities. In some cases, there are also limits based on regional electricity service provider capacity constraints.

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

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

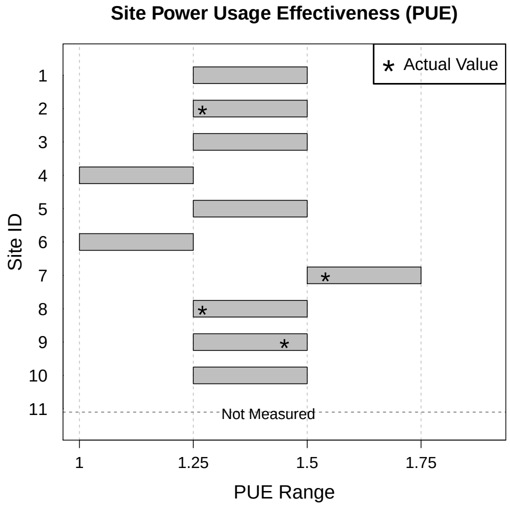

Figure 2: Site Power Usage Effectiveness

and impact for responding to requests from electricity service providers. 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. 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.”

LLNL reports that they âĂIJtook 3 years to raise the temperature in their center by 18 degrees F. It was done in conjunction with a failure rate analysis of the systems as well as a measurement of the electrical savings prior to moving to the next set pointâĂİ. LLNL is currently operating in the ASHRAE recommended range, but expresses concerns with increasing temperature as a grid-integration response. The concerns include hardware failures, tape storage read/write errors and compromising dew point requirements where liquid and air-cooling are co-located.

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 electricity 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 electricity service providers. In contrast, San Diego Supercomputer Center (SDSC) stands out as a leader in integrating with their electricity 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 that are willing to pursue this degree of integration.

The responses to the questionnaire also suggest that some electricity 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 electricity 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 electricity 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 **??** 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 electricity service provider and maintained on an ongoing basis, or, perhaps more likely, a power draw level that is requested by the electricity service provider to handle unanticipated loads somewhere else in the electricity 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 electricity service provider is distributed 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 electricity service provider on an ongoing basis or could be in response to on-demand requests made by the electricity 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 electricity 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 primarily due to the fact that the questionnaire presented in Section 4 was distributed to supercomputer centers in the United States, not to electricity service providers. That said, opportunities do exist for electricity 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 electricity service providers. At a higher level, electricity 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 power demand, 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 had 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 [20], 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 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 are becoming increasingly important. Electrical grid programs may react in different ways to the timescale of a supercomputer center's load response. 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] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] G. Ghatikar, D. Riess, and M. A. Piette. Analyis of open automated demand response deployments in california and guidelines to transition to industry standards. Technical Report LBNL-6560E, Lawrence Berkeley National Laboratory, 1 Cyclotron Rd, Berkeley, CA 94720, January 2014.

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

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

[23] 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.

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

[25] 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.

[26] 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.

[27] D. A. Lifka. The ANL/IBM SP scheduling system. In *In Job Scheduling Strategies for Parallel Processing*, pages 295–303. Springer-Verlag, 1995.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] J. Niccolai. New data center survey shows mediocre results for energy efficiency. April 2013.

[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, pages 255–266, New York, NY,

USA, 2003. ACM.

[45] 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.

[46] 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.

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

[48] 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.

Appendix

For the purposes of this paper, this appendix contains a summary of the questionnaire. The complete text of our questionnaire, which includes some background and explanatory material, is available at EEHPCWG questionnaire. Note that the questionnaire at that site is no longer active; any responses are not saved.

The questionnaire is divided into the following three sections:

- Facility Energy. The total facility energy and the total HPC load should be the same number that you use when calculating PUE, as defined by the Green Grid Whitepaper #49.
- Management and Control. Please answer whether or not you employ any of the strategies described below for managing and controlling total facility energy in response to a request from your Electrical Utility/Provider. You may use some of these same strategies for improving energy efficiency. Answer "Yes" only when the strategy is used at least in part for grid response. Answer "No" if the strategy is only used for improving energy efficiency.
- Electrical/Utility Provider Information. Answers to these questions help us understand the nature of any relationship you might have between your HPC facility and your site's electric utility/provider. Please answer "Yes" if you have had any communication about the following programs and methods with your site's electric utility/provider. For each program and/or method for which there has been communication, please describe the nature of that communication in the comments.

Facility Energy

1. What is your "total facility energy?"
2. What is your total HPC load?
3. What is your facility PUE?
4. What is your facility's theoretical peak energy, as the infrastructure is currently fit up.
5. What is the maximum variation in total facility energy that is likely to re-occur?
6. How often does this variation occur?
7. If there is any regular pattern to this variation, please describe the circumstances. Include the reason for the variation, the magnitude and duration if possible. For example, "There is a 5MW drop every two weeks for a 6 hour period during Preventative Maintenance periods."

Management and Control

8. COARSE-GRAINED POWER MANAGEMENT: manage power for the HPC system or subsystem (could include storage, networking as well as compute subsystems). Example: power capping.
9. FINE-GRAINED POWER MANAGEMENT: intelligent built-in power management. Examples: voltage and frequency governors, hibernation.
10. LOAD MIGRATION: shift computing loads to a different electrical grid.
11. JOB SCHEDULING: Job shifting or queuing (scheduling) has historically been used as a strategy for managing CPU utilization, but could also be used to manage the energy utilization of IT equipment.
12. BACK-UP SCHEDULING: Defer data storage processes to off-peak periods
13. SHUTDOWN: Graceful shutdown of idle HPC equipment loads. Usually applies when there is redundancy
14. LIGHTING CONTROL: With advance warning, datacenter lights could be shutdown completely.
15. TEMPERATURE ADJUSTMENT: Widen acceptable (ASHRAE Thermal Conditions) temperature setpoint ranges and humidity levels for short periods.
16. BACK-UP RESOURCES: Using generators and other electrical storage devices.
17. Are there any other strategies that you use to manage and control your total facility energy in response to a request from your energy/utility provider. Please describe.
18. Please evaluate as high, medium or low the MW impact of each of these strategies as a response to a grid request.
    - Power capping
    - Load migrations
    - Temperature adjustments
    - Clock speeds
    - Lighting control
    - Job scheduling
    - Back-up scheduling
    - Idle management
    - Shutdown
    - Back-up resources

Electrical Utility/Provider Information

19. PEAK SHEDDING: Utility provider arrangements used to reduce peak load, where the reduced load is not shifted to another time.
20. PEAK SHIFTING: Utility provider arrangements where the load during peak times is moved, typically to non-peak hours.
21. DYNAMIC PRICING: Time varying pricing arrangements 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.

22. GRID SCALE STORAGE: Methods used to store electricity on a large scale. Pumped-storage hydroelectricity is the largest-capacity form of grid energy storage.
23. RENEWABLES: Variability in the electric power generation from renewable resources and the methods used to respond to that variability.
24. 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.
25. 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, congestion), distinguished by who controls them and what they are used for.
26. 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. Or, methods used to resolve congestion that occurs when the distribution control system is overloaded. It generally results in deliveries that are held up or delayed.
27. Is there information you would like from your provider that you are not getting? If yes, please describe what you would like to know.
28. Is your provider asking for information from you that you are not able to provide? If yes, please describe what they are asking for.
29. Do you experience any power quality issues at your HPC facility? If yes, please describe.