Energy Efficiency Considerations for HPC Procurement Documents: 2017

(revision 1.4)

Energy Efficient High Performance Computing Working Group (EE HPC WG)

Contact: Natalie Bates, natalie.jean.bates@gmail.com

November 10, 2017

Table of Contents

1.0 Introduction 1
2.0 High Level Objectives 3
2.1 Energy Related Total Cost of Ownership (TCO) 3
2.2 Power Usage Effectiveness (PUE) 3
2.3 Total Usage Effectiveness (TUE) 4
2.4 Energy Re-Use Effectiveness (ERE) 4
2.5 Power Distribution 4
3.0 Benchmarks 5
4.0 Usage Cases for Power, Energy and Temperature Management and Control 6
4.1 Data Center Infrastructure 6
4.2 System Hardware and Software 7
4.3 Applications, Algorithms, Libraries 7
4.4 Schedulers, Middleware, Management 8
5.0 Cooling 9
5.1 Liquid Cooling (new) 9
5.2 Air Cooling 10
6.0 Measurement Requirements for Power and Energy 12
6.1 System, Platform and Cabinet Level Measurements 14
6.2 Node-Level Measurements 15
6.3 Component-Level Measurement 16
7.0 Timestamps and Clocks 17
8.0 Temperature Measurements 18
8.1 Cabinet Level Temperature 18
8.2 Node Level Temperature 19
8.3 Component Level Temperature 20
Appendix A: List of Items to Consider for Next Major Version of This Document 21
A.1 Liquid Cooling Commissioning 21
A.2 Power API support 21
A.3 Measurement accuracy 21
A.4 Power Bounding requirements 21
A.5 Further guidance with respect to continuous vs. on-demand data collection 21
A.6 Sub-component power/energy measurement requirements 21
A.7 Workload metrics like energy to solution and time to solution 21
Appendix B: References and Links 22
B.1 EE HPC WG 22
B.2 2013 Document 22
B.3 ANSI C12.20 22
B.4 EnergyStar 22
B.5 Firestarter 22
B.6 ASHRAE Air Cooling 22
B.7 ASHRAE Liquid Cooling 22
B.8 TUE 22
B.9 ERE 22

List of Figures

Figure 1: IT Equipment Environmental Classes 11
Figure 2: Reported Values vs. Internal Samples 13

List of Tables

Table 1: System/Platform/Cabinet Internal Sampling Frequency 14
Table 2: System/Platform/Cabinet External Power/Energy Reported Value Frequency .15
Table 3: Node Internal Sampling Frequency..15
Table 4: Node External Power/Energy Reported Value Frequency 16
Table 5: Component Internal Sampling Frequency 17
Table 6: Component External Power/Energy Reported Frequency.................................17

1.0 Introduction

This document is written by the Energy Efficient High Performance Computing Working Group (EE HPC WG) to encourage the adoption of energy conservation measures and energy-efficient design in high performance computing (HPC). It is intended that this document encourage dialogue in the entire community about priorities and specific requirements for HPC system energy-efficient features and capabilities. It captures energy efficiency requirements that the Energy Efficient High Performance Computing Working Group (EE HPC WG) recommends with varying degrees of importance as considerations when writing procurement documents for supercomputer acquisitions. It draws upon recent procurement documents created and used by major supercomputing sites, and also draws upon content experts in energy-efficient HPC to modify and supplement the material from these documents. The document incorporates feedback by reviewers in the vendor community.

This document is not a primer or 'how-to' on writing procurement documents. The requirements described in this document are intended to be vendor and technology neutral. They are intended to be high level and encourage dialogue, not to set guidelines or define standards. They should encourage innovation and not pick a particular vendor, architecture, technology, product or specific implementation.

The energy efficiency of HPC systems is improving, but is far from optimal. This document is based on a 2014 document that incorporated improvements over an original 2013 version. The updates reorder the document and add new liquid cooling information. The document is intended to provide vision for systems to be delivered and accepted in about three to five years, meaning there are requirements that may not be achievable with currently available products. The goal is to identify priorities and set an immediate bar. It is expected that the priorities will change and the bar will rise over time. The next update to this document is expected in 2018–2019 and should be a major update to include information from recent procurements.

Most of the document describes requirements that could be used to specify system features and capabilities. These requirements are categorized as Baseline, important or enhancing. In addition to these requirements, the document includes informational content that could be used to set the context for the acquisition, but not be used as a requirement.

Each HPC center has its own unique mission, and priorities may differ greatly between users. The requirements are intended to draw lines in the sand that can be easily re-drawn, not to build isolating fences. Some of these requirements, especially those that are enhancing, may drive up product cost beyond the value of the feature or capability to the user. The authors recognize that there may be trade-offs, but also want to encourage the dialogue that helps to communicate requirements as well as costs. The HPC center has the exclusive responsibility for managing its procurement processes. It is hoped that this document will encourage a consideration of energy efficiency during the execution of those processes.

1. In the 2017 version of the document, the order of the sections was rearranged so that general sections come first and more specialized sections later. There is also new material in Section 5 on liquid cooling.
2. Section 2 describes requirements for benchmarking power and energy.
3. Section 3 describes requirements for high level objectives, like Total Cost of Ownership and other metrics. Many of these are more specific to the data center than to the computer system.
4. Section 4 describes usage cases for management and control, but doesn't define requirements. These are suggestive examples that serve to help clarify the requirements set forth in sections above.
5. Section 5 describes requirements for cooling, both air and liquid. This section covers both the computer system and the data center. This section has new material compared to the 2014 version of the document. The new material covers both liquid cooling contols and liquid cooling commissioning.
6. Section 6 describes power and energy measurement requirements. The measurement requirements span from a high level view of the entire system to a low level view of individual components.
7. Section 7 describes requirements for timestamps and clocks.
8. Section 8 describes requirements for temperature measurements.

Conventions

Information: info

Requirements: enhancing, important, Baseline

2.0 High Level Objectives

Info	The vendor shall provide equipment, services and/or resources that among other objectives establish a highly energy-efficient solution at justifiable cost. The proposed solutions should demonstrate net benefits under normal production conditions.

2.1 Energy Related Total Cost of Ownership (TCO)

Enhancing	It is an objective of the 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 the 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 the Leibniz Supercomputing Center (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.

2.2 Power Usage Effectiveness (PUE)

Info	It is an objective of the 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 U.S. Federal Data Center Consolidation Initiative has set a requirement to lower average annual PUE. As a result, the vendor is encouraged to qualitatively describe their support for helping the Customer to meet this requirement.

2.3 Total Usage Effectiveness (TUE)

Info	Total Power Usage Effectiveness (TUE) and IT Power Usage Effectiveness (ITUE) account for infrastructure elements that are a part of the HPC system (like cooling and power distribution). TUE allows for inter-site comparison and, as such, is an improvement over PUE. ITUE is not only a metric that is necessary for calculating TUE but stands on its own as a metric for a site to use for improving infrastructure energy efficiency. For more information, see: https://www.brighttalk.com/webcast/679/96847
Enhancing	The vendor is encouraged to qualitatively describe their support for measuring ITUE and TUE.

2.4 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 would be an objective of the Customer to achieve an ERE < 1.0.
Enhancing	The vendor is encouraged to qualitatively describe their support for helping the Customer to achieve an ERE < 1.0.

2.5 Power Distribution

Important	The vendor is encouraged to describe energy-efficient and innovative solutions that help to: • Optimize connection to electrical supply (e.g. electrical grid or on-site generation) • Optimize electrical distribution within the data center and the HPC equipment. This will consider electrical equipment and conductor sizing, as well as back and redundancy configurations (e.g. dual power supplies) to minimize electrical power conversion losses by considering the entire distribution chain to the processing components within the HPC system.

3.0 Benchmarks

Info	Since power and energy costs, both operational and capital, are increasingly significant, it is 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 subsystems: compute (CPU, GPGPU, etc.), I/O, networks (internal, facility and WAN). They may focus on applications that are based on integer as well as floating point computations.
Baseline	The customer will specify the set of benchmarks they want. Vendors shall provide the power and energy-efficiency requirements and run times of a set of benchmarks.
Baseline	The benchmarks shall cover compute problems, memory problems, networking problems, idle and sleep system state.
Important	Benchmarks shall also cover dynamic power consumption. By regularly alternating between high and low power consumption, the measurement for accuracy for dynamic power consumption can be verified. This exposes aliasing issues in the measurement. Different rates for alternating the workload shall be tested, with a focus on rates around the measurement sampling rate and other measurement processing rates.
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 and FIRESTARTER developed at TU Dresden (http://tu-dresden.de/zih/firestarter/).
Baseline	Customers will specify the run rules and the measurement quality. Each benchmark must be measurable using the Green500 run rules and attain Level 2 or 3 measurement quality.
Important	Vendors shall provide the power and energy-efficiency requirements of a set of site-supplied workloads. These workloads will reflect the typical case, not the extremes, so that vendors can design around the typical case.
Important	Customers may also require application power profiles with power and energy requirements.

4.0 Usage Cases for Power, Energy and Temperature 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 the Customer utilize early capabilities in this area and start defining and developing advanced capabilities and integrating them into a user-friendly production environment.
Info	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 six-sigma continuous improvement), and even to autonomic controls.
Info	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. Furthermore, it is recognized that many of these solutions would be provided by a third party, not by the system vendor.

4.1 Data Center Infrastructure

Info	Respond to utility requests or rate structures. For example, cut back usage during high load times or limit power during expensive utility rate times. “Power capping” may have multiple uses, including one that 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.
Info	Respond to demand requests, including increases in load to accommodate waste heat recovery and renewable energy.
Info	Manage rate of power changes, e.g. avoid spikes. Another example is that 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.
Info	Provide an integrated view for the system and the building in a building management system.

4.2 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 shutdown that responds to environmental temperature excursions that are outside of the facility design envelope by reducing system loads.
Info	Identify higher than normal power draw components needing maintenance and/or replacement. Also to identify higher than normal power, draw usage from software perhaps "stuck" in an infinite loop.
Info	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.
Info	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.

4.3 Applications, Algorithms, Libraries

Info	Provides programming environment support that leads to enhanced energy efficiency. Some examples are reducing wait states and reducing the power draw in wait states.
Info	Reduce wait-states examples: • 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.
Info	Reduce the power draw in wait-states examples: • 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.
Info	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 critical path, i.e. those with wait-states before global synchronization points. • Enable network bandwidth tuning for performance and energy efficiency.
Info	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.
Info	Programming and performance analysis tools. An example is the following: • 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.

4.4 Schedulers, Middleware, Management

Info	Putting hardware into the lowest reasonable power state or switching off idle resources (nodes, storage, etc.) when job scheduling cannot allow for full utilization.
Info	Different power states. Careful about how we switch it off. Can't affect reliability. Sleep states is probably the best direction. Response time is much better.
Info	Energy-aware scheduling. Develop mechanism to automatically select processor frequency for which energy to solution is minimized for a specific application.
Info	Demand response. As in the ability to react to electrical grid based incentives. Requires enhanced scheduling tools.
Info	Evolving hardware features will likely require enhanced system software and scheduling tools with control at all levels of the hierarchy, from the system down to the components. An example might be a scenario where you have a high priority job, and/or there are available nodes to run the job, but if run at the desired P-state, the system would exceed some notion of a power cap. In this situation, can one dynamically alter the P-state of lower priority jobs to allow

	them to continue, perhaps at a slower rate, while also accommodating the new high priority job?

5.0 Cooling

5.1 Liquid Cooling (bolded material is new)

Info	For systems designed to be liquid-cooled, there is an opportunity for large energy savings compared to air-cooled systems. 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 can be produced without the need for compressor-based cooling. Higher return liquid temperatures increase opportunities for recovery of waste heat, thereby increasing system efficiency.
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 to W2. Systems designed to operate in these ranges will have limited energy efficiency.
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.
Info	Include a description of the facility and encourage vendors to participate in a site survey.

Info	Describe the computing equipment and its method(s) of cooling (e.g. direct liquid cooling, indirect liquid cooling, air cooling, rear-door heat exchangers, hybrid systems, etc.). Reference ASHRAE Liquid Cooling Guidelines book, chapter 4-5.
Info	**Describe how controls, including settings for temperature, water flow and differential pressure will interface with facilities and computer control systems. Reference "EE HPC WG Liquid Cooling Controls Team Whitepaper. June 11, 2017" https://eehpcwg.llnl.gov/pages/infra_ctrls.htm**
Info	Describe water chemistry and treatment, quality of water in cooling towers and cooling loops (all systems). Some computing equipment vendors could specify special water quality requirements, so the design must include strategies to provide the required water quality and materials compatible with the water treatment specified. The water quality of both open and closed loop systems should be defined in the cooling design phase and implemented and validated during commissioning.
Info	Describe use of cooling fluids such as refrigerants or dielectric fluids if these are required for the liquid-cooled HPC system.
Info	**Ensure that the vendor participates in the commissioning of the liquid cooling system, both for the facility and the HPC system. Commissioning documents will include specifying the liquid cooling system design, a commissioning plan and an acceptance test plan. Reference "Systematic Approach For Universal Commissioning Plan For Liquid-Cooled Systems" https://eehpcwg.llnl.gov/pages/infra_lcc.htm**

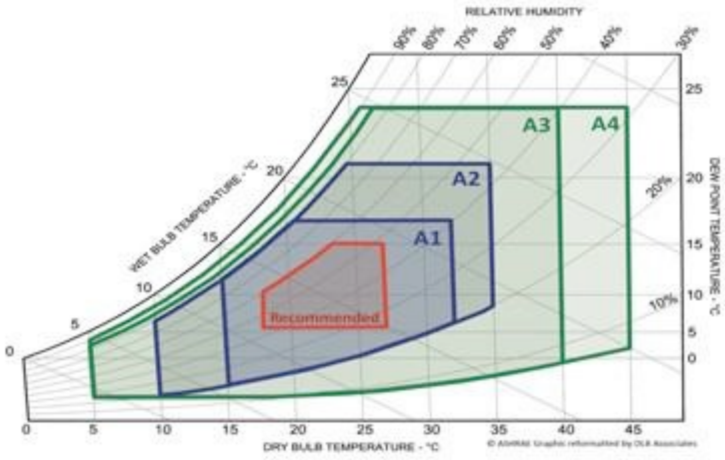
5.2 Air Cooling

Info	Ensure that fan power does not rise to extremes, thereby countering potential data center level savings.
Info	Refer to ASHRAE book – Thermal guidelines for data processing environment, 3rd edition. It describes all the trade-offs.
Baseline	The system must be able to operate in at least a class A1 environment. It is better to operate in a class A2 environment.
Enhancing	All other things equal, it is highly desirable to operate in a class A3 environment.
Baseline	The system must support full performance throughout the allowable range while operating in redundant mode (if applicable). Note: most redundantly cooled systems can operate with a single cooling component failure.
Enhancing	The system must support full performance throughout the allowable range while operating in non-redundant mode (if

	applicable).

AHRAE Thermal Guidelines (2011) define environmental classes that allow temperatures up to 40°C and 45°C. These allowable environmental temperature and humidity limits along with the recommended limits are shown in the psychrometric chart below (see Figure 1). Past IT equipment and some IT equipment manufactured today have aligned with operation within the A1 to A2 classes. Some present equipment as well as future equipment will certainly enable operation within classes A3 and A4 to aid the industry in increasing energy savings. Generally, performance trade-offs are made to enable operation in A3 or A4 environments. This must be balanced with the potential for energy savings.

Figure 1: IT Equipment Environmental Classes



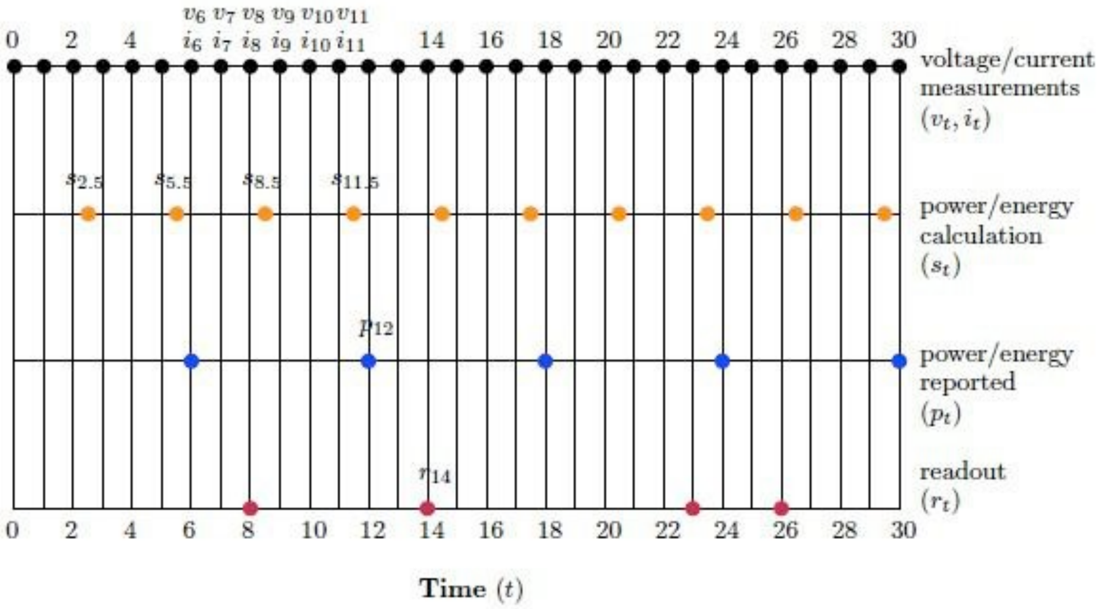
6.0 Measurement Requirements for Power and Energy

Info	Power and energy measurement capabilities are necessary to understand the energy demand of the HPC systems in order to properly size support systems and plan for future growth. These mechanisms may differ in implementation and purpose, and include capabilities for measuring the energy consumption of entire systems, platforms (subsystems), cabinets, nodes and components.
Info	This section is primarily focused on measuring the system power and energy, which includes system hardware and software.
Info	Section 8 describes usage cases for power and energy management and control.
Baseline	The vendor shall provide the mechanism, interface, hardware, firmware, software and any other elements necessary to capture the individual power and energy measurements.
Baseline	This capability should have no (or minimal and defined) impact on the computation, security and energy consumption of the equipment. The vendor shall describe the impact, preferably in quantitative terms.
Baseline	Scalable tools to extract, accumulate, and display power, energy and temperature information. Accumulated energy and peak, instantaneous, as well as average power between any two points in time, should be delivered.
Baseline	The power and energy data must be exportable with at least a comma-separated value (CSV) or user-accessible application programming interface (API).
Baseline	For power, energy and discrete current and voltage measurements if available, a detailed description of the measurement capabilities must be provided. Should include a specified value for measurement precision, accuracy, and how data samples are time-stamped. Reference ANSI C12.20. The data can be based on real physical measurements or heuristic event-based models.

A number of terms are used in this document to describe measurement capabilities. It is important to understand the context in which the terms are used. Figure 2 illustrates these terms. The x-axis of Figure 2 is Time (in generic units). Note that Figure 2 represents a range of possible capabilities that are useful for this discussion; it does not imply that these specific capabilities are a requirement.

- The top horizontal line represents points in time when discrete internal current and voltage measurements are sampled at the device level. These samples are not exposed externally. At each time interval a voltage and current sample is internally measured (v6, i6 pair for example).

Figure 2: Reported Values vs. Internal Samples


v_6 v_7 v_8 v_9 v_{10} v_{11}
i_6 i_7 i_8 i_9 i_{10} i_{11}
0 2 4 14 16 18 20 22 24 26 28 30
voltage/current measurements (v_t, i_t)
$s_{2.5}$ $s_{5.5}$ $s_{8.5}$ $s_{11.5}$
power/energy calculation (s_t)
p_{12}
power/energy reported (p_t)
r_{14}
readout (r_t)
0 2 4 6 8 10 12 14 16 18 20 22 24 26 28 30
Time (t)

- The second line down represents the points in time when an internal power and/or energy calculation is performed. Again, this is not exposed externally.
- The third line down represents the points in time a reported value is available to be read, externally. Each reported value could represent an average power, an instantaneous power, or an accumulated energy value, depending on the device capabilities. For example, point P12 could simply be the power value calculated at S8.5, or S11.5. P12 could also be the average power of points S8.5 and S11.5, or all of the calculated power samples prior to P12. P12 could likewise be an accumulated energy value representing any range of power samples up to that point in time. The important distinction is the difference between the device's internal sampling capability (frequency of and what the samples represent) and the external reported value capability of the device.
- Finally, the fourth line down represents when the user actually obtains the reported value readout. It is critical that the timestamp of the reported value represents the time, as accurately as possible, of the measurement. Notice that the actual readout takes place at various time intervals following availability of the reported value. This emphasizes the importance of time stamping at the time of measurement, not at the time of reading the value.

For example, a measurement device may be capable of producing 100 discrete power samples per second (internally). The power calculation (sample) and availability of the reported value of this same device may be equivalent to the lowest level sampling frequency, but no greater. Both are typically less than the internal sampling frequency. For example, the same device may have the ability of producing a reported value at 1 time per second. This reported value could be a power value averaged over 1 second, an accumulated energy value over the past 1 second, or simply a discrete power value for that moment in time.

Generally speaking, the requirements for the frequency of the reported value depend on what the reported value represents. If the reported value is a discrete power value, then a higher frequency of reporting is desired. If the reported value represents an average power or accumulated energy value, reported frequency is less important than the internal sampling frequency that is used to derive the reported average power or energy value.

6.1 System, Platform and Cabinet Level Measurements

Info	The system level may vary by site and architecture, but could be so as to include all of the parts of the system that explicitly participate in performing any workload(s). This might include supporting internal and external power and cooling equipment as well as internal and external communication and storage subsystems.
Info	The platform is distinguished from the system so as to differentiate compute from other subsystem equipment (such as external storage) that may be managed distinctly, but together comprise a system.
Info	The cabinet (or rack) is the first order discretization of the platform measurement. The cabinet may be part of the compute, storage or networking platform.
Baseline	Must be able to measure system, platform, and cabinet power and energy.
Important	The vendor shall assist in the effort to collect these data in whatever other subsystems are provided (e.g. another vendor's storage system).
Important	Those elements of the system, platform and cabinet that perform infrastructure-type functions (e.g. cooling and power distribution) must be measured separately with the ability to isolate their contribution to the power and energy measurements.

Table 1 lists the requirements for the internal device sampling frequency. The internal samples may be individual current and voltage samples or combined into a discrete power sample (see Figure 2).

Table 2 lists the requirements for the external reported value frequency. This is the data that is exposed externally for consumption (or readout, see Figure 2). The external reported values can represent a discrete or average power value, or an energy value. The details of the time period represented by the average power and energy values, how power and energy are calculated and time-stamped, must be specified. Note that reported rate might differ from readout rate. Readout is when a user chooses to consume the reported value and is limited by the reported rate.

Table 1: System/Platform/Cabinet Internal Sampling Frequency

Category	Internal Sampling Frequency
Baseline	≥ 10 per second
Important	≥ 100 per second

Enhancing	≥ 1000 per second

Table 2: System/Platform/Cabinet External Power/Energy Reported Value Frequency

Category	Unit of Measure	External Reported Value Frequency
Baseline	Discrete Power (W)	≥ 1 per second
	Average Power (W)	≥ 1 per second
	Energy (J)	≥ 1 per second
Important	Discrete Power (W)	≥ 10 per second
	Average Power (W)	≥ 1 per second
	Energy (J)	≥ 1 per second
Enhancing	Discrete Power (W)	≥ 100 per second
	Average Power (W)	≥ 1 per second
	Energy (J)	≥ 10 per second

6.2 Node-Level Measurements

Info	A node level measurement shall consist of the combined measurement of all components that make up a node for the architecture. For example, 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 and energy characteristics 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 power and energy of any and all nodes shall be provided.

Table 3 lists the requirements for the internal device sampling frequency. The internal samples may be individual current and voltage samples or combined into a discrete power sample (see Figure 1).

Table 4 lists the requirements for the external reported value frequency. This is the data that is exposed externally for consumption (or readout, see Figure 1). The external reported values can represent a discrete or average power value, or an energy value. The details of the time period represented by the average power and energy values, how power and energy are calculated and time-stamped, must be specified. Note that reported rate might differ from readout rate. Readout is when a user chooses to consume the reported value and is limited by the reported rate.

Table 3: Node Internal Sampling Frequency

Category	Internal Sampling Frequency
Baseline	≥ 100 per second
Important	≥ 1000 per second
Enhancing	≥ 10000 per second

Table 4: Node External Power/Energy Reported Value Frequency

Category	Unit of Measure	External Reported Value Frequency
Baseline	Discrete Power (W)	≥ 10 per second
	Average Power (W)	≥ 10 per second
	Energy (J)	≥ 1 per second
Important	Discrete Power (W)	≥ 100 per second
	Average Power (W)	≥ 100 per second
	Energy (J)	≥ 10 per second
Enhancing	Discrete Power (W)	≥ 1000 per second
	Average Power (W)	≥ 1000 per second
	Energy (J)	≥ 10 per second

6.3 Component-Level Measurement

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. Counters are special purpose registers built into CPUs to store the counts of activities and are used for low-level tuning. Components can be any devices that are part of a node for a particular architecture.
Enhancing	The ability to measure the power and energy of each individual component should be provided.

Table 5 lists the requirements for the internal device sampling frequency. The internal samples may be individual current and voltage samples or combined into a discrete power sample (see Figure 1).

Table 6 lists the requirements for the external reported value frequency. This is the data that is exposed externally for consumption (or readout, see Figure 1). The external reported values can represent a discrete or average power value, or an energy value. The details of the time period represented by the average power and energy values, how power and energy are calculated and time-stamped, must be specified. Note that reported rate might differ from readout rate. Readout is when a user chooses to consume the reported value and is limited by the reported rate.

Table 5: Component Internal Sampling Frequency

Category	Internal Sampling Frequency
Baseline	≥ 1000 per second
Important	≥ 10000 per second
Enhancing	≥ 1000000 per second

Table 6: Component External Power/Energy Reported Frequency

Category	Unit of Measure	External Reported Value Frequency
Baseline	Discrete Power (W)	≥ 100 per second
	Average Power (W)	≥ 10 per second
	Energy (J)	≥ 1 per second
Important	Discrete Power (W)	≥ 1000 per second
	Average Power (W)	≥ 100 per second
	Energy (J)	≥ 10 per second
Enhancing	Discrete Power (W)	≥ 10000 per second
	Average Power (W)	≥ 1000 per second
	Energy (J)	≥ 10 per second

7.0 Timestamps and Clocks

Info	For any post-mortem analysis, measured values need to be associated with a specific time or timeframe. Having this time information allows system administrators to recall measured values in the past and correlate them to system events, configuration changes or batch jobs. Similarly, users can correlate energy consumption to application progress in order to improve the application's energy efficiency. All this requires meaningful timestamps to be associated with the measurement values.
Important	The vendor shall provide a mechanism to associate a timestamp with each measured value reported by the vendor infrastructure. The timestamp shall indicate the time at which the measurement value is derived and will indicate known accuracy. Any measured value and its associated timestamp shall be provided automatically by the vendor infrastructure.
Baseline	The vendor shall provide documentation that enables quantification or limits on the "age" of measured values. This shall include network latencies and jitter, filter delay, processing times and the update rate.
Info	Each timestamp is with respect to a reference clock. Possible reference clocks include the compute node clocks that are used in recording application progress and management clocks that are used

	in recording system events.
Important	The vendor shall provide information that allows for quantifying the accuracy of timestamps. To that end, the vendor shall describe the applicable factors that have significant impact. They can include: • On which component and clock the timestamps are generated. • Which clock is used to compute energy from power. • The drift of the used clocks. • A description of the synchronization mechanisms that are in place between the involved clocks. • How the delay between data acquisition and timestamp generation can be quantified. • The delay of analog filters or A/D conversion.

8.0 Temperature Measurements

Baseline	The system must operate safely under all conditions. This includes when thermal emergencies are detected and when thermal sensors are faulty. The system must operate safely even with faulty sensors. Faulty sensors should be identified at the lowest possible level of sensor hierarchy.
Baseline	The data on temperature must be physically accurate, reported in real-time, and provide sufficient detail. The accuracy must be ± 0.5°C or better. Measurements must be sampled no slower than the quickest thermal response time expected. They must be accurately time-stamped. The vendor shall provide a detailed explanation of the location in the system where temperature is being measured.
Info	There is no consensus yet on which sampling rates are considered sufficient. This bears more study. For now, whatever sampling rate is used should be stated explicitly.
Info	Generally, referencing existing standards is preferred to creating new ones. The U.S. Energy Star Program for computer servers can be used as a reference, although it does not completely capture the requirements for HPC.

8.1 Cabinet Level Temperature

Important	The temperature measurements must characterize the range of operating temperatures within the system by type of device (component, node, cabinet, etc.) as well as supply and return temperatures for each coolant. These temperature measurements must include uncertainty bounds and be reported faster than the shortest thermal response time (e.g. every second).
Important	Dew point temperature of the air supplied to cabinets must be

	measured and reported to the cooling control system in charge of prevention of condensation.
Info	Temperature data are more valuable at the platform and cabinet levels than at the system level. Node and component level temperature measurements are also important but for different reasons. These temperatures are monitored to make sure the silicon remains within bounds.

8.2 Node Level Temperature

Baseline	The node level temperature measurements must be representative of temperatures within the node, which must be physically described and justified. Uncertainty in measured temperatures must be stated, and measurements must be delivered faster than the shortest thermal response time of the chosen measurement location. Support must be present for the safety of the node when thermal emergencies are detected, and system management must be notified in a timely fashion with a detailed account of the incident.
Baseline	The type of temperature being measured (for example, average or peak) shall be explicitly stated because node-level temperatures vary across the device.
Info	The information about the correlation between temperature and power is more critical in an air-cooled environment.
Info	The following lists some envisioned use cases of node-and component-level temperature measurement data exposed outside the node/component: • A better understanding of how the power-consumption behavior of a device is influenced by its surrounding temperature. This also reveals the trade-off between leakage power and temperature. Higher temperature can reduce power on cooling but increase leakage power. But this may still be advantageous if compressor-based cooling can be eliminated. • A better modeling of device failure due to thermal effects as well as the development of mechanisms for short-term and long-term failure prediction. Measuring and constraining processor temperature can improve application performance in a faulty environment. However, different applications have different optimal temperatures (A cool way of improving reliability of HPC machines – SC'13). • A better understanding of the thermal distribution within the machine and across machines to optimize the power cost for thermal management. • Temperature aware job scheduling: Different applications heat up CPUs to different temperatures. CPU temperature distribution is not homogeneous throughout the data center (i.e.

	same workload would heat up different CPUs to different temperatures). An intelligent job scheduler can take CPU-temperature and application-temperature profiles into account while assigning resources. Temperature aware scheduling could be useful in heat re-use as well. • A better understanding of the influence of temperatures on turbo mode. Different temperatures can result in different maximal frequencies. therefore creating an imbalance in computational capability that could have a negative impact on parallel applications but also create a potential for improving the scheduling of load-imbalances.

8.3 Component Level Temperature

Baseline	The temperature data must include specified uncertainty bounds and be sampled faster than the quickest thermal response time expected. Support must be present for the safety of the component when thermal emergencies are detected, and system management must be notified in a timely fashion with a detailed account of the incident.
Enhancing	The temperature of each individual component should be able to be measured.

Appendix A: List of Items to Consider for Next Major Version of This Document

A.1 Liquid Cooling Commissioning

A.2 Power API support

A.3 Measurement accuracy

A.4 Power Bounding requirements

A.5 Further guidance with respect to continuous vs. on-demand data collection

A.6 Sub-component power/energy measurement requirements

A.7 Workload metrics like energy to solution and time to solution

Appendix B: References and Links

B.1 EE HPC WG

https://eehpcwg.llnl.gov/

B.2 2013 Document

https://eehpcwg.llnl.gov/pages/compsys_pro.htm

B.3 ANSI C12.20

https://www.nema.org/Standards/ComplimentaryDocuments/ANSI-C12-20-Contents-and-Scope.pdf

B.4 EnergyStar

https://www.energystar.gov/products/spec/enterprise_servers_specification_version_3_0_pd

B.5 Firestarter

https://tu-dresden.de/zih/forschung/projekte/firestarter?set_language=en

B.6 ASHRAE Air Cooling

https://tc0909.ashraetcs.org/index.php

B.7 ASHRAE Liquid Cooling

https://tc0909.ashraetcs.org/index.php

B.8 TUE

https://eehpcwg.llnl.gov/pages/infra_itue.htm

B.9 ERE

https://www.thegreengrid.org/en/resources/library-and-tools/242-WP