Energy Efficient Considerations for HPC Procurement Documents

(version 0.45)

The Energy Efficient High Performance Computing Working Group

Contents

List of Tables

List of Figures

1 Introduction

This document captures some best practices to consider when writing procurement documents for supercomputer acquisitions. These best practices concern energy efficiency, especially capabilities to measure and manage both power and energy consumption. The document draws upon recent procurement documents created and used by two major supercomputing sites. In addition the document modifies and supplements the material from these procurement documents with input from experts in energy efficient HPC.

The team that wrote this draft consists exclusively of members from the user community, mostly from US DOE Labs. General publiction will include review and feedback from vendors.

Although progress has been made, there remains much room for improvement in HPC energy efficiency. This document sets this year's vision (2013) for systems to be delivered and accepted in two years (2015). It identifies priorities and sets an immediate goal. Becasue it is expected that these priorities will change and that the bar will rise over time, this document will be refreshed on a yearly basis.

Some of the content below is informational and as such is intended to set the context for the acquisition, but not to be used as a requirement. Additional content reflects requirements and is intended to specify system features and capabilities. These requirements are categorized as mandatory, important, or enhancing.

The intent is that this document encourage dialogue about priorities and requirements for HPC system energy efficient features and capabilities while recognizing that each HPC center has its own unique mission with differing priorities. The requirements discussed here are intended to draw lines in the sand that can be easily re-drawn, not to build isolating fences.

Finally, this document is intended to be high level while remaining vendor and technology neutral. It should encourage innovation and not pick a particular vendor or solution.

The content is organized into five categories, all focused on energy efficiency.

- Measurement, benchmarks and management focus almost exclusively on system hardware and system software, but also span applications.

- The other two categories are general objectives and cooling. These two areas span infrastructure and the supercomputer system itself (mostly system hardware, but some aspects of system software as well).

Conventions:

Information: info

1. enhancing
2. important
3. mandatory

2 Measurements

(info) Power and energy measurement capabilities are necessary to meet the needs of future supercomputing power and energy constraints. These mechanisms may differ in implementation and purpose and include capabilities for measuring the energy consumption of entire systems, platforms (subsystems), cabinets, node, and components.

(info) This section is primarily focused on measuring the system power and energy, which includes system hardware and software.

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

(mandatory) This capability should have no (or minimal and defined) impact on the computation, security, and energy consumption of the equipment. The vendor must describe the impact, preferably in quantitative terms.

(mandatory) 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.

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

(mandatory) For power, energy (and discrete current and voltage measurements if available) a detailed description of the measurement capabilities must be provided, including a specified value for measurement precision, accuracy, and how data samples are time-stamped. WE HAD A LOT OF DISCUSSION ABOUT PRECISION and ACCURACY, SHOULD WE EXPAND ON WHAT WE WANT HERE FOR THE ENTIRE DOCUMENT? CAN WE PROVIDE A SPECIFIC REQUIREMENT OR BE MORE DESCRIPTIVE ABOUT WHAT WE WANT HERE JHL? Reference ANSI C12.1

(info) Why hierarchy?

The document is formatted in somewhat of a hierarchical fashion. The purpose of this is to address the various current and anticipated future use cases related to this topic. Component level measurement, for example, is required for fine-grained application power and energy analysis; likewise, component level control could be used to shift power from one component to another based on specific application requirements. Measurement at node level granularity is necessary for understanding the power and energy characteristics of a multi-node application, for example. While cabinet level measurement might have fewer current use cases, cabinet level power capping, as well as node level, are emerging as important requirements in recent procurements. Platform level measurement and control has many facility inspired use cases and is a critical piece of overall platform management.

(info) Reported Values versus Internal Samples

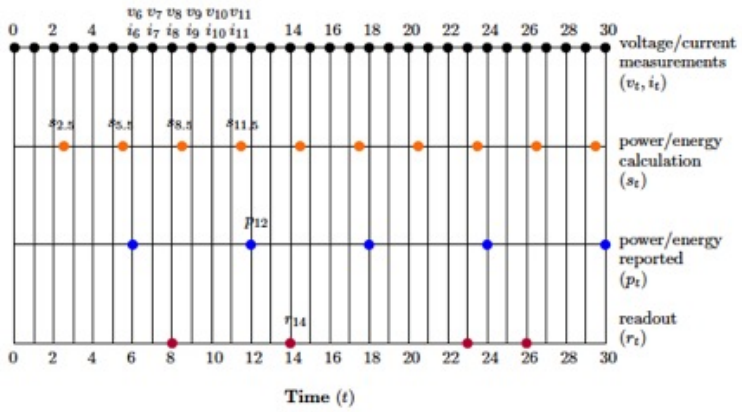

Figure 2.1: Power profile HPL run

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 2.1 illustrates these terms. The x-axis of Figure 2.1 is Time (in generic units). Note that Figure 2.1 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 necessarily exposed externally. At each time interval a voltage and current sample is internally measured (v6, i6 pair, for example).

- The second line down represents the points in time when an internal power and/or energy calculation is performed. Again, the result is not necessarily exposed externally.

- The third line down represents the points in time when 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 (again, frequency of and what the values represent).

- Finally, the forth 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 of the measurement as accurately as possible. Note that the actual readout takes place at various time intervals following the availability of the reported value. This emphasizes the importance of timestamping 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 a value at 10 times per second. This reported value could be a power value averaged over 10 seconds, an accumulated energy value that includes 10 additional seconds with each new reported value, or simply a discrete (instantaneous) 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 reported value is typically 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.

2.1 System, Platform, and Cabinet Level Measurements

(info) The system level may vary by site and architecture, but could be so broad as to include all the parts of the system that explicitly participate in performing any workload(s). This might include supporting internal and external power and

Table 2.1: System, platform, and cabinet requirements for internal and reported frequency

Requirements	Internal Sample Frequency	Measured Value	External Reported Value Frequency
Mandatory	≥10 per second	Discrete Power (W) Average Power(W) Energy(J)	≥1 per second ≥1 per second ≥1 per second
Important	≥100 per second	Discrete Power (W) Average Power(W) Energy(J)	≥10 per second ≥=1 per second ≥1 per second
Enhancing	≥1000 per second	Discrete Power (W) Average Power(W) Energy(J)	≥100 per second ≥1 per second ≥10 per second

cooling equipment as well as internal and external communication and storage sub-systems.

(info) The platform is distinguished from the system so that compute equipement is differentiated from other system-level equipment (such as external storage) that may be managed distinctly, yet together make up a system.

(info) The cabinet (or rack) is the first order discretization of the platform level measurement. The cabinet may be part of the compute, storage or networking platform.

(mandatory) Must be able to measure system, platform, and cabinet power and energy.

Table 2.1 lists the mandatory, important and enhancing requirements for the internal sampling frequency (internal device capability) and the external reported value frequency (data available to the consumer) at the system, platform and cabinet level. Figure 2.1 should be referenced in conjunction with Table 2.1 to help clarify these requirements. Note that Figure 2.1 also depicts readout, i.e. when the consumer chooses to read the data. Readout rate will not be addressed in the requirements since readout rate is driven by the computer and limited by the reported value frequency. The details describing the internal sampling frequency (voltage and current or power) and how the average power and/or energy value is calculated **must** be provided.

(mandatory) The power and energy values must be based on electrical measurements (e.g., based on shunts or Hall effect sensors). Values that are derived from heuristic models based on architectural events or system state (e.g., RAPL) can complement but not replace them.

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

2.2 Node Level Measurements

Table 2.2: Node Level Requirements for internal and reported frequency

Requirements	Internal Sample Frequency	Measured Value	External Reported Value Frequency
Mandatory	≥100 per second	Discrete Power (W) Average Power(W) Energy(J)	≥10 per second ≥10 per second ≥1 per second
Important	≥1000 per second	Discrete Power (W) Average Power(W) Energy(J)	≥100 per second ≥=100 per second ≥10 per second
Enhancing	≥10000 per second	Discrete Power (W) Average Power(W) Energy(J)	≥1000 per second ≥1000 per second ≥10 per second

(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 must/should be provided. NOTE: Should this be mandatory? Change must/should appropriately.

Table 2.2 lists the mandatory, important, and enhancing requirements for the internal sampling frequency (internal device capability) and the external reported value frequency (data available to the consumer) at the node level. Figure 2.1 should be referenced in conjunction with Table 2.2 to help clarify these requirements. Note that Figure 2.1 also depicts readout, i.e., when the consumer chooses to read the data. Readout rate will not be addressed in the requirements since readout rate is

driven by the computer and limited by the reported value frequency. The details describing the internal sampling frequency (voltage and current or power)and how the average power and/or energy value is calculated **must** be provided.

(mandatory) The power and energy values must be based on electrical measurements (e.g., based on shunts or Hall effect sensors). Values that are derived from heuristic models based on architectural events or system state (e.g., RAPL) can complement but not replace them.

2.3 Component Level Measurement

Table 2.3: Component Level Requirements for internal and reported frequency

Requirements	Internal Sample Frequency	Measured Value	External Reported Value Frequency
Mandatory	≥1000 per second	Discrete Power (W) Average Power(W) Energy(J)	≥100 per second ≥10 per second ≥1 per second
Important	≥10000 per second	Discrete Power (W) Average Power(W) Energy(J)	≥1000 per second ≥=100 per second ≥10 per second
Enhancing	≥100000 per second	Discrete Power (W) Average Power(W) Energy(J)	≥10000 per second ≥1000 per second ≥10 per second

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

(enhancing) The ability to measure the power and energy of each individual component must (or should if this is enhancing?) be provided.

Table 2.3 lists the mandatory, important, and enhancing requirements for the internal sampling frequency (internal device capability) and the external reported value frequency (data available to the consumer) at the node level. Figure 2.1 should be referenced in conjunction with Table 2.3 to help clarify these requirements. Note

that Figure 2.1 also depicts readout, i.e., when the consumer chooses to read the data. Readout rate will not be addressed in the requirements since readout rate is driven by the computer and limited by the reported value frequency. The details describing the internal sampling frequency (voltage and current or power) and how the average power and/or energy value is calculated **must** be provided.

(mandatory) The power and energy values must be based on electrical measurements (e.g., based on shunts or Hall effect sensors). Values that are derived from heuristic models based on architectural events or system state (e.g., RAPL) can complement but not replace them.

3 Management and Control

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

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

These usage models are not requirements for the vendor, but rather suggestive examples that serve to help clarify the requirements for measurement capabilities described in Section 2.3 above. Furthermore, it is recognized that many of these solutions would be provided by a third party, not by the system vendor.

3.1 Datacenter/Infrastructure

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

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

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

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

3.2 System Hardware and Software

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

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

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

In addition to traditionally 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 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.

3.3 Applications, Algorithms, Libraries

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

Reduce wait-states. Examples are the following:

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

Reduce the power draw in wait-states. An example is the following:

- Attain energy reduction for task-parallel execution of dense and sparse linear algebra operations on multi-core and many-core processors, when idle periods are leveraged by promoting CPU cores to a power-saving C-state.

Scale resources appropriately. Examples are the following:

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

Select appropriate energy-performance trade-off. An example is the following:

- Optimize the power profile of a dense linear algebra algorithm (PLASMA) by focusing on the specific energy requirements of the various factorization algorithms and their stages.

Programming and performance analysis tools. An example is the following.

- Counters, accumulators, in-band support.

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

3.4 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.

Different power states. Careful about how we switch a power state off. Cannot affect reliability. Sleep states are probably the best direction. Response time is much better.

Energy-aware scheduling. Develop mechanism to automatically select processor frequency for which energy to solution is minimized for a specific application.

Demand response (as in the ability to react to electrical grid based incentives) requires enhanced scheduling tools.

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 there are available nodes to run the job; but if the job runs 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.

4 Benchmarks

(info) Since power and energy costs, both operational and capital, are increasingly significant, it is very important to understand the power and energy efficiency requirements of the system. This is best understood when running workloads, either applications or benchmarks. Each site will have to select the workloads to run as part of the procurement and acceptance process. These workloads may differentially exercise or stress various 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.

(mandatory) [Customer] shall specify the set of benchmarks [Customer] wants. Vendors shall provide the power and energy efficiency requirements as welll as the run times of a set of benchmarks.

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

(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.

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

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

(important) Vendors shall work with Customers to provide the power and energy effiency 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.

5 General 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.

5.1 Energy-related Total Cost of Ownership (TCO)

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

(info) An example of an innovative program for bringing the energy/power element of TCO to the front was used by LRZ. LRZ's 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.

5.2 Power Usage Effectiveness (PUE)

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

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

5.3 Total Usage Effectiveness (TUE)

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

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

5.4 Energy Re-Use Effectiveness (ERE)

(info) Some sites have the ability to utilize the heat generated by the datacenter 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 datacenter, but it can reduce the energy requirements for the surrounding environment. For those sites, it would be an objective of [Customer] to achieve an ERE <1.0.

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

5.5 Power Distribution

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

6 Cooling

6.1 Liquid Cooling

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

(info) [Customer] will specify the type of liquid cooling systems contained within the datacenter. 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 datacenter is cooled using compressor-based cooling (i.e., chillers or CRAC units) and additional heat rejection equipment such as cooling towers or dry coolers. These liquid-cooled systems operate within ASHRAE-recommended ranges W1 and W2. Systems designed to operate in these ranges will have limited energy efficiency capability.

(important) For improved energy efficiency and reduced capital expense, many datacenters can be operated without compressor-based cooling, by using cooling towers or dry coolers combined with water-side economizers. These datacenters 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 its procurement documents. ASHRAE provides guidance on these parameters, although they are not defined in this guideline.

6.2 Air Cooling

(info) ASHRAE Thermal Guidelines (2011) define environmental classes that allow temperatures up to 40°C and 45°C.

Figure 6.1 is a pschometric chart illustratiing these new environmental temperature and humidity limits along with the recommended limits.

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

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

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

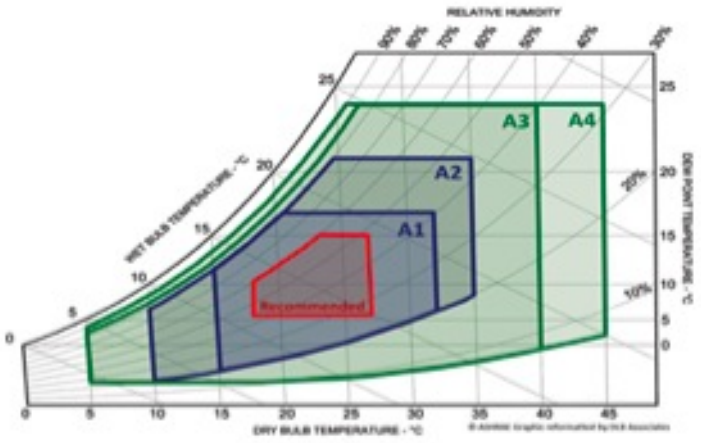

Figure 6.1: IT equipment environmental classes