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

Fname Lname[1], Fname Lname[2] Fname Lname, Fname Lname, Fname Lname, Fname M. Lname, and Fname Lname

[1] Institute Name, City State Zip, Country,
Name@email.adr,
WWW home page: http://web/page.html
[2] Another Institute, Institute Department, address1, address2, Country

Abstract. Some of the largest supercomputer centers in the United States are developing new relationships with their electricity providers. These relationships, similar to other commercial and industrial facilities, are driven by mutual interest to save energy costs and improve electricity grid reliability. Supercomputer centers are concerned about electricity price, quality, environmental impact and availability. Electricity providers are concerned about supercomputer center's impact on the electrical grid reliability, both for energy consumption, peak power and fluctuations in power. Supercomputer center power demand can be greater than 20 megawatts (MW) or more, theoretical peak power requirements greater than 45MW and re-occurring intra-hour variability can exceed 8MW. As a consequence, there are some supercomputer centers whose electricity providers are asking for hourly forecasts of power demand, a day in advance. This paper explores today's relationships, potential partnerships and possible integration between supercomputer centers and their electricity providers and its value. It develops a model for possible integration between supercomputer centers and the electrical grid. It then explores the utility of this model based on feedback from a questionnaire of Top 100 List sized supercomputer centers in the United States.

1 Introduction and Background

Supercomputer centers with petascale systems for high-performance computing (HPC) are realizing the large impact they will be putting on their energy service providers with peak power demands of 20MW and instantaneous power fluctuations of 8MW.

The Energy Efficient HPC Working Group (EE HPC WG) has been investigating opportunities for large supercomputer sites to more closely integrate with their energy service providers. This paper documents the results of this investigative activity.

Leveraging prior work on data center and grid integration opportunities done by Lawrence Berkeley National Laboratory's (LBNL) Demand Response Research Center (http://drrc.lbl.gov/publications), this paper takes as a starting point LBNL's model for integrating data centers and the electrical grid. The model describes programs that are used by the energy service providers to integrate with their customers (such as demand response) and methods used to balance the grid supply and demand of electricity. It also describes strategies that data centers might employ for managing their electricity and power requirements. This paper tuned this model's data center strategies for supercomputer centers. As opposed to data centers, supercomputer centers have very high system utilization and are not likely to use virtualization as a strategy. Also, supercomputer applications are generally not easily portable between geographic locations for a variety of reasons; security, data-locality, system tuning. Therefore, although included in the model for data centers, both virtualization and geographic load shifting were eliminated as potential supercomputer center strategies.

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

2 Supercomputer Centers and Electrical Grid Integration

The EE HPC WG Team took as their starting point a model developed by LBNL's Demand Response Research Center [3] that describes challenges and opportunities in which data centers and energy service providers may interact and how this integration can advance new market opportunities. This integration model describes programs that are used by the energy service providers to encourage particular behaviors by their customers and methods used to balance the grid supply and demand of electricity. It also describes strategies that data centers might employ for utility programs to manage their electricity and power requirements to lower costs and benefit from utility incentives. The EE HPC WG Team adopted this model with slight tweaks to reflect the HPC environment (versus the general data center).

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

2.1 Electricity Supply and Demand Side Management

Energy providers support demand-side management for both energy efficiency and to balance the electricity supply (e.g., peak summer afternoons). Demand side management programs are used to manage the electricity consumption on the consumers side of the meter. Demand side management can include energy conservation, energy efficiency and peak load management, and responses to changing supply conditions (e.g., demand response). Some of these consumer actions are in response to notifications (day-ahead or day-of) from the energy providers. The focus for this paper is on programs that are targeted at load management to improve electric grid reliability. Management of electricity supply are used to ensure the efficient and reliable generation, transmission and distribution of electricity. These methods by the energy provide are intended to address the changing demand conditions and any supply contingency management.

There are many different ways that the end-users of electricity or consumers can modify or change their electricity use. The following is a key list and brief definitions of key supply-side programs and demand-side management strategies that HPC data centers can utilize. Consumers' strategies are typically in response to support one or many of these supply-side programs.

- SUPPLY SIDE PROGRAMS
- System Peak Load: Programs designed for responses during peak hour events and focus on reducing peaks on forecasted high-system load days.
- Continuous Grid Reliability: These programs are designed for fastest, shortest duration responses. Response is only required during power system and intra-hour variability events that typically cannot be forecasted.
- Regulation Management (Up or Down): Programs designed to follow minute-to-minute commands from the grid to balance the aggregate system load and generation. These programs can be expanded to manage the variable and uncertain generation nature of many renewable resources, supplement the methods for grid scale storage, which is used to store electricity on a large scale to meet the system ramping (up or down) requirements. Pumped-storage hydroelectricity is currently one of the largest forms of grid energy storage. Such responses can also be used for frequency response, which are methods used to keep grid frequency constant and in-balance.
- Congestion Management: Methods used to resolve congestion that occurs when there is not enough transmission capability to support all requests for transmission services. Or, methods used to resolve congestion that occurs when the distribution control system is overloaded. Peak response can be expanded for congestion management.
- Dynamic Pricing: Day-ahead and day-off pricing programs that are designed to change dynamically in response to system load and generation conditions.
- DEMAND SIDE MANAGEMENT STRATEGIES
- Load Shed: Strategies used to reduce the load consumption during system peaks, where the reduced load is not used at a later time.
- Load Shift: Strategies where the load during peak times is moved to, typically, non-peak hours.

- Price Response: Strategies for time varying and real-time pricing programs used to motivate modification to electricity consumption.
- Continuous or Real-time response: Emerging strategies that modify the load consumption (increase or decrease) based on intra-hour changing supply conditions on daily basis.

To support the reliable functioning of today's electric grid and as it evolves to support more distributed genertation, its integration with and responses from HPC centers can be used to support better grid management.

2.2 Supercomputer Center Response Strategies

One dimension of the model to support better grid management in supply-side programs are demand-side management strategies that a supercomputer site might use to manage power in response to a request from their electric service provider.

Although these strategies can be used to temporarily modify loads in response to a request from an electric service provider, some of strategies them could also be used at all times to improve energy efficiency. It is the former that is of primary interest to this investigation - what HPC systems can do in response to a grid request that they cannot do all the time? Two examples may help to clarify this distinction. Temporary load migration is an example of a strategy that is well suited to responding to an electric service provider request, but is not likely to improve energy efficiency (lowering aggregate energy use). Fine grained power management at all times, on the other hand, is more likely to be used for improving energy efficiency, unless the strategy is specifically used in response to a service provider's request.

Below is a list of strategies:

- Fine grained power management refers to the ability to control HPC system power and energy with tools that are high resolution control and can target specific low level sub-systems. A typical example is voltage and frequency scaling of the Central Processing Unit (CPU).
- Coarse grained power management also refers to the ability to control HPC system power and energy, but contrasts with fine grained power management in that the resolution is low and it is generally done at a more aggregated level. A typical example is power capping.
- Load migration refers to temporarily shifting computing loads from an HPC system in one site to a system in another location that has stable power supply. This strategy can also be used in response to change in electricity prices.
- Job scheduling refers to the ability to control HPC system power by understanding the power profile of applications and queuing the applications based on those profiles.
- Back-up scheduling refers to deferring data storage processes to off-peak periods.

- Shutdown refers to a graceful shutdown of idle HPC equipment. It usually applies when there is redundancy.
- Lighting control allows for data center lights to be shutdown completely.
- Thermal management is widening temperature set-point ranges and humidity levels for short periods.

It should be noted that these strategies are intended to temorarily reduce HPC service levels witout impacting the facility operations or the equipment.

3 Prior Work

The prior work described in this paper addresses strategies that HPC centers can take to manage power. A lot of work has been done on energy efficiency, some of which has an element of power management; but, there is not a lot work that is specifically focused on data center power management (both cooling systems and IT equipment) in response to a request from an electrical service provider (cite Ghatikar et al., 2012a).

Most of the prior work mentioned in this section is focused on HPC systems. Since SC cooling requirement is a derivative of IT equipment, any strategy that lowers the amount of heat generated could be compensated by lowered cooling.

3.1 Power Management

DVFS and power capping are two popular ways to manage node power. Prior work in the HPC domain looked at analytical models to understand energy consumption [36,18,25] and at trading execution time for lower power/energy [2,20]. Several DVFS algorithms have also been proposed, such as CPUMiser [18] and Jitter [23]. Varma et al, 2003 [41] 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. [37] analyzed HPC applications under an energy bound. Rountree et al. used linear programming to find near-optimal energy savings without degrading performance [34] and implemented a runtime system based on this scheme [33].

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[22,35,38,39]. Other real-time approaches looked at saving energy [29,27,28,45,43].

In addition, there has been active research in the domain of virtual machines. Von Laszewski et al. [42] presented an efficient scheduling algorithm to allocate virtual machines in a DVFS-enabled cluster by dynamically scaling the supplied voltages. Dhiman et al. designed vGreen [8], which is a system for energy efficient computing in virtualized environments. They linked linking online workload characteristics to dynamic VM scheduling decisions and achieved better performance, energy efficiency and power balance in the system. Curtis-Maury et. al

[4,6,5] introduced Dynamic Concurrency Throttling, which is a technique to dynamically optimize for power and performance by varying the number of active threads in parallel codes.

Chip power measurement and capping techniques were initially introduced with the Running Average Power Limit (RAPL) interface on Intel Sandybridge processors [21,7]. In the HPC domain, Rountree et al. [32] 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. [31] used power capping techniques to demonstrate how hardware overprovisioning can improve HPC application performance under a global power bound significantly. Overprovisioning was also explored in the data center community [15].

3.2 Job Scheduling

The problem of scheduling jobs has been extensively studied. In general, most of the schedulers implement the First Come First Served (FCFS) policy as a simple but fair strategy for scheduling jobs. But this policy suffers from low system utilization. The most commonly used optimization is backfilling [26] [30] [14] [Lif95, Mua95, Fei04], which is proposed to improve system utilization. By identifying free capacities backfillin allows smaller jobs that fit those capacities to move forward and run on idle processors.

In [Yang ref is not in Zotero] [44] [Yang13] and [Zhou13] , **job scheduling** as a DR strategy and **dynamic pricing** as a grid integration program have been used to propose **a power-aware job scheduling** approach to reduce **electricity costs** *without degrading system utilization.* The novelty of the proposed job scheduling mechanism is its ability to take *the variation of the price of electricity* into consideration as a means to make better decisions of the timing of scheduling jobs with diverse power profiles. Experiments on an IBM Blue Gene/P and a cluster system as well as a case study on Argonne's 48-rack IBM Blue Gene/Q system have demonstrated the effectiveness of this scheduling approach. Preliminary results show a **23%** reduction in the cost of electricity for HPC systems.

Fan et al. [13] discussed power-aware job scheduling in the data center domain. They discussed implementation of power capping with a **power monitoring system** based on a power estimation method or direct power sensing, and a **power throttling mechanism**. Power throttling generally works best when there 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. They suggested that power consumption can be reduced simply by de-scheduling tasks or by using any available component-level power management knobs.

Etinski et al. [10,9,11,12] explored scheduling under a power budget in supercomputing and analyzed bounded slowdown of jobs. In their series of papers, they introduced three policies. Their first policy is based looks at current system utilization and uses DVFS during job launch time to meet a power bound. Their

second policy meets a bounded slowdown condition without exceeding a job-level power budget. Their third policy improves upon the former by analyzing job wait times and adding a reservation condition.

A grid computing infrastructure with large amount of computations normally contains parallel machines (a supercomputer cluster) as main computational resources. [16] [Fos01] Incoming jobs to Grid's local resources are scheduled by local scheduling system. Local scheduling system for parallel machines typically use batch queued space-sharing and its variants as scheduling policies. Most current local schedulers use backfilling strategies with FCFS queue-priority order as policy for parallel job scheduling.

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, which would be impossible without a start time guarantee. Furthermore, providing a time guarantee enables the job to be coordinated with other activities, essential for co-allocation and workflow applications. Advance reservation is a guarantee for the availability of a certain amount of resources to users and applications at specific times in the future [17] [Fos99]. The advance reservation feature requires local scheduling systems to support a reservation capability beside a batch-queued policy for local and normal jobs. In load migration, we encounter the need to deliver resources at specific times in order to accept jobs from other HPC centers to respond to their demand enforced by the electricity grid. This requirement can be achieved by advance reservations [17] [Fos99]. Modern resource management and scheduling systems such as Sun Grid Engine, PBS, OpenPBS, Torque, Maui, and Moab support backfilling and advance reservation capabilities.

3.3 Load Migration

In order to balance the electrical grid, [3] [Chiu12] proposes a low-cost **geographic load migration** to match electricity supply. In addition, authors present a real grid balancing problem experienced in the Pacific Northwest. They propose a symbiotic relationship between datacenters and electrical grid operators by showing that **mutual cost benefits** can be accessible.

(Ghatikar et al., 2012b: include citation) looks at two applied cases of distributed data centers. The results show that load migration is possible in both homogenous and heterogenous systems. Although the migration strategies were through a manual process, the responses could benefit from automation.

3.4 Thermal Management

Thermal and cooling metrics are becoming important metrics in scheduling and resource management of HPC centers. 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 reactively migrate/load-balance jobs from high temperature servers to servers with lower temperatures. ***T****

[24] [Kau12] takes a data-centric thermal- and energy-management approach and does proactive, thermal-aware file placement which allows cooling energy costs savings without performance trade-offs. T* is cognizant of the uneven thermal-profile of the servers, differences in their thermal-reliability-driven load thresholds, and differences in the data-semantics, i.e., computation job rates, sizes, and evolution life spans, of the big data placed in the cluster.

In this paper, we assume that the grid is a given constant as a fundamental property. But, grid integration solutions may take into consideration that it isn't a given as electrical grid infrastructures will evolve in the future [19] [He08] . Thus, changes in the grid could make grid integration more or less difficult.

In [1] [Aik11] , authors explored the potential for HPC centers to adapt to dynamic electrical prices, variation in carbon intensity within an electrical grid, or availability of local renewables. Through simulations experiments on workloads from the Parallel Workloads Archive alongside real-world pricing data, they demonstrate potential savings on the cost of electricity ranging typically between 10-50%. Nonetheless, adaptation to the variation in the electrical grid carbon intensity was not as successful, but adaptation to the availability of local renewables showed potential to significantly increase their use.

4 Questionnaire

We used a questionnaire to understand the current experiences of a supercomputer center's interaction with their energy service 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. [40] [Tor10].

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). Eight sites didn't respond (National Center for Atmospheric Research, IBM Corporation, National Energy Technology Laboratory, Indiana University, Texas Advanced Computing Center, Sandia National Laboratory, National Renewable Energy Laboratory, National Aeronautics and Space Administration). 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).

The total power load as well as the intra-hour fluctuation of these sites varied significantly. 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. We chose less than 3MW intra-hour variability as the bottom of the scale because we assumed that the electrical service providers would not be affected by 3MW (or less) fluctuations. 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. The rest of the sites were all less than 3MW. Most of the intra-hour variability was due to preventative maintenance.

Table 1. Caption Number 1

Total Load	**Variability**	**Frequency**
16-17MW	5MW	weekly
13-14MW	8MW	monthly
10-11MW	Less than 3MW	weekly
10-11MW	7MW	weekly
4-5MW	Less than 3MW	weekly
4-5MW	4MW	weekly
1-2MW	Less than 3MW	weekly
1-2MW	140kW	daily
1-2MW	Less than 3MW	yearly
1-2MW	200kW or less	daily
1-2MW	Less than 3MW	daily

We asked if the supercomputer centers had talked to their energy service 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 demand side and not supply-side programs.

Table 2. Caption Number 2

Discussions with Energy Providers	% Answered 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

More than half of the respondents are not interested in shedding or shifting load during peak demand. LANL reports that the ”technical feasibility” and

"business case has yet to be developed." For the sites where there is interest, shifting is more attractive than shedding load. SDSC is an exception to this trend because of a site-wide program. "University of California San Diego generates 30-35MW of power yet still imports 5-10MW. As a large generation source the utility providers see the campus as a highly attractive partner for offloading grid stress. Automatic load shedding is being explored/deployed today."

Responding to pricing incentive programs is also not considered interesting, 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 service 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 been many more respondents who have had discussions with their electricity service 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 "working on obtaining additional data from them and a means of sharing data between them and us" and has been requested to provide "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 service 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."

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.

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% .

Table 3. Summary of aspects and quality levels

HPC strategies for responding to Energy Provider requests (listed from highest to lowest interest + impact)	**% Interested**	**% High Impact**	**% Medium Impact**
Course grained power management	64	46	27
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

5 Opportunities/Solutions and Barriers

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

The responses to the questionnaire also suggest that some energy service providers are requesting that their supercomputer center customers develop ca-

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

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

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

On the other hand, the job scheduler has knowledge of and control over the computational resources within the supercomputer center, giving the job sched-

uler 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 datacenters 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). 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.

Several of the ideas described in the preceding paragraphs assume that the supercomputer center environment has some amount of instrumentation and metering that allows for the collection of power telemetry data. Not only is this telemetry necessary for the job scheduler to be able to make decisions about the workflow and resources it is to schedule, the telemetry is also important to the datacenter facility manager in order to understand how the power supplied by the energy service provider is distributed to resources within the center. 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, datacenter facility managers perceive this accounting data as distinct from the per-user or per-job accounting data described above and indicate that this data should be retained for electricity provisioning planning purposes. At a very high level, this detailed knowledge of where electricity is being used in a supercomputer center is an important piece in capabilities such as power capping, where the overall consumption of power is maintained at or below some maximum threshold level. Power capping may be accomplished either manually by the facility manager or automatically by the job scheduler, but both approaches require detailed knowledge that come from monitoring and accounting.

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). 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 capabili-

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

6 Conclusions and Next Steps

1. Potential HPC-specific value proposition for active DR engagement
2. Based on Grid Integration solutions – local and system-wide impacts
3. Next steps – specific directions or target areas to focus

In past, electricity providers have viewed the hourly, daily, and seasonal fluctuations of demand as facts of life. These fluctuations required additional generating capacity, particularly peaking plants that were needed only a few hours per year.

However, with increasing adoption of Smart Grid, information technology, communications, and a more dynamic and reslient grid, the value proposition of replacing the expensive and fossil-based peaking plants with more predictable demand with forecasting and demand response has the potential for new value proposition to the HPC data centers. The increase in renewable generation and its intermittency offers opportunities for large and flexible loads that can adapt to changing electricity generation and sources.

DR adoption: Automation: As the previous research as shown with data centers, HPC can also be considered as a resource to the grid and provide different services. To enable this, automation technologies, which can link the HPC data centers with the electric grid, and on-site power management strategies for different timescales for grid services will play a key role in ease of adoption and lowering the participation costs.

Electriicty markets: One of the key enabler for HPC data centers to participate in electriicty markets (e.g., demand response, electricity prices) is the markets that value their participation. In other areas of commercial buildings and select industrial facilities, benefits to both electricity service providers and customers are well documented. However, as the electric grid and new dynamic loads such as HPCs evlove, the markets need mechanisms to identify and provide

value of participation (e.g., cost, energy, carbon). (LIST SOME KEY FINDINGS FROM QUESTIONNAIRE)

value proposition, (INCLUDED IN THE ELECTRICITY MARKETS ABOVE)

the measurement and verification models, (LISTED BELOW)

patents, (NOT SURE WHAT THIS IS AND WHY THIS IS RELEVANT)

intellectual property (NOT SURE WHAT THIS IS AND WHY THIS IS RELEVANT)

lighting, and heating, ventilation, and air conditioning (HVAC) It is clear from the HPC operation that the largest opportunity for load management exists in the IT equipment. However, the previous work suggests that the opportunity in HVAC loads may also be an immediate opportunity. The opportunity depends on the ratio of cooling loads to the IT equipment load (e.g., PUE). Considering that the lighting loads are a small percentage of overall HPC load, their participation alongside other strategies (IT equipment and HVAC) may prove beneficial. The advancement in technologies and vendor recognition to provide more dynamic power management capabilities in the IT equipment offers larger opportunity in consolidated IT and cooling load reductions.

electricity-price markets (INCLUDED IN THE ELECTRICITY MARKETS ABOVE)

interoperability

The grid integration need to be standardized and provide interoperable interfaces to be interoperable. Interfaces, communication infrastructure, data, information exchange, agreement should be based on standards. Communication with grid providers need to be standardized. grid request/response messages. Requests include DR event, price, renewable generation

How is architected an accounting system (energy and utilization) of an HPC center? based on sensor systems like in [Hay09] . Sensor systems for an HPC center to report real time power consumption of various components such as cooling, compute systems, storage, networks, racks, etc.

[Hay09] S. Hay and A. Rice, "The case for apportionment," in Proceedings of the First ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Buildings, New York, NY, USA, 2009, pp. 13–18.

Apportioning the total energy consumption of a building or organization to individual users may provide incentives to make reductions. We explore how sensor systems installed in many buildings today can be used to apportion energy consumption between users. We investigate the differences between a number of possible policies to evaluate the case for apportionment based on energy and usage data collected over the course of a year. We also study the additional possibilities offered by more fine-grained data with reference to case studies for specific shared resources, and discuss the potential and challenges for future sensor systems in this area.

– If accounting data can be used to forecast and model future energy usage of an HPC center? so this can be communicated and be in integrated with electricity grid.

- If/how electricity grid providers can use energy and usage accounting data to plan electricity provisioning of an HPC center?
- user-specific accounting data versus workload-specific accounting data.
- accounting data in terms of HPC center components, cooling, systems, lighting, etc.

These are excellent questions. What you've outlined below is a set of value of real-time data (the term "accounting" confused me earlier) of energy and utilization for HPC systems. Some of these values are for EE and the rest is how the electric grid service providers can benefit from it. For example, telemetry data for wholesale DR markets and M&V.

M&V or Measurement and Verification refers to quantification of load shed that a particular load is participating in. Typically, there are many baseline methodologies that the utilities and ISOs use to calculate the amount of DR a particular load/facility is providing through real-time and day-ahead metered data. The metering and telemetry to provide the M&V is key in determining if a particular resource can participate in a DR market and validate its performance for settlement (economics).

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):4852, Jan. 2012.
4. M. Curtis-Maury, F. Blagojevic, C. D. Antonopoulos, and D. S. Nikolopoulos. Prediction-based power-performance adaptation of multithreaded scientific codes. *IEEE Trans. Parallel Distrib. Syst.*, 19(10):1396–1410, Oct. 2008.
5. M. Curtis-Maury, J. Dzierwa, C. D. Antonopoulos, and D. S. Nikolopoulos. Online power-performance adaptation of multithreaded programs using hardware event-based prediction. In *International Conference on Supercomputing*, New York, NY, USA, 2006. ACM.
6. M. Curtis-Maury, A. Shah, F. Blagojevic, D. S. Nikolopoulos, B. R. de Supinski, and M. Schulz. Prediction models for multi-dimensional power-performance optimization on many cores. In *International Conference on Parallel Architectures and Compilation techniques*, New York, NY, USA, 2008. ACM.
7. H. David, E. Gorbatov, U. R. Hanebutte, R. Khanna, and C. Le. RAPL: Memory Power Estimation and Capping. In *Proceedings of the 16th ACM/IEEE international symposium on Low power electronics and design*, ISLPED '10, pages 189–194, New York, NY, USA, 2010. ACM.
8. G. Dhiman, G. Marchetti, and T. Rosing. vGreen: a system for energy efficient computing in virtualized environments. In *Proceedings of the 14th ACM/IEEE international symposium on Low power electronics and design*, ISLPED '09, page 243248, New York, NY, USA, 2009. ACM.
9. M. Etinski, J. Corbalan, J. Labarta, and M. Valero. Optimizing Job Performance Under a Given Power Constraint in HPC Centers. In *Green Computing Conference*, pages 257–267, 2010.
10. M. Etinski, J. Corbalan, J. Labarta, and M. Valero. Utilization driven power-aware parallel job scheduling. *Computer Science - R&D*, 25(3-4):207–216, 2010.
11. M. Etinski, J. Corbalan, J. Labarta, and M. Valero. Linear Programming Based Parallel Job Scheduling for Power Constrained Systems. In *International Conference on High Performance Computing and Simulation*, pages 72–80, 2011.
12. M. Etinski, J. Corbalan, J. Labarta, and M. Valero. Parallel job scheduling for power constrained hpc systems. *Parallel Computing*, 38(12):615–630, Dec. 2012.
13. X. Fan, W.-D. Weber, and L. A. Barroso. Power provisioning for a warehouse-sized computer. In *The 34th ACM International Symposium on Computer Architecture*, 2007.
14. D. G. Feitelson, U. Schwiegelshohn, and L. Rudolph. Parallel job scheduling - a status report. In *In Lecture Notes in Computer Science*, page 116. Springer-Verlag, 2004.
15. M. E. Femal and V. W. Freeh. Safe overprovisioning: using power limits to increase aggregate throughput. In *International Conference on Power-Aware Computer Systems*, Dec 2005.
16. I. Foster. The anatomy of the grid: enabling scalable virtual organizations. In *First IEEE/ACM International Symposium on Cluster Computing and the Grid, 2001. Proceedings*, pages 6–7, 2001.

17. I. Foster, C. Kesselman, C. Lee, B. Lindell, K. Nahrstedt, and A. Roy. A distributed resource management architecture that supports advance reservations and co-allocation. In *1999 Seventh International Workshop on Quality of Service, 1999. IWQoS '99*, pages 27–36, 1999.
18. 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. M. M. He, E. M. Reutzel, X. Jiang, Y. H. Katz, S. R. S, and D. E. Culler. An architecture for local energy generation, distribution, and sharing. In *IEEE Energy2030 Conference Proceedings*, Atlanta, Georgia, USA, Nov. 2008.
20. C.-H. Hsu and W.-C. Feng. A power-aware run-time system for high-performance computing. In *Supercomputing*, Nov. 2005.
21. Intel. Intel-64 and IA-32 Architectures Software Developer's Manual, Volumes 3A and 3B: System Programming Guide. 2011.
22. 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.
23. N. Kappiah, V. W. Freeh, D. K. Lowenthal, and F. Pan. Exploiting slack time in power-aware, high-performance programs. In *Supercomputing*, Nov. 2005.
24. R. T. Kaushik and K. Nahrstedt. T*: a data-centric cooling energy costs reduction approach for big data analytics cloud. SC '12, page 52:152:11, Los Alamitos, CA, USA, 2012. IEEE Computer Society Press.
25. 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.
26. D. A. Lifka. The ANL/IBM SP scheduling system. In *In Job Scheduling Strategies for Parallel Processing*, page 295303. Springer-Verlag, 1995.
27. 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.
28. 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.
29. 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.
30. 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):529543, June 2001.
31. 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.
32. 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.
33. 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.
34. 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.

35. 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.
36. 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.
37. 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.
38. V. Swaminathan and K. Chakrabarty. Real-time task scheduling for energy-aware embedded systems. In *IEEE Real-Time Systems Symposium*, Nov. 2000.
39. 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.
40. J. Torriti, M. G. Hassan, and M. Leach. Demand response experience in europe: Policies, programmes and implementation. *Energy*, 35(4):1575–1583, Apr. 2010.
41. A. Varma, B. Ganesh, M. Sen, S. R. Choudhury, L. Srinivasan, and B. Jacob. A control-theoretic approach to dynamic voltage scheduling. In *Proceedings of the 2003 international conference on Compilers, architecture and synthesis for embedded systems*, CASES '03, page 255266, New York, NY, USA, 2003. ACM.
42. G. von Laszewski, L. Wang, A. Younge, and X. He. Power-aware scheduling of virtual machines in DVFS-enabled clusters. In *IEEE International Conference on Cluster Computing and Workshops, 2009. CLUSTER '09*, pages 1–10, 2009.
43. 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.
44. Z. Zhou, Z. Lan, W. Tang, and N. Desai. Reducing energy costs for IBM blue Gene/P via power-aware job scheduling.
45. 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.

7 Appendices

7.1 Background

Over the last few years, load growth, increases in intermittent generation, declining technology costs and increasing recognition of the importance of customer behavior in energy markets have brought about a change in the focus of DR in Europe [Tor10]. The long standing programs involving large industries, through interruptible tariffs and time of day pricing, have been increasingly complemented by programs aimed at commercial and residential customer groups.

[40] Tor10] examines experiences within European countries as well as at European Union (EU) level. While business programs, technical and economic potentials vary across Europe, there are common reasons as to why coordinated DR policies have been slow to emerge: the limited knowledge on DR energy saving capacities; high cost estimates for DR technologies and infrastructures;

and policies focused on creating the conditions for liberalizing the EU energy markets.

advances in DR It describes initiatives, studies and policies of various European countries, with in-depth case studies of the UK, Italy and Spain.

Spees, K., & Lave, L. B. (2007). Demand Response and Electricity Market Efficiency. *The Electricity Journal*, *20*(3), 69–85. doi:10.1016/j.tej.2007.01.006

Interruptible Programs represent 6.5% of peak power and Load Shedding Programs initiate automatic load shedding in emergency situations [30]. With Interruptible Programs participants are required to reduce their load to predefined values. With Load Shedding Programs utilities have the possibility to remotely shutdown participants' equipment at short notice. One significant difference between these two programs is that for Interruptible Programs participants who do not respond can face penalties.

Load Shedding Programs are divided into real time programs (without notice) and 15 min notice programs. The size ranges from 1200 MW for real time programs to 1750 MW for notice programs. Participants in these programs have to install and maintain Load Shedding Peripheral Units and will be compensated according to a non-market price defined in regulation. The size of curtailable power is of 10 MW for programs without notice and 3 MW for programs with notice.

Load forecasting is very important for power system operation and planning. Traditional load forecasting tools have limitations to reflect DR customer behaviors into load predictions. In [Zhou12], existing DR contracts are reviewed for both wholesale and retail markets. In this study, an illustrative example is provided to explore the impact of these contracts on load forecasting. In conclusion, a concept of proactive load forecasting based on contract types is proposed for forecasting loads in a smart grid environment.

Modern hardware components such as processor, memory, disk and network offer feature sets (Burd and Brodersen, 1995) to support energy aware operations. Exploiting these feature sets in order to be more energy efficient is a very important and challenging task in modeling cost/performance trade-offs, in designing algorithms, and in defining policies. Today processors offer two power-aware features, i.e., cpuidle and Dynamic Voltage and Frequency Scaling (DVFS). The cpuidle feature offers a number of CPU power states (C-states) in which they could reduce power when CPU is idle by closing some internal gates. The CPU C-states are C0, C1, ..., Cn. C0 is the normal working state where CPU will execute instruction, and C1, ...,Cn are sleeping state where CPU stops executing instruction and power down some internal components to save power. The DVFS is another power-saving method especially when CPUs are in load line, allowing quick adjustment to frequency/voltage upon demand in small interval. The key idea behind DVFS techniques is to dynamically scale the supply voltage level of the CPU so as to provide just-enough circuit speed to process the system workload, thereby reducing the energy consumption.

We received eleven responses from the following sites: LLNL, ANL, Intel, SDSC, Illinois, NOAA, ORNL, Purdue, AF, NERSC, LANL. The table below

shows the ranks of the responding organization in the top 100 list. 80% of the top 10, 60% of the top 50, and 30% of the top 100 participated in the survey.

Table 4. Participating organizations and their rank in the top 100.

Organization	ORNL	LLNL	ANL	LANL	NERSC	Purdue	AF	NOAA	Intel	SDSC	Illinois
Rank	2	3	5	22	24	28	40	48	71	102	

The survey had a total of 29 questions, eight of the questions were skipped 16 times by the 11 participants. This is only 5% of the total questions answered by the eleven participants. The graph below shows the distribution of skipped questions.

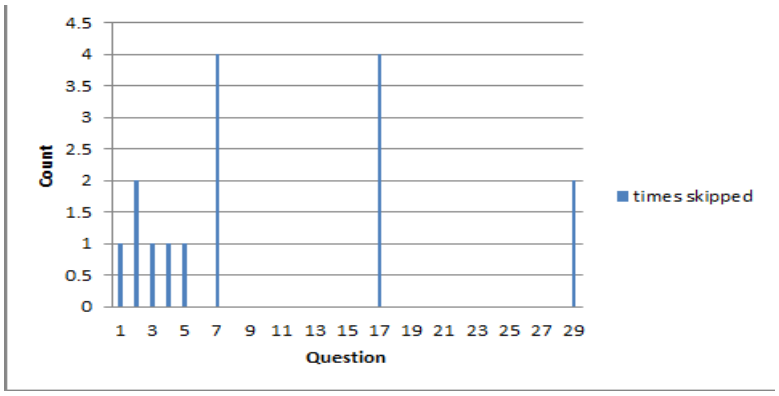

Fig. 1. Figure 2

Question 7 and 17 were skipped 4 times each, these two questions are follow up questions to questions 6 and 16 and we were trying to collect more details about user's responses in questions 6 and 16.