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

Natalie Bates[1], Girish Ghatikar[2], Ghaleb Abdulla[3], Gregory A. Koenig[4], Sridutt Bhalachandra[5], Mehdi Sheikhalishahi[6], Tapasya Patki[7], Barry Rountree[3], Stephen Poole[4]

[1]Energy Efficient HPC Working Group
[2]Lawrence Berkeley National Laboratory
[3]Lawrence Livermore National Laboratory
[4]Oak Ridge National Laboratory
[5]University of North Carolina
[6]University of Calabria
[7]University of Arizona

Some of the largest Supercomputing Centers (SCs) in the United States are developing new relationships with their Electricity Service Providers (ESPs). These relationships, similar to other commercial and industrial partnerships, are driven by mutual interest to reduce energy costs and improve electrical grid reliability. While SCs are concerned about electricity quality, cost, environmental impact and availability, ESPs are concerned about electrical grid reliability, particularly in terms of energy consumption, peak power and power fluctuations. The power demand for SCs can be 20 MW or more—the theoretical peak power requirements are greater than 45 MW—and recurring intra-hour variability can exceed 8 MW.

This paper evaluates today's relationships, potential partnerships and possible integration between SCs and their ESPs. The paper uses feedback from a questionnaire submitted to supercomputer centers on the Top100 List in the United States to describe opportunities for overcoming the challenges to HPC-Grid integration.

1. INTRODUCTION

Supercomputing centers (SCs) with petascale[1] systems for high-performance computing (HPC) can have an outsized impact on their Electricity Service Providers (ESPs), with peak power demands exceeding 20 MW and instantaneous power fluctuations of up to 8 MW. As the HPC community moves towards exascale computing[2], we anticipate that a growing number of facilities will be reaching or exceeding these service levels, with significant potential effect on electrical grid reliability. 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 in order to manage this risk.

Being a "good citizen" on the electrical grid has several historical precedents. In the past, electrically-intensive industries such as aluminum 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 F and the residential power usage in the area surges. Other SCs are exploring the benefits of predicting hour-ahead and day-ahead use in concert with their ESPs. A mutual understanding of concerns between SCs and ESPs can produce a symbiotic relationship that goes beyond the current producer-consumer paradigm, paving the way for possible integration of SCs with the electrical grid. HPC-Grid Integration in the context of this study refers to the dynamic interaction and value between the demand-side resources (SCs) and the supply-side resources (ESPs) as well as the relationship between the electricity grid and its markets.

The Energy Efficient HPC Working Group (EE HPC WG) investigates opportunities for large supercomputing sites to integrate more closely with their ESPs. 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 distributed it to the Top100 SCs in the United States.

This paper leverages prior work on datacenter and grid integration opportunities done by Lawrence Berkeley National Laboratory's (LBNL) Demand Response Research Center [**?**]. This prior work describes the challenges and opportu-

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

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

SC14 '14 New Orleans, Louisiana USA

nities for datacenters and ESPs to interact with each other and how this integration can advance new market opportunities [**?**, **?**]. This integration model describes programs that are used by some of the ESPs to encourage particular responses by their customers and methods used to balance the electrical grid supply and demand. This is referred to as *Demand Response (DR)*.

Eleven sites responded to the aforementioned questionnaire. Based on these responses, we noted a few primary observations:

- Only 20% of SCs currently communicate with their ESPs about DR issues.
- SC managers believe that the candidate solutions most likely to be effective for responding to ESP requests involve coarse-grained power management techniques, job scheduling techniques, and shutting down computing resources.
- A stronger relationship, including DR capabilities, between SCs and ESPs can lead to both energy savings and cost savings over time, and in some cases such capabilities might become a requirement for large SCs located in energy-challenged locations.

One of the most straightforward ways that SCs can begin the process of engaging in integration is by participating in efforts to develop software infrastructure to manage their electricity requirements in a tightly coupled manner with their ESPs, facilitating both energy efficiency and grid reliability. In addition, this will provide for extensive funding and cost analysis and help the community base future requirements for SCs and ESPs on facts and a proven set of measurements.

Our analysis in this paper focuses on SCs in the United States. However, the findings can be extended to and may relate to SCs in other countries with similar practices. Electric grid infrastructure and market design are highly dependent on governmental regulations that vary across geographies. We restricted the initial analysis to the understanding of electricity markets in the United States. Future work can extend the analysis to the electricity markets in Europe and other geographies.

The paper is organized as follows. Sections 2 and 3 of this paper describe in greater detail the model for integrating SCs and the electrical grid. Section 3.1 reviews prior work in SC 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 **??** and the Appendix summarizes the survey questions.

2. ELECTRICITY SERVICE PROVIDERS AND HPC-GRID INTEGRATION

The EE HPC WG took as their starting point a model developed by LBNL's Demand Response Research Center. This model 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 HPC WG adopted this model to reflect the supercomputing environment focus (as opposed to the datacenter focus described by LBNL's Demand Response Research Center).

It is important to highlight the differences between SCs and datacenters. Unlike datacenters, SCs are more performance oriented, have significantly higher system utilization, and use little or no virtualization. Additionally, supercomputing applications are distinguished by their lack of geographical portability due to security concerns, data size and machine-specific optimizations.

We also note that SCs tend to be more energy efficient than datacenters. Power Usage Effectiveness (PUE) is a good measure for energy efficiency. PUE is the ratio of the total energy supplied for the facility to the amount of energy that actually reaches the IT infrastructure. A PUE of 1.0 is ideal. In our survey, none of the SCs exceeded a Power Usage Effectiveness (PUE) of 1.53, while the average PUE for a datacenter falls in the range of 1.91 and 2.9 [**?**].

2.1 Electricity Service Provider Methods and Programs

An ESP seeks to provide efficient and reliable generation, transmission, and distribution of electricity. *Methods* and *programs* employed by ESPs and their consumers are key to managing and balancing the supply and demand of electricity. While the *methods* describe how ESPs manage supply, the *programs* describe the activities that the ESPs can offer to their consumers in order to balance demand with supply.

Although critical to ESPs, 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.

ESP programs encourage customer responses to target both energy efficiency and real-time management of demand for electricity. Real-time programs can be *day-ahead* or *day-of*. Day-ahead programs refer to timescales of notification and responses from the customer that are determined based on advanced forecasting and capacity planning (for example, day-ahead 24-hour wholesale electricity prices). The programs are day-of are the ones when the notification and responses are based on same-day capacity planning or emergency response.

An example of an ESP program that encourages energy efficiency would be to provide home consumers a financial incentive for replacing single pane windows 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 the 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 (for example, operating reserves, 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.

- 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 SCs, the following example illustrates their potential relevance. The generation capacity requirements and response timescales vary across the country for ESPs 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—such as SCs 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 some of the largest SCs.

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 that 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" and "power-aware computing". The new model has been exemplified by the US Department of Energy issuing guidance that the first DOE exascale machine should not exceed 20 MW; effectively a $1000x$ performance improvement with only a $3x$ increase in power.

However, the problem is not as simple as provisioning 20 MW. 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 energy may also lead to cheaper but less predictable sources of power. In addition, as described in Section 2, ESPs may request a change in timing and/or magnitude of demand by SCs. To adapt to this new landscape, SCs may employ one or more strategies to control their electricity demand. We describe some of these strategies below.

- **Fine-grained Power Management** refers to the ability to control SC system power and energy with tools that offer high resolution control and can target specific low level sub-systems. A typical example is CPU voltage and frequency scaling.

- **Coarse-grained Power Management** also refers to the ability to control SC 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 SC 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 SC 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 SC equipment. It usually applies when there is redundancy.

- **Lighting Control** allows for datacenter lights to be shutdown completely.

- **Thermal Management** is widening temperature setpoint ranges and humidity levels for short periods.

These strategies can be used temporarily to modify loads in response to a request from an ESP. Additionally, some of these strategies could eventually be used at all times to improve overall energy efficiency if the SC sees no operational issues. 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 ESP request, but is not likely to improve energy efficiency (lowering aggregate energy use). Fine-grained power management, on the other hand, can be used at all times and is more likely to be used for improving overall energy efficiency, unless the strategy is specifically used in response to an ESP's request.

SC system power management has a very broad range of implementations and warrants greater exploration. For example, the coarse-grained and fine-grained strategies described above can be implemented at many levels of the system hierarchy—from node-level to site-level. We discuss these implementation approaches below.

- **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 a 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 components such as NVIDIA's Kepler GPGPU. Other efforts reduce DRAM power by optimizing 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. SC 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 degradation.

- **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 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

This paper pulls together several diverse research domains. In this section, we provide an overview of prior work in these areas.

3.1.1 Power Management

Processor power management can be divided into two distinct eras. First, with the introduction of Dynamic Voltage Frequency Scaling, users were able to change the CPU clock speed of their processors, lowering both voltage and, in most cases, energy: the workload used less power and ran longer, but the quadratic relation of power to frequency biased the results towards overall energy savings. Early work included several modeling efforts focused on the effects of CPU- and memory-boundedness on delay and energy in MPI programs [**?**, **?**, **?**, **?**, **?**]. This work led to the CPUMiser [**?**] and Jitter runtime systems, which were designed to maximize energy saving consistent with a user-specified delay [**?**]. Treating energy savings as an optimization problem led to a linear programming solution [**?**]. The follow-on Adagio runtime system slowed only computation that could be proven to be off the critical path, leading to significant energy savings with only negligible slowdown [**?**]. These techniques were also applied to non-MPI datacenter workloads [**?**].

Other power saving approaches were attempted that did not use DVFS, but most were not deemed relevant to the supercomputing environment. A notable exception is Dynamic Concurrency Throttling, where energy savings are realized by varying the number of threads at runtime [**?**, **?**, **?**, **?**].

The research landscape changed considerably with the introduction of Intel's Sandy Bridge processor. Turbo mode allowed higher clock frequencies to be reached so long as

fewer cores were in use, making for a nontrivial power- performance tradeoff calculation. The Running Average Power Limit (RAPL) technology provided an onboard power model that allowed the processor to both estimate power and, using rapid dithering of CPU clock frequencies, enforce a user-specified power bound across a short time window [**?**, **?**]. For the first time, users were able to ask questions about performance under power bounds. This new capability arrived concurrent with Department of Energy guideline that exascale machines would be subject to power (as opposed to energy) bounds.

Initial work showed that while processor performance at a fixed frequency was reproducible across processors, execution in turbo mode or under a power bound revealed significant performance variation [**?**]. Further work demonstrated a $2x$ performance improvement between conservative and optimal processor configurations while executing under a power bound [**?**].

3.1.2 Thermal Management

Thermal management is a key driver for improving energy efficiency of datacenters as well as SCs. 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) datacenter Technical Committee TC9.9 published guidelines that expanded the environmental range for datacenters and SCs [**?**]. 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 SC might be able to increase temperature as a response to a request from an ESP. The guideline defines both recommended and allowable environmental ranges. It also specifies a maximum rate of change, which is most stringent for tape drives. For SCs, 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 SCs 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. At this point, however, the guidelines do not document rate of change for liquid temperature. Although it is not 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 [**?**] describe field studies on using thermal management as a grid-integration strategy. They demonstrate increasing "facility 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 [**?**] 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 [**?**] designed a runtime system that does temperature-aware load balancing in datacenters using DVFS and task migration. They also discussed how hotspots could be avoided in datacenters, 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* [**?**, **?**, **?**]. Backfilling improves system utilization by executing jobs with small resource requests out of order on idle nodes.

Fan et al. [**?**] discussed power-aware job scheduling in the datacenter 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 a 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. [**?**, **?**, **?**, **?**] 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 [**?**, **?**] 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 [?] discussed an 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. [?] looked at two applied cases for distributed datacenters. 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 Datacenter Participation in Smart Grid Programs

Aikema et. al [?] 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 [?, ?]. 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.

Chen et al. [?] studied the potential of datacenter participation in the demand side regulation services. They proposed a dynamic control policy that modulates the datacenter power consumption in response to independent service operator (ISO) requests by leveraging server power capping techniques and various server power states. Results show that datacenters can decrease their energy costs around 50% by providing regulation service reserves, without a major deterioration in quality of service.

Liu et al. [?] introduced a way to reduce cost and environmental impacts using a holistic approach that integrates energy and cooling supply control with IT workload planning to improve the overall attainability of datacenter operations. The results demonstrated a reduction of the recurring power costs and the use of non-renewable energy by as much as 60% compared to existing techniques, while still meeting Service Level Agreements.

Aikema et al. [?] also analyzed a number of different potential advanced power markets for datacenters to participate in, and showed energy cost reduction by up to 12% with only a small impact on the quality of service provided to users. Ghamkhari et al. [?] built an analytical profit model to show that datacenters can noticeably increase their profit by participating in voluntary load reduction to offer ancillary services, and help the grid achieve better service quality and reliability.

4. QUESTIONNAIRE

We used a questionnaire to understand the current experiences of interaction between SCs and their ESPs. We restricted the analysis to sites in the United States because the results of the survey and practices of DR are are highly correlated and driven by energy policies in the country. [?].

Nineteen Top100 List sized sites in the United States were targeted for the questionnaire. Eleven sites responded—Oak Ridge National Laboratory (ORNL), Lawrence Livermore National Laboratory, Argonne National Laboratory (ANL), Los Alamos National Laboratory (LANL), Lawrence Berkeley National Laboratory (LBNL), Wright Patterson Air Force Base, National Oceanic Atmospheric Administration (NOAA), National Center for Supercomputing Applications (NSCA), 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 (One 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 NOAA).

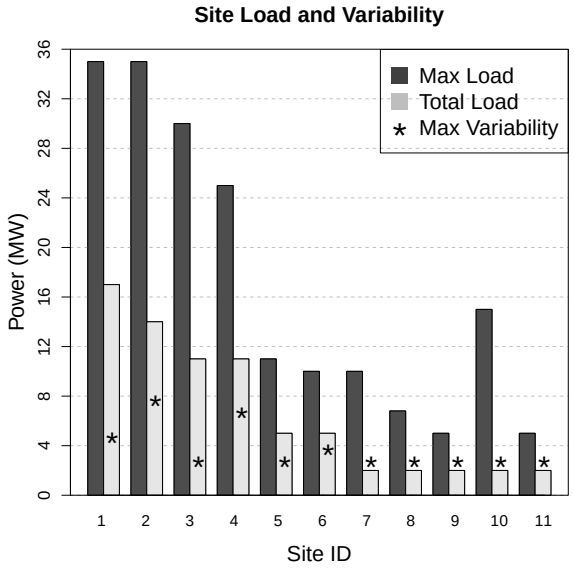

Figure 1: Site Load and Variability

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 10 MW, two sites with ˜5 MW total power load and five sites with less than 2 MW of total power load. For those with total power load greater than 10 MW, the intra-hour fluctuation (maximum variability) varied from less than 3 MW to 8 MW. One of ˜5 MW sites said that they experienced 4 MW variability. We chose less than 3 MW intra-hour variability as the bottom of the scale because we assumed that the ESPs would

not be affected by 3 MW (or less) fluctuations. The rest of the sites all reported less than 3 MW intra-hour fluctuation. Most of the intra-hour variability was due to preventative maintenance.

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 SCs. 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 ESP capacity constraints.

We asked if the SCs had talked to their ESPs about programs and methods used to balance the grid supply and demand of electricity (see Table 1). 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: Discussions with ESPs

Discussions with ESPs	% 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 SC.

Responding to pricing incentive programs is also not considered currently interesting by approximately half of the respondents, 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.

Only twenty percent of the respondents have had discussions with their ESPs about congestion, regulation and frequency response. LANL is one of the 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 ESPs 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 SC.

An open-ended question was posed as to whether or not there was information either requested of the SCs by their ESPs or, conversely, requested of the ESPs by the SCs. In both cases, well over 75% of the respondents answered no. 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 ESPs, which would provide automated additional detailed forecasting and ultimately real time data." LANL has also been requested to provide average "power projections, hour by hour, for at least a day in advance." Additionally, LANL has asked their ESP for 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)."

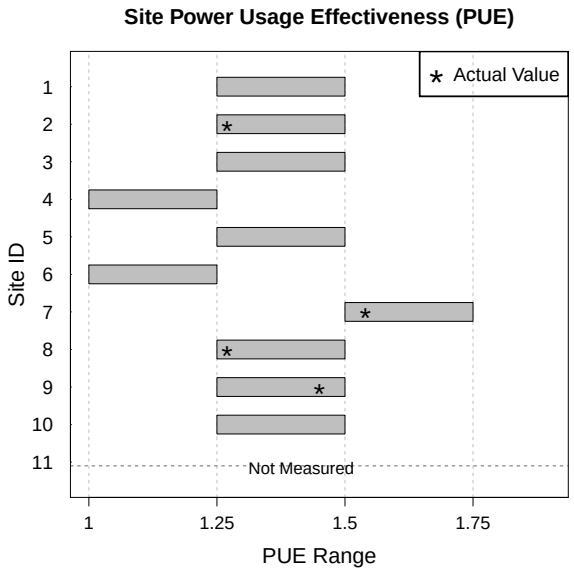

Figure 2: Site Power Usage Effectiveness

Given the low levels of current engagement between the ESPs and the SCs, it is not surprising that none of the SCs are currently using any power management strategies to respond to grid requests by their ESPs. 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. That is, 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

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

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 (see Table 2).

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 ESPs. 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 "took 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 point." 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 HPC 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 SCs and ESPs. For example, the responses from the two SCs with the largest power draws, LLNL and ORNL, diverge in several areas. This divergence is perhaps primarily due to characteristics of their respective ESPs. In contrast, SDSC stands out as a leader in integrating with their ESP on a site-wide level. To that end, the responses from SDSC may exemplify some of the opportunities available to other SCs that are willing to pursue this degree of integration.

The responses to the questionnaire also suggest that some ESPs are requesting that their SC 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 SCs of delivering as many uninterrupted computational cycles as possible to their users. In some extreme cases, SCs may not have a choice in the matter as the size and energy requirements of supercomputers increase; indeed, some ESPs may *require* large centers to develop a DR capability. However, a direct business case may exist to encourage SCs to develop this negotiation capability on their own. For example, if ESPs were to offer electricity at a significantly reduced rate on the condition that the SC customer develop DR capabilities, the long-term cost savings to the center could make undertaking such a project worthwhile.

Perhaps one of the most straightforward ways that SCs can begin the process of developing a DR 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 3.1 both strongly support the idea that the greatest opportunities for SCs 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 man-

agement (such as uniform processor power capping across the cluster), job scheduling, load migration, rescheduling backups, and fine-grained power management (such as dynamic, per-processor power capping).

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 SC. By attenuating resources, the SC manager can ensure that power consumption stays below some defined level. This defined level may be a pre-arranged power cap negotiated between the SC and the ESP and maintained on an ongoing basis, or, perhaps more likely, a power draw level that is requested by the ESP to handle unanticipated loads somewhere else in the ESP's system. Note 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 SC, which may allow significant power savings in the cooling system as well.

The coarse-grained power capping technique described above assumes that the SC environment has some amount of instrumentation and metering that allows for the collection of power telemetry data. This telemetry is necessary for the SC facility manager in order to understand how the power supplied by the ESP 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 supercomputing 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, SC 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 SC, 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 SC 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 pre-arranged blocks of time in which jobs with high power loads will run. These blocks of time could be negotiated with the ESP on an ongoing basis or could be in response to on-demand requests made by the ESP. 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 SC's resources under traditional circumstances, but they are also a vital piece of successfully implementing a DR capability for at least two reasons. First, they provide a set of metrics by which the SC can estimate the cost in terms of the "lost opportunity" of responding to an ESP's request to run with attenuated resources. Second, they allow the SC 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 DR scenario.

At a lower level, schedulers and runtime systems can exercise fine-grained, dynamic DR capability. 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 supercomputing 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, fine-grained power scheduling will be directed by the runtime system. For example, in the presence of load imbalance within a job, traditional applications may rely on periodically moving data around the allocated nodes to ensure all processors are performing a roughly equal amount of work. This load-balancing process is both time- and energy-intensive. By relocating power instead of data, processors with lighter loads can surrender power and run slower, allowing more heavily-loaded processors to use additional power to run faster. Combining both techniques should lead to improved execution time as well as more efficient power utilization.

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 SC. These scenarios are most appropriate when the supercomputing scenario contains a pervasively heterogeneous mix of computational resources. For example, many contemporary SCs 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 (that is, 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 (that is, 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 SC'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 DR capabilities.

Opportunities may also exist for SCs 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 SC 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 SC to cooperate with other centers would need to be structured differently than they are currently.

In a very broad sense, DR 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 SC. According to the questionnaire results presented in the previous Section, facilities managers view these approaches as the most likely candidates for creating effective DR capabilities.

Finally, this Section has focused heavily on the opportunities available to SCs that come from developing DR capabilities. This notion is primarily due to the fact that the questionnaire presented in Section 4 was distributed to SCs in the United States, not to ESPs. That said, opportunities do exist for ESPs that develop DR 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 ESPs. At a higher level, ESPs 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 DR strategy in conjunction with SCs, 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 ESPs and SCs with increased communication and engagement from both parties.

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

We focused our attention on the largest SCs in the United States. The two SCs 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 ESP. LLNL's experience is opposite to that of ORNL. Because of large swings in power usage, the LLNL SC was approached by their ESP with a request for daily predictable demand forecasts. That request began an ongoing relationship.

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

As previous research with datacenters has shown [**?**], SCs can serve as resources to the grid. To enable this, automation technologies and data communication standards, which can link the SCs 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 ESPs.

Nonetheless, the business case for the grid integration of SCs remains to be demonstrated. SCs have concerns that deploying these strategies might have an adverse impact on their primary mission. One of the key enablers for SCs to participate in electricity markets (for example, DR, electricity prices) is having markets that value their participation. In other areas like commercial buildings and select industrial facilities, benefits to both ESPs and customers are well documented. However, as the electrical grid and new dynamic loads such as SCs evolve, the markets need mechanisms to identify and provide value of participation (for example, 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 (for example, India).

We plan on following-up with the ESPs that support these US-based SCs. We note that this work's focus was from the perspective of the SC, 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 SC's load response. The trends in intra-hour fluctuation patterns need to be studied and analyzed further.

In addition, we need to understand communication components to be exchanged between SCs and ESPs. For instance, as part of this communication SCs may provide accounting data on their power usage; ESPs can use this data to forecast and model future energy usage of SCs for providing better power quality and power provisioning purposes.

Appendix

For the purposes of this paper, this appendix contains a summary of the questionnaire.

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.