Energy Efficiency Considerations for HPC Procurement Document: 2019

(revision 2.0)

FINAL

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

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

DATE: December 17, 2019

Table of Contents

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 supercomputing and, more generally, 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 with varying degrees of importance that the EE HPC WG recommends considering 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.

Each HPC center has its own unique mission, and priorities may differ greatly between centers. Some of the requirements, especially those that are Innovative, may drive up product cost beyond the value of the feature or capability to the center. 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 consideration of energy efficiency during the execution of those processes.

Also, the requirements recommended and the considerations described in this document are intended to be vendor and technology neutral.

The energy efficiency of HPC systems is improving but is far from optimal. This document is based on earlier versions, but incorporates some major updates for information from more recent procurements.

The document is intended to provide a vision for systems to be delivered and accepted in about three to five years (2022-2025), 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.

Most of the document describes requirements to consider that could be used to specify system features and capabilities. These requirements form tables in the document and are categorized as:

Basic Requirement: A requirement that is possible to achieve with today's technology.

Advanced Requirement: A requirement that stretches today's technology somewhat.

Innovative Requirement: A requirement that is hard to achieve today but would be nice to aim for in the near future.

Request for Information. Can be used in a special document 'Request for Information' or in the 'Request for Proposal'. The 'Requirements for Information' do not contain any other conditions that need to be met but the information provided can be used in the evaluation of the proposal if that has been stated in the procurement documents. In the ‘Request for Proposal’, they can be considered as requirements since the vendor needs to provide the requested information.

Information to the vendor Not formally a requirement but information that is given to the vendor so it can be considered in the procurement process and be used in other requirements.

In addition to these requirements, the document includes **background information** that could be used to set the context for the reader, but not be used as a requirement. The background information is not part of any table.

The following describes the content of the rest of this document.

- Section 2 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.
- Section 3 describes requirements for benchmarking power and energy.
- Section 4 describes use cases for management and control. These are suggestive examples that serve to help clarify the requirements set forth.
- 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 controls and liquid cooling commissioning.

- 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.
- Section 7 describes requirements for timestamps and clocks.
- Section 8 describes requirements for temperature measurements.
- Section 9 describes considerations for interfacing with existing facilities

2.0 High-Level Objectives

BACKGROUND INFORMATION:

This section is focused on high-level objectives for energy efficiency in HPC procurements.

Table 1: High-Level Requirements

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

2.1 Energy-Related Total Cost of Ownership (TCO)

BACKGROUND INFORMATION:

This section is focused on the Total Cost of Ownership for HPC procurements. The Total Cost of Ownership is the total cost to purchase and operate a computer system for a specified time, for example, 5 years. The TCO includes the purchase price, cost of installation, cost of maintenance and power and cooling costs for the specified period with a defined workload. Also, the costs for the adaptation of the existing infrastructure can be included. This gives a more accurate estimate of the real cost to operate the system for its planned lifetime. The intention is to provide a clear incentive for the vendor to deliver a solution that would yield low operational costs. In the selection of the system, one factor is the TCO per workload performed by the system. Other factors - like available software or how easy the system is to program - can be weighted in. Procurements based on Total Cost of Ownership has been used with excellent results by several sites in Europe; for example, by the Leibniz Supercomputing Center (LRZ) and by CINES in France.

Table 2: Total Cost of Ownership Requirements

Request for Information	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 should describe their support for these programs in qualitative as well as quantitative terms.

2.2 Power Usage Effectiveness (PUE)

BACKGROUND INFORMATION:

This section focuses on Power Usage Effectiveness in HPC procurements. 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.

Table 3: Power Usage Effectiveness Requirements

Basic requirement	The vendor should deliver a system that can run with high energy efficiency in the data center that is existing or will be built.
Request for Information	The U.S. Federal Data Center Consolidation Initiative as well as the European Code of Conduct for Computing Centers have set the 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)

BACKGROUND INFORMATION:

This section focuses on Total Usage Effectiveness in HPC procurements. 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). 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://eehpcwg.llnl.gov/infra_itue.html

Table 4: Total Usage Effectiveness Requirements

Basic requirement	The vendor should qualitatively describe their support for measuring ITUE and TUE.

2.4 Energy Re-use Effectiveness (ERE)

BACKGROUND INFORMATION:

This section focuses on Energy Re-use Effectiveness in HPC procurements. 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 and hence save on those costs. For those sites, it would be an objective of the Customer to achieve an ERE < 1.0.

Table 5: Energy Re-use Effectiveness Requirements

Advanced requirement	The vendor should qualitatively describe their support for helping the Customer to achieve an ERE < 1.0.

2.5 Power Distribution

BACKGROUND INFORMATION:
This section focuses on power distribution considerations for HPC procurements.

Table 6: Power Distribution Requirements

Information to the vendor	Include a description of the facility electrical systems showing component detail to include electrical supply and distribution design. See Section 9.4 below for more suggestions about what to include.
Request for Information	The vendor is encouraged to describe energy efficient and innovative solutions that help to:

	• Optimize connection to the 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.
Basic Requirement	The vendor shall provide a power distribution system that optimizes energy efficiency.

3.0 Benchmarks

BACKGROUND INFORMATION:

This section focuses on information and requirements for benchmarking related to the energy efficiency in HPC procurements. Since both power and energy related operational and capital costs 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. 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/).

Dynamic power consumption can cause aliasing issues in the measurement. By regularly alternating between high and low power consumption, the measurement for accuracy for dynamic power consumption can be verified.

Table 7: Benchmark Requirements

Basic Requirement	The customer has specified a set of benchmarks to be run by the vendor. Areas can be from compute problems, memory problems, networking problems, idle and sleep system state. This workload should reflect the typical use case, not the extremes so that vendors can design around the typical case. Vendors shall provide the commitments for the maximum power and energy, and run times for this set of benchmarks.
Basic Requirement	Vendors shall follow the specified run rules and measurement quality. For example, each benchmark must be measurable using the Green500 rules and attain Level 2 or 3 measurement quality.
Advanced Requirement	The customer provides benchmarks that show dynamic power consumption, alternating between high and low power consumption. Different rates for alternating the workload should be a parameter to the benchmark so the vendor can demonstrate that measurements are accurate and not cause aliasing issues.
Innovative Requirement	Vendors shall provide telemetry and tools to enable characterization of applications, e.g,, application power profiles (power over time measurements) for a set of benchmarks specified by the customer.

4.0 Usage Cases for Power, Energy and Temperature Management and Control

BACKGROUND INFORMATION:

This section concentrates on power, energy and temperature management and control related to energy efficiency of HPC procurements. As with measurement capabilities, 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 important that the Customer utilizes early capabilities in this area and starts defining and developing advanced capabilities and integrating them into a user-friendly production environment. The mechanisms to manage and control the power and energy consumption of the system may differ in implementation and purpose. 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.

Table 8: Power, Energy, and Temperature Management & Control Requirements

Basic requirement	The vendor should provide access to all collected system data via a documented programmable interface (API).
Advanced Requirement	The vendor shall provide mechanisms to manage and control the power and energy consumption of the system.

4.1 Data Center Infrastructure

BACKGROUND INFORMATION:

This section provides information on data center infrastructure power, energy and temperature management and control related to energy efficiency in HPC procurements.

Table 9: Data Center Infrastructure Requirements

Basic requirement	The vendor should provide interfaces that give an integrated view of the system and the building in an operational data analytics system. Examples can include providing information from integrated cooling or electrical systems (e.g., CDU, fans or power supplies).
Advanced Requirement	Managing system-level power may be done to respond to utility requests or rate structures. For example, cut back usage during high load times or limit power during expensive utility rate times. It may also be done to provision the data center infrastructure for closer to average power usage which could lead to substantial capital expense savings. The vendor shall provide the tools and capabilities to enable/support the ability to manage system-level power,

Advanced Requirement	The vendor shall provide a solution capable of responding to demand requests; including increases in load to accommodate waste heat recovery and renewable energy.
Innovative requirement	Sudden and short-duration power swings in modern super-computers can have a challenging impact on the voltage of the adjacent power grids. The vendor should provide the capability to measure and control these power swings.
Innovative requirement	Large variations of harmonic current produced by computer loads should be balanced in the data center as well as the site's broader infrastructure and even the grid. The vendor should enable the capability to measure and control these variations in harmonic current.

4.2 System Hardware and Software

BACKGROUND INFORMATION:

This section focuses on system hardware and software power, energy and temperature management and control for energy efficiency in HPC procurements.

Table 10: System Hardware & Software Requirements

Innovative Requirement	The vendor shall provide interfaces that allow for reducing power utilization during when the environmental conditions exceed design tolerances for the cooling system (e.g., do not enable the use of free cooling without backup chillers). For example, providing the ability to alarm and/or automatically shut down parts of the system to reduce system loads.
Advanced Requirement	The vendor shall provide a capability to identify components that are operating out of designed or normal power range and flag them for maintenance or replacement.
Advanced Requirement	The vendor shall provide measurements and control of platform components in addition to processors, such as memory, network interfaces, accelerators and storage. Control shall include, where applicable, setting both lower and upper limitations of power use.
Advanced Requirement	The vendor shall provide the user with mechanisms to compile for power and or energy efficiency. For example: • Compiler flags for specifying trade-offs between performance and point in time power and or energy use, for example, levels of power or energy optimization similar to performance optimization. • Programming directives that allow the user to convey opportunities where the compiler can target power and or energy optimizations. • Compiler-based tools for reporting analyzed results regarding the power and energy efficiency of applications.

Advanced Requirement	The vendor shall provide automated mechanisms to enable exchange of information regarding power and energy efficiency among multiple system software components such as a job schedulers, resource managers, and cluster management tools.

4.3 Applications, Algorithms, Libraries

BACKGROUND INFORMATION:
This section concentrates on applications, algorithms, and libraries power, energy and temperature management and control related to energy efficiency in HPC procurements.

Some examples on what can be done can be found in the following examples:

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.

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.

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 the 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 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.

Table 11: Application, Algorithm, Library Requirements

Advanced	The vendor shall provide programming environment support that

Requirement	leads to enhanced energy efficiency. For example, reducing wait states and reducing the power draw in wait states.

4.4 Schedulers, Middleware, Management

BACKGROUND INFORMATION:

This section focuses on power, energy and temperature management and control for schedulers, middleware, and management in HPC procurements. 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, 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.

Table 12: Scheduler, Middleware, Management Requirements

Basic Requirement	The vendor shall put hardware into the lowest reasonable power state or switching off idle resources (nodes, storage, etc.) when job scheduling cannot allow for full utilization.
Basic Requirement	The vendor shall support different power states.
Basic requirement	The vendor should be careful about how power is switched off so as not to adversely affect reliability. Sleep states are preferred due to superior response times.
Advanced Requirement	The vendor shall provide energy-aware scheduling. For example, automatically selecting processor frequency for a specific application to minimize the energy required to complete the scheduled request.
Innovative Requirement	The vendor shall support a demand response to react to electrical grid based incentives.
Advanced Requirement	The vendor shall provide some functionalities in system-level software (e.g. job schedulers) that can interact with job-level runtime to monitor current job / node level behavior (e.g. signals and telemetry describing application and platform characteristics).
Advanced Requirement	The vendor shall provide automated mechanisms to leverage monitored information to drive changes in application and node performance for achieving system-level constraints as dictated by site policies.

5.0 Cooling

BACKGROUND INFORMATION:
Modern high-performance computing (HPC) systems consume large amounts of power and produce a large amount of heat, and consequently, the efficiency and reliability of the

cooling system are important. When an organization is buying a new HPC system, choosing the right cooling system – and hence selecting the appropriate requirements for the procurement process – can be a difficult and tedious task. However, it is possible to save substantial amounts of energy with an efficient cooling solution.

When looking into the possibilities for an energy-saving cooling system and selecting requirements based on this document, it is important to be aware that the specifics depend to a large extent on local conditions, such as the type of cooling system already utilized in the HPC center, local climate, and goals for heat disposal or heat re-use. Below are some strategic-level questions that are intended to stimulate broad thinking about environmental and other site-specific conditions that might affect the design and selection of an energy-efficient and economical cooling solution. These questions are not meant to be a systematic list or guideline, nor are they meant to be answered exhaustively, but rather they are meant to prompt considerations and discussion by the site.

- What type of cooling exists in the building today? Is it sufficient to cool the HPC system and economical to use?
- What is the local climate like? Can the heat from computers be disposed of in open air without using compressors?
- Would free cooling be sufficient for the warmest periods of the year or will other methods be needed for some periods?
- Can cooling towers or heat exchanger and fans be placed outdoors close to the building with the computers?
- Do lakes or streams with cool water that can be used for cooling of the computers exist in the area near the building?
- Can the heat be used for other purposes, like heating building heating a pool or heating greenhouses?
- Does district heating or cooling exist in the area? Can the heat from the computers be disposed of or even sold to these facilities
- Is it feasible to produce cooling for other computers or for offices with adsorption chillers with the heat from the computers?

5.1 General Cooling requirements

BACKGROUND INFORMATION:

Requirements in this section apply generally to all types of computer cooling systems and environments.

Table 13: General Cooling Requirements

Information to the vendor	Include a description of the facility cooling systems showing component detail to include design capacity for how heat is transferred from the compute platform to the facility environment.
Information to the	The site must work with the vendor to define the methodology that will be used for calculating and measuring 'maximum power load'

vendor	supported by the facility cooling system. The site may chose to design the facility infrastructure (cooling and/or electrical systems) to support less than the potential power load of the HPC system. This is not the usual case, but there have been instances of sites that have chosen this approach. It is anticipated that this might be a more common occurance in the future. This design decision is based on information that a) the average load of an HPC system running applications in a production environment is consistently less than the maximum power load as specified by the vendor and/or b) the maximum power load of even the most stressful application (HPL or firestarter, for example) is also less than the maximum power load as specified by the vendor. Hence, maximum power load might be defined by running tuned HPL or firestarter application on all nodes of the system.
Basic requirement	The tenderers should 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.).
Basic requirement	The cooling infrastructure provided by the HPC system vendor should be sufficient to cool the proposed system under maximum power load.
Basic requirement	The above cooling system may be able to operate with some component failures (e.g., with N+1 redundancy). If this is the case, in this mode, it should also be able to cool the HPC system under maximum power load.
Basic requirement	The reliability, availability and serviceability of the cooling infrastructure provided by the HPC system should be described by the tenderers.
Basic requirement	The vendor will provide external access to all cooling system monitoring information. Integration into the site's monitoring system(s) may be done by the vendor or by the customer (with support from the vendor).
Basic requirement	The vendor should describe how controls, including settings for temperature, water flow, and differential pressure will interface with facilities and computer control systems.
Basic requirement	The contractor should participate in a site survey.

5.2 Liquid Cooling Requirements

BACKGROUND INFORMATION:
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 (maximum temperature is dependent on the used components). The higher temperature liquid can be realized without the need for compressor based cooling. Higher return liquid temperatures increase opportunities for free cooling and also for recovery of waste heat, thereby increasing overall system efficiency.

For different classes of infrastructure cooling design, ASHRAE liquid cooling guidelines

- Class W1/W2: Typically a data center that is traditionally cooled using chillers and a cooling tower but with an optional water side economizer to improve energy efficiency depending on the location of the data center. W1= 2°C – 17°C, W2 = 2°C – 27°C
- Class W3: For most locations these data centers may be operated without chillers in a water-side economizer mode. Some locations may still require chillers to meet facility water supply temperature guidelines during peak (i.e., design) ambient conditions for a relatively short period of time. W3 = 2°C – 32°C
- Class W4: To take advantage of energy efficiency and reduce capital expense, these data centers are operated in a water-side economizer mode without chillers. Heat rejection to the atmosphere can be accomplished by either a cooling tower or a dry (closed-loop liquid-to-air) cooler. W4 = 2°C – 45°C
- Class W5: To take advantage of energy efficiency, reduce capital and operational expense with chillerless operation, and make use of the waste energy. The facility water temperature is high enough to make use of the water exiting the IT equipment for heating local buildings. W5 > 45°C

define inlet water temperatures to the IT equipment (W1 through W5). Refer to bullets and ASHRAE book Liquid Cooling Guidelines for Datacom Equipment Centers, Second Edition.

Please note that cooling is very site specific and requirements might vary according to many things, including the size of the HPC datacenter and HPC system being cooled.

Table 14: Liquid Cooling Requirements

Information to the vendor	Customer will specify the type of liquid cooling systems contained within the data center. The range of liquid supply temperatures available in the center is provided by the customer. Liquid cooling circuits should be described by the flow rate, available pressure, and temperature range.
Information to the vendor	Describe water chemistry and treatment, quality of water in cooling towers, and cooling loops (all systems).

Basic requirement	The cooling design of the tender must be able to cope with the water treatment and the piping material and the material of the cooling towers 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.
Basic requirement	The system should have provisions for leak detection on cooling systems with liquids. It can be used to alert system staff and/or shut down the affected part of the system possibly with some delay.
Basic requirement	The vendor should describe the use of cooling fluids such as refrigerants or dielectric fluids and provide the fluid's Safety Data Sheet (SDS – formally MSDSs or Material Safety Data Sheet) if these are required for the liquid-cooled HPC system.
Basic requirement	The vendor should participate in the commissioning of the liquid cooling system, both for the facility and the HPC system. Commissioning documents should 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" .
Basic requirement	Size and location of all connectors should be provided by the vendor.
Advanced requirement	For IT equipment located in facilities with inlet water temperature not exceeding 35°C, at least 90% of the energy consumed by the system can be cooled by water.
Advanced requirement	For inlet water temperatures in the range between 35°C and 45°C and a room temperature of 25°C, 85% of the energy consumed by the system can be cooled by water.
Innovative requirement	For inlet water temperatures in the range between 35°C and 45°C and a room temperature of 25°C, 95% of the energy consumed by the system can be cooled by water.
Innovative requirement	For warm water cooling a cooling efficiency of 100% is physically only reachable when the room temperature is higher than the water outlet temperature. For cases in which the room temperature should stay below the water inlet temperature, the racks must be thermally insulated, and a little bit of air cooling is necessary. For inlet water temperatures in the range between 35°C and 45°C and a room temperature below 35°C, >98% of the energy consumed by the system can be cooled by water.
Innovative Requirement	Processors and other components should be able to be cooled with liquids with a temperature higher than 32°C and up to 45°C (ASHRAE W4).
Basic requirement	The vendor should specify if higher water temperatures than ASHRAE W4 are allowed. In that case, the vendor should give the resulting effect on performance and reliability.

Advanced requirement	For a procurement that uses a Total Cost of Ownership method, the vendor should define the outgoing water temperature for a typical workload of the system. A formula given in the procurement defines the cost to produce this cooling with possible income from heat re-used taken into account. Don't forget the cost to cool the residual power (heat not going to water).
Basic requirement	For liquid cooling systems with heat re-use, with a specified incoming liquid temperature to the HPC system the outgoing liquid temperature should be at least 5 to 10 degrees C warmer than the inlet for a defined workload of the system. Alternatively, ask the vendor what the outgoing temperature will be.

5.3 Air Cooling

BACKGROUND INFORMATION:

ASHRAE thermal guidelines define environmental classes (A1 through A4) for temperature and humidity limits. Refer to ASHRAE book - Thermal Guidelines for Data Processing Environments, Fourth Edition

Table 15: Air Cooling Requirements

Basic requirement	Ensure that the fan power of the HPC system does not rise to extremes thereby countering potential data center level savings.
Basic requirement	The vendor should define the amount of heat exchanged with the ambient air within the data center as a percentage of overall rack load at idle, normal, and benchmark power levels. This applies to HPC systems that have direct or no direct air exchange with the data center. Direct air exchange can occur at any opening at the cabinet boundary at which a pressure differential exists. This includes cabinet enclosures designed without air transfer perforations.
Basic requirement	The system must be able to operate in at least a class A1 environment. It is better to operate in a class A2 environment.
Advanced requirement	All other things equal, it is highly desirable to operate in a class A3 or A4 environment.

6.0 Measurement Requirements for Power and Energy

BACKGROUND INFORMATION:

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. Note that many of the requirements below are most suitable for large sites with a research interest in how the system is utilized in more detail. The amount of data collected can be substantial and have an associated high cost to tranfer and store. For smaller sites and sites with just the interest to operate the system with reasonable efficiency, the requirements could be scaled down.

This section is primarily focused on measuring the system power and energy, which includes system hardware and software. 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 1 illustrates these terms. The x-axis of Figure 1 is Time (in generic units). Note, Figure 1 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, a voltage and current sample are internally measured (v6, i6 pair for example).

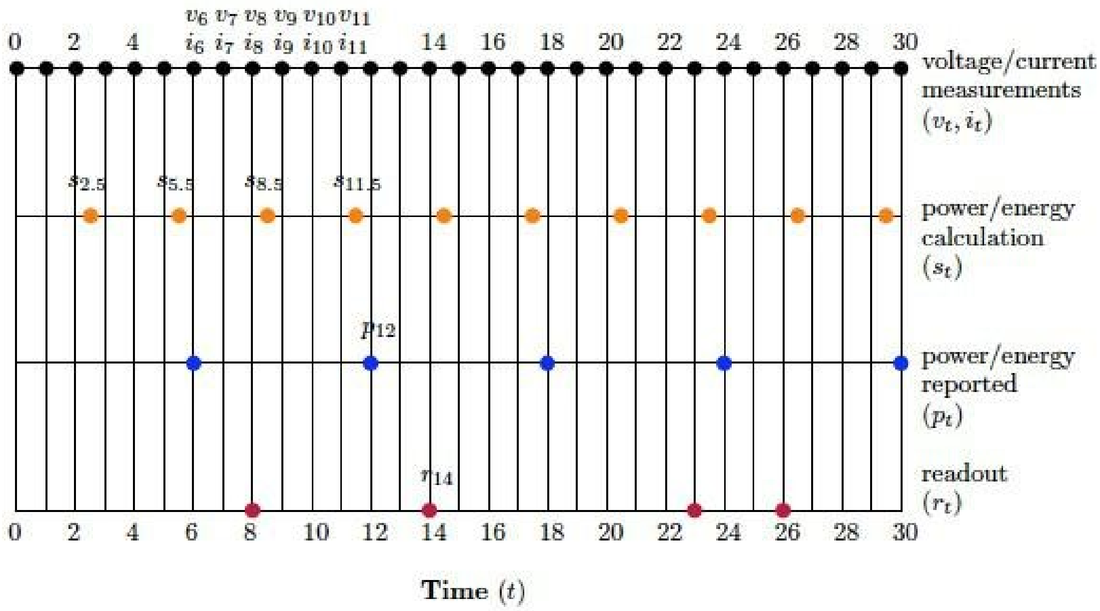

Figure 1 Reported Values vs. Internal Samples

- 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 externally 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 the 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 to produce 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 number reported is a discrete power value, then a higher frequency of reporting is desired.

Table 16: Measurement for Power & Energy Requirements

Basic Requirement	The vendor shall provide the mechanism, interface, hardware, firmware, software, and any other elements necessary to capture the individual power and energy measurements.
Basic Requirement	The vendor-provided capability to measure the power and energy should have a minimal and defined impact on the computation, security and energy consumption of the equipment.
Advanced Requirement	The vendor shall provide 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.
Basic Requirement	The vendor shall ensure power and energy data is exportable with at least a comma-separated value (CSV) or user-accessible application programming interface (API).
Basic Requirement	The vendor shall provide a detailed description of the measurement capabilities for power, energy and discrete current and voltage measurements. 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.

6.1 System, Platform and Cabinet Level Measurements

BACKGROUND INFORMATION:
This section contains information and requirements on distinguishing the location or level at which the measurements are taken.

Table 17 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 18 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 the reported rate might differ from the readout rate. The readout is when a user chooses to consume the reported value and is limited by the reported rate.

Table 17: System, Platform, Cabinet Internal Sampling Frequency

Category	Internal Sampling Frequency
Basic	≥ 10 per second
Advanced	≥ 100 per second
Innovative	≥ 1000 per second

Table 18: System, Platform, Cabinet External Power Reported Value Frequency

Category	Unit of Measure	External Reported Value Frequency
Basic	Discrete Power (W)	≥ 1 per second
	Average Power (W)	≥ 1 per second
	Energy (J)	≥ 1 per second
Advanced	Discrete Power (W)	≥ 10 per second
	Average Power (W)	≥ 1 per second
	Energy (J)	≥ 1 per second
Innovative	Discrete Power (W)	≥ 100 per second
	Average Power (W)	≥ 1 per second
	Energy (J)	≥ 10 per second

Table 19: System, Platform, Cabinet Requirements

Information to the vendor	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.
Information to the vendor	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.
Information to the vendor	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.
Basic Requirement	The vendor shall be able to measure system, platform, and cabinet power and energy.
Basic Requirement	The vendor shall assist in the effort to collect these data in whatever other subsystems are provided (e.g. another vendor's storage system).

Basic Requirement	The vendor shall measure power and energy separately for those elements of the system, platform, and cabinet that perform infrastructure-type functions (e.g. cooling and power distribution).

6.2 Node-Level Measurements

BACKGROUND INFORMATION:
This section contains information and requirements specific to measurements taken at the node-level of a computer system.

Table 20 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 21 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. The readout is when a user chooses to consume the reported value and is limited by the reported rate.

Table 20: Node Internal Sampling Frequency

Category	Internal Sampling Frequency
Basic	≥ 100 per second
Advanced	≥ 1000 per second
Innovative	≥ 10000 per second

Table 21: Node External Power Reported Value Frequency

Category	Unit of Measure	External Reported Value Frequency
Basic	Discrete Power (W)	≥ 10 per second
	Average Power (W)	≥ 10 per second
	Energy (J)	≥ 1 per second
Advanced	Discrete Power (W)	≥ 100 per second
	Average Power (W)	≥ 100 per second
	Energy (J)	≥ 10 per second
Innovative	Discrete Power (W)	≥ 1000 per second
	Average Power (W)	≥ 1000 per second
	Energy (J)	≥ 10 per second

Table 22: Node Level Measurement Requirements

Information to the vendor	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.
Advanced Requirement	The vendor shall provide the ability to measure the power and energy of any and all nodes.

6.3 Component-Level Measurement

BACKGROUND INFORMATION:
This section contains information and requirements specific to measurements taken within a node of a computer system for a particular component.

Table 23 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 24 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. The readout is when a user chooses to consume the reported value and is limited by the reported rate.

Table 23: Component Internal Sampling Frequency

Category	Internal Sampling Frequency
Basic	≥ 1000 per second
Advanced	≥ 10000 per second
Innovative	≥ 1000000 per second

Table 24: Component External Power Reported Frequency

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

Advanced	Discrete Power (W)	≥ 1000 per second
	Average Power (W)	≥ 100 per second
	Energy (J)	≥ 10 per second
Innovative	Discrete Power (W)	≥ 10000 per second
	Average Power (W)	≥ 1000 per second
	Energy (J)	≥ 10 per second

Table 25: Component Requirements

Information to the vendor	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.
Innovative Requirement	The vendor shall provide the ability to measure power and energy of each individual components defined by the customer (like CPU, GPU, memory etc., I/O-devices, etc.)..

7.0 Timestamps and Clocks

BACKGROUND INFORMATION:

This section concentrates on timestamps and clocks related to energy efficiency with HPC procurements. 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.

Table 26: Timestamp & Clock Requirements

Basic Requirement	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.
Basic Requirement	The vendor shall provide documentation that enables quantification or limits on the error in time between when the measured value is obtained and the timestamp on the value.. This may include network latencies, and jitter, filter delay, processing times and the update rate.

Information to the vendor	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.
Advanced Requirement	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 a 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

BACKGROUND INFORMATION:

This section focuses on temperature measurements for focusing on the energy efficiency of HPC systems in procurements.

Table 27: Temperature Measurement Requirements

Basic requirement	Conditions of interest for safe operation include thermal emergencies and thermal sensors faults. The system must operate safely even with faulty sensors. Faulty sensors should be identified at the lowest possible level of sensor hierarchy.
Basic Requirement	The vendor shall provide real-time, accurate, and sufficiently detailed data on temperature.
Information to the vendor	The accuracy of temperature measurements 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.
Basic Requirement	The vendor shall explicitly state the selected sampling rate used for temperature measurements.
Information to the vendor	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

BACKGROUND INFORMATION:

This section focuses on cabinet level temperature measurement for energy efficiency in HPC procurements.

Table 28: Cabinet Level Temperature Requirements

Basic Requirement	The vendor shall supply temperature measurements characterizing the range of operating temperatures within the system by type of device (component, node, cabinet, etc.), in addition to supply and return temperatures for each coolant.
Advanced Requirement	The vendor shall supply temperature measurements with uncertainty bounds.
Advanced Requirement	The vendor shall report temperatures faster than the shortest thermal response time (e.g. every second).
Basic Requirement	If there is a risk for condensation, the vendor shall provide dew point temperature measurements of the air supplied to cabinets, reporting those measurements to the cooling control system in charge of prevention of condensation.
Information to the vendor	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

BACKGROUND INFORMATION:

This section concentrates on node level temperature measurement for energy efficiency of HPC systems in procurements.

Table 29: Node Level Temperature Requirements

Basic Requirement	The vendor shall provide accurate node level temperature measurements for points of interest on the nodes.
Advanced Requirement	The vendor shall provide the uncertainty in the measured temperatures.
Advanced Requirement	The vendor solution shall deliver measurements faster than the shortest thermal response time of the chosen measurement location.
Advanced Requirement	The vendor guarantees the safety of the node when thermal emergencies are detected.
Innovative Requirement	The vendor’s system notifies system management in a timely fashion with a detailed account of the incident.

Basic Requirement	The vendor shall explicitly specify the type of temperature being measured (for example, average or peak).
Information to the vendor	The information about the correlation between temperature and power is more critical in an air-cooled environment.
Request for Information	The following list 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 (Coles H, Ellsworth M, Martinez D., et. al., “Hot for Warm Water Cooling”. SC11 International Conference for High Performance Computing, Networking and Storage. https://eehpcwg.llnl.gov/pubs.html). • 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. The 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 creates a potential for improving the scheduling of load-imbalances.

8.3 Component Level Temperature

BACKGROUND INFORMATION:

This section focuses on component level temperature measurement for energy efficiency of HPC systems.

Table 30: Component Level Temperature Requirements

Basic Requirement	The vendor shall provide temperature data for components within the system.
Advanced Requirement	The vendor shall explicitly specify uncertainty bounds for temperature measurement.
Advanced Requirement	The vendor shall ensure temperature measurements are sampled faster than the quickest thermal response time expected.
Advanced Requirement	Support must be present for the safety of the component when thermal emergencies are detected.
Innovative Requirement	The vendor shall provide system management with a detailed account of the incident in a timely fashion when thermal emergencies are detected.
Innovative Requirement	The temperature of individual components defined by the customer (like CPU, GPU, memory etc., I/O-devices, etc.). should be able to be measured.

9.0 Interfacing with Facilities

BACKGROUND INFORMATION:
Existing or planned facilities around the procured system form natural prerequisites that have to be described in the procurement document. Depending on how the procurement is designed there are several alternatives to how upgrades or changes can be handled:

- Upgrades cannot be done
- Upgrades or changes are paid by the vendor
- Upgrades or changes are paid by the customer
- Upgrades or changes are included in the total cost of ownership calculation
- The vendor should indicate that a change or upgrade is required

Different options can be applied to different prerequisites. Of course, this has to be specified in the procurement text.

9.1 Facilities overview

- Give an overview of the facilities.
- Describe if it is existing, new or refurbished.

9.2 Structural

- Give a map of the facilities and indicate where the system should be placed.
- Indicate columns and other obstructions.
- Don't forget the height of the room and possible obstructions in the ceiling.
- Indicate if there is a raised floor. In that case, give the height of the raised floor.
- Give the point load, rolling point load, safety point load, and uniform load rating for the floor.
- Indicate the way the system will be brought into the computer hall. Give the allowed load for the entryway.
- Describe and give size and weight capabilities of any ramps or freight elevators used in the entryway.

9.3 Cooling

Give an overview of how the cooling principal required for system i.e. air, liquid cooling, direct liquid cooling.

9.3.1 Air cooling

- How is the cooled air coming to the system: from the raised floor, air ducts, back-door cooling, in-row cooling, etc.
- What is the total cooling power available?
- What is the maximum power per rack or area in the computer hall?
- What temperature is expected in the room? Often expressed in ASHRAE environmental classes.

9.3.2 Liquid cooling

- Where are the pipes located? Under the raised floor, on top of the racks, etc.?
- What is the cooling media? Water, glycol-water?
- How is the liquid filtrated?
- What is the temperature of the liquid? Does it vary with outdoor conditions?
- Is there an outgoing temperature or delta-T that should be kept/guaranteed?
- What is the maximum pressure drop of the liquid?
- What is the maximum flow of the water?
- What is the maximum cooling power available?
- How are the parameters of the cooling circuit controlled?
- If there is more than one cooling loop, please specify the above data for each loop.

9.4 Power

- Which safety standards do you have to adhere to (i.e. NFPA 496, NFPA 70, NRTL, CE)?

- What Voltage and what configuration do you use? Three-phase delta (3PH + G) or three-phase wye 380 V (3PH + N + G)?
- Describe if there are several systems (A and B) or UPS/diesel generator or not and which parts of the system which should be connected to what feed.
- Describe where existing Power Distribution Units (PDUs) are located and what available breakers they have. Are additional PDUs required?
- Where should the power feeds be placed? Above the racks or under the raised floor?
- Describe if the computer hall has a busbar system. Include a drawing of the busbars and what available power they have. Indicate what tap-off boxes that are available for the busbar system.

9.5 Building Management System

- Describe the current Building Management System and how more components can be integrated into the system?
- How are the components connected? Modbus, Modbus over IP, BACNET, Ethernet, SNMP.

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

A.1 Liquid Cooling Standards

A.2 PowerStack and PowerAPI support

A.3 Measurement accuracy

A.4 Power Bounding requirements, including over-provisioning

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

A.8 A perspective on 5 years ahead and more

Appendix B: References and Links

B.1 EE HPC WG

https://eehpcwg.llnl.gov/

B.2 Previous Versions of this 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