The 4th Workshop on HPC Best Practices: Power Management

Held September 28–29, 2010, San Francisco

Workshop Report

March 2011

Compiled by Kim Cupps and Mary Zosel
Lawrence Livermore National Laboratory

LLNL-AR-472771

Workshop Steering Committee

Anna Maria Bailey (LLNL), Kathye Chavez (SNL), Susan Coghlan (ANL), David Cowley (PNNL), James Craw (NERSC), Kim Cupps, workshop chair (LLNL), Sander Lee (NNSA/ASC/DOE HQ), Dave Martinez (SNL), Nick Nagy (LANL), James Rogers (ORNL), Yukiko Sekine (SC/ASCR/DOE HQ), and Mary Zosel, host organizer (LLNL)

Workshop Group Chairs

Anna Maria Bailey (LLNL), Kathye Chavez (SNL), Susan Coghlan (ANL), David Cowley (PNNL), Jim Laros (SNL), Dave Martinez (SNL), Nick Nagy (LANL), Jim Rogers (ORNL), Mark Seager (LLNL), and Bill Tschudi (LBNL)

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

Contents

Executive Summary

At the request of the Department of Energy (DOE) Office of Science (SC), Lawrence Livermore National Laboratory (LLNL) hosted the *4th Workshop on HPC Best Practices: Power Management* workshop held September 28–29, 2010, in San Francisco.

The purpose of the workshop, which was sponsored by the SC/Advanced Scientific Computing Research (ASCR) and the National Nuclear Security Administration (NNSA)/Advanced Simulation and Computing (ASC), was to identify best practices related to power management at high performance computing (HPC) centers. Cost of power—and therefore, power management—has been identified as a key issue for future systems. This workshop addressed the current practices and issues related to controlling and reducing power required by facilities and systems, specifically, whether power challenges can be met by evolving current practices, facilities, and systems or if major new efforts must be undertaken now to prepare for the systems expected later in the decade. Participants from 19 HPC organizations, six HPC major platform vendors, and agency representatives from the DOE, the Department of Defense (DoD), and the National Science Foundation (NSF) attended the workshop. Overall workshop findings coalesced around the major themes of requirements, integrated system and facility monitoring, and future system and facility designs.

Several requirements issues were discussed. First, the system power utilization specification from vendors is generally much higher than is actually achieved on standard machine workloads. This power is plumbed to the machine and some fraction of it sits unused, yet captive, for the life of the machine. More discussion is needed to explore alternative approaches to reducing this problem. Second, there was agreement that vendors need to publish more accurate/realistic temperature limits. Many felt these limits were set arbitrarily low, eliminating efficient cooling technology options unnecessarily.

Integrated facility and system monitoring was the top area of concern for workshop participants, independent of which breakout track (facility/system software) the participant attended. None of the centers represented in the room felt they had made much progress on the issue and only a couple had begun projects to implement automated integration of facility and system monitoring. The concern was that as systems and facilities become larger and more complex, it becomes increasingly difficult to diagnose facility/system dependencies and effects. HPC leaders are all aware of these issues today—system hot spots, related localized failures—yet all of the diagnosis between facility data and system data is painstakingly manual. This manual correlation does not scale to petascale- and exascale-class machines. A final important finding was that workshop participants identified flexibility and expandability of both system and facility design as crucial to moving forward cost-effectively through the next decade. This implies that as much as possible, designs should allow for alternative cooling and power technologies that might be brought in later.

After two days of breakout sessions, themes from the preliminary reports were distributed and participants voted on those they thought most important. The results of the voting are contained in Appendix C. Results of Workshop Questionnaires on page 51 of this document. Finally, workshop participants identified future areas of power management collaboration, as listed in Final Workshop Session on page 9. It is expected that DOE will review the aforementioned topic areas and others and with input from the labs, decide whether any of the topics warrants a workshop in 2011.

Introduction

The *4th Workshop on HPC Best Practices: Power Management* workshop held September 28–29, 2010, at the Hotel Nikko in San Francisco, convened to assess current and emerging techniques, practices, and lessons learned for dealing with power requirements at HPC centers (HPCCs). Sponsored jointly by the DOE SC and NNSA and hosted by LLNL, the workshop was targeted at HPCC managers and key staff responsible for HPC facilities and system software. The areas of concentration for the workshop were facilities and system software. Eight breakout discussion topics were developed (four from each area) to address specific power management issues. The workshop was attended by 70 HPCC representatives invited from the HPC community of DOE, the NSF, the DoD, six HPCC representatives from Europe and Japan, and major HPC platform vendors. For a complete list of workshop attendees, see Appendix D. Workshop Attendees on page 54. The workshop steering committee, comprised of individuals from the major DOE computing centers and DOE headquarters (HQ) agreed on the abstract and specific goals for the workshop and provided the leadership for the breakout sessions.

The following abstract was submitted prior to the workshop:

> *Power management has been identified as a key issue for future systems. This workshop will address the current practices and issues related to controlling and reducing power required by facilities and systems. An important question is whether the power challenges can be met by evolving current practices, facilities and systems, or if major new efforts must be undertaken now to prepare for the systems expected later in the decade.*
>
> *This workshop is intended to facilitate collaborative progress on questions such as:*
>
> - *Planning and monitoring the various power aspects of HPC facilities*
> - *Metrics we (should) collect to improve our understanding*
> - *Power-aware reliability, availability, and serviceability (RAS) activities*
> - *Feasibility of power-down, or "sleep," of some system components*
> - *System software features needed to enable power conservation*
> - *Hardware features to expose*
> - *Improvements in power distribution and cooling configurations*
> - *Power-aware, system-wide scheduling techniques and incentives*

The *Power Management* workshop was a continuation of a series of workshops[1] that have been termed Best Practices. The first workshop was the Petascale System

[1] Links to the previous workshops may be found on the Power Management web site at https://outreach.scidac.gov/pmbp

Integration Workshop hosted by the National Energy Research Scientific Computing Center (NERSC)[2] in 2007. The second workshop addressed Risk Management Techniques and Practice and was hosted by LLNL in 2008. The third in the series was titled HPC Center Software Lifecycles and was hosted by NERSC in 2009. All of the workshops have been held at the Nikko hotel in San Francisco, CA.

Workshop Goals

The organizing committee agreed on the following goals for the workshop:

- Foster a shared understanding of power management issues in the context of HPCCs
- Identify top challenges and open issues
- Share best practices and lessons learned
- Establish communication paths for managerial and technical staff at multiple sites to continue discussion on these topics
- Discuss roles and benefits of HPCC stakeholders
- Present findings to DOE and other stakeholders

Workshop Format and Plenary Sessions

The Best Practices Power Management workshop agenda (see Appendix A. Workshop Agenda on page 16) was a combination of plenary sessions to provide an update on community activities and vendor perspectives, along with breakout sessions for detailed interactive discussion of different aspects of power management. The sections below describe the morning plenary sessions and the activities related to the breakout sessions.

Summary of Plenary Session 1[3]

The impetus for increased community interest in power consumed by computing systems comes from two major sources:

- The major HPC system/facility power requirements have been increasing during a heightened public awareness of energy efficiency
- The current estimates for the power requirements to house and operate exascale systems projected for the end of the decade are so high that the number of sites able to afford the power bill, to say nothing of the computing system, will be severely limited if the power trends cannot be altered

[2] Information about NERSC can be found at http://www.nersc.gov/.

[3] For the talks presented in this session, see https://outreach.scidac.gov/pmbp/

Mark Seager (LLNL) gave the featured talk[4] for plenary session 1: an overview of the planning efforts for exascale applications technologies for DOE mission needs. To understand the requirements for future systems, there have been a number of planning activities, both within DOE and internationally, to understand the applications of the systems, as well as the architecture. Two different technology paths (informally called swim lanes) have been proposed. Seager emphasized that there needs to be buyers at the end of the process. Energy efficiency efforts, especially as related to the system itself, are crucial to ensuring economic viability of the systems. To this end, early prototypes have been proposed. If a reduction in the system power can be achieved, it will have a dramatic impact on the community for the exascale systems.

Following Seager's talk, there were overviews from a number of different perspectives of the DOE and other community activities related to future systems and energy efficiency. Ken'ichi Itakura (JAMSTECH) described the architecture, facilities, and usage for the second Earth Simulator system (ES2), which began operation in 2009. Ladina Gilly (CSCS) gave an overview of the European Union Partnership for Advanced Computing in Europe (PRACE) organization and objectives. She also presented results from a recent survey of infrastructure planning, including some of the innovative activities related to facility power and cooling. In October 2010, the PRACE organizations held the second in a series of annual workshops to address the infrastructure facilities[5]. Natalie Bates (Energy Efficient HPC Working Group) gave an overview of the Energy Efficiency HPC working group activities[6]. For the final overview, Erich Strohmaier (LBNL) gave an overview of the effort to create energy efficient system metrics and their application in the Green 500 list, associated with the TOP500 computing systems. Comparisons between systems are often complicated by the different generations of processor technologies, as well as kinds of interconnects. For the top 10 systems, power efficiency has increased but not as much as the increase in power consumption, as measured by tracking Linpack performance/power consumption. Systems are getting bigger more quickly than the ability to improve power consumption, thereby increasing total cost of ownership (TCO). Linpack is one of the applications with the highest power consumption profile. There are outstanding questions about how one ought to model power consumption of the more general scientific workloads.

Summary of Plenary Session 2

Jim Rogers (ORNL) led the plenary session that featured a panel with presentations from system vendors Appro, Cray, HP, IBM, and SGI to address novel and emerging methods for managing the significant heat loads of the increasingly dense HPC server designs. Most of the companies discussed current power management trends and their own product lines and innovations in power management. John Lee, leader of Appro's hardware product development engineering team, represented the commodity space and discussed both increases in rack density and the challenges of air-cooling techniques.

[4] Presentation can be found at https://outreach.scidac.gov/pmbp/seager.pdf

[5] See http://www-hpc.cea.fr/en/events/Workshop-HPC-2010.htm.

[6] See http://eehpcwg.lbl.gov.

Doug Kelly, leader of Cray's mechanical design team, addressed the question of reducing non-essential power loads, including configuration and efficiency changes. Alan Goodrum, Fellow in HP's Industry Standard Server division and involved with architectural and technology planning, discussed opportunities for leveraging the commercial market. A key issue for HPC is the rising central processing unit (CPU) utilization rates that are higher than the commercial space. HP is addressing the facility issues by providing modular pod facilities that can be expanded as needed. Mike Ellsworth, a senior technical staff member in the Advanced Thermal Laboratory for IBM, emphasized the greater efficiencies achieved with water-cooling. Tim McCann, chief engineer with SGI, gave an overview of options for both air and water-cooled products and the use of their modular data center product to reduce TCO.

Final Workshop Session

In the final workshop session, the issues and findings reported by the breakout sessions were listed and presented for a vote to identify the top findings, as identified by the workshop attendees. The results of the vote are given in Appendix C. Results of Workshop Questionnaire on page 51.

The final session also addressed opportunities for further collaborations. There was general agreement that facilities should share their experiences with modifying cooling temperatures. Among the collaborations opportunities identified, several found potential volunteers for follow on discussions. Jim Laros (SNL) was interested in what hardware interfaces need to be available to extract monitoring information. Natalie Bates (Energy Efficient HPC Working Group) would like to work with a group to look at using the operating system (OS) to monitor and control power. She is also interested in an activity to define an energy efficient utilization metric that is more meaningful than power usage effectiveness (PUE).

The last topic discussed was the workshop's future. Yukiko Sekine (SC) emphasized that the workshops are held for the HPCCs and can continue as long as the centers find them useful. The participants suggested the following ideas as topics that might be considered for further workshops:

- Collaboration best practices
- Exascale best practices
- Monitoring best practices
- Input/output performance
- Benchmarking general purpose graphics processing units
- Performance tuning
- Many core in general, extreme parallelism
- Optimizing facility management
- Optimizing utilization/job scheduling
- Programming models (beyond message passing interface)

- Resiliency for large systems: hardware, system software, applications, file systems, correctness
- Cyber security: what are we doing, what should we be doing, what are the threats (inside and outside)

Workshop Breakout Topics and Crosscut Questions

The remainder of the workshop was organized around eight breakout sessions. The topics were chosen for a balance related to facilities-operations, planning, and system software. In addition to the specific topic of the breakout session, each session was asked to address a series of crosscut questions as an organizing factor for out-brief reports. The following topics were addressed on the first day:

- Facilities: power distribution and cooling configurations from facility to racks
- Facility metrics: metering and monitoring the computer center
- Power-aware OS features and scheduling
- Leveraging and encouraging power and cooling innovations in the commodity ecosystem

The topics addressed on the second day were:

- Power-related facility and equipment standards, rating, and certifications
- Alternative energy solutions
- Power-aware system monitoring
- Integrated (power-related) facility planning for system and network upgrades

The crosscut topics and/or questions given to each breakout sessions were:

- Experience: novel/interesting approaches (summarize the notable experiences/approaches that came up in the breakout discussion)
- Best practices (list things the breakout agrees can be called a “best practice”)
- Gaps looking forward to new systems (what are the major power-management-related challenges in this area)
- Evolve or start over for future systems (is there a natural evolutionary path for this area to support future systems or are there issues and projected requirements such that a complete new start is needed)
- Issues shared with large commercial centers (are the problems in this area shared by the large commercial centers, and are there opportunities for collaborations)
- Hardware/facility/system interfaces to influence (considering the things bought instead of built, what things can the DOE HPC community work together to influence so they better fit current and coming requirements)

- Status of (de facto) standards (are there standards (formal or de facto) that need to be improved/developed)
- Other key findings (did the group identify additional key findings/issues/action items)

Following each group of breakout sessions, the leads reported their findings back to the full group of attendees. They also provided a written report of their discussions. These summary presentations for each breakout are included on the workshop Web site[7] and the detailed written reports are in Appendix B. Breakout Sessions and Reports on page 18.

Workshop Findings

This section summarizes highlights from the breakout sessions as well as the voting done at the workshop. Workshop findings coalesced around the major themes of requirements, integrated system and facility monitoring, and future system and facility designs. Detailed written reports from each breakout session are in Appendix B. Breakout Sessions and Reports on page 18.

Requirements

In the area of requirements, several issues were discussed. First, the system power utilization specification from vendors was generally much higher than what was actually achieved on standard machine workloads. Session attendees noted that this power was provided to the machine and some fraction of it sat unused, yet captive, for the life of the machine. More discussion is needed to explore alternative approaches to reducing this problem. Next, there was agreement that vendors need to publish more accurate/realistic temperature limits. Many feel these limits are set arbitrarily low, eliminating efficient cooling technology options unnecessarily. Some operators further aggravate the situation by further overcooling the machine.

Other findings in the requirements area included:

- From the system software side, there are promising power measurement experiments to develop application power signatures and understand the tradeoffs between power reduction and performance; however, better collection interfaces are needed, especially to get data per node. HPCCs need the platform vendors to provide OS hooks to deliver power consumption to application software.
- The requirements to house the new systems make it highly beneficial to have facility and energy management personnel integrated into decision making about future systems and layouts.
- Power distribution and quality, transformer locations, and power safety issues are all a concern.
- The requirement for providing uninterruptible power supply (UPS) for facilities was questioned, with many sites reducing or eliminating UPS.

[7] See https://outreach.scidac.gov/pmbp.

- There currently is an effort to have existing power ratings and power requirements better defined for HPC facilities and systems. Energy parameters should also be articulated in request for proposals (RFPs) and request for quotes (RFQs). Requestors are interested in understanding potential energy gains from vendor offerings.
- An important step in adopting best practices should be to establish a comprehensive energy and environmental monitoring system.
- A recurring theme during the workshop was the need for more realistic temperature requirements and thresholds from system providers.
- Future power requirements should drive a movement to 480V three-phase and similar high-voltage circuits and/or use of 380V DC circuits.
- More "knobs" are needed in various system components to measure and monitor power usage, but these need standardization so that they are not completely different on every system.
- As more system and facility monitoring is performed, there should not be an expectation that all of the data is managed by a single point, but rather, it will be distributed. However, it is preferable to standardize as much as possible on interface mechanisms, otherwise each center spends their valuable dollars doing similar, overlapping work. A variety of standards and mandates for American Society of Heating, Refrigerating and Air-Conditioning Engineers (ASHRAE), fire protection, and reducing electrical intensity needs to be defined or reviewed for how they apply to HPC data centers. Some standardization and requirements are missing, such as emergency power off (EPO) systems and electrical safety.
- As monitoring data is increased, security standards for the data need to be applied. There is also a need for open standards for access to power data and a standard for aggregating related by disparate data (for example, weather data and power data). A common "dashboard" for reporting energy monitoring at individual sites and DOE headquarters would be valuable.
- Existing OS power standards such as ACPI and Intelligent Platform Management Interface (IPMI) are not really a basis for what HPCCs need.
- The PUE metric is used, but it should be improved. Better consistency in how PUE is determined is needed (for example, include fuel for standby generation). Other implications of the use of PUE need to consider shifting of power requirements from the facility to the computing equipment or vice versa (for example, eliminating one set of fans or consolidating power supplies). Overall efficiency could improve but not be reflected in PUE.

Integrated System and Facility Monitoring

Integrated facility and system monitoring was the top area of concern for workshop participants—independent of which breakout track (facility/system software) the participant attended. None of the centers represented in the room felt they had made much progress on the issue and only a couple had begun projects to implement automated integration of facility and system monitoring. The concern is that as systems and facilities

become larger and more complex, it will become increasingly difficult to diagnose facility/system dependencies and effects. HPCC managers are all aware of these issues today—system hot spots, related localized failures—yet all of the diagnosis between facility data and system data is painstakingly manual. This manual correlation does not scale to petascale- and exascale-class machines. One of the glaring holes in system monitoring is the inability to monitor power consumption at the chip level in almost all systems. Vendors must provide hooks to this information so that global information can be used to make good choices by all facets of the system and by facility software.

Other findings in the integrated system and facility monitoring area included:

- There is wide interest in monitoring and correlating data from multiple sources (for example, environment, mechanical data, and system data), but some experience has shown that correlation is difficult because of the wide variety in areas such as formats and sampling rates.
- In the future, monitoring sensors within the HPC could be used to control building systems, thus reducing or eliminating the need for separate facility and HPC monitoring.
- From the system software side, there are promising power measurement experiments to develop application power signatures and understand the tradeoffs between power reduction at performance, but better collection interfaces are needed—especially to get data per node.
- With respect to goals to use less power, a better understanding of power management costs is needed. Frequent power cycling of cooling equipment or system components may have an adverse impact on equipment lifetime. In addition, system operations to switch power modes may introduce operating system "noise" known as jitter that reduces application efficiency.

Future Facility and System Designs

A final important finding is that many workshop participants identified flexibility and expandability of both system and facility design as crucial to moving forward cost-effectively through the next decade. This implies that as much as possible, designs should allow for alternative cooling and power technologies that might be brought in later. Participants also noted that machine racks should not be packed so tightly that additional racks cannot easily be added later for expansion. Most participants indicated they were already doing this planning, and identified it as a best practice.

Other findings in the future facility and system designs area included:

- The majority of the HPCCs currently have raised floor but would consider a move to overhead air cooling and cable management in new data center designs. In both cases, possible congestion from the various types of pipes, conduits, and networking must be considered.
- Some data centers are experimenting with higher inlet/ambient temperatures but in one case, this resulted in the computers ramping up fan speed to maintain constant temperature. The energy implications of this need to be weighed with reduced cooling

plant energy use. There are also efforts to achieve better results with approaches such as hot and cold aisle containment.

- To achieve greater energy savings, non-conditioned power is considered an option.
- There are a variety of novel energy reuse ideas and experiences to mitigate high-power requirements, such as making heated water available for other community purposes, use of ground water cooling, and grey water reuse.
- A local climate study is an important planning tool to understand the strengths and weaknesses of the center location, for example, the impact of humidity and leveraging the outside environment for cooling. Additional planning activities should include the use of modeling techniques, such as a computational fluid dynamics (CFD) study of the projected layout. It is useful to involve a systems or energy engineer.
- Within the center, the hot and cold environments need to be contained. There may also be a requirement for a mixture of air-cooled systems with liquid-cooling solutions. Facilities are experimenting to find the real air and water temperature bounds and run as warm as possible.
- Automated monitoring and control systems, including wireless sensors, are recommended, as is sharing of monitoring data and experiences between sites (for example, thermal history experiences). The ability to obtain node data and current draw off of individual components and subsystems is a best practice that should be made available on all systems. Such data can help identify opportunities from both applications and system scheduling knowledge to reduce power.
- It is not surprising that many of the gaps identified for new systems center around better understanding of power, from proactive monitoring of external power quality with event notifications for transient voltage deviations to better interfaces for power monitoring at the node level with standard application programming interface (API) for power-related data. Methods for dynamic control of monitoring sampling rates are desired so that high-frequency sampling can be avoided. There should also be an increased effort to integrate platform-level information with facility information.
- Whether future systems can be architected to run in warmer environments, concerns about impact of changes in structural/weight constraints, and an issue of how increased network requirements coexist with the power and cooling without causing congestion were also identified.
- Clearly, both HPC and the commercial centers have a collective interest in running their centers more economically and hence collaboration opportunities exist based on previous experience. However, the large commercial centers tend to have different loads and requirements in terms of number of users, duration, and size of applications, so the priorities they have in driving change may be different from the HPC priorities.

Top Issues—Based on Questionnaire Voting

During the final session of the workshop, the attendees were given two lists of items identified by the breakout groups and invited to vote for those items they felt were most important. One list was related to system software (five votes per person) and the other list was related to the facilities (eight votes per person).

System Software Top Issues

- Create more interfaces to power measurement and control from systems (for tools and applications), especially high-power activities such as memory access and data movement. Currently, operating system interfaces that track system power usage are limited and the interfaces passed on for use by tools and applications are almost non-existent. With interfaces defined, there is opportunity for both the systems and applications to understand the power usage and to identify opportunities to take advantage of some low-power features the hardware may provide.
- Integrate facility and system management. In today's centers, there are a multitude of interfaces to monitor and manage the center effectively. Systems are constantly inspected for failures. What does not currently exist is an integration of these systems that ties specific temperature events in the facility to corresponding failures in the compute systems. This integration will be key to effectively managing multi-million core exascale systems in large facilities.
- Implement features for power savings during idle time. The most obvious source of power savings that can be introduced is to put the system into a power-saving mode when the system is idle, including periods when nodes are being held to make room for a large job.
- Establish a computing metric (*x* per watt). The Green 500 list has been established to sort out systems that use the least power to achieve Linpack performance numbers. There was interest in finding a better benchmark for the purposes of characterizing the relative power cost of systems. The HPC user group organized by FEMP is working to develop a metric.

Facilities Top Issues

- Implement more effective power connections such as higher voltage direct to computers
- Create metering and monitoring from rack to utility and correlate with system data
- Establish a smart design for the future; build flexibility and expandability into designs
- Publish more accurate/realistic temperature limits by vendors
- Implement wireless sensor networks in open and secure facilities

Appendix A. Workshop Agenda

Day 1: September 28, 2010

7:30–8:15	Registration and Continental Breakfast
8:15–8:30	Welcome: Kim Cupps (LLNL) and Yukiko Sekine (DOE SC)
8:30–9:00	The Exascale Initiative, Mark Seager (LLNL)
9:00–10:15	Overview of planning and activities: NNSA Facility Planning, Sander Lee (DOE NNSA) Office of Science Facility Planning, Dan Hitchcock (DOE SC) Facilities and Plan for the Japanese Earth Simulator II, Ken'ichi Itakura (JAMSTEC) European Activities, Ladina Gilly (CSCS) Energy Efficient HPC Working Group Activities, Natalie Bates (Energy Efficient HPC Working Group) Update on Green 500 Activity, Erich Strohmaier (LBNL)
10:15–10:20	Instructions for breakout sessions
10:20–10:45	*Break*
10:45–12:15	Day 1 breakouts: 1a: Facilities—Power distribution and cooling configurations from facility to racks 1b: Facility Metrics—Metering and monitoring the computer center 1c: Power-aware OS features and scheduling 1d: Leveraging and encouraging power and cooling innovations in the commodity ecosystem
12:15-1:15	*Lunch – Peninsula Room*
1:15–2:45	Day 1 breakouts (cont.)
2:45–3:15	*Break*
3:15–3:30	Report from Best Practices Third Workshop, David Skinner (LBNL)
3:30–5:30	Day 1 breakout reports and discussion
5:30–6:30	*Break before dinner*
6:30	Dinner—The Challenge of the Barcelona HPC Facility, Sergio Girona (BSC)

Day 2: September 29, 2010

7:30–8:15	Continental Breakfast
8:15–9:30	Plenary Panel: Unique Cooling Solutions for Dense HPC Systems, Jim Rogers (ORNL), chair Participants: John Lee (Appro), Doug Kelly (Cray), Alan Goodrum (HP), Mike Ellsworth (IBM), Tim McCann (SGI)
9:30–12:30	Day 2 breakouts: 2a: Power-related facility and equipment standards, ratings, and certifications 2b: Alternative energy solutions 2c: Power-aware system monitoring 2d: Integrated (power-related) facility planning for system and network upgrades
12:30–1:30	*Lunch – Peninsula Room*
1:30–3:30	Day 2 breakout reports and discussion
3:30–3:45	*Break*
3:45–4:45	Plenary workshop summary and next steps

Appendix B. Breakout Sessions and Reports

Session 1a: Power Distribution and Cooling Configurations from Facility to Racks

Session Leaders: David Martinez (SNL) and Kathye Chavez (SNL)

Participants: Helmut Breinlinger (Leibniz Supercomputing Centre), Jason Budd (ANL), Matt Campbell (San Diego Supercomputer Center), James Craw (LLNL/NERSC), Chris DePrater (LLNL), Thomas Durbin (NCSA, University of Illinois), Ladin Gilly (CSCS, Swiss National Supercomputing), Alan Goodrum (HP), Richard Griffin (ORNL/UT-Battelle), Mark Hartzell (PNNL), John Hutchings (LBNL), Doug Kelly (Cray), Peter Kulesza (ORNL), Michael Luzius (DOD), Justin Mann (Defense Department), John Parks (NASA/Ames Research Center), David Prucnal (DOD), Greg Rottman (DOD HPC Modernization Program), Bryan Webb (Pittsburgh Supercomputer Center), and Ryan Wright (PNNL)

Session Description:

This breakout session focused on the two main problems facing data centers in terms of delivering power and providing adequate cooling while also running an energy efficient data center. For power to the racks, discussion centered on 1) types of voltage, for example, supplying higher voltage directly to the racks, and 2) conditioned power (PDUs) versus non-conditioned power. The objective was to garner participant experiences with what has been done in the past, where data centers are today, and what is seen for the future. Subtopics included discussion of experiences with 1) power overhead versus under floor, and 2) different PDUs. For example, is a higher energy efficient transformer to prevent power losses worth the higher price? On the cooling side, the discussion focused on direct (water or refrigerant-based) versus indirect (air) cooling and the utilization of hot or cold aisle containment. Was utilization of tower water/air-side economizers effective in the data center and what was the limitation (if any) of kW to rack where it is no longer effective. This session also touched on the risk/reward associated with the costs of converting data centers from current state.

Session Process and Discussion:

Session was led into a general open discussion focused on the session abstract contents. Participants were invited to include any additional topics or concerns related to data center power and cooling.

Crosscut Topic 1: Experiences—Novel or Interesting Approaches

It would be novel to have segregated raised and non-raised floors for indirect and direct cooled machines. The majority of centers currently have 100% raised floors but would consider both in new data center designs. One caution in a non-raised floor space is the potential to overcrowd the ceiling with, for example, chilled water piping and conduit.

Facilities are employing techniques to utilize blended tower water and campus water. Almost everyone is trying to actively raise the temperature of the data center. In some data centers experience, the higher temperature resulted in the computers ramping up fan speed to maintain constant temperature. This poses the questions: 1) if air/water temperatures are raised but flow speeds need to be increased, what is the gain in terms of saved energy? Does anyone think data centers can go chiller-less?

Experience with overhead versus under-floor cabling has had mixed results. Under floor provides a cooling advantage and electrical code allows to run at 105% of the conductor rating. Reducing cabling under floor to power only offers some advantage from a safety perspective. Where all cabling is co-located, data centers have seen it advantageous to run power in conduit for separation and shielding. Three factors driving the decision are: 1) what else will be in same space (for example, chilled water piping), 2) how much air per cubic feet per minute has to be delivered, and 3) aesthetics.

For power configurations, some facilities have placed transformers outside to keep heat out of the data center, while others with transformers on the floor isolated the transformers with panels. Future scalability leads some data centers to place panels on exterior walls. A consideration for placement of critical equipment is harmonics. Different components in a chilled water plant (for example, variable frequency drives (VGDs), pumps, and motors) could be put on different transformers to alleviate harmonics issues. A harmonics baseline prior to compute system install would provide the data needed to make those decisions. To err on the side of safety, most data centers installed conditioned power because they did not know the shape of the power curve. Currently, the curve and vendor specifications are more defined, and it is accepted practice to run 480V under the floor. To achieve greater energy savings, non-conditioned power is considered an option.

The idea of geothermal heat to manage the room is being explored to reduce carbon emissions versus energy savings. For most data centers, the return on investment (ROI) does not support it.

Crosscut Topic 2: Best Practices

Operations may be optimized by taking gain cooling from the local climate and utilizing waste heat to preheat air for makeup or building water. In some locations, humidity may be a problem, thus making the data center too hot/dry and resulting in static issues. A local climate study is beneficial as a planning tool.

UPS should be offered only at an absolute minimum. UPS can create a false sense of "data security." If the cooling side goes down but racks remain on with UPS, a delayed system shutdown increases the risk of "frying" the machine. A best practice when designing a data center or upgrading an existing facility would be to consider a flywheel/UPS system.

Automated controls of building systems are an essential tool to management and scalability. CFD modeling is beneficial, however, piping done after the modeling, in most cases, changes the results. Three-dimensional modeling via scanners, radar, or GPS has had a favorable impact on data center operations in the areas of safety and planning. Modeling tools provide confidence in proposed solutions for layouts, designs, upgrades,

or additions. Visible marking of pipes, valves, and connections is a valuable and common practice.

Hot work is generally not permitted, so power has been configured to a mixture of outlets/transformers for change outs without losing power to equipment. The installation of PDUs/other equipment designed with hot-swappable components is increasing.

Data centers have deployed hot and/or cold aisle containment with a reasonable ROI, but success is linked to the volume of air required. Cooling is put exactly where needed, thus enabling some air handlers to be turned off; during a power loss, cooling can be pulled from a cool room. Fire codes may be a constraint.

Primary infrastructure must be in place so required secondary infrastructure (for example, PDUs) can readily be installed and/or powered on. Centers use modular connections, which enables them to swiftly and safely remove/deploy equipment and incorporate building tie-ins for current and projected systems.

Crosscut Topic 3: Gaps Looking Forward to New Systems

There is controversy regarding DC-power readiness. Use of DC has the potential to save energy due to minimization of transformations; however, a variety of equipment, industry, and safety regulations would have to be created and implemented. The session attendees felt that the power industry is using legacy equipment, which reduces the pressure on vendors to build better products.

HPCCs need to be able to more proactively monitor power quality and obtain real-time data on state of the grid. The Information Technology Industry Council (ITIC) curve is the power quality curve developed for networks but is now being applied to data centers. Power quality meters can measure if an event went outside the range or remained inside the curves. Power quality management would enable data center managers to find where problems occur and address the issues. However, power quality management is not typically done today due to cost constraints. Smart PDUs would be costly to deploy but would provide power usage at the breaker level.

Crosscut Topic 4: Evolve or Start Over for Future Systems

Should a new energy efficient data center be built or should existing data centers be retrofit to capture more energy efficiencies? For most facilities, the cost of a new data center is prohibitive. However, data center managers today have used new technologies and deployed existing equipment in new ways to attain energy efficient operations. Environmental factors inherent in the data center location also play a key role in determining what changes (for example, plate frame heat exchangers) to an existing center will produce the greatest efficiencies while yielding an acceptable ROI.

One option might be to tell vendors what power and cooling will be provided and require vendors to deliver a machine that runs efficiently in that environment. Most vendor specifications are extremely broad, and facilities end up designing to the specification high end. In reality, most equipment runs well below the high end of specifications. Vendors feel if the bid specifications are too restrictive, they will simply opt out.

Should an HPCC “go modular” versus building or maintaining fixed environments (static data centers)? The maturity of HPC may prevent this from being effective. A vendor

suggested that in the future, a "farm" of connected computers should be considered versus one large computer, as done today. Vendor specifications may prevent this from being a viable solution. One vendor may have a four-row configuration that works while another vendor may only have a two-row solution. This results in a center being tied to one solution. For blended data centers, the modular option may be difficult to implement and not render a good ROI.

Crosscut Topic 5: Issues Shared with Large Commercial Centers

What is the commonality in general terms? What is the future compute environment going to look like? For commercial data centers (for example, banking) UPS is a requirement. Currently, DOE's centers provide UPS for enterprise application, but, by and large, HPC has found raw power to be sufficient (power loss recovery is driven by data point restarts). For both environments, what can be done to make UPS more efficient and cost effective (UPS/generator or UPS/flywheel)?

Cooling is one factor that can be easily influenced to reduce energy consumption and/or costs. However, are centers just trading pump energy for fan energy? Cooling the entire volume of under-floor air has been the predominant method, but it is much too inefficient and does not address increasing machine densities. The response has been to use temperature sensors to control drives that modulate the air handlers in response to the need for cooling, but it will be offset if fans rev up to keep constant computer temperature.

Direct cooling appears to be the only feasible method in the near term. The data center community does not see any vendors leading the market nor do they see any cooperation to build machines to suit the facility. Third party vendors have come up with solutions such as direct cooled chips and cooling doors, but these all impact the facility infrastructure and are, generally, an "add-on" cost to the rack. Are there any benefits to room neutral solutions? The general community feels it is reasonable to ask that vendors package a rack (air or liquid cooled) that takes ambient air in and sends ambient air out (vendors should shoulder some responsibility to determine how efficient centers can operate). The vendor response is "we would be happy to do that if it was marketable."

Crosscut Topic 6: Hardware/Facility/System Interfaces to Influence

Future power requirements should drive a movement to 480V three-phase and similar high-voltage circuits. The general consensus is that there are a variety of standards both within the U.S. and European communities and many differences among those, but a unified standard would provide the opportunity to work towards common goals.

Most customers demand UPS power until asked to pay for it, then they can do without it. Data centers can see a roughly 30% increase in costs to keep UPS running (maintenance/cooling). There should be a push to ensure only critical equipment has a UPS feed, which possibly could lead to smaller UPS systems. The European community has had success with UPS/flywheel systems and, in general, UPS is strictly managed.

It is vital that chillers are run as efficiently as possible. Raising chilled-water temperatures and achieving higher delta T will help improve efficiency. Variable speed chillers, pumps and cooling tower fans can also help to improve chilled water plant efficiency. Chiller controls could be modified to allow for ride through when power

quality dips and allow auto restart to decrease wait time to bring machines back online. Can computer equipment be improved so that it can withstand a two-three-cycle event?

Crosscut Topic 7: Status of de Facto Standards

The standard may possibly become 480V to racks and utilization of 240V to disks and other equipment (maybe less copper/bulk). Industrial systems call for industrial power. Data centers will be better able to work towards common goals when one can determine common elements and ASHRAE standards and guidelines have been defined specifically for data centers. It is critical to run chiller plants as efficiency as possible; data centers must ensure they get as much capacity as possible. Think of the chiller plant as a processing plant and expect real-time efficiency. Most data center managers feel raw power has been reliable but still manage UPS because the information technology experts do not share that opinion. It is considered a standard to provide UPS; should this be reconsidered since UPS has such a big impact on costs and energy consumption?

Crosscut Topic 8: Other Key Findings

No further findings were identified.

Session 1b: Facility Metrics—Metering and Monitoring the Computer Center

Session Leaders: Nicholas Nagy (LANL) and Anna Maria Bailey (LLNL)

Participants: Bill Allcock (ANL), Marc Berman (PNNL), Susan Coghlan (ANL), Thomas Davis (LBNL), Sergi Girona (BSC), Sander Lee (NNSA HQ), Josip Loncaric (LANL), Richard Rivera (LANL), Jim Rogers (ORNL), David Skinner (LBNL), Erich Strohmaier (LBNL)

Session Description:

This breakout session focused on the information collected in the data center to improve its effectiveness and efficiency. PUE has become a "buzz word" in the industry and although it is simple in concept, it can be difficult to accurately measure. Total facility power is usually a straightforward measurement, but total computing equipment power can be much more difficult to accurately determine. Participants in this session discussed new and interesting approaches they are employing or developing at their sites, including their experience with various commercial products. The discussion included the participant's experience with air-side and water-side economizers, as well as temperature set points and humidity controls. Instrumentation and graphical displays were of considerable interest, and the cost trade-offs associated with improving PUE were considered. The discussion included how this technology will facilitate the integration of higher-density racks into the computing center and how the real-time data compares to the thermodynamic predictive models.

Session Process and Discussion:

The participants in this breakout session were all knowledgeable about the operation of the data centers at their site, and each had experience with various techniques to improve efficiency. Nagy and Bailey facilitated the discussion by asking questions about various

aspects of metering and monitoring at the individual sites and by guiding the discussion through the eight crosscutting topics.

Crosscut Topic 1: Experiences—Novel and Interesting Approaches

There was overwhelming agreement that hot-aisle and cold-aisle containment techniques were highly effective. These techniques are typically accomplished with a combination of chimneys, hoods, ducts, and/or vents that eliminate the mixing of the cold supply air with the hot return air. However, they can be expensive to implement.

One novel and interesting approach to this problem that has been employed at only a few sites thus far is to contain only the bottom seven feet of the cold aisle and thus eliminate the "wrap-around effect" of hot exhaust air being sucked into the supply side of the computer racks. This approach suggests that only a minimal amount of hot air will wrap over the top of the racks (because of the supply-side air pressure), and it saves the expense of containment hoods and the subsequent fire protection modifications. A modest frame with a Plexiglas door at each end of the computer rows could do the job.

Crosscut Topic 2: Best Practices

Another topic receiving general agreement among the break-out participants was the value of automated monitoring and control. Several sites reported that their monitoring systems had the ability to control environmental aspects of the computing room such as temperature and humidity. Such systems can start or turn off facility components such as chillers and air-handling units in response to an increase in the computing workload or the weather. Concern was expressed about the possibility of "hackers" being able to capture such a system and possibly alter the settings, but thus far, this has not been a problem.

Wireless temperature and pressure sensors within the data centers have become a useful tool for many of the represented sites. (One site in the DOE complex has recently received security approval to use the sensors in a classified environment.) The wireless feature allows the sensors to be deployed in a variety of locations within the data centers (for example, ceilings, top of racks, and under floor). The appropriate number of sensors and their placement has been the subject of several recent studies. This is clearly an innovation that data center managers will find extremely useful in helping them operate their centers in the future.

Many of the represented sites are now putting energy-saving requirements into their procurement documents (RFPs and RFQs). This forces the computing equipment vendors to be responsive. Requirements like "must be Energy-Star rated" are helping to drive down the power consumption within data centers and are already in use at several of the represented sites.

Crosscut Topic 3: Gaps Looking Forward to New Systems

Because this topic focused on "gaps," there were several suggestions, but the concepts are not well developed. Some of these ideas included:

- Power monitoring at the node level is often not available, but it may be useful for decision making (for example, scheduling).

- Better analytics are needed to enable monitoring (for example, trends, correlations, and model validation)
- A standard API for power-related data would greatly facilitate software development
- More analysis is needed to determine the effect of power quality (for example, sags, dropouts, and harmonics) on computer hardware performance

Crosscut Topic 4: Evolve or Start Over for Future Systems

The committee agreed that there is no need to "start over" with measurement capabilities, but it was recommended that sites evolve, expand, and work to integrate "lessons learned" from other industries.

Crosscut Topic 5: Issues Shared with Large Commercial Centers

Although many of the sites represented in this breakout session run HPC systems that employ unique architectures, the participants generally agreed that on significant overlap for many issues with large commercial data centers. Sites need better facility automation systems and increased facility availability. Running the centers more economically is important to both entities. Numerous collaborative opportunities exist in setting standards, raising the temperatures in the facilities without adversely impacting the hardware, improving cooling techniques (for example, hot aisle/cold aisle arrangements) and sensor placement. Security for building automation systems is a shared issue.

Crosscut Topic 6: Hardware/Facility/System Interfaces to Influence

Computer centers need to convince the vendor community to publish accurate and realistic temperature requirements for their products, including the optimal operating range, warning thresholds, and shutdown limits. In addition, the support equipment that manufacturers employ in their computer hardware (for example, sensors and fans) should be of the appropriate quality and design commensurate with the impact of a failure of such equipment.

Standards and protocols should be open and public. Metering and monitoring systems need to employ standards-based security, and committee members endorse improvements in interoperability of various data center systems (such as, utilities, hardware, and systems).

Crosscut Topic 7: Status of de Facto Standards

Many of the existing standards do not scale to the requirements of high performance data centers (the notable exception is ASHRAE, whose standards have been significantly broadened). Security is lacking in building automation and control networks (for example, modbus and BACnet), and applications should employ standard security systems such as PAM and LDAP.

Crosscut Topic 8: Other Key Findings

Some concern was expressed regarding various government energy savings plans as they relate to the data centers. Certainly, this type of legislation will be a catalyst for increased and improved metering and monitoring applications. However, with the potential of exascale computing becoming a reality within this decade, and if such machines actually do require 15 to 20 megawatts of power, the thought of reducing the amount of power in

the data centers is not realistic. Computer centers will probably be able to use that power more efficiently (more flops per watt), but total usage will not go down.

Session 1c: Power—Aware Operating System Features and Scheduling

Session Leaders: James Laros (SNL) and Marcus Epperson (SNL)

Participants: Natalie Bates (Energy Efficient HPC Working Group), Myra Branch (LANL), Kim Cupps (LLNL), Marcus Epperson (SNL), Jim Garlick (LLNL), Mark Grondona, EEHPCWG (LLNL), Michael Knobloch (Julich), Michael Lang (LANL), James Laros (SNL), Jacques Noe (CEA), Tisha Stacey (ANL), Mary Zosel (LLNL)

Session Description:

This breakout session focused on both hardware and software issues related to achieving power efficiency. Example issues included:

- Advanced Power Management (APM) features available on current and future architectures (frequency scaling, sleep/low power states, dynamic voltage transitions)
- Available OS interfaces to APM features
- OS techniques to leverage APM features (independent of applications)
- OS interfaces exposed to enable higher level exploitation of APM features
- OS abstraction of underlying APM features
- What, if any, features to expose directly to the application
- Power/performance trade-offs
- Power aware scheduling
- Scheduling benefits and impacts of power aware scheduling

These issues are largely interdependent and were considered from the system perspective. In addition, power efficiency issues and techniques necessary for HPC-class platforms likely differ greatly from commodity approaches developed for PC and enterprise-class platforms. The committee set as a goal to identify obstacles and opportunities specific to HPC in this emerging area.

Session Process and Discussion:

The breakout was organized to optimize discussion. This is a relatively new concentration for HPC and has quickly become one of the fundamental concerns for next-generation and exascale platforms. The session began with introductions and descriptions of current work and individual or site interest in the topic. The following is a brief summary of each member's introductory comments.

Jacques Noe discussed some of the currently deployed platforms at his site. Jacques mentioned an observation that the difference in power consumption on one of their platforms from running idle to running Linpack is approximately 1MW. Jacques is

interested in scheduling jobs with different power consumption characteristics such that total power consumption is controlled and or limited. One motivation is removing power spikes and unexpected fluctuations. If one part of the system is stopped, how does this affect facilities?

Michael Lang discussed his work in the application performance area and how it relates to systems software at large scale. He is interested in optimizing power/performance for large-scale systems. Michael gave a presentation to the group, which is included in its entirety in the breakout slide presentation.

Trisha Stacey discussed her role as lead systems and network administration and the effects/impacts changes in this area will have on her areas of responsibility. ANL is currently preparing for the next generation Blue Gene platform.

Michael Knobloch briefly discussed his broad interest in this topic. Michael gave the group a presentation later in the discussion, which is included in the breakout slides.

Kim Cupps discussed her interest related to her role as computing division leader at LLNL. She related power savings efforts to our "fiduciary responsibility" to use as little as possible to get the job done. Kim also expressed great interest in novel cooling and power techniques. Powering HPC platforms is becoming a huge percentage of their TCO. Kim also pointed out the importance of identifying what we can do versus what the vendors can contribute.

Mary Zosel discussed the increase in power budgets and interest in power aware scheduling. Will policy, priority, or power use decide what gets run and when? For example, if a site had a power score, what would be done with it and how would it affect behavior? She also expressed interest in what hooks exist in current architectures that have or have not been accessed with current OSs.

Myra Branch introduced herself as a team leader of system administrators for large cluster systems. She believes that her site will be dealing with power allocation and moving towards keeping applications running within a power budget, even to the extent of re-writing code.

Jim Garlick expressed his growing interest in the committee topic. His current work involves the cluster utility *powerman,* and he is interested in possibly expanding that utility to incorporate findings in this area.

Mark Grondona introduced himself as a colleague of Garlic's. Grondona works on systems software and is interested in this area specifically in exposing hooks to *slurm* and scheduling in general.

Marcus Epperson commented that power has become a recent interest during his involvement in the integration of the Red Sky cluster at SNL. Marcus pointed out the possible environmental restrictions of manipulating platform power from his experiences with the cooling range of the water-cooled Red Sky platform.

James Laros noted that he works in the area of leveraging architectural power saving features with operating systems and monitoring features of HPC platforms. Laros felt that it is important to socialize this topic and get on the same page. He is interested in whether different approaches will have to be taken on capability versus capacity platforms,

different OSs, and different approaches for different workloads. Laros gave a presentation of his past and current research at SNL.

The session moved immediately into project presentations (included in breakout slides) and discussion. The following are condensed unattributed thoughts resulting from the group discussion.

- Sites need to understand the affect of their actions on the system (including facilities).
- The information needed is not widely available on today's platforms, for example, voltage and current draw. Granularity and frequency of samples are also important.
- Contemporary work in the commercial sector focuses on laptop and enterprise-class systems. This approach could actually be detrimental to HPC. The effect could be equivalent to OS jitter.
- Some architectural features expected on new platforms could be detrimental, for example, the automatic reduction and increase of frequency per core based on micro-code heuristics. Sites need to be able to turn off features that hurt the site.
- Is there an acceptable power/performance trade-off? The consensus (unanimous) was yes, but the committee is not sure the application community has the same view. The committee felt this would be different if users were accountable for total costs rather than just node hours.
- Question posed: Could some applications run just as or almost as fast while saving power? Recent work at SNL seems to suggest yes. The less efficient a parallel application is, the more advantage there is to gain in power saving approaches.
- Power management on clusters seems to typically only be enabled during idle periods due to the affect on running applications.
- Some of the hardware that is presently capable of providing some of the information necessary was discussed. This included the Cray XT architecture, which seemed to have the most extensive capability. IBM Power architectures and Blue Gene were also discussed. It was pointed out that the collection capability on Blue Gene was not scalable.
- The necessary sampling rate was discussed. While more seemed better, one sample per second per sensor seemed to be a good target. Depending on the need, the frequency requirement would be different. The committee felt that if more could be done, less could also be done…if that was what was needed.
- Interconnect power efficiency was discussed. The committee did not have enough information but suggested that SerDes[8] spin constantly, whether used or not. Will future Network Interface Card (NIC) hardware have similar features to CPUs? Would it be possible to turn off channels/lanes when not in use? Would this be practical?

[8] A **Serializer/Deserializer** (**SerDes** pronounced sir-dees) is a pair of functional blocks commonly used in high-speed communications to compensate for limited input/output. These blocks convert data between serial data and parallel interfaces in each direction.

- The committee discussed the performance counter approach to measuring power efficiency. This is an interesting approach but effects at scale have not been validated. Can a combination of approaches be useful?
- Power consumption by an application seems to be significantly different at scale, which makes *in-situ* at scale monitoring more important.
- Where do hooks to control and monitor features belong—OS, middle-ware, application, or all levels?

Crosscut Topic 1: Experiences—Novel and Interesting Approaches

The committee felt at this early stage that most efforts were at least somewhat novel. It is important to continue to extend monitoring at the hardware level in an out-of-band manner to further this work. This ties in with "to have an affect you must see the effect" concept that appeared repeatedly during discussions. It was clear to the group that continuing to extend and make power and frequency manipulation features available to at least the OS level if not higher levels is a must.

Crosscut Topic 2: Best Practices

Sites need to find efficiencies and exploit them. There are savings during idle times, the so-called "low-hanging fruit," that needs to be picked. Savings during application execution is more challenging, with many factors to consider, but this is likely worth the extra effort. Application power signatures (SNL) is an example of a standard way of collecting and quantifying application power use (directly observe effect). The trade-off between performance and power, and implications of each, still needs to be decided. Who decides—policy, user, or other? The cost model should include power. Paying for what is used influences how it is used.

Crosscut Topic 3: Gaps Looking Forward to New Systems

The capability to monitor voltage and current draw (to name the two most important) on a per-component level was a re-occurring theme in the breakout. The committee also recognized that sites need to drastically increase the ability to control power management features whether sites wish to employ them or not (recall that turning them off might be critical to the usage model). The integration of platform-level information with facility-level information was also pointed out as important and currently rare if non-existent. Often these two areas evolve separately and in this case, seem more tightly tied. Understanding the overhead of power management is also critical—there is no free ride.

Crosscut Topic 4: Evolve or Start Over for Future Systems

It was again recognized that this area is just emerging. Linux and lightweight kernels can provide a basis for what sites need but might be inadequate. In the case of Linux, it might be poor for site needs (directed towards PC and enterprise). A lightweight kernel approach has the advantage of being lighter, making it easier to accomplish what sites need directly. Lightweight kernels are more deterministic. In the area of scheduling, there exists a basis for evolution. Some, but unfortunately few, platforms provide a basis for evolving hardware interfaces for monitoring. There is a growing list of architectural features for controlling power management but if and how they are exposed might be a problem.

Crosscut Topic 5: Issues Shared with Large Commercial Centers

The committee certainly shares the desire to take actions to be more power efficient with commercial centers. The question about the difference in approach between capability and capacity is important in this area. There is certainly overlap at the low level (chip architecture, for example). How to leverage or manage these architectural features might certainly diverge. In the area of monitoring, the areas likely overlap more. Sites might differ in required frequency of samples or scale but from the computing center perspective, companies like Google have as many or more devices to monitor. The potential that the overlap changes in time was recognized. This overlap could increase or decrease. There are certainly opportunities to collaborate with vendors in this area. It seemed that requirements were not orthogonal with their primary business, which is one of the obstacles sites have in influencing other architectural features. The committee saw this as an opportunity.

Crosscut Topic 6: Hardware/Facility/System Interfaces to Influence

The committee felt it important to influence the addition of "knobs" interfacing chip and component architectures and board designs. By knobs, the committee means ways to control and monitor any feature that is power-management related. It is also important to expose these knobs. The level of exposure was recognized to be hard to define and might be dependent on what the knob does. These knobs should be standardized, both control and monitoring. The software interfaces to knobs existing or in the future is largely unimplemented and sites stand a good chance of influencing or driving how this is accomplished. The HPC community could have a very positive impact on the development of these standards.

Crosscut Topic 7: Status of de Facto Standards

While Advanced Configuration and Power Interfaces Specifications (ACPI)[9] is somewhat of a standard, it was not clear that it provides the basis for what sites need going into the future. Again, it was designed for the PC and enterprise space and will likely not be realistic for exascale. The committee felt that there are many opportunities for standardization in this area, including OS and application interfaces (see Topic 6 above). Deciding what areas to expose will need to be a community discussion. Critical for HPC, any standards developed must consider scale. Many standards that sites currently leverage have been developed by "other" communities with no concept of HPC scaling needs.

Crosscut Topic 8: Other Key Findings

Other key findings were incorporated into previous topics, where appropriate.

References:

Topics on Measuring Real Power Usage on High Performance Computing Platforms, James H. Laros Kevin T. Pedretti, Suzanne M. Kelly, John P. Vandyke, Kurt B. Ferreira,

[9] For more information, see http://www.acpi.info, http://www.intel.com/technology/iapc/acpi/, and http://developer.amd.com/cpu/apml/Pages/default.aspx.

Courtenay T. Vaughan, Mark Swan, IEEE International Conference on Cluster Computing, September 2009.

Analysis of Dynamic Voltage Scaling for System Level Energy Management, Gaurav Dhiman, Kishore Kumar Pusukuri, Tajana Rosing, HotPower'08 Proceedings of the 2008 conference on power aware computing and systems.

Implications of Historical Trends in the Electrical Efficiency of Computing, Jonathan G. Koomey, Stephen Berard, Marla Sanchez, Henry Wong, IEEE Annals of the History of Computing, 2010.

Memory-aware Scheduling for Energy Efficiency on Multicore Processors, Andreas Merkel, Frank Bellosa, HotPower'08 Proceedings of the 2008 conference on power aware computing and systems.

Compiler-Directed Dynamic Voltage/Frequency Scheduling for Energy Reduction in Microprocessors, Chung-Hsing Hsu, Ulrich Kremer, Michael Hsiao, Proceedings of the 2001 international symposium on Low power electronics and design 2001.

Semantic-less Coordination of Power Management and Application Performance, Aman Kansal, Jie Liu, Abhishek Singh, Ripal Nathuji, Tarek Abdeizaher, ACM SIGOPS Operating Systems Review January 2010.

Energy-Efficient Processor Design Using Multiple Clock Domains with Dynamic Voltage and Frequency Scaling, Greg Semeraro, Grigorios Magklis, Rajeev Balasubramonian, David H. Albonesi, Sandhya Dwarkadas, and Michael L. Scott, Proceedings of the Eighth International Symposium on High-Performance Computer Architecture, 2002.

Power and Performance Trade-Offs in Contemporary DRAM System Designs for Multicore Processors, Hongzhon Zheng, Zhichun Zhu, IEEE Transactions on Computers, August 2010.

Empirical Analysis on Energy Efficiency of Flash-based SSDs, Euiseong Seo, Seon Yeong Park, Bhuvan Urgaonkar, Proceedings of the 2008 conference on power aware computing and systems, 2008.

Session 1d: Leveraging and Encouraging Power and Cooling Innovations in Progress the Commodity Ecosystem

Session Leaders: Mark Seager, lead (LLNL), Buddy Bland, co-lead (ORNL)

Participants: Bryan Biegel (NASA/Ames), Jeff Broughton (NERSC), Dave Cowley (PNNL), Pam Hamilton, note taker (LLNL), Ken'ichi Itakura (JAMSTEC), Anthony Kenisky (Appro), John Lee (Appro), Tim McCann (SGI), Michel McCoy (LLNL), Tommy Minyard (TACC), Bill Tschudi (LBNL), Ash Vadgama (AWE)

Session Description:

This breakout session reviewed current trends in power/cooling innovations in the commodity ecosystem. The session considered these trends and discussed the impacts to DOE HPC facilities planning.

Session Process and Discussion:

A formal presentation was made on industry trends, followed by open discussion stimulated by the presentation and the crosscutting topics.

Crosscut Topic 1: Experiences—Novel and Interesting Approaches

Several novel approaches were discussed, including:

- Use free air cooling. For example, 85% of the year LLNL could utilize free air cooling. For a $4.5M modification, free air cooling could be added to the LLNL computer facility and drop the PUE to 1.15 from 1.34; this is a very different direction from the modular/container trend. Containers are VERY expensive (upwards of $500k for site prep).
- Use DC power to the rack and in rack. The advantages include the ability to plug renewables in easier and a higher reliability. This also offers a unique opportunity to standardize globally on 380V. The same codes/regulations for AC power apply to DC. T12
- he disadvantage is cost of the DC infrastructure, which is not readily available.
- Use immersed cooling (mineral oil). The advantages include being able to over clock processors without a need for a raised floor. A disadvantage is that drives must be sealed.
- Remove fans and replace with conductive ribbon (Clustered Systems). This would require server manufacturers to adapt.
- Move away from raised floors. More facilities are moving away from raised floors.
- Minimize layers of fans down to one layer.

Crosscut Topic 2: Best Practices

The best practices discussed included utilizing the RFP process to ask industry for improved power efficiency and HPC-wide standards or guidelines. Facilities should also raise the temperature of inlet air and water with the goal of getting to the upper end of the ASHRAE range (80°F/27°C). This may lead to locating equipment with different temperature requirements in different rooms. Other related best practices included reuse of heat, free cooling (air and water), liquid cooling (rack or even down to the chip), and separating hot and cold air. Another best practice is to increase/improve power and cooling efficiency through higher level engineering. With regard to high-voltage power distribution, minimizing I*R drop and the cost of cables would be a best practice to consider.

Crosscut Topic 3: Gaps Looking Forward to New Systems

The gaps identified when looking forward to new systems included not having standards for HPC facilities and having each lab come up with their own requirements. A concern was raised regarding being careful with standards—that they do not restrict innovation.

HPC centers need to figure out how to leverage rack and higher level designs coming out of large data center deployments. That being said, HPC and large data centers are on different trajectories, for example, water (HPC) versus air (IT) and modular (IT) versus

consolidated facilities (HPC). It may be wise to think about broader guidelines rather than deciding a winner between, for example, liquid cooling versus containers.

HPC facilities need to develop ways to calculate component carbon footprint from cradle to grave, also taking into account any savings gained from recycling/reuse. The Industrial Technologies Program within DOE has a tool, DC Pro Software Tool Suite that HPC sites can use to identify and evaluate energy efficiency opportunities in data centers. Having HPC sites use common methodologies for evaluating energy efficiency would be beneficial.

A final gap identified was the raising of the ASHRAE temperature limits.

Crosscut Topic 4: Evolve or Start Over for Future Systems

Within the limitations of a facility, evolution is possible but a new facility or a major renovation is required for disruptive changes. Free air cooling is one example of an evolutionary approach.

Cross-Cut Topic 5: Issues Shared with Large Commercial Centers

Several issues were identified that HPC facilities share with large commercial centers including:

- Multi-MW data centers
- High power density configurations
- 1800 to 2500 watts per sq. ft.
- Very heavy floor loading
- Current cost of power plus uncertainty of future cost of power
- Use same component technology
- Security (cyber and physical)
- Capital, operating, and facility budgets are often separated, leading to miscalculations of TCO benefits

Crosscut Topic 6: Hardware/Facility/System Interfaces to Influence

Two interfaces identified to influence this topic were broadening humidity tolerances and raising the ASHRAE temperature limits.

Crosscut Topic 7: Status of de Facto Standards

No status of de facto standards were identified.

Crosscut Topic 8: Other Key Findings

How can industry be incentivized? Different ways to incentivize industry include developing an R&D agenda for power and cooling improvement and efficiencies. If the DOE HPC facilities can demonstrate big budgets, then this will garner the attention of industry. DOE could subsidize the R&D and collaborate on design/demonstrations. The follow-on to this idea would be to then align procurements with the R&D.

Session 2a: Power-Related Facility and Equipment Standards, Ratings, and Certifications

Session Leaders: Bill Tschudi, lead (LBNL), Bob Schroeder, co-lead (Glumac)

Participants: Jim Craw, note taker (LBNL), Natalie Bates, note taker (EE HPC WG), Buddy Bland (ORNL), Kathye Chavez (SNL), Chris DePrater (LLNL), Alan Goodrum (HP), Bryan Webb (PSC), Sam Graves (Glumac), Doug Kelly (Cray), Tom Durbin (NCSA)

Session Description:

This breakout session gave an overview of a number of standards, rating programs, training, and federal requirements that impact HPC. The breakout session considered the impact of these regulations on current HPC centers and projected the impact on facilities housing the next generation of systems. For example, energy-efficiency standards and Federal mandates are becoming more aggressively stringent while power and cooling requirements continue to grow. These mandates could be considered a barrier or an opportunity for a paradigm shift that could radically alter the way systems are designed and deployed.

Session Process and Discussion:

Formal presentations were made on the topics below, followed by open discussion stimulated by the presentations and the crosscutting topics.

- Federal requirements for data centers
- DOE programs (Save Energy Now and Federal Energy Management Program)
- EPA Energy Star for products (servers, storage, UPS) and for buildings (data centers)
- California Energy Commission
- ASHRAE standards, training, and publications
- The Green Grid
- LEED™ Certification for data centers
- Federal regulations for carbon measurement and carbon measurement tools

Crosscut Topic 1: Experiences—Novel and Interesting Approaches

The presentation and ensuing discussion focused on experience with novel and interesting approaches to standards, ratings, and certifications for energy efficiency.

LBNL proposed LEED™-type criteria for data centers to the U.S. Green Building Council (USGBC) because the current certification criteria were primarily for commercial office space and of limited utility to data centers.

There is collaboration between the Top500, Green500, EE HPC WG, and Green Grid to come up with a widely adopted standard metric for measuring supercomputer energy efficiency based on computational output.

There is a training course jointly developed by ASHRAE and the DOE's Save Energy Now program that provides information on energy efficiency strategies to improve data center energy performance. This course provides tools that are more for cross-organizational self-assessment than applying a standard, rating, or certification.

There was some discussion about the potential problems that might arise from driving PUE as a key metric for data center efficiency. One participant noted that there should be a focus on rewarding excellence. Over emphasis on metrics can cause irrational behavior, like rewarding improvements over excellence. A new metric was developed by the Green Grid organization with input from LBNL and that deals with the beneficial use of waste heat from data centers. This metric, termed Energy Reuse Effectiveness (ERE), is described in a Green Grid white paper.

Bill Tschudi noted that DOE's commitment to exceed minimum requirements for ASHRAE Standard 90.1/Energy Standard for Buildings may have helped to influence tightening of requirements in the 90.1 standard. DOE's goal is to be 30% better than this standard.

Crosscut Topic 2: Best Practices

Existing energy standards for buildings exclude data centers. At the June, 2010 ASHRAE meeting, the ASHRAE 90.1 standards committee voted to eliminate the exclusion for data centers; however, the proposed addendum to give guidance on how data centers should comply is being developed. In California, there is a building standard called Title 24, which similarly excluded data centers. There currently is a committee developing language for how to include data centers in the standard. However, building standards such as ASHRAE 90.1 or California's Title 24 do not represent best practices. They only set the minimum performance allowed by law. So every data center would need to comply with the standards and best practice would be significantly better than the standard. In the past, DOE has adopted goals to exceed the ASHRAE 90.1 standard by 30%.

Similarly, ASHRAE standard 127 provides methods for testing and rating computer room air conditioners and air handlers.

Local jurisdictions can adopt the ASHRAE standards in their building codes. Many, but not all, jurisdictions adopt the ASHRAE standards for their minimum performance requirements. DOE and the HPC community can adopt efficiency goals that far exceed the ASHRAE minimums through adoption of best practices.

The EPA Energy Star Program collects energy use data on products or buildings with the goal of awarding an Energy Star Label to the top 25% performers. EPA Energy Star currently ranks servers for some minimum requirements (for example, efficient power supplies or power management features). The goal of future server specifications from EPA is to include compute performance metrics (computational work/watt). Additional specifications covering storage and UPS are in development. Although these rating specifications are currently under development rather than fully defined and deployed, the HPC community can begin planning for including Energy Star requirements into future procurements. Procuring Energy Star Equipment does not necessarily ensure

achieving best practice energy consumption, but it can help to raise performance. DOE currently does require purchasing of Energy Star products.

The LEED™ rating system developed by the USGBC is gaining popularity in the commercial building market. Unfortunately, achieving a LEED™ certification does not necessarily ensure best practice energy performance in any type of building. For data centers in particular, the rating criteria are not specific to the most important aspects of HPC centers—energy use and water consumption. Alternative criteria that are more heavily weighted to these areas have been developed by LBNL and submitted for consideration by USGBC. However, achieving a certification under current LEED™ criteria does not correlate to best practices.

Industry and DOE have partnered to develop a certificate process to qualify energy practitioners to evaluate energy efficiency opportunities in existing data centers. The key objective of this certification is to raise the standards of those involved in energy assessments to accelerate energy savings in data centers by providing repeatability and credibility of recommendations.

A first step in adopting best practices should be to establish a comprehensive energy and environmental monitoring system—itself a best practice. This will enable establishing a baseline and the ability to track performance as best practices are implemented. Many HPC centers are implementing monitoring for this purpose. Chris DePrater (LLNL) mentioned that they are implementing an energy monitoring dashboard to provide a whole picture of LLNL's data center for measuring, verifying, and improving operations.

Crosscut Topic 3: Gaps Looking Forward to New Systems

It becomes increasingly compelling as exascale computing is approached to recognize and define supercomputer centers as scientific instruments rather than data centers. Should supercomputer centers' standards or ratings be different from other commercial/more standard data centers? There were several who felt that the mission of supercomputer centers is very different from the mission of data centers. For example, the Uptime Institute's focus on availability and reliability is not as critical for supercomputer centers. This provides an immediate energy efficiency advantage for HPC facilities by minimizing redundancy and back up requirements.

That segued into a discussion about an HPC-specific tier structure to capture energy efficiency best practices based on the mission. The HPC tier structure was envisioned to be a checklist of items (versus a single metric such as PUE) that collectively help guide the management team towards energy efficiency.

It was noted that DOE can influence industry to allow for a broader range of environmental conditions to cool the IT equipment and to expand temperature and humidity operating limits. This can directly benefit HPC centers and will also affect the data center industry at large. Generally, these requirements are established through recommendations developed by ASHRAE's data center committee, TC 9.9 in collaboration with the IT equipment manufacturers.

The procurement process and RFP requirements can help influence the HPC market. Collectively, DOE HPC centers represent a large market share—if there is a consistent message to the HPC manufacturers that efficiency is not only a part of the selection

process but a main requirement to bid, then the market will respond. A large group of end users requiring higher efficiency systems will help to drive the market.

There was a concern that Energy Star ratings for servers may not apply to HPC equipment, and that there should not be a requirement to buy only Energy Star rated servers if it keeps data centers from meeting their missions.

There was a general consensus that development of energy efficiency metrics (computing per watt) should be a priority on the R&D agenda. Development of metrics relating computing and energy use would enable differentiation of efficient computing platforms.

Crosscut Topic 4: Evolve or Start Over for Future Systems

There is a need to allow higher temperatures and wider humidity ranges for air-cooled equipment. Likewise, there is a need to encourage the use of higher maximum temperatures for liquid cooling and to improve the delta T. It was noted that ASHRAE could develop recommended and allowable ranges for liquid cooling. Following the workshop, Bill Tschudi asked the ASHRAE committee to consider liquid cooling temperature recommendations.

Crosscut Topic 5: Issues Shared with Large Commercial Centers

One of the big differences between Federal facilities and large commercial centers is that Federal regulations target Federal facilities. However, there are opportunities for very productive collaborations between Federal agencies and industry groups such as the Green Grid, ASHRAE, and the Silicon Valley Leadership Group.

Crosscut Topic 6: Hardware/Facility/System Interfaces to Influence

Some members felt that a "standard" or template request for information (RFI) that sets minimum energy performance or minimum supporting infrastructure capability could be very useful in this regard. If the vendors started seeing such a document from multiple customers, it would undoubtedly have a big impact. DOE could require this for DOE procurements.

Crosscut Topic 7: Status of de Facto Standards

High Performance Linpak is the de facto standard to measure absolute performance and has been extended to include energy efficiency. The measurement methodology is not consistent between sites using the metric. Many sites have not provided energy and power data for their machines. This needs to be resolved through development of common computational metrics.

PUE is a well-defined metric and the Green Grid organization has several white papers detailing its use. The new Energy Reuse metric will also be useful in quantifying the amount of energy that is reused.

Crosscut Topic 8: Other Key Findings

No further findings were identified.

Session 2b: Alternative Energy Solutions

Session Leaders: David Cowley (PNNL) and Marc Berman (PNNL)

Participants: Helmut Breinlinger (LRZ), Ladina Gilly (CSCS), Sam Graves (GLUMAC), Patricia Kovatch (NICS/UTK), Pete Kulesza (ORNL), Dave Martinez (SNL), Tommy Minyard (TACC), Dave Prucnal (DoD), Mark Seager (LLNL), Ash Vadgama (AWE)

Session Description:

How can HPC centers reduce cost and environmental impact by making creative use of local natural resources? Energy efficiency inside the data center is only part of the story. In keeping with the principle of reduce, reuse, recycle, HPC centers should be able to take advantage of local resources to increase efficiency either at new or existing locations. Are there creative ways to reduce PUE? Is a more meaningful way needed to express and measure the environmental effects of operating HPC centers? This session explored approaches such as sustainable energy sources, use of ambient external air or water temperatures, and reuse of "waste" heat.

Session Process and Discussion:

The session began with short informal presentations and discussions of ways to reuse waste heat from HPC centers or to generate power using local sustainable resources. Green Grid's proposed new Energy Reuse Factor (ERF) was presented and discussed briefly. The significance of ERF is that it can express reuse of energy that would otherwise be wasted, which the accepted PUE metric cannot do. Finally, the group fit its raw comments and ideas from discussion to the crosscutting questions.

Specific themes of discussion in this breakout session included:

- Ways to power and cool systems with minimal environmental impact and minimal energy use/waste
- Ways to take advantage of local natural resources (while observing the first point)

It is advantageous to locate an HPC center where there is cheap power or a hospitable climate. It is less obvious that there may be other local resources to draw upon, such as methane from a landfill that can be burned or aquifer water that can be used for cooling. It may be advantageous to use waste heat to warm nearby facilities. On the other hand, there may be serious local issues associated with operating HPC centers. One participant found it was politically unacceptable to exhaust heat into the air locally. Another found it was less expensive to bring in a supply of gas and burn it to power the facility than it was to buy electricity from the local power grid.

It is clear that there are no cut and dried approaches to alternative energy solutions that work across a majority of sites. Some common principles may be applied, but local decisions have to be made, taking into account the local situation at each site.

Crosscut Topic 1: Experiences—Novel and Interesting Approaches

Several principles were used to guide the breakout group's discussion, including attempting to reuse all waste heat, considering ways that heat could be used to generate

power, considering local sources of sustainable power, and considering local sources of cooling.

The breakout group considered multiple approaches:

- Reuse of waste data center heat to provide heat to other facilities or perhaps to generate electricity
- Recovery of methane from landfills or other local sources
- Use of solar or wind energy, possibly in conjunction with some form of energy storage system
- Use of geothermal sources for cooling, heat, or energy
- Use of cool outside air and local lakes and water sources
- Use of waste heat and/or cooling water to grow plants or algae as a potential fuel source
- Use of site resources to return electricity to the grid

Several issues were noted with these approaches. It is appealing at first glance to try to reuse waste HPC center heat for purposes such as generating power. However, the temperatures are generally too low for efficient conversion with today's technologies. Two possible solutions to that problem were suggested. The first was to further heat cooling water with solar energy, and the second was to raise temperatures (inlet and outlet) across the board so that wastewater would be in a useful temperature range for power generation. The first proposal clearly requires a more elaborate and expensive infrastructure, and it is not immediately clear that this is beneficial to HPC sites and funding agencies. The second proposal may be feasible if temperatures can be raised enough. This requires vendors to engineer systems so that they can operate at elevated inlet and ambient temperatures. Waste heat is then at a higher temperature and can be used for power generation. The only additional energy required will be to overcome system inefficiencies. This is preferable to dumping all energy used in computing to the environment.

Sustainable energy sources were considered, but a hallmark of these energy sources is that they only provide power some of the time. To provide consistent system availability, they must be buffered somehow so that power is continuously available. If center-level battery or capacitor technologies were available and affordable, this might be feasible, but they are not even on the horizon at this point. The best option currently available in the U.S. is to use local sustainable power sources and sell them back to the power grid to help offset the site's HPC consumption.

Geothermal sources of power or cooling may seem appealing. This approach is, however, only sustainable if the thermal equation is balanced (such that, what gets taken out of the earth must be put back for the method to be sustainable). Geothermal power requires substantial up-front investment, constant maintenance, and runs out quickly if not operated in a closed circuit. Geothermal sources absorb heat much more readily than they give it up, so this can make them highly efficient sources of cooling water if bodies of water (above or below ground) are available. One site exhausts heat into a local lake,

another pumps aquifer water through heat exchange for cooling. These approaches can be highly effective, but return on investment time frames, the local regulatory situation, and environmental impact must be considered.

Crosscut Topic 2: Best Practices

Several points were made concerning best practices:

- Use the most current facility control technology available
- Reduce/replace UPSs; recognize that 20-minute battery capacity represents power wasted due to conversion losses; for many sites, the ability to ride out 20-second power interruptions is sufficient
- Make the modest investments in quality equipment and instruments (for example, control dampers and humidity sensors) that can make big differences in energy efficiency over time (payback in a year)
- Experiment and find the upper bound on inlet cooling air or water for a facility and run it as warm as possible
    - Higher temperatures in the facility are usually resisted out of inertia rather than having basis in actual performance
    - Cold air is usually unnecessary; cool air is good enough if the facility is well designed and maintained
- Employ a systems engineer (or energy engineer) to ensure that initial systems operate as designed, instruments remain calibrated, data is continuously collected, data is analyzed, and efficiency improvements are developed/implemented

One site uses waste heat from HPC to help heat the building and biology labs. Another site uses land it owns as a test site for companies developing photovoltaic technologies. They provide the land, pay the cost of installing the arrays, and sell the resulting half-megawatt of power to the local utility, offsetting some of the power consumption. Yet another site pumps warm water through the local town, allowing the local utility company to make use of it before it is exhausted to a lake. Finally, the kinetic energy of the falling water is used to offset power use by the pumping system via a micro turbine.

Crosscut Topic 3: Gaps Looking Forward to New Systems

It was noted that waste heat temperatures are not high enough for efficient reuse. Raising inlet and outlet cooling temperatures would help, but it may be necessary for vendors to reengineer their systems so that they could operate in the higher heat range. This also pushes facilities further towards use of water cooling for reasons of both heat exchange efficiency and maintaining livable air temperatures in the facility.

Many sustainable power sources do not have constant enough output to be reliable, and battery or capacitor technologies do not seem to be available at the scale that could support the power needs of an HPC site. Current funding models for HPC sites incentivize reduction of cost, which is not the same as reduction of impact on the environment. Sites will improve efficiency in order to save money but are unlikely to spend money to increase efficiency unless there is return on investment or a larger

agenda. Investment and research would be required to run HPC sites on renewable power sources and turn waste heat into meaningful quantities of power.

Crosscut Topic 4: Evolve or Start Over for Future Systems

Enterprises are clearly making the decision to create new data centers in locations where power is abundant, cheap, and relatively green; witness Google and Microsoft data centers that have been built in rural portions of the Pacific Northwest. Some are taking this a step further by using containerized systems that can be relocated to take advantage of shifting power costs or locality to a population or event. HPC sites could adopt a similar approach for new facilities. However, existing HPC sites usually have political and other ties to the nations, states and, communities in which they exist. Moving them could prove a politically complicated operation. It is important for all centers to make the best of the natural resources they have around them. For existing facilities, it must be decided case by case whether it is better to retrofit or build new.

Crosscut Topic 5: Issues Shared with Large Commercial Centers

This breakout group was concerned mostly with energy, efficiency, and sustainability at the facility level, rather than the nature of the computation. It is expected that exactly the same issues would be of concern to commercial centers. Carbon tax, overall cost and efficiency, environmental and regulatory concerns, and PUE are of vital concern both to commercial centers and HPC sites.

Crosscut Topic 6: Hardware/Facility/System Interfaces to Influence

If good use is to be made of waste heat, it will be necessary to influence the vendors to have higher inlet/outlet temperatures. Recent procurements have proved that vendors can be induced to do this. Higher temperature chips are expected to consume more power due to leakage, so the tradeoff between cooling savings and higher power consumption must be carefully considered.

Renewable power sources would require power storage and release at the megawatt level to buffer its variability. It is not clear that a solution just for the HPC space would be economical even if it were feasible with today's technology.

Site electrical supply and cooling currently need to be customized to a vendor's particular system, which introduces uncertainty and probable delay when sites are conducting competitive procurements between vendors. DOE would be well served if its HPC sites could at least define a common convention (if not a standard) for HPC inlet and ambient temperatures. During the workshop, vendors repeatedly said that they could engineer their systems to meet requirements if requirements were articulated and agreed upon and the numbers were sufficient for economical production.

Crosscut Topic 7: Status of de Facto Standards

The de facto standard for measuring data center efficiency is PUE, as defined by the Green Grid (http://www.thegreengrid.org). It expresses the ratio of total power used by the data center to the power used by the computing equipment. By this measure, an ideal facility would have a PUE of 1, meaning all the power went to computing and none of it went to cooling or power conversion. In the real world, best-in-class facilities have PUE measurements approaching 1.1 This can be accomplished by following best practices in

provisioning power with minimal conversion losses and expending minimal energy on cooling (for example, using water cooling and avoiding use of compressors and computer room air conditioners (CRACs).

One shortcoming of the PUE metric is that it does not give any credit for energy that is reused for purposes other than IT. Some sites have reused energy and tried to claim a PUE less than 1.0. While such reuse of energy is a good thing, PUE by definition cannot be less than 1.0. To address this problem, Green Grid is developing a new metric, ERF. As proposed, sites are allowed to account for energy (heat) generated in the data center that is reused *outside* the perimeter of the data center. ERF incorporates and extends upon PUE. While PUE cannot be less than 1.0, ERF can be, and this new metric, if adopted, could be more even powerful than PUE in incentivizing energy savings and reuse.

Crosscut Topic 8: Other Key Findings

There is tension between policy and DOE HPC objectives. Current guidance (as exemplified by EO 13514 and O 430.2B) calls for increasing sustainable energy use, reducing greenhouse gas production, and reducing absolute energy consumption. Yet growing demand for HPC means that energy density and energy consumption by HPC sites will continue to rise unless there is a fundamental breakthrough in computing technology that radically reduces power consumption and heat production.

From an energy reuse standpoint, higher exhaust temperatures are desirable. So are ranges of inlet temperatures that fit the efficiency curves of a site's cooling capacity. Vendors are likely to be able to do the engineering to accommodate such requirements if the incentives are right. For example, one participant's center created an RFP that charged vendors for power costs if their solution was above a set PUE. All of the vendors committed to the lower PUE rather than bear the power costs.

Although "free" air or water for cooling has no financial cost attached, it is not always a given that it is sustainable and hence may have a cost to the wider community. Regulatory or political concerns might preclude use of an otherwise attractive resource. Potable water for cooling (or equipment that can use non-potable water) can be a considerable expense. Long-term environmental impacts must be considered as well. HPC must not only be efficient but sustainable and beneficial to the community.

Session 2c: Power-Aware Systems Monitoring

Session Leaders: Susan Coghlan, co-chair (ANL) and Bill Allcock, co-chair (ANL)

Participants: Jeff Broughton (LBNL), Matthew Campbell (San Diego Supercomputer Center), Kim Cupps (LLNL), Thomas Davis (LBNL), Marcus Epperson (SNL), Mark Grondona (LLNL), Michael Knobloch (Juelich Supercomputing Centre), Mike Lang (LANL), Jim Laros (SNL), Josip Loncaric (LANL), Jacques Noe (CEA/DAM), Jim Rogers (ORNL), Greg Rottman (DoD), David Skinner (LLNL), Tisha Stacey, note taker (ANL), Ryan Wright (PNNL), Mary Zosel (LLNL)

Session Description:

This breakout session considered how system monitoring can provide useful data for making improvements and managing utilization as data centers become more power and cooling constrained. The committee discussed what data is available and useful, how sites are managing the high volume of data, what data correlation sites are doing today, and potential useful correlations, along with challenges that exist in this area. One might correlate environmental data, such as power draw and rack temperatures, with science application running on the system. For example, you might develop an application "power score" that could be used for scheduling higher power score applications during lower cost power periods. In addition, with the sophisticated RAS systems available on the large HPC systems, it is possible to correlate error data (rates, types) with environmental data and the science applications. Finally, tying this system monitoring data together with facility data could bring additional insights and management techniques.

Session Process and Discussion:

The session began with introductions including name, job function, and how power monitoring was of interest. The remainder of the session was organized around group discussion of the crosscut topics. A specific issue for people to keep in mind as we went through the discussion was called out: How is power-aware monitoring different than other monitoring that might be done?

A couple of ways that it was different came out of the discussion:

- Power-aware monitoring goes well outside the facility, adding a layer of complexity
- Power-aware monitoring is hierarchical

Crosscut Topic 1: Experiences—Novel and Interesting Approaches

The items discussed ranged from approaches already implemented and automated by people in the room, to approaches that people have read and thought about but have not yet implemented:

- Correlation of applications with hardware events/environmental data. It is not uncommon to do this for system events, but doing it for applications is a new approach. Examples of this include determining the “power signature” of applications and utilizing that information in scheduling decisions, such as running power-intensive applications in the evening and correlating voltage transients to application failures.
- Collecting current draw on a per-node, per-second basis and using the data to validate operating system power saving techniques. Higher frequencies and current measurements at the sub-system level (for example, CPU and read-only memory) is desirable, but, at this time, as far as anyone in the room knew, can only be done by adding external instrumentation.
- Correlating data from multiple sources, such as building mechanical data, system data, and sensor data. This has proved to be more difficult than expected because of

disparate formats and protocols, time correlation, different time scale, and sampling times. Several of the people in the room are just now beginning to explore this.

- Allowing data from the job scheduler to act as a "feed forward" controller to the facilities. Specifically, if the scheduler knows there will be a surge in power due to a change in the application load, it can notify the chiller plant so that it can proactively ramp up the chiller.
- Correlation of CPU performance counters with energy draw to increase the frequency at which power data can be acquired. It is not clear how accurate this approach is, or if a similar technique could be used for temperatures.
- Correlation of changes in temperatures or other environmental factors to predict equipment failure.
- Proactively and automatically trigger equipment shutdown in the event of a cooling failure. Most equipment has temperature sensors that will shut the equipment down if it gets too hot. However, that will be an abrupt, uncoordinated, and probably incomplete shutdown. Having an automated system that would execute a script to do a staged and orderly shutdown minimizes the likelihood of loss of data, failed hardware, and associated difficulties and delays when coming back up.

Crosscut Topic 2: Best Practices

The group chose to interpret best practices in a broad sense that included not only practices that were in place at one or more DOE facilities, but also practices known from elsewhere or even projected based on current research. Following are the best practices to which the committee agreed:

- Ability to get current draw off individual components/sub-systems. Some systems (Cray) have this now. In the future, hopefully this will be ubiquitous and at a much higher frequency.
- The management and monitoring of the compute hardware and the facility should be fully integrated into one unified system.
- Sharing of such monitoring data and site experiences between different sites and sites and vendors should be easy and become the norm. This is generally true, but in terms of power-aware monitoring, things such as machine "thermal history" and experiences when increasing data center operating temperatures were discussed as areas where sharing of data would be particularly useful.
- Historical data should be analyzed for correlation to failure modes in the system.
- CFD simulations should be run before every major system change.
- Facilities should have baseline data for all their data streams.
- The new features and capabilities requested from this, and other, sessions should be subjected to a "return on investment" analysis to ensure the increased system cost is justified.

- Non-critical monitoring must not be in the critical path for system operation. The failure of a sensor should not impact system operation unless that sensor is critical to safe operation of the system.

Crosscut Topic 3: Gaps Looking Forward to New Systems

The following capabilities were seen as probable gaps in future systems:

- Monitoring and event notification for voltage deviations beyond specified limits. It was felt that having the system monitor the voltages internally and then issue a fault notification if it went out of specification may be easier/preferable/more practical than sampling at a high enough rate to catch the transients. This can then be used to correlate to system issues such as job failures, as well as being available for correlation with external environmental variables.
- The ability to get an instantaneous and simultaneous measure of any two of power, voltage, or current. This would allow the third to be calculated and provide an accurate snapshot of the power state of the system. All issues about frequency, accuracy, and precision apply.
- A facility that was described as a monitoring "oscilloscope" with the ability to rapidly and dynamically change the frequency and level of monitoring data. During normal operations, the frequency and amount of data may be small but during troubleshooting, the ability to get more detail, including past data (up to some limit on the order of hours), would be extremely helpful. This would probably require some sort of intelligent local caching on the monitored device to minimize data volumes to the central repository.
- Appropriate control systems that can integrate a wide range of data from disparate sources and make adjustments in environmental or other controls based on this data. In an ideal world, this might include things such as ozone action days, automated utility load shedding requests, weather, site-wide power and cooling state rather than just facility state.
- Common and useful standards of measure across the DOE HPC centers.
- Software tools for facilities and applications to use the available data for better power-aware operation.
- High reliability, user programmable baseboard management controllers on systems.

Crosscut Topic 4: Evolve or Start Over for Future Systems

The consensus within the group was that this had to be evolutionary. There is too much invested in existing power monitoring to start over. The two places where starting over, or at least significant changes, might make sense were the IPMI standard and general monitoring infrastructure to deal with the significant volumes of data generated by exascale machines.

Crosscut Topic 5: Issues Shared with Large Commercial Centers

The consensus within the group was that there was significant overlap between the large DOE HPC centers and large commercial centers. However, the following significant differences were pointed out:

- In some areas, such as power savings, the large commercial centers push the envelope more because they are profit driven and it can significantly impact their bottom line.
- A generalization of the above is that commercial centers are appropriately much more driven by ROI than DOE HPC centers.
- Commercial centers are more likely to be able to use a single node as a model and then just scale up based on the number of nodes because they tend to run many independent processes. This often does not work well in the big HPC centers because, in that environment, one tends to see large non-linearities as one scales up to the big parallel machines.
- The high frequency/fine grained resolution data discussed above is, at least for now, more research related, and the large commercial customers likely would not support it, particularly if it significantly increased costs. However, it was hypothesized that with energy costs increasing, a cost model could emerge in the future where users were charged for specific power consumption. At that time, the additional instrumentation would become more valuable.

Crosscut Topic 6: Hardware/Facility/System Interfaces to Influence

Most of the issues that the group felt might fall under this section had already been covered in other discussions. For instance, the discussions about the power and temperature sensors, signals for power out of range, and IPMI could also fall under this classification. The one additional topic that did come up in this section was having watt-hour data, similar to the data from the power meter on the average home, by subsystem. This was seen as a possible alternative to very-high-frequency sampling and the associated data handling issues. Instead, with a watt-hour meter, one could zero out, could let it do the accumulation, and take the result at the end of a job.

Crosscut Topic 7: Status of de Facto Standards

Standards were seen as an area of need for power-aware monitoring. There are desired interfaces where no standard exists, and there are existing standards that are inadequate for the DOE HPC needs. Some of the specific needs and issues called out included:

- A need for an open standard and open-source implementation of that standard, to allow access to power data across the facility. It was felt that there were significant opportunities for improvement if it was possible to access power data all the way from a subsystem on a node (how much power is the RAM consuming) to the sub-station connection to the power grid. The closest existing standard the group was aware of was BACNet.
- Though more related to monitoring in general than power-aware monitoring specifically, there was interest in a standard API for accessing monitoring and RAS data. In the discussion, this was referred to as “PAPI for power”, such that PAPI

provides a consistent interface for getting performance data across platforms, something that provided similar functionality for accessing power data. Better yet, RAS and monitoring data in general, would be useful.

- Related to the above, is an open, standard method for aggregating related but disparate data. An example is weather data and power data. There could be substantial benefits that could be gained by real-time response to weather or even a type of feed-forward control loop based on weather forecasts.
- Monitoring software in general was seen as an issue. There is no dominant player in the field, suggesting that no one has really come up with the right answer. There was also concern over the volume and rate of monitoring data (both power-aware and in general) that will be required on exascale machines.
- IPMI is the most ubiquitous method for obtaining environmental data from a host. The general consensus was that the protocol is poor and the implementations are worse. Some specific concerns were that it is not reliable. It is not uncommon to have to execute a command multiple times before it works or to have the IPMI system hang when the system hangs, which prevents utilizing one of the key features, the ability to remote power cycle a node that is hung. It was also felt that the security should be pluggable and even have the option to be disabled. As IPMI is often routed over a heavily secured, internal, non-routable, administrative network, it was felt that running it with no additional security could be acceptable. System Management Architecture for Server Hardware (SMASH) is an alternative to IPMI, but it was also seen as too heavyweight and cumbersome.
- Advanced Message Queuing Protocol (AMQP) is widely used in the financial sector and is highly scalable. It was suggested this would be worth looking at to see if it could form the basis for an exascale monitoring system.

Session 2d: Integrated Facility Planning for System and Network Upgrades

Session Leaders: Anna Maria Bailey (LLNL) and Nicholas Nagy (LANL)

Participants: Bryan Biegel (NASA Ames), Myra Branch (LANL), Jason Budd (ANL), Sergi Girona (BSC), Rick Griffin (ORNL), Mark Hartzell (PNNL), Ken'ichi Itakura (JAMSTEC), Sander Lee (DOE/NNSA), Tim McCann (SGI), Terri Quinn (LLNL), Richard Rivera (LANL)

Session Description:

This breakout session focused on the necessary integrated facility planning required to meet the demands of future systems. With the increasing demands in power and cooling, the solutions for the upgrades will spread across the entire spectrum of the facility, from the system layout to the facility infrastructure improvements and upgrades required in active HPC environments. Participants in this session discussed new and interesting approaches that they are employing or developing at their sites, including their experience with various solutions in power distribution, free cooling, liquid cooling,

networking, and environmental monitoring while maintaining flexibility and expandability of their current sites. The discussion also included the participants' experience with integrated approaches in meeting the future demands while still serving the current demands of existing active systems.

Session Process and Discussion:

The breakout session was led in a general open discussion fashion focusing on the session abstract contents. The key elements were grouped into facility and system layouts, facility requirements, and environmental monitoring. Integrated facility planning, from the rack to the utility level with emphasis on scalability, expandability, reliability, and flexibility was applied across each of the breakout topics.

Crosscut Topic 1: Experiences—Novel and Interesting Approaches

The discussions on novel and interesting approaches centered on achieving exascale. The initiative requires that sites prepare for it. The questions arise, "How do we prepare for it?" "Will it be water cooled?" The challenge is to achieve exascale in the infrastructure and facility in the most reliable and efficient manner.

Nick Nagy (LANL) is coordinating with an outside specialty engineering firm to look at ways to increase efficiency and reliability within the areas of free cooling, water cooling configuration, and chiller configuration. LANL is also looking at using reclaimed water instead of groundwater. The grey water would be cleaned to the point where it could be used within the cooling tower loop. If it works well, they will have a larger supply of water available.

Anna Maria Bailey (LLNL) is in the process of developing a HPC Master Plan addressing many core competencies to achieve exascale, such as power management, free cooling, and sustainability.

PNNL is looking into ground water cooling. There is a large source at this site so limitations are not currently identified as an issue. This is being looked into as a way to increase cooling capacities.

The general consensus of the group is that there needs to be a focus on building what is required now without investing in larger infrastructure that can be deferred. Without fully understanding the requirements, making large modifications is risky. Scalability, expandability, reliability, and flexibility are all key features that need to be applied across all decision points when siting future platforms. In the area of flexibility, some installations include the specification of overhead cabling interconnections to reduce the densities beneath the floor. In the area of scalability, all users need to have the ability to site on and off raised floor systems in the same room.

It was recommended that facility personnel have more input into the exascale development as decisions are being made about structure and layout. Facility personnel could assist with the influencing the design to be more beneficial and efficient. This is not the case at LLNL, where the facility personnel, operations staff, and system administrators are integrated into the same department to improve communication and ensure issues are addressed across all disciplines. Weekly meetings assist everyone to focus on the same goals. This common management structure provides a unified direction.

Crosscut Topic 2: Best Practices

Understanding regional strengths and weaknesses is key to push the envelope of the facility and get as much out of the existing infrastructure as possible. Is the outside temperature and humidity conducive to the platform requirements? Flexibility and expandability are not only a novel and interesting approach, they are also a best practice that needs to be applied to all decision points when siting platforms or building facilities. It was the general consensus that the infrastructure of the facility should follow the "pay-as-you-go" model for growth.

Some future best practices are to prepare for water and liquid cooling solutions while relying on air cooled solutions in the same space, apply higher voltages directly to the racks, and ensure that the integration of highly skilled facility staff within the programs be a priority.

Crosscut Topic 3: Gaps Looking Forward to New Systems

It was discussed that a wider range of cooling temperatures is required from the vendors in order to push the envelope of the current facilities. This will allow the bypassing of the chilled water completely if higher temperatures are allowed in the indirect cooling solutions at the rack level.

There is also enormous amount of uncertainty in power densities of future systems and cooling approaches. For the energy densities stated for exascale machines (>100kW/rack), air cooling is not a viable option. It will require a combination of both air and liquid cooling.

The existing networks are limited in capability and will they be able to support exascale? How do networks play into power and cooling? It is not clear how this will integrate. LLNL is already seeing difficulties and congestion with all the power, cooling, and network space requirements while they develop and plan for the installation of Sequoia.

There will be structural and weight constraints associated with exascale. Currently, petascale machines with water weigh approximately 4,000 lbs. There is a lot of support and structural remediation issues with supporting these weights on a standard stinger system.

Crosscut Topic 4: Evolve or Start Over for Future Systems

For exascale software and hardware, the system infrastructure will evolve and be new. From the facilities perspective, it has been developed in an additive pattern based on the capacity increases over time. For the future, a start over may be required to achieve the exascale capacities. This will require new or modified switchgear, transformers, chillers, and cooling towers.

It is also expected that in the future, real-time data will be a requirement to get an overall view of the HPC center energy usage. The current monitoring and SCADA systems cannot provide this type of detail; an open protocol system will be ideal. Migrating SCADA and information systems for the future will be the key to achieve power management and ultimately exascale.

Containerized systems and warehouse structures will need to be considered as options to site platforms due to the high cost of new facilities. The question is, "Can exascale

systems be deployed in these configurations?" Other questions are, "Can the system be spread out? Can it be arranged in a cube?" The days of the "show place" HPC center may be over for exascale. The configuration of the facility may not lend itself to the glass viewing window approach.

Crosscut Topic 5: Issues Shared with Large Commercial Centers

There are some common issues between HPC and enterprise centers, but the largest difference is how load is established. HPC desires larger computing capability that is shared under a smaller user base. The enterprise industry is flipped with a larger user base with a smaller computing capability. The commercial industry provides the power requirements and questions the performance for the power budget. HPC operates in the opposite manner, specifying performance and then developing power and cooling requirements and budgets.

The question arises, "How do we see achieving exascale under our current power budgets?" The costs to operate such systems will be $15M for the electrical bill alone, based on a 20MW machine. Discussions of 40MW or 100MW would be unaffordable. If sites direct the manufacturers to meet a power efficiency standpoint instead of speed baseline, innovations might occur and power budgets might be driven down.

Power and cooling is always an issue for both HPC and enterprise data centers. The commercial industry will always drive the market. Currently, the commercial industry has limited water-cooled solutions or requirements. Most likely, commercial data center innovations in free cooling for the air-cooled side can be incorporated into HPC. The commercial market will always try to do things as economically as possible, so there needs to be a continuing collaboration between HPC centers and large enterprise data centers.

Crosscut Topic 6: Hardware/Facility/System Interfaces to Influence

Standardization on purchases should be addressed by purchasing more off-the-shelf products for energy management and power management tools, to ultimately aggregate all racks to the utility data for the HPC centers. Being able to understand how the power is distributed into and across the center will be key to achieving exascale. All of the data is not centralized into a common database managed and owned by a single point.

Going forward, specifying redundancy will be cost prohibitive in exascale systems due to the scale of systems required. How will reliability be balanced for performance?

As higher voltages are available to the equipment, the RFP process will need to ensure that these are a requirement and not an option. The use of DC distribution is not readily available anywhere. The electrical distribution systems would have to be retrofitted to accommodate the use of DC; this is not a trivial or inexpensive solution.

Crosscut Topic 7: Status of de Facto Standards

Standards are needed for the following areas in HPC:

- Computational metrics for computational efficiencies. What is the metric? FLOPS/watt? Square foot/FLOP? PUE?

- DC electrical distribution systems. What are the requirements for DC in relation to NFPA 70E for electrical safety? What are the arc flash requirements?
- Power quality. Will redundancy be available for exascale systems?
- Broader liquid cooling temperature ranges.
- Network connections.
- EPO systems. The EPO is required by the National Electrical Code (NEC) but is not standard from facility to facility, and each authority having jurisdiction (AHJ) interprets Article 645 of the NEC differently.
- Fire protection standards for HPC centers. Is the fire suppression local to the rack or are overhead fire sprinklers sufficient? The fire codes need to be updated to reflect a data center and not the general building codes. This is something the commercial industry deals with as well.

Mandates need to be resolved. DOE 430.2b requires that the site-wide electrical intensity be reduced 30% by 2015 from the baseline year of 2003. Some of the power-hungry systems installed were installed after the baseline year. How is this to be accomplished and achieve exascale?

Crosscut Topic 8: Other Key Findings

The group expressed concerns that decisions provided by vendors and the industry are driving the infrastructure of the facility with a greater impact than previously seen. Sites need to establish HPC working groups with vendors who present their latest developments. This was recently done at a Webinar with the EE HPC Working Group. Sites need to encourage more of these types of engagements.

Appendix C. Results of Workshop Questionnaires

In the workshop questionnaire, each attendee had the option of casting eight votes from a list of facility-related breakout findings and five votes from system-related findings.

The following lists show the ranking of each issue and/or practice that workshop attendees found most important, from most votes to least votes.

Facilities

Votes	Issue and/or Practice
31	Higher voltage direct to computers
29	Metering and monitoring from rack to utility
21	Monitor and aggregate rack to utility data and correlate with system data
18	Flexibility and Expandability in overall infrastructure design
17	Vendors need to publish accurate and realistic temperature limits
15	Finding better ways to quantify other than PUE, something related to the computational output
14	Wireless sensor networks in open and secure facilities
14	Move toward upper end of ASHRAE range
12	Cold isn't necessary—cool is good enough
11	Control cold air flows—(avoid cold/hot air mix)
11	Separate types of equipment by environments or power or space density
11	Preparing for more water/liquid cooling requirements
10	Baseline data center metrics
10	Measure and verify, dashboard
10	Integration of highly skilled facility staff within programs
10	Explore raising ASHRAE limits
9	Consider DC Power
8	More energy efficient chillers, tower fans, CRAC units with plug fans (instead of centrifugal fans)
7	Reduce/repace UPS's: More centers realizing that 20 minute batteries aren't helpful
7	Use data center heat in another location that needs heat

6	Understanding regional strengths/weakness
6	Energy Reuse metric
6	Minimization of fans
5	Use CFD modeling to simulate airflow before each new system is installed in facility
5	Move away from raised floors
5	Segregate mechanical and computer loads at the transformer level
3	ASHRAE Standars-90.1, 127
3	Utilizing ground water/grey water for cooling loops/sources
2	Pay –as-you-go for growth
2	Use cold side control based on load (design facility to have range of input temperatures)
1	LEED Rating for Data Centers
1	DOE Data Center Energy Practitioner program certification
1	Don't mix power and cold water pipes under floor
0	California Title24-possible candidate
0	EPA Energy Star

Systems

Votes	Issue and/or Practice
21	Need more interfaces to power measurement and control from systems, from tools, and applications; especially think of things like memory usage; and standardize these interfaces
20	Integrated facility and system management
19	Tools for applications to monitor their component power usage, particularly high power activities like memory access and data movement
16	Establish a computing metric(s) (xx per watt)
15	Tools for better power-aware system management
13	Standardizing on measurement requirements (not on specific methods)
13	Implement features for saving power during idle time (but issue about how fine-grain the "idle time" is
11	Baseline
9	Need an API that exists on the edge of the RAS system – this is the PAPI for

Votes	Issue and/or Practice
	the Power (and Environmentals)
8	Ability to get current draw off individual components, want to get multiple samples per sec.
8	Add watt/hour data output, maybe accumulators, for the hardware
8	Replace IPMI with something that works consistently and supports the HPC community
7	Analyze your historical data for failure correlation
7	Share data from sites increasing their temps
7	Do ROI analysis
6	Need control of chip power features to so that those that inject system noise and cause jitter effects can be disabled as needed
6	Consider overhead of power management features
5	Cost models for application runs should include power considerations
3	Share thermal histories with vendors
3	Run simulations before every major system change
0	Programmable BMC (baseboard management control)

Appendix D. Workshop Attendees

William Allcock, Argonne National Laboratory

Anna Maria Bailey, Lawrence Livermore National Laboratory

Natalie Bates, Energy Efficient HPC Working Group

Marc Berman, Battelle – Pacific Northwest National Laboratory

Bryan Biegel, NASA Ames Research Center

Arthur (Buddy) Bland, Oak Ridge National Laboratory

Myra Branch, Los Alamos National Laboratory

Helmut Breinlinger, Leibniz Supercomputing Centre (LRZ)

Jeff Broughton, Lawrence Berkeley National Lab/NERSC

Jason Budd, Argonne National Laboratory

Matt Campbell, San Diego Supercomputer Center

Kathryn Chavez, Sandia National Laboratories

Susan Coghlan, LCF/Argonne National Laboratory

David Cowley, Pacific Northwest National Laboratory

James Craw, Lawrence Berkeley National Lab/NERSC

Kimberly Cupps, Lawrence Livermore National Laboratory

Thomas Davis, Lawrence Berkeley National Lab/NERSC

Chris DePrater, Lawrence Livermore National Laboratory

Thomas Durbin, NCSA, University of Illinois

Michael Ellsworth, Jr., IBM Corporation

Marcus Epperson, Sandia National Laboratories

Jim Garlick, Lawrence Livermore National Laboratory

Ladin Gilly, CSCS - Swiss National Supercomputing Centre

Sergi Girona, Barcelona Supercomputing Center

Alan Goodrum, Hewlett-Packard Company

Richard Griffin, ORNL/UT-Battelle

Mark Grondona, Lawrence Livermore National Laboratory

Pam Hamilton, Lawrence Livermore National Laboratory

Mark Hartzell, Pacific Northwest National Laboratory

Daniel Hitchcock, DOE/ASCR

John Hutchings, UC Lawrence Berkeley Lab

Ken'ichi Itakura, JAMSTEC

Douglas Kelley, Cray Inc.

Anthony Kenisky, Appro International, Inc.

Brent Kerby, AMD

Michael Knobloch, Juelich Supercomputing Centre, Forschungszentrum Juelich

Patricia Kovatch, NICS/UTK

Peter Kulesza, Oak Ridge National Laboratory

Michael Lang, Los Alamos National Laboratory

James Laros, Sandia National Labs

Sander Lee, Department of Energy / NNSA

John Lee, Appro International, Inc.

Josip Loncaric, Los Alamos National Laboratory

David Martinez, Sandia National Laboratories

Timothy McCann, SGI

Tommy Minyard, Texas Advanced Computing Center

Nicholas Nagy, Los Alamos National Laboratory

Jacques Noé, CEA/DAM

John Parks, NASA Ames Research Center

Rob Pennington, NSF

David Prucnal, Department of Defense

Terri Quinn, Lawrence Livermore National Laboratory

Richard Rivera, Los Alamos National Lab

James Rogers, Oak Ridge National Laboratory

Greg Rottman, DoD High performance Computing Modernization Program

Mark Seager, Lawrence Livermore National Laboratory

Yukiko Sekine, Office of Science/DOE

David Skinner, Lawrence Livermore National Laboratory

Tisha Stacey, Argonne National Laboratory/LCF

Erich Strohmaier, Lawrence Berkeley National Laboratory

William Tschudi, Lawrence Berkeley National Laboratory

Ash Vadgama, AWE (UK)

Bryan Webb, Pittsburgh Supercomputing Center
Ryan Wright, Pacific Northwest National Laboratory
Mary Zosel, Lawrence Livermore National Laboratory

Administrative Support
Lori McDowell, LLNL
Valorie McFann, LLNL