Energy Efficiency Considerations for HPC Procurements

Vendor Forum

12 September, 2013

Contributing Authors: Jim Laros (SNL), Chung Hsing Hsu, (ORNL), Bill Tschudi (LBNL), Natalie Bates (EE HPC WG)

With help from: Steve Poole (ORNL), Jim Rogers (ORNL), Robin Goldstone (LLNL), Anna Maria Bailey (LLNL), Josip Loncaric (LANL), Parks Fields (LANL), Susan Coghlan (Argonne), Jeff Broughton (LBNL), Rod Mahdavi (LBNL), Steve Hammond (NREL), Herbert Huber (LRZ), Jean-Phillipe Nomine (CEA), Francis Belot (CEA), Ralph Wescott (PNNL), Greg Rottman (ERDC), David Martinez (SNL), Ted Kubaska (IEEE), Dale Sartor (LBNL)

Agenda

- Goal and process
- Requirements and Priorities
- Expectations and next steps

- OBJECTIVES:
  - Informational presentation
  - Opportunity to ask clarifying questions
  - Level set on next steps

Goal

- Encourage dialogue about requirements and priorities for HPC energy efficiency features and capabilities
- Provide language that can be inserted in anyone's RFP regarding capabilities that vendors should provide to measure, monitor and manage energy use and efficiency, from both the system and facility perspective
  - Not how to write RFP
  - Nor suggest specific technologies
- Sets this year's vision (2013) for systems to be delivered and accepted in two years (2015)
  - Identifies priorities and sets an immediate bar

Process

- Initial team mostly DOE/government agencies, no vendor involvement
- Collect and analyze existing RFP documents for best practices
- Draw upon content experts on HPC energy efficiency to modify and supplement best practices
- Share with vendor community and solicit feedback
- Promote and encourage use in procurement
- Raise the bar over time

Measurements

(mandatory) The vendor shall provide the mechanism, interfaces, hardware, firmware, software, and any other elements that are necessary to capture the individual measurements.

(mandatory) This capability shall have minimal impact on the computation, security, as well as the energy consumption of the equipment. The vendor shall describe the impact, if any, in both quantitative and qualitative terms.

(mandatory) Scalable tools to extract, accumulate and display energy and power information (peak, instantaneous as well as average power between any two points in time) shall be delivered.

(mandatory) The energy and power data shall be exportable with at least a comma separated value or a user-accessible API.

(mandatory) For energy, power, current and voltage measurements, a detailed description of the measurement capabilities shall be provided, including a specified value for measurement precision and accuracy.

Measurements: System, Platform and Cabinet

(mandatory) Shall be able to measure the current and voltage of the system, platform(s) and cabinet(s).

The current and voltage measurements shall provide a readout capability of
(mandatory) ≥ 1 per second
(important) ≥ 50 per second
(enhancing) ≥ 250 per second

(mandatory) The current and voltage data shall be real electrical measurements, not based on heuristic models

(important) The vendor shall assist in the effort to collect these data in whatever other subsystems are provided (e.g., another vendor's back-end storage system).

(important) Those elements of the system, platform and cabinet that perform infrastructure-type functions (e.g., cooling and power distribution), shall be measured separately with the ability to isolate their contribution to the energy and power measurements.

Measurements: Nodes

(Info) A node level measurement shall consist of a combined measurement of all components that make up a node in the architecture. For example, these components may include the CPU, memory and the network interface. If the node contains other components such as spinning or solid state disks they shall also be included in this combined measurement. The utility of the node level measurement is to facilitate measurement of the power or energy profile of a single application. The *node* may be part of the network or storage equipment, such as network switches, disk shelves and disk controllers.

(important) The ability to measure the current and voltage of any and all nodes shall be provided.

The current and voltage measurements shall provide a readout capability of:
(mandatory) ≥ 1 per second
(important) ≥50 per second
(enhancing) ≥ 250 per second

(mandatory) The current and voltage data must be real electrical measurements, not based on heuristic models.

Measurements: Components

(Info) Components are the physically discrete units that comprise the node. This level of measurement is important to analyze application energy performance trade-offs. This level is analogous to performance counters and carries many of the same motivations. Components may not only be silicon devices. For example, it would be useful to know how much fan energy is being used by the Muffin fans at the back of the rack or by some active rear door cooling methodology. Also, some systems may have a CDU. How much energy is being used by the CDU for motors, fans.

(enhancing) The ability to measure the current and voltage of each individual component must be provided.

The measurement sampling frequency should be:
(mandatory) 10 samples per second
(important) 100 samples per second
(enhancing) 1000 samples per second

(mandatory) The current and voltage data shall be both real electrical measurements and based on heuristic models.

Benchmarks, Info

(Info) Power and energy costs, both operational and capital, are increasingly significant, it is very important to understand the power and energy efficiency requirements of the system. This is best understood when running workloads; either applications or benchmarks. Each site will have to select the workloads to run as part of the procurement and acceptance process. These workloads may differentially exercise or stress various sub-systems; compute (CPU, GPGPU, Memory, etc.), I/O, Networks (Internal, facility and WAN). They may focus on applications that are based on integer as well as floating point computations.

(Info) Suggested examples: HPL (compute problem), Integer-dominant codes (compute problem), Graph500 (memory/networking problem), GUPS, GUPPIE, mySQL and non-mySQL database applications, and SystemBurn developed at ORNL.

Benchmarks, Requirements

(mandatory) Customers shall specify the set of benchmarks they want. Vendors shall provide the power and energy efficiency requirements, and run times of a set of benchmarks.

(mandatory) The types of problem in the benchmarks shall cover compute problems, memory problems, networking problems, idle and sleep system state

(mandatory) Customers shall specify the run rules and the measurement quality. Each benchmark must be measurable using the Green500 run rules and attain a Level 2 measurement quality.

(important) Customers shall specify the run rules and the measurement quality. Each benchmark must be measurable using the Green500 run rules and attain a Level 3 measurement quality.

(important) Vendors shall work with Customers to provide the power and energy efficiency requirements of a set of site-supplied workloads. These workloads shall reflect the typical case, not the extremes so that vendors can design around the typical case.

(important) Customers may require application power profiles with power and energy requirements.

Energy Related Total Cost of Ownership (TCO)

(enhancing) It is an objective of [***Customer***] to encourage innovative programs whereby the vendor and/or [***Customer***] are incentivized to reduce the costs for energy and/or power related capital expenditures as well as the operational costs for energy. This may be for the system, data center and/or broader site. By doing this, the vendor would be reducing the energy-related TCO for [***Customer***]. The vendor is encouraged to describe their support for these innovative programs in qualitative as well as quantitative terms.

(Info) An example of an innovative program for bringing the energy/power element of TCO to the front was used by LRZ. Their procurement was based on TCO whereby the budget covered not just investment and maintenance, but operational costs as well. The intent was to provide a clear incentive for the vendor to deliver a solution that would yield low operational costs and, thereby, lower TCO.

Power Usage Effectiveness (PUE)

(Info) It is an objective of [***Customer***] to run a highly energy efficient data center. One measure for data center efficiency is PUE. It is recognized that the metric PUE has limitations. For example, solutions with cooling subsystems that are built into the computing systems will result in a more favorable PUE than those that rely on external cooling, but are not necessarily more energy efficient. In spite of these limitations, PUE is a widely adopted metric that has helped to drive energy efficiency.

(enhancing) The US Department of Energy Office of the Chief Information Officer has set a requirement to achieve an average PUE of 1.4 by 2015. As a result, the vendor is encouraged to qualitatively describe their support for helping [Customer] to meet this requirement.

Total Usage Effectiveness (TUE)

(Info) TUE is another metric that has been developed to overcome the limitations of the PUE metric. Specifically, it resolves the issue of PUE differences due to infrastructure loads moving from inside to outside the box. TUE is the total energy into the data center divided by the total energy to the computational components inside the IT equipment.

(enhancing) The vendor is encouraged to qualitatively describe their support for measuring TUE.

Energy Re-Use Effectiveness (ERE)

(Info) Some sites have the ability to utilize the heat generated by the data center for productive uses, such as heating office space. Energy re-use is not strictly adding to the energy efficiency of either the computing system or the data center, but it can reduce the energy requirements for the surrounding environment. For those sites, it shall be an objective of [***Customer***] to achieve an ERE < 1.0.

(enhancing) The vendor is encouraged to qualitatively describe their support for helping [***Customer***] to achieve an ERE < 1.0.

Power Distribution

(important) The vendor is encouraged to qualitatively describe energy efficient and innovative solutions that help to reduce conversion losses in the data center.

Expectations and Next Steps

- System Integrator and Component Provider Presentations on Energy Measurement Capabilities and Roadmaps
  - Webinars through October: Thurs. at 9AM PT
  - Watch for Ready-Talk Announcements
  - First come first served scheduling
- Session at EE HPC WG SC13 Workshop
  - http://eehpcwg.lbl.gov/conferences

Questions and Thank you!

Additional Material

Liquid Cooling

(Info) For systems designed to be liquid-cooled, there is an opportunity for large energy savings compared to air-cooled designs. Since liquids have more heat capacity than air, smaller volumes can achieve the same level of cooling and can be transported with minimal energy use. In addition, if heat can be removed through a fluid phase change, heat removal capacity is further increased. By bringing the liquid closer to the heat source, effective cooling can be provided with higher temperature fluids. The higher temperature liquid cooling can be produced without the need for compressor based cooling.

(Info) [***Customer***] will specify the type of liquid cooling systems contained within the data center. The range of liquid supply temperatures available in the center corresponding to ASHRAE recommended classes (W1-W4) will be provided to the vendor.

(Info) A traditional data center is cooled using compressor-based cooling (i.e., chillers or CRAC units) and additional heat rejection equipment such as cooling towers or dry coolers. These liquid-cooled systems operate within ASHRAE recommended ranges W1 and W2. Systems designed to operate in these ranges will have limited energy efficiency capability.

(important) For improved energy efficiency and reduced capital expense, many data centers can be operated without compressor based cooling, by using cooling towers or dry coolers combined with water-side economizers. These data centers can operate within the ASHRAE W3 range and accordingly, systems should be requested to operate in this range.

(enhancing) In most locations liquid cooling of up to 45° C can be provided using dry coolers. The ASHRAE W4 classification was defined to accommodate this low energy form of cooling. For this type of infrastructure, ASHRAE W4 class should be requested.

(Info) Parameters like pressure, flow rate and water quality may also be specified by each site in their procurement documents. ASHRAE provides guidance on these parameters, although they are not defined in this guideline.

Air Cooling

(Info) ASHRAE Thermal Guidelines (2011) define environmental classes that allow temperatures up to 40°C and 45°C. These new environmental temperature and humidity limits along with the recommended limits are shown in the psychometric chart below. Most IT equipment manufactured today fall within the A1 and A2 classes, while future equipment will most certainly fall within classes A3 or A4 to aid the industry in increasing energy savings.

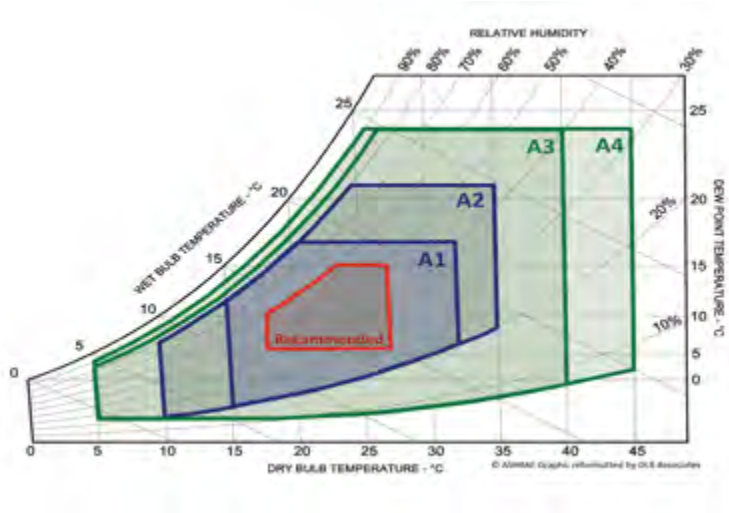

Figure 1: IT Equipment Environmental Classes

(mandatory) The system must be able to operate in a Class A1 environment.

(important) It is better to operate in a Class A2 environment (important, 2)

(enhancing) All other things equal, it is best to operate in a Class A3 environment.

Management and Control

(Info) As with the measurement capabilities described above, power and energy management and control capabilities (hardware and software tools and application programming interfaces (APIs)) are necessary to meet the needs of future supercomputing energy and power constraints. It is extremely important that [***Customer***] utilize early capabilities in this area and start defining and developing advanced capabilities and integrating them into a user friendly, production environment.

The vendor shall provide mechanisms to manage and control the power and energy consumption of the system. These mechanisms may differ in implementation and purpose. Below are envisioned usage models for these management capabilities. They are categorized loosely by where the management occurs. It is envisioned that this capability will evolve over time from initial monitoring and reporting capabilities, to management (including activities like 6-sigma continuous improvement), and even to autonomic controls.

These usage models are not requirements for the vendor, but rather suggestive examples that serve to help clarify the requirements for measurement capabilities described in section 4 above. Furthermore, it is recognized that many of these solutions would be provided by a third party, not by the system vendor.

Management and Control: Data Center/Infrastructure

(Info) Respond to utility requests or rate structures. For example, cut back usage during high load times, limit power during expensive utility rate times of the day.

“"Power capping” the system allows for provisioning the infrastructure for closer to average usage, leading to substantial infrastructure savings compared to those centers which are designed for theoretical peak usage.

Respond to demand requests; including increases in load to accommodate waste heat recovery, renewable energy, etc.

Manage rate of power changes; e.g., avoid spikes. Another example, the large variations of harmonic current produced by computer loads may need to be balanced in the data center as well as the site’s broader infrastructure and even the grid.

Management and Control: System Hardware and Software

(Info) Reduce power utilization during "design days" so as to enable use of free cooling without backup chillers. Alarm and/or automatic shut-down that responds to environmental temperature excursions that are outside of the facility design envelope by reducing system loads.

Identify higher than normal power draw components needing maintenance and/or replacement. Or, also to identify higher than normal power draw usage from SW- perhaps that is “stuck” in an infinite loop-back mode.

Proliferate power scaling and management beyond computation, to memory, communication, I/O and Storage. For example, under and overclocking, OS/hardware control of the total amount of energy consumed

Besides the traditional compiling for performance, the compiler vendor may want to provide the user with mechanisms to compile for energy efficiency. The possible mechanisms may include the following.

Compiler flags for specifying performance-energy trade-offs or regarding energy efficiency as an optimization goal or a constraint.

- Programming directives for conveying user-level information to the compiler for a better optimization in the context of energy efficiency.
- Program constructs to promote energy as the first-class object so that it can be manipulated directly in source code.
- Compiler-based tools for reporting analyzed results regarding the energy efficiency of applications.

Management and Control: Applications, Algorithms, Libraries

(Info) Provides programming environment support that leads to enhanced energy efficiency

Reduce wait-states. Examples are the following: Schedule background I/O activity more efficiently with I/O interface extensions to mark computation and communication dominant phases. Use an energy-aware MPI library which is able to use information of wait-states in order to reduce energy consumption.

Reduce the power draw in wait-states. An example is the following: Attain energy reduction for task-parallel execution of dense and sparse linear algebra operations on multi-core and many-core processors, when idle periods are leveraged by promoting CPU cores to a power-saving C-state.

Scale resources appropriately. Examples are the following: Apply the phase detection procedure to parallel electronic structure calculations, performed by a widely used package GAMESS. Distinguishing computation and communication processes have led to several insights as to the role of process-core mapping in the application of dynamic frequency scaling during communications. Analyze the energy-saving potential by reducing the voltage and frequency of processes not lying on a critical path, i.e. those with wait-states before global synchronization points. Enabling network bandwidth tuning for performance and energy efficiency.

Select appropriate energy-performance trade-off. An example is the following: Optimize the power profile of a dense linear algebra algorithm (PLASMA) by focusing on the specific energy requirements of the various factorization algorithms and their stages.

Programming and performance analysis tools. An example is: Counters, accumulators, in-band support

Open up control of these policies so that we can turn them on and off. Zero setting if it is detrimental to our applications at scale.

Management and Control:, Schedulers, Middleware, Management

(Info) Reduce power utilization during "design days" so as to enable use of free cooling without backup chillers. Alarm and/or automatic shut-down that responds to environmental temperature excursions that are outside of the facility design envelope by reducing system loads.

Identify higher than normal power draw components needing maintenance and/or replacement. Or, also to identify higher than normal power draw usage from SW- perhaps that is “stuck” in an infinite loop-back mode.

Proliferate power scaling and management beyond computation, to memory, communication, I/O and Storage. For example, under and overclocking, OS/hardware control of the total amount of energy consumed

Besides the traditional compiling for performance, the compiler vendor may want to provide the user with mechanisms to compile for energy efficiency. The possible mechanisms may include the following.

Compiler flags for specifying performance-energy trade-offs or regarding energy efficiency as an optimization goal or a constraint.

- Programming directives for conveying user-level information to the compiler for a better optimization in the context of energy efficiency.
- Program constructs to promote energy as the first-class object so that it can be manipulated directly in source code.
- Compiler-based tools for reporting analyzed results regarding the energy efficiency of applications.