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 Lname,
Fname M. Lname, and Fname Lname

[1] Princeton University, Princeton NJ 08544, USA,
I.Ekeland@princeton.edu,
WWW home page: http://users/~iekeland/web/welcome.html
[2] Université de Paris-Sud, Laboratoire d'Analyse Numérique, Bâtiment 425,
F-91405 Orsay Cedex, France

Abstract. Some of the largest supercomputer centers in the United States are developing new relationships with their electricity providers. These relationships are driven by mutual interest. 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, both for energy consumption, peak power and fluctuations in power. Supercomputer center power demand can be greater than 20 megawatts (MW), 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. 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

Supercomputer centers with petascale systems for high-performance computing (HPC) are realizing the large impact they will be putting on their electricity service providers as they bring on (and perhaps turnoff or idle) megawatt scale (some double digit) supercomputers.

The Energy Efficient HPC Working Group (EE HPC WG) has been investigating opportunities for large supercomputing sites to more closely integrate with their electricity 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 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 electricity 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.

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 electricity 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

2 Supercomputing Centers and Electrical Grid Integration

The EE HPC WG Team took as their starting point a model developed by Lawrence Berkeley National Laboratory's Demand Reponse Research Center http://drrc.lbl.gov/publications that describes ways in which data centers and electricity service providers may interact. This model describes programs that are used by the electricity 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 managing their electricity and power requirements. The EE HPC WG Team adopted this model with slight tweaks to reflect the HPC environment (versus the general data center).

2.1 Electricity Provider Programs and Methods

Programs are used by the electricity service providers to encourage particular behaviors by their customers. Methods used to balance the grid supply and demand of electricity.

Below is a list of programs and methods:

- Energy Efficiency: Programs used to reduce overall electricity consumption, generally but not always at times of peak demand.
- Peak Shaving (shed): Programs used to reduce load during peak times, where the reduced load is not used at a later time.
- Peak Shaving (shift): 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.
- Grid Scale Storage: Methods used to store electricity on a large scale. Pumped-storage hydroelectricity is the largest-capacity form of grid energy storage.
- Renewable (off-site): Methods used to manage the variable uncertain generation nature of many renewable resources.
- Frequency response: Methods used to keep grid frequency constant and in-balance. Generators are typically used for frequency response.
- 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.
- Congestion: 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 distributio

2.2 Supercomputer Center Strageties

Another dimension of the model is a list of strategies that a supercomputer site might use for managing power in response to a request from their electric service provider.

Although these strategies can be used for managing power in response to a request from an electric service provider, many of them could also be used for improving energy efficiency. It is the former that is of primary interest to this investigation. Two examples may help to clarify this distinction. 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. Fine grained power management, on the other hand, is more likely to be used for improving energy efficiency than for responding to electric service provider requests.

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 CPU.
- Course 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.
- Job scheduling refers to the ability to control HPC system power by understanding the power profile of applications and queuing the applications based on those profiles.
- Back-up scheduling refers to deferring data storage processes to off-peak periods.

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

3 Prior Work

The prior work described in this paper is that which 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 power management in response to a request from an electrical service provider.

3.1 Fine Grained Power Management

In (Varma et al., 2003) system level DVFS techniques demonstrated. They monitor CPU utilization at regular intervals and then perform dynamic scaling based on their estimate of utilization for the next interval.

Authors in (von Laszewski et al., 2009) present an efficient scheduling algorithm to allocate virtual machines in a DVFS-enabled cluster by dynamically scaling the supplied voltages.

vGreen (Dhiman et al., 2009) is a system for energy efficient computing in virtualized environments by linking online workload characterization to dynamic VM scheduling decisions to achieve better performance, energy efficiency and power balance in the system.

3.2 Power Capping

Power capping techniques set a value below the actual peak power and preventing that number from being exceeded through some type of control loop [Fan07]. There are numerous ways to implement this, but they generally consist of a **power monitoring system** such as a power estimation method or one based on 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 data center is approaching **the power cap value**. Power consumption can be reduced simply by **de-scheduling tasks** or by using any available component-level power management **knobs**, such as DVFS [Fan07] .

3.3 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 backfill [Lif95, Mua95, Fei04]. Backfilling is proposed to improve the system utilization. Backfilling by identifying free capacities allows the smaller jobs fit those capacities to move forward and run on idle processors.

In [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 the system utilization.* The novelty of the proposed job scheduling mechanism is its ability to take *the variation of electricity price* into consideration as a means to make better decisions of the timing of scheduling jobs with diverse power profiles. Experimentations 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 electricity cost of HPC systems.

A grid computing infrastructure with large amount of computations normally contains parallel machines (supercomputers cluster) as main computational resources [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. In the US, supercomputer centers are connected via grid computing infrastructures such as TeraGrid, Open Science Grid. Grid computing's protocols, interfaces, and standards can facilitate the execution of DR strategies, as a result grid computing may increase the interest level and/or the impact level of DR strategies.

There are many use cases in 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 time guarantee enable the job to be coordinated with other activities, essential for co-allocation and workflows 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 [Fos99]. Advance reservation feature requires local scheduling systems to support a reservation capability beside 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 electricity grid. This requirement can be achieved by advance reservations [Fos99]. Modern resource management and scheduling systems such as Sun Grid Engine, PBS, OpenPBS, Torque, Maui, and Moab support backfilling and advance reservation capabilities.

By using *advance reservation capabilities* of schedulers (within local resource managers) of HPC centers, we facilitate *the execution of load migration strategy* between HPC centers (e.g., in terms of automation); as a result we increase the interest level and to some extent the impact level of load migration strategies.

3.4 Load Migration

In order to balance the electrical grid, [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 data centers and electrical grid operators by showing that **mutual cost benefits** can be accessible.

3.5 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**** [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 [He08] . Thus, changes in the grid could make grid integration more or less difficult.

In [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 Questionaire

We used a questionnaire in order to understand the current experiences of supercomputer centers with respect to interacting with their electricity service providers. We restricted the analysis to sites in the United States since the results of the survey and practices of demand response is highly correlated and driven by energy policies in the country. [Tor10].

Nineteen Top100 List sized sites in the United States were targeted for the questionnaire. Eleven sites responded (ORNL, LLNL, ANL, LANL, LBNL, WPAFB, NOAA, NCSA, SDSC, Purdue, and Intel) and eight sites didn't respond (NCAR, IBM, NETL, Indiana University, TACC, SNL, NREL, NASA).

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 NOAA).

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 that magnitude of fluctuation. 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.

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 electric 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 programs and not methods.

More than half of the respondents are not interested in shedding or shifting load during peak demand. There is some indication that this low interest is primarily due to the lack of a clear business case. For the sites where there is interest, shifting is more attractive than shedding load. SDSC is an exception to this trend, but because of a site-wide program. “UCSD 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

Discussions with Electricity Providers	% Answered Yes
Programs	
Shedding load during peak demand	54
Responding to pricing incentive programs	45
Shifting load during peak demand	36
Methods	
Enabling use of renewables	36
Congestion, Regulation, Frequency Response	18
Contributing to electrical grid storage	10

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

HPC strategies for responding to Electricity Provider requests (listed from highest to lowest interest + impact)
Course grained power management
Facility shutdown
Job scheduling
Load migration
Re-scheduling back-ups
Fine grained power management
Temperature control beyond ASHRAE limits
Turn off lighting
Use back-up resources (e.g., generators)

5 Opportunities/Solutions and Barriers

[koenig – 01-NOV-2013] These are things I think this section should include:

1. Link the beginning of this section to the end of the previous (Survey Results) section by suggesting that the biggest “opportuni ty” is to start a process of negotiation/interaction between utility providers and HPC centers; the survey data seems to indicate that this is being asked for by providers, at least in some small way already
2. opportunities related to system software
   - If negotiation starts happening between utility providers and HPC center operators, the system software (i.e., job schedu ler) is a key component in order to ensure that this happens as efficiently as possible in order to keep high utilization / business ut ility going at the HPC center (consider here things such as fluctuations in HPC use; e.g., things like large-scale acceptance / Top500 style runs)
   - just lowering the power consumption of a batch job (i.e., by using fine grained power management techniques) does not ens ure that the overall energy consumption is reduced; need some kind of knowledge about the workflow in the organization to make these ki nds of determinations
   - in addition to the advanced reservation capabilities discussed in the three paragraphs above, other areas where system so ftware can participate are
     - power capping

  - temperature adjustment within the datacenter (make reference to discussion on Page 6)
  - load migration - requires cooperation by HPC centers; also, migrating large datasets is hard
- – need to expose "knobs" into the system software so that HPC facility managers can easily adjust the objectives that the system software is using to make decisions because the overall number of ways of scheduling a workflow makes the problem too hard to r eadily solve by hand

3. if there is some kind of automated "negotiation" process that takes place between utility providers and HPC centers, it's likel y that the utility providers will need to improve their capabilities to be able to participate in this process (e.g., probably need to solve some kind of weighted optimization problem in near real time in order to know where the most important places are to ensure unint errupted service); there is an opportunity to the utility provider in this, however, in that these advancements in their technology for monitoring and adjusting their infrastructure might be leveraged toward other ends that are not related to HPC centers specifically

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

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.

DR adoption: automation

automation technologies

DR market:

value proposition,

the measurement and verification models,

patents,

intellectual property

lighting, and heating, ventilation, and air conditioning (HVAC)

electricity-price markets

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 organisation 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 egrated 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. He, M. M., Reutzel, E. M., Jiang, X., Katz, Y. H., S, S. R., Culler, D. E.: An Architecture for Local Energy Generation, Distribution, and Sharing. IEEE Energy2030 Conference Proceedings, Atlanta, Georgia, USA., (2008)
2. Foster, I.: The Anatomy of the Grid: Enabling Scalable Virtual Organizations. *First IEEE/ACM International Symposium on Cluster Computing and the Grid, 2001. Proceedings*, 6–7, 10.1109/CCGRID.2001.923162 (2001).
3. Lifka, David A.: The ANL/IBM SP Scheduling System. *In Job Scheduling Strategies for Parallel Processing*, 295–303. Springer-Verlag, 1995.
4. Mu'alem, Ahuva W., and Dror G. Feitelson: Utilization, Predictability, Workloads, and User Runtime Estimates in Scheduling the IBM SP2 with Backfilling. *IEEE Trans. Parallel Distrib. Syst.* 12, no. 6 (June 2001): 529–543. doi:10.1109/71.932708.
5. Feitelson, Dror G., Uwe Schwiegelshohn, and Larry Rudolph: Parallel Job Scheduling - A Status Report." *In Lecture Notes in Computer Science*, 1–16. Springer-Verlag, 2004.
6. X. Fan, W. Weber, and L. A. Barroso: Power Provisioning for a Warehouse-sized Computer. *In Proceedings of ISCA*, 2007.
7. Foster, I., C. Kesselman, C. Lee, B. Lindell, K. Nahrstedt, and A. Roy: A Distributed Resource Management Architecture That Supports Advance Reservations and Co-allocation. *1999 Seventh International Workshop on Quality of Service, 1999. IWQoS '99*, 27–36, 1999. doi:10.1109/IWQOS.1999.766475.
8. Czajkowski, K., I. Foster, and C. Kesselman: Resource Co-allocation in Computational Grids. *The Eighth International Symposium on High Performance Distributed Computing, 1999. Proceedings*, 219–228, 1999. doi:10.1109/HPDC.1999.805301.