Energy Efficient High Performance Computing Power Measurement Methodology

(version 1.2RC2)

Contents

List of Tables

List of Figures

1 Introduction

This document recommends a methodological approach for measuring, recording, and reporting the power used by a high performance computer (HPC) system. The document also discusses auxiliary data, such as environmental conditions, related to running a workload.

This document is part of a collaborative effort between the Green500, the Top500, the Green Grid, and the Energy Efficient High Performance Computing Working Group. While it is intended for this methodology to be generally applicable to benchmarking a variety of workloads, the initial focus is on High Performance LINPACK (HPL), the benchmark used by the Top500.

This document defines four aspects of a power measurement and three quality ratings. All four aspects have increasingly stringent requirements for higher quality levels.

The four aspects are as follows:

1. Granularity, time span, and type of raw measurement
2. Machine fraction instrumented
3. Subsystems included in instrumented power
4. Where in the power distribution network the measurements are taken

The quality ratings are as follows:

- Adequate, called Level 1 (L1)
- Moderate, called Level 2 (L2)
- Best, called Level 3 (L3)

The requirements for all four aspects for a given quality level must be satisfied to grant that quality rating for a submission.

2 Checklist for Reporting Power Values

When you are ready to make a submission, go to the Top500 (`http://www.top500.org/`) and Green500 (`http://www.green500.org/`) sites for more information.

This section contains a checklist for all the items of information you need to consider when making a power measurement.

Read through the list and ensure that you can record the needed information when you run your workload.

[ ] **Quality Level**

Choosing a quality level is the first important decision a submitter must make. Refer to Section 3.3 Aspect and Quality Levels for general information about the three quality levels. Sections 3.4 through 3.8 describe the details of the three quality levels

[ ] **Power Measurement Locations**

Measurements of power or energy are often made at multiple points in parallel across the computer system. A typical location might be the output of the building transformer. Refer to Section 3.8 Aspect 4: Point where the Electrical Measurements are Taken for more information about power measurement locations.

Note that in some cases, you may have to adjust for power loss. For information about power loss, refer to Section 3.8.1 Adjusting for Power Loss. If you adjust for power loss, how you determined the power losses must be part of the submission.

[ ] **Measuring Devices**

Specify the measuring device or devices used. A reference to the device specifications is useful.

Refer to Section 3.2 for some terminology about the measuring device specific to the power submissions described in this document. This section describes the difference between power-averaged measurements and total energy measurements.

Refer to Section 3.1 for information about the required measuring device.

If multiple meters are used, describe how the data aggregation and synchronization were performed. One possibility is to have the nodes NTP-synchronized; the power meter's controller is then also NTP-synchronized prior to the run.

[ ] **Workload Requirement**

The workload must run on all compute nodes of the system. Level 3 measures the power for the entire system. Levels 1 and 2 measure the power for a portion of the system and extrapolate a value for the entire system.

[ ] **Level 3 Power Measurement**

Level 3 submissions include the average power during the core phase of the run and the average power during the full run.

The core phase is usually considered to be the section of the workload that undergoes parallel execution. The core phase typically does not include the parallel job launch and teardown.

Level 3 measures energy. Power is calculated by dividing the measured energy by the elapsed time . The measured energy is the last measured total energy within the core phase minus the first measured total energy within the core phase.

Refer to Section 3.2.2 Power-Averaged and Total Energy Measurements for information about the distinction between energy and power.

The complete set of total energy readings used to calculate average power (at least 10 during the core computation phase) must be included, along with the execution time for the core phase and the execution time for the full run.

Refer to Section 3.4 Aspect 1: Granularity, Timespan and Reported Measurement for more information about the Level 3 Power Submission.

Refer to Section 3.5 for more information about the format of reported measurements.

For Level 3, all subsystems participating in the workload must be measured. Refer to Section 3.7 Aspect 3: Subsystems Included in Instrumented Power for more information about included subsystems.

With Level 3, the submitter need not be concerned about different types of compute nodes because Level 3 measures the entire system.

[ ] **Level 2 Power Measurement**

Level 2 submissions include the average power during the core phase of the run and the average power during the full run.

The complete set of power-averaged measurements used to calculate average power must also be provided. Refer to Section 3.2.2 Power-Averaged and Total Energy Measurements for the definition of a power-averaged measurement and how it differs from a total energy measurement.

For Level 2, the workload run must have a series of equally spaced power-averaged measurements of equal length. These power-averaged measurements must be spaced close enough so that at least 10 measurements are reported during the core phase of the workload. The reported average power for the core phase of the run is the numerical average of the 10 (or more) power-averaged measurements collected during the core phase.

Each of the required equally spaced measurements required for L2 must power-average over the entire separating space.

Refer to Section 3.4 Aspect 1: Granularity, Timespan and Reported Measurement for more information about the Level 2 Power Submission.

Refer to Section 3.5 for more information about the format of reported measurements.

For Level 2, all subsystems participating in the workload must be measured or estimated. Level 2 requires that the greater of $\frac{1}{8}$ of the compute-node subsystem or 10 kW of power be measured. It is acceptable to exceed this requirement.

The compute-node subsystem is the set of compute nodes. As with Level 1, if the compute-node subsystem contains different types of compute nodes, you must measure at least one member from each of the heterogeneous sets. The contribution from compute nodes not measured must be estimated. Refer to Section 3.7 Aspect 3: Subsystems Included in Instrumented Power for information about heterogeneous sets of compute nodes.

[ ] **Level 1 Power Measurement**

Level 1 requires at least one power-averaged measurement during the run. Refer to Section 3.2.2 Power-Averaged and Total Energy Measurements for the definition of a power-averaged measurement and how it differs from a total energy measurement.

The total interval covered must be at least 20% of the core phase of the run or one minute, whichever is *longer*.

If the choice is one minute (because it's longer than 20% of the core phase) that minute must reside in the middle 80% of the core phase. If the middle 80% of the core phase is less than one minute, the measurement must include the entire middle 80% and overlap equally on both sides.

If the choice is 20% of the core phase (because this 20% is greater than one minute), this 20% must reside in the middle 80% of the core phase.

Refer to Section 3.4 Aspect 1: Granularity, Timespan and Reported Measurement for more information about the Level 1 power submission.

For Level 1, the only subsystem included in the power measurement is the compute-node subsystem. The compute-node subsystem is the set of compute nodes. Measure the greater of $\frac{1}{64}$ of the compute-node subsystem or 1kW of power.

List any other subsystems that contribute to the workload, but do not provide estimated values for their contribution.

For some systems, it may be impossible not to include a power contribution from some subsystems. In this case, list what you are including, but do not subtract an estimated value for the included subsystem.

If the compute node-subsystem contains different types of compute nodes, measure at least one member from each of the heterogeneous sets. The contribution from compute nodes not measured must be estimated. Refer to Section 3.7 Aspect 3: Subsystems Included in Instrumented Power for information about heterogeneous sets of compute nodes.

[ ] **Idle Power**

Idle power is defined as the power used by the system when it is not running a workload, but it is in a state where it is ready to accept a workload. The idle state is not a sleep or a hibernation state.

An idle measurement need not be linked to a particular workload. The idle measurement need not be made just before or after the workload is run. Think of the idle power measurement as a constant of the system. Think of idle power as a baseline power consumption when no workload is running.

For Levels 2 and 3, there must be at least one idle measurement. An idle measurement is optional for Level 1.

[ ] **Included Subsystems**

Subsystems include (but are not limited to) computational nodes, any interconnect network the application uses, any head or control nodes, any storage system the application uses, and any internal cooling devices (self-contained liquid cooling systems and fans).

- For Level 1, all subsystems participating in the workload must be listed.

  Only the compute-node subsystem must be measured. Not every compute node belonging to the compute node subsystem must be measured, but the contribution from those compute nodes not measured must be estimated. Measure the greater of at least $\frac{1}{64}$ of the compute-node system or at least 1kW of power.

- For Level 2, all subsystems participating in the workload must be measured and, if not measured, their contribution must be estimated. The Measured % and the Derived % must sum to the Total %.

  In the case of estimated measurements for subsystems other than the compute-node subsystem, the submission must include the relevant manufacturer specifications and formulas used for power estimation.

  Measure the greater of at least 10KW of power or $\frac{1}{3}$ of the compute-node subsystem.

- For Level 3, all subsystems participating in the workload must be measured.

Include additional subsystems if needed.

Refer to Section 3.7 Aspect 3: Subsystems Included in Instrumented Power for more information about included subsystems.

Refer to Section 3.6 Aspect 2: Machine Fraction Instrumented for information about % requirements for Levels 1 and 2.

[ ] **Tunable Parameters**

Listing tunable parameters for all levels is optional. Typical tunable values are the CPU frequency, memory settings, and internal network settings. Be conservative, but list any other values you consider important.

A tunable parameter is one that has a default value that you can easily change before running the workload.

If you report tunable parameters, submit both the default value (the value that the data center normally supplies) and the value to which it has been changed.

[ ] **Environmental Factors**

All levels require information about the cooling system temperature. Reporting other environmental data (such as humidity) is optional.

Submissions require both the in and the out temperature of the cooling system. For air-cooled systems, these are the in and out air temperatures. For liquid-cooled systems, these are the in and out temperatures of the liquid.

Refer to Section 3.9 Environmental Factors for more information.

3 Reporting Power Values (Detailed Information)

Refer to this section for detailed information about the elements of a power submission.

This section describes the information that must be included with a power measurement submission. It also describes some optional information that submitters may decide to include.

The section contains definitions of the terms used to describe the elements of a power submission, some background information, motivation about why the list contains the elements it does, and any other details that may be helpful.

3.1 Measuring Device Specifications

Measuring devices must meet the Level requirements as defined in Sections 3.3 and 3.4. This section lists resources for finding and evaluating meters.

The ANSI specification for revenue-grade meters is ANSI C12.20.

Also, refer to the *Power and Temperature Measurement Setup Guide* and the list of accepted power measurement devices from the Standard Performance Evaluation Corporation.

`http://www.spec.org/power_ssj2008/`
`http://www.spec.org/power/docs/SPECpower-Device_List.html`

3.2 Measuring Device Terminology

Levels 1 and 2 specify power measurements. Level 3 specifies an energy measurement, but reports a power value.

3.2.1 Sampling

For Levels 1 and 2, power measurements must be sampled at least once per second. The actual measurements that constitute a sample may be taken much more frequently than once per second.

Sampling in an AC context requires a measurement stage that determines the true power delivered at that point and enters that value into a buffer where it is then used to

calculate average power over a longer time. So "sampled once per second" in this context means that the times in the buffer are averaged and recorded once per second. Sampling delivered electrical power in a DC context refers to a single simultaneous measurement of the voltage and the current to determine the delivered power at that point. The sampling rate in this case is how often such a sample is taken and recorded internally within the device.

If the submitter is sampling in a DC context, most likely it will be necessary to adjust for power loss in the AC/DC conversion stage. Refer to Section 3.7 Aspect 3: Subsystems Included in Instrumented Power for details.

3.2.2 Power-Averaged and Total Energy Measurements

The reported power values for Levels 1 and 2 are power-averaged measurements. A power-averaged measurement is one taken by a device that samples the instantaneous power used by a system at some fine time resolution for a given interval. The power-averaged measurement for the interval is the numerical average of all the instantaneous power measurements during that interval and constitutes one reported measurement covering that interval.

Consider Level 1, which requires only one reported power value. This reported power value may consist of several power measurements taken at a frequency of at least once per second and averaged over an interval. That interval must be at least 20% of the run or one minute, whichever is longer. For example, if power measurements may be taken once per second for one minute for a total of 60 power measurements and then averaged, Level 1 reports one power-averaged value.

Level 2 also requires that power measurements be taken at least once per second. Level 2 requires that power values be reported for both the core phase of the workload and the total workload. The reported power value for the core phase must be the result of at least 10 power measurements. These 10 power measurements may themselves be power-averaged measurements.

Each of the required equally spaced measurements required for L2 must power-average over the entire separating space. All the values reported by the meter must be used in the calculation.

For example, the meter may sample at one-second intervals and report a value every minute. Assume that the core phase is 600 minutes long. Assume further that the requirement for 10 equally spaced measurements can be satisfied with 10 measurements spaced 50 minutes apart. Using just those 10 measurements does not conform to this specification because all the values reported by the meter during the core phase are not used.

Although those two measurements were equally spaced over 500 minutes , each averaged only over a minute. So some (the majority) of the separating space between the measurements was not included in the average.

For Levels 1 and 2, the units of the reported power values are watts.

Level 3 specifies a total energy measurement that, when divided by the measured time, also reports power. An integrated measurement is a continuing sum of energy measurements. Typically, there are hundreds of measurements per second. Depending on the local frequency standard, there must be at least 120 or 100 measurements per second. The measuring device samples voltage and current many times per second and integrate those samples to determine the next total energy consumed reading.

Level 3 reports an average power value for the core phase, an average power value for the whole run, at least 10 equally spaced energy values within the core phase, and the elapsed time between the initial and final energy readings in the core phase. The average power value for the core phase is the difference between the initial and final energy readings divided by the elapsed time.

3.3 Aspect and Quality Levels

Table 3.1 summarizes the aspect and quality levels introduced in Section 1 Introduction

3.4 Aspect 1: Granularity, Timespan and Reported Measurements

Aspect 1 has the following three parts. Levels 1, 2, and 3 satisfy this aspect in different ways.

- The granularity of power measurements. This aspect determines the number of measurements per time element.
- The timespan of power measurements. This aspect determines where in the time of the workload's execution the power measurements are taken.
- The reported measurements. This aspect describes how the power measurements are reported.

For all required measurements, the submission must also include the data used to calculate them. For Level 2 and Level 3 submissions, the supporting data must include at least 10 equally spaced points in the core of the run.

Levels 2 and 3 require a number of equally spaced measurements to be reported for two reasons.

Table 3.1: Summary of aspects and quality levels

Aspect	Level 1	Level 2	Level 3
1a: Granularity	One power sample per second	One power sample per second	Continuously integrated energy
1b: Timing	The longer of one minute or 20% of the run	Equally spaced across the full run	Equally spaced across the full run
1c: Measurements	Core phase average power	• 10 average power measurements in the core phase • Full run average power • idle power	• 10 energy measurements in the core phase • Full run average power • idle power
2: Machine fraction	The greater of 1/64 of the compute subsystem or 1 kW	The greater of 1/8 of the compute-node subsystem or 10 kW	The whole of all included subsystems
3: Subsystems	Compute-nodes only	All participating subsystems, either measured or estimated	All participating subsystem must be measured
4: Point of measurement	Upstream of power conversion **OR** Conversion loss modeled with manufacturer data	Upstream of power conversion **OR** Conversion loss modeled with offline measurements of a single power supply	Upstream of power conversion **OR** Conversion loss measured simultaneously

- One is that facility or infrastructure level power measurements are typically taken by a system separate from the system OS and thus cannot be easily synchronized with running the benchmark.
- Secondly, with multiple periodic measurements, more reporting points are included before and after the benchmark run to ensure that a uniform standard of "beginning" and "end" of the power measurement can be applied to all the power measurements on a list.

There is no maximum number of reported points, although one reported measurement per second is probably a reasonable upper limit. The submitter may choose to include more than 10 such points.

The number of reported average power measurements or total energy measurements is deliberately given large latitude. Different computational machines will run long or short benchmark runs, depending on the size of the machine, the memory footprint per node, as well as other factors. Typically the power measurement infrastructure is not directly tied to the computational system's OS and has its own baseline configuration (say, one averaged measurement every five minutes). These requirements are specified not only to give a rich data set but also to be compatible with typical data center power measurement infrastructure.

All levels specify that power measurements be performed within the core phase of a workload. Levels 2 and 3 specify that a power measurement for the entire application be reported. Consequently, these levels require measurements during the run but outside of the core phase.

3.4.1 Core Phase

All submissions must include the average power within the core (parallel computation) phase of the run.

Every workload has a core phase where it is maximally exercising the relevant component(s) of the system. More power is often consumed in a workload's core phase than in its startup and shutdown phases. The core phase typically coincides with maximum system power draw.

For example, the core phase of the HPL workload is the portion of the core that actually solves the matrix. It is the numerically intensive solver phase of the calculation. Note that HPL now contains an HPL_timer() routine that facilitates power measurements.

3.4.2 The Whole Run

Level 2 and Level 3 submissions must also include the average power for the whole run, from the initiation of the job to its completion.

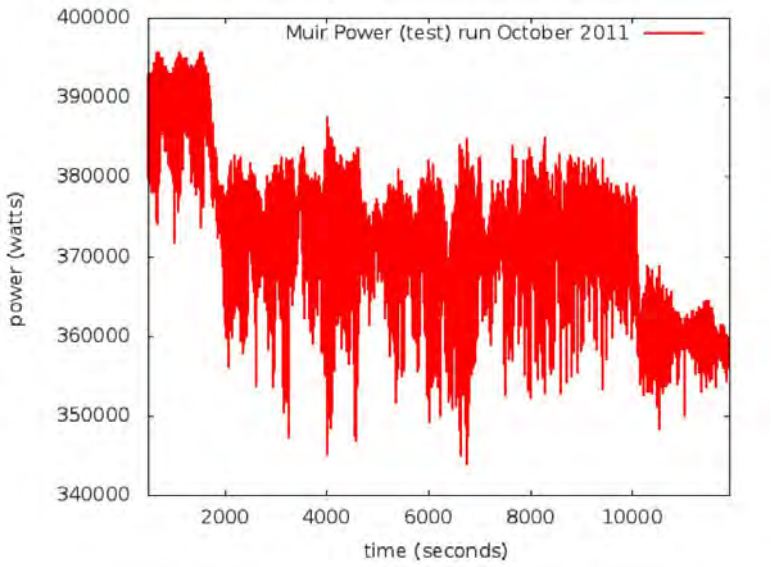

Figure 3.1: Power Profile HPL Run

Since HPL only reports the time spent executing the core phase of the workload, the time for the total run must be measured and reported separately by the submitter, for example by prepending the UNIX time command to the job invocation or the parallel application launch, whichever is earlier.

Levels 2 and 3 require the entire run to be measured and reported because HPL (the default workload at the time this document was written) drops significantly in power consumption during the course of a computation, as the matrix size being computed gets smaller. Requiring the entire run eliminates systematic bias caused by using different parts of the run for the measurement. Figure 3.1 shows an example of a power graph taken from an HPL run on the LLNL Muir system in the fall of 2011.

Note that the power drops by 8% or so during the computational phase of the run. This graph also illustrates the need for the device measurement to have as high a time resolution as possible. The spread of power measurements even within a small time span is very likely caused by the sampling going in and out of phase with the AC input power. Figure 3.2 shows a smaller time slice that clearly illustrates this.
The boxes are individual one-second power samples. This fast up-down fluctuation is not caused by the behavior of an individual power supply; the power being sampled is over the entire computer system doing the HPL run. The reason that L3 requires integrating total energy meters is because they measure at a high enough frequency to not be subject to these sampling artifacts.

L2 and L3 each decrease the device's inherent time measurement granularity. L2 allows 1-

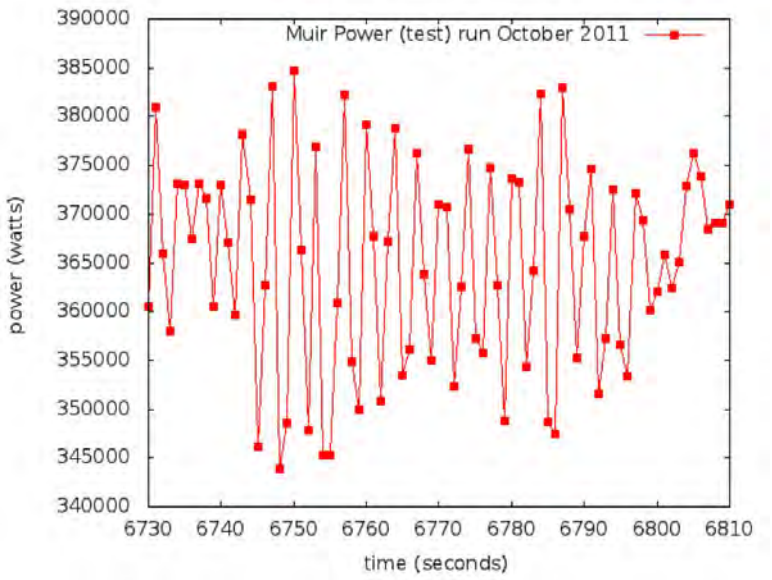

Figure 3.2: Spread of Power Measurements

second intervals of the input power. L3 requires an integrating total-energy meter, which samples the input power multiple times per AC cycle and so is much less susceptible to sampling artifacts caused by the AC waveform.

3.4.3 Level 1

The device measurement granularity must be at least one instantaneous measurement of power per second. This requirement holds whether the measurement is DC or AC.

There must be at least one power-averaged measurement during the run. The total interval covered must be at least 20% of the core phase of the run or one minute, whichever is longer. The power-averaged measurement must be taken within the middle 80% of the core phase of the benchmark.

Figure 3.3 illustrates Aspect 1 Level 1 power measurement. If the minimum timespan turns out to be one minute, that minute must reside in the middle 80% of the core phase. If, as required, the instantaneous measurement must be at least once per second and you are measuring for one minute, you have taken at least 60 instantaneous measurements and averaged them.

3.4.4 Level 2

Level 2 submissions include a measurement of the average power during the core phase of the run and the average power during the full run. The workload run must have a

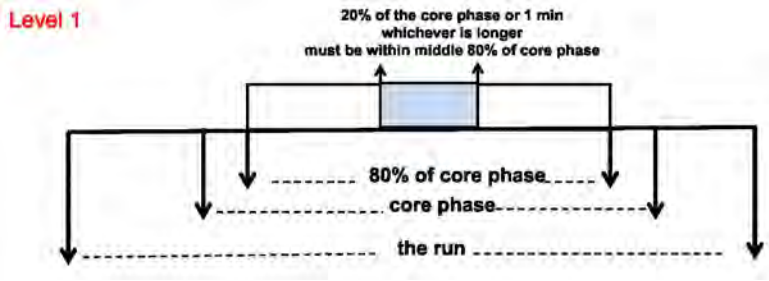

Figure 3.3: Aspect 1 Level 1 Power Measurements

series of equally spaced power-averaged measurements of equal length. These power-averaged measurements must be spaced close enough so that at least 10 measurements are reported during the core phase of the workload. The reported average power for the core phase of the run will be the numerical average of the 10 (or more) power-averaged measurements collected during the core phase.

The complete set of power-averaged measurements used to calculate average power must also be provided. The device measurement sampling granularity must be at least one instantaneous measurement of power per second.

There is some unspecified number of power-averaged measurements during the workload but outside of the core phase. The reported average power for the whole run will be the numerical average of the power measurements for the whole run.

Figure 3.4 illustrates Aspect 1 Level 2 power measurement. Each measurement is an average of instantaneous power measurements, and these instantaneous measurements are taken once per second. As an example, the figure shows 10 power-averaged measurements within the core phase and four power-averaged measurements outside the core phase, two before the core phase and two after the core phase.

3.4.5 Level 3

Level 3 submissions include a measurement of the average power during the core phase of the run and the average power during the full run.

The complete set of total energy readings (at least 10 during the core computation phase) must be included, along with the execution time for the core phase and full run.

Level 3 requires continuously integrated total energy measurements rather than power-averaged measurements. The readings must begin before the start of the run and extend to when it is finished.

The measuring device must sample voltage and current, whether AC or DC, at least 120

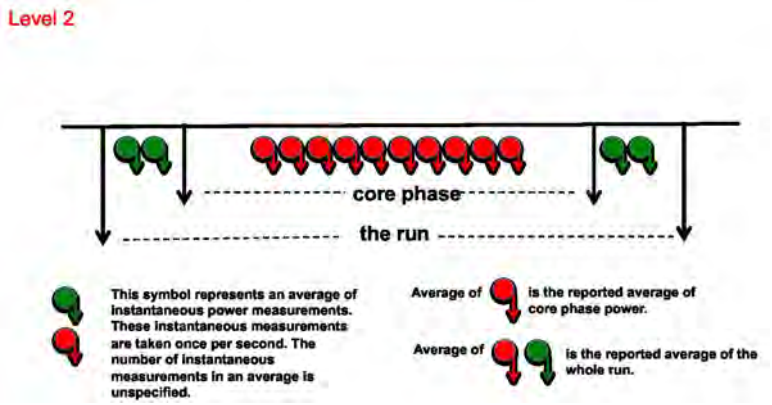

Figure 3.4: Aspect 1 Level 2 Power Measurements

times or 100 times per second (depending on what the frequency standard in the current location is) and integrate those samples to determine the next total energy consumed reading. Sampling at a greater rate is permitted.

The reported total energy readings must be spaced so that at least 10 reported readings fall within the core phase of the workload. Note that each reported reading is the average of many samples.

Figure 3.5 illustrates Aspect 1 Level 3 power measurement. The figure shows 10 readings in the core phase of the workload. Note that these are integrated readings. To obtain a power reading, one must subtract two integrated readings and divide by the time between the readings.

3.5 Format of Reported Measurements

Levels 2 and 3 require the complete set of measurements. The submitter may choose to provide these values in a CSV file. Do not provide scans of paper documents.

The submitter may find it useful to create a graph showing the power and energy during the workload as shown in Figure 3.6. Keep this graph for reference, but do not provide it as part of the submission.

3.6 Aspect 2: Machine Fraction Instrumented

Aspect 2 specifies the fraction of the system whose power feeds are instrumented by the measuring equipment.

Level 3 requires that the entire machine be measured. Level 2 requires a higher fraction

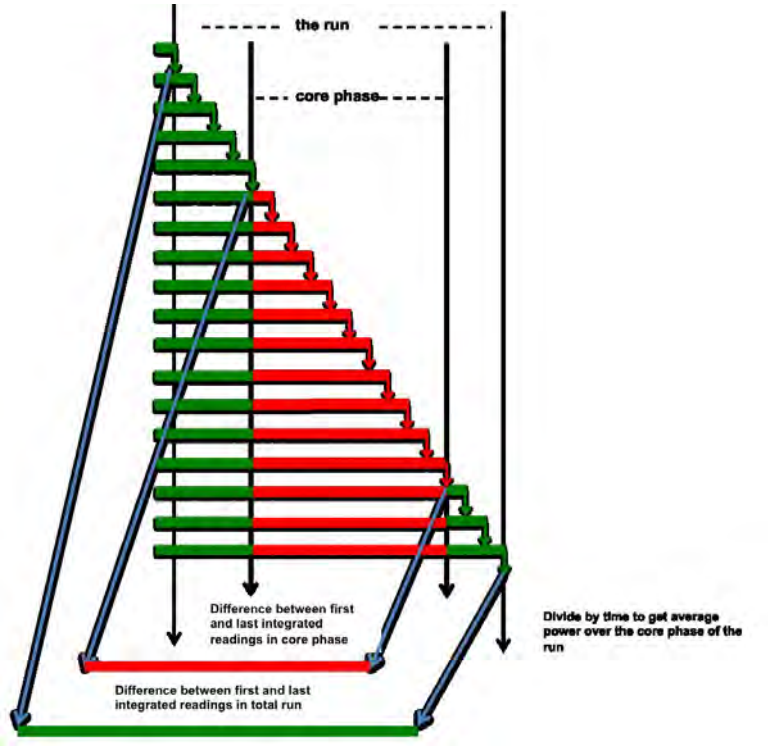

Figure 3.5: Aspect 1 Level 3 Power Measurements

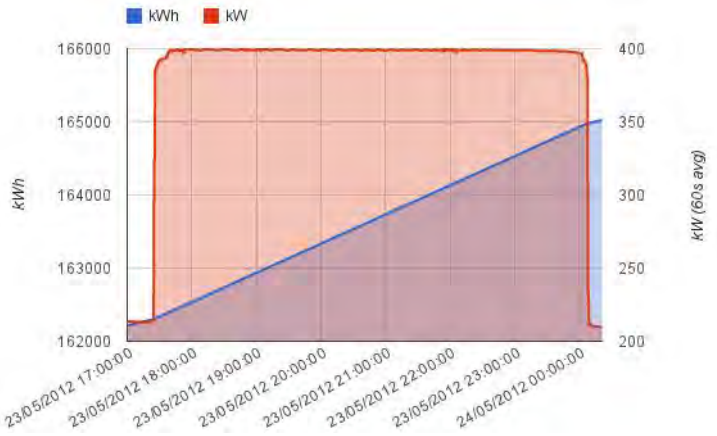

Figure 3.6: Power and Energy During the Workload
(used with permission from Universit Laval, Calcul Qubec, Compute Canada)

than Level 1.

When calculating the average power of the full machine for Levels 1 and 2, the measured power must be divided by this fraction to estimate the average power drawn by the whole machine. For example, if the submitter measures the power delivered to $\frac{1}{4}$ of the machine, the submitter must then multiply the measured power by 4 to estimate power for the whole machine. The higher machine fractions are required at the higher quality levels to reduce the effects of random fluctuations and minor differences in hardware influencing the power measurements. The larger the sample, the more transients will tend to cancel out.

The requirements for each quality level are as follows.

- L1: $\frac{1}{64}$ of the compute-node subsystem or 1 kW, whichever is greater
- L2: $\frac{1}{8}$ of the compute-node subsystem or at least 10 kW, whichever is greater
- L3: the power use of the whole machine must be measured

3.7 Aspect 3: Subsystems Included in Instrumented Power

Aspect 3 specifies the subsystems included in the instrumented power.

Subsystems in the context of this document are power subsystems. A power subsystem is that part of a supercomputer which can be measured in isolation for power consumption while the supercomputer is performing a task.

Subsystems include computational nodes, any interconnect network the application uses, any head or control nodes, any storage system the application uses, and any internal cooling devices (self-contained liquid cooling systems and fans).

If some subsystems are part of the measured power, their power may not be subtracted out after the fact. The explicitly measured value must be used as is.

For Level 1, only include the compute-node subsystem. For some systems, it may be impossible not to include a power contribution from some subsystems. In this case, provide a list of the measured subsystems, but do not subtract an estimated value for the included subsystem.

For Level 2, include the compute node subsystem. Other subsystems participating in the workload must be measured or estimated. The estimations must be based on derived numbers from the equipment manufacturer's specifications.

For Level 3, all power going to the parts of a computer system that participate in a workload must be included in the power measurement.

For Level 3, the reported power measurement must include all computational nodes, any interconnect network the application uses, any head or control nodes, any storage system the application uses, all power conversion losses inside the computer, and any internal cooling devices (self-contained liquid cooling systems and fans).

For Levels 2 and 3, the reported power measurement may exclude storage subsystems that don't participate in the workload. It is not required to exclude such storage subsystems. However, if these storage subsystems are part of the cabinet or rack being measured, they may not be excluded even if they are not used. That is, the submitter cannot calculate their contribution and subtract that contribution. If the storage subsystem is not part of the rack or cabinet being measured and it does not participate in the workload, it need not be measured.

In some cases, the submitter may be measuring power that the application doesn't actually use. If the submitter can exclude the unused subsystems from the measurement or easily turn off the power to the unused subsystems, then the submitter can choose not to include those subsystems in the measurement.

For example, the node board may include compute nodes and GPUs, and the application may not actually use the GPUs. If you cannot easily shut down the GPUs (say with an API), you must still include the power that they use. It is not acceptable to measure the power for both the compute nodes and the GPUs and then subtract the GPU power from the measurement.

A site may include more subsystems than are strictly required if it chooses or if it is advantageous from a measurement logistics point of view.

A particular system may have different types of compute nodes. The system may have compute nodes from different companies or even compute nodes with different architectures. These compute nodes are said to belong to different heterogeneous sets.

With Level 3, the submitter need not be concerned about heterogeneous sets of compute nodes because Level 3 measures the entire system.

Levels 1 and 2, however, measure a portion of the compute-node subsystem and estimate contributions from unmeasured portions. With Levels 1 and 2, the submitter must measure at least one member of each heterogeneous set. The submitter must include a power measurement from at least one compute node in each heterogeneous set and then estimate the contribution from the remaining members of the set.

For example, assume there exist two sets of compute nodes, a set called A and another called B. The submitter is able to measure the power consumed by $\frac{1}{2}$ of the A compute nodes and $\frac{1}{4}$ of the B compute nodes.

The total power measurement reported for compute nodes would then be
$Total\ power = 2 * (power\ from\ compute\ nodes\ A) + 4 * (power\ from\ compute\ nodes\ B)$

The assumption of Levels 2 and 1 is that all the compute nodes in a set react identically to the workload.

3.8 Aspect 4: Point where the Electrical Measurements are Taken

Aspect 4 specifies where in the power distribution system the power delivery is measured. For all quality levels, the submission indicates where power is measured and the quantity of parallel measurements points.

Measurements of power or energy are typically made at multiple points in parallel across the computer system. For example, such locations can be at the entrance to each separate rack, or at the exit points of multiple building transformers.

All the reported measurements taken in parallel at a given instant in time are then summed into a total measurement for that time. The total measurement for a given moment in time constitutes one entry in the series of measurements that becomes part of the submission.

AC measurements are upstream of the system's power conversion. If the measurements are in a DC context, the submitter may have to take into account some power loss. Refer to Section 3.8.1 Adjusting for Power Loss.

Electrical power or energy measurements shall be taken in one of the following locations.

A) At a point upstream of where the electrical supply from the data center is delivered to the computer system

   OR

B) At the point of the electrical supply delivery to the computer system from the data center

   OR

C) At a point just inside of the computer system that is electrically equivalent to location B) above. This includes the following.

   - At any point within a passive PDU, at the input to the PDU, at the exit point(s) of the PDU, or anywhere in between
   - At the entry point to the first power-modifying component (for example, the Blue Gene bulk power module, Cray AC/DC converters, and possibly the input point to one or more crate power supplies)

If the measuring device or devices used to satisfy Aspect 1 also meet the ABC location requirements just specified above, then those devices are sufficient to obtain the measurements needed for submission.

3.8.1 Adjusting for Power Loss

If the measurement device(s) that satisfy Aspect 1 are downstream of the ABC locations specified above, then two sets of measurements must be taken in order to determine the power loss between the required and the actual measurement location.

- For a Level 1 measurement, the power loss may be a load-varying model based on specifications from the manufacturer.
- For a Level 2 measurement, the power loss may be a load-varying model based on an actual physical measurement of one power supply in the system.
- For a Level 3 submission, the power loss must be measured simultaneously by a device at the required point (one of the ABC locations) measured at least once every five minutes, averaged long enough to average out the AC transients.

For all three levels, the power losses used and how they were determined must be part of the submissions.

3.8.2 Data Center Schematic

Figure 3.7 is an example of a simple power measurement schematic. This example shows only one power measurement point.

Submitters may find it useful to create such a schematic to identify the power measurement locations. Keep this schematic for reference, but do not provide it as part of the submission.

3.9 Environmental Factors

Reporting information about the cooling system temperature is required. All Levels require both the in and the out temperature of the cooling system. For air-cooled systems, these are the in and out air temperatures. For liquid-cooled systems, these are the in and out temperatures of the liquid.
Even though large systems often have multiple levels of heat energy exchange, the cooling method here refers to how the CPU is cooled. If the CPU has a heat sink and air blows over it, then it is air-cooled. If the CPU is embedded in water or some other refrigerant, then it is liquid-cooled.

For air-cooled systems, the air temperature both at the entrance to the computer system and at the exhaust must be measured sometime during the core phase of the computational run. These temperatures must be reported as part of the submitted data.

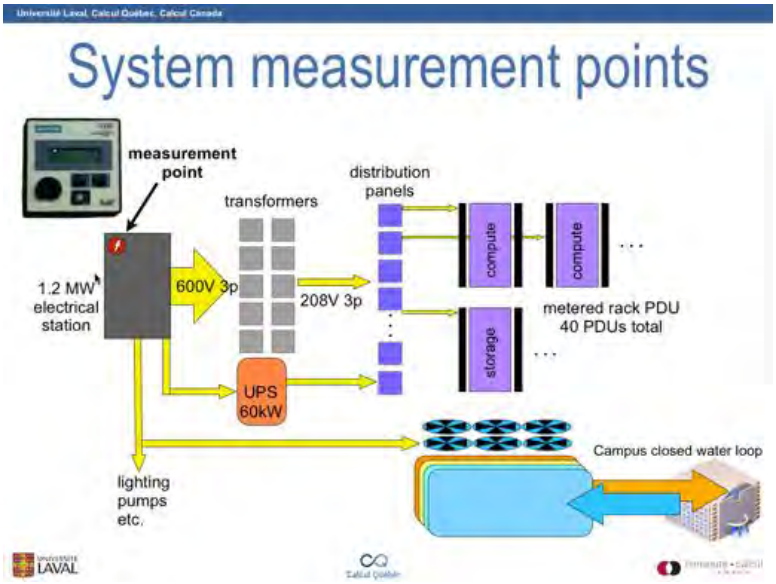

Figure 3.7: Example of a Power Measurement Schematic
(used with permission from Universit Laval, Calcul Qubec, Compute Canada)

For liquid-cooled systems, the temperature of both the incoming and outgoing cooling liquid must be measured sometime during the core phase run, and that temperature must be reported in the submission.

All other environmental data is optional. Other environmental data may include factors such as:

- % deviation between supply and rated voltage and frequency (recommended +/- 5%)
- % total harmonic distortion (recommended <2% THD)
- line impedance (recommended <0.25 ohm)
- relative humidity

4 Change Notices

A Change Management process establishes an orderly and effective procedure for tracking the submission, coordination, review, and approval for release of all changes to this document. For a description of this process and a list of change notices, refer to `https://sites.google.com/a/lbl.gov/eehpcwg/documents`.

5 Conclusion

This document specifies a methodology for measuring power that is used in conjunction with workload(s) to support the use of metrics that characterize the energy efficiency of high performance computing (HPC) systems as a function of both computational work accomplished and power consumed. It reflects a convergence of ideas and a reflection of best practices that form the basis for comparing and evaluating individual systems, product lines, architectures and vendors.

This document is intended for those involved in the HPC computer system architecture design and procurement decision-making process including data center and facilities designers/managers and users.

This document was a result of a collaborative effort between the Green500, the Top500, the Green Grid and the Energy Efficient High Performance Working Group.

6 Definitions

CSV file

A comma-separated values (CSV) file stores tabular data (numbers and text) in plain-text form. A CSV file is readable by most spreadsheet programs.

Metric

A basis for comparison; a reference point against which other things can be evaluated; a measure.

Methodology

The system of methods followed in a particular discipline; a way of doing something, especially a systematic way; implies an orderly logical arrangement (usually in steps); a measurement procedure.

Network Time Protocol (NTP)

Time Protocol (NTP) is a networking protocol for clock synchronization between computers.

Power-Averaged Measurement

A measurement taken by a device that samples the instantaneous power used by a system at some fine time resolution (say, once per second) for a given interval (say, 10 minutes). The power averaged measurement for the interval is the numerical average of all the instantaneous power measurements during that interval, and constitutes one reported measurement covering that interval.

Sampling

Sampling delivered electrical power in a DC context refers to a single simultaneous measurement of the voltage and the current to determine the delivered power at that point. The sampling rate in this case is how often such a sample is taken and recorded internally within the device. Sampling in an AC context requires a measurement stage, whether analog or digital, that determines the true power delivered at that point and enters that value into a buffer where it is then used to calculate average power over a longer time. So "sampled once per second" in this context means that the times in the buffer are averaged and recorded once per second.

Workload

The application or benchmark software designed to exercise the HPC system or subsystem to the fullest capability possible.