Energy Efficient High Performance Computing Power Measurement Methodology

(version 2.0 RC 1.0)

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 while 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 (EEHPC WG). 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).

This document defines four aspects of a power measurement and three quality levels. All four aspects have requirements that become increasingly stringent at 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. Location of measurement in the power distribution network and accuracy of power meters.

The quality ratings are as follows:

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

To grant a given quality level for a submission, the submission must satisfy the requirements of all four aspects at that quality level or higher.

2 Checklist

This section contains a checklist with an overview of the information you need to consider when making a power measurement. Section 3 then provides a more detailed description of all these items.

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 3.2 Aspect and Quality Levels for general information about the three quality levels. Sections 3.3 through 3.7 describe the details of the three quality levels.

[ ] **Power Measurement Locations**

Measurements of power or energy are often made at multiple locations in parallel across the computer system. A typical location might be the output of the building transformer. Refer to Section 3.7 Aspect 4: Instrumentation Location 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 3.7.1 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 3.1 for some terminology about the measuring device specific to the power submissions described in this document. That 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 1 Power Measurement Summary**

Level 1 submissions include the average power over the entire core phase of the run (see 3.3.1 for definition of core phase).

For Level 1, the power during the entire core phase must be measured. The submitted value must be the average of all power readings taken during the core phase of the run. Refer to Section 3.1.3 for information on power-averaged measurements. The core phase is required to run for at least one minute.

Refer to Section 3.3 Aspect 1: Granularity, Timespan and Reported Measurements for more information about the Level 1 power submission.

For Level 1, both the compute-node subsystem and the interconnect power must be reported. The compute-node subsystem power must be measured. For the compute subsystem, measure one of the following:

- The entire machine
- At least 40 kW
- Whichever is largest of: a minimum of 2 kW of power, $\frac{1}{10}$ of the system, or 15 nodes

The power of interconnect subsystem participating in the workload must also be measured or estimated. Estimation is performed by substituting the measurement by an upper bound derived from the maximum specified power consumption of all hardware components. Include everything that you need to operate the interconnect network that is not part of the compute subsystem. This may include infrastructure that is shared, but excludes parts that are not servicing the system under test.

For some systems, it may be impossible to avoid including a power contribution from certain subsystems that are not used for the benchmark run. In this case, list what you are including, but do not subtract an estimated value for the subsystems that are not needed.

If the compute-node subsystem contains different types of compute nodes, measure at least $\frac{1}{10}$ of each of the heterogeneous sets, and extrapolate these measurements to the full system. Refer to Section 3.6 Aspect 3: Subsystems Included in Instrumented Power for information about heterogeneous sets of compute nodes.

[ ] **Level 2 Power Measurement Summary**

Level 2 submissions include the average power during the core phase of the run and the average power during the full run (see Section 3.3.1 for definition of core phase).

For Level 2, the power during the core phase and during the full run must be measured. As with Level 1, the submitted value for the core phase must be the average of all power readings taken during the core phase of the run. In addition, the average of all power readings during the full run must be submitted. Refer to Section 3.1.3 for information on power-averaged measurements. The core phase is required to run for at least one minute.

On top of these full measurements, Level 2 also requires a set of intermediate measurements in order to see how the power consumption varies over time. For this purpose, a series of power-averaged measurements of equal length at regular intervals must be submitted for the full run. These intervals must be short enough that at least 10 measurements are reported during the core phase of the workload. This series of measurements in total must cover the full run.

Refer to Section 3.3 Aspect 1: Granularity, Timespan and Reported Measurements for more information about the Level 2 Power Submission.

Refer to Section 3.4 for more information about the format of reported measurements.

For Level 2, the compute node subsystem must be measured and all other subsystems participating in the workload must be measured or estimated. As for Level 1, estimation is performed by substituting the measurement by an upper bound derived from the maximum specified power consumption of all hardware components. Level 2 requires that the largest of $\frac{1}{8}$ of the compute-node subsystem, or 10 kW of power, or 15 compute nodes be measured. It is acceptable to exceed this requirement or to measure the whole machine.

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 a fraction of each heterogeneous set. For Level 2, this fraction must be at least $\frac{1}{8}$ of each set. These measurements are then extrapolated to the full system. Refer to Section 3.6 Aspect 3: Subsystems Included in Instrumented Power for information about heterogeneous sets of compute nodes.

[ ] **Level 3 Power Measurement Summary**

Level 3 submissions include the average power during the core phase of the run and the average power during the full run (see 3.3.1 for definition of core phase).

Level 3 measures energy. The measured energy is the last measured total energy within the core phase minus the first measured total energy within the core phase. The final power is calculated by dividing this energy by the elapsed time between these first and last energy readings. These last and first measurements in the core phase must be timed such that no more than a total of ten seconds (five each at begin and end) of the core phase are not covered by the total energy measurement.

Refer to Section 3.1.3 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 in the submission, along with the execution time for the core phase and the execution time for the full run.

Refer to Section 3.3 Aspect 1: Granularity, Timespan and Reported Measurements for more information about the Level 3 Power Submission.

Refer to Section 3.4 Format of Reported Measurements for more information about the format of reported measurements.

For Level 3, all subsystems participating in the workload must be measured. Refer to Section 3.6 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.

[ ] **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; that is, 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, both the compute-node subsystem and the interconnect must be reported. The compute-node subsystem power must be measured. The interconnect subsystem participating in the workload must also be measured or, if not measured, the contribution must be estimated.

  Measure one of the entire machine; 40 kW or more; or the greatest of at least 2 kW of power, $\frac{1}{10}$ of the compute-node subsystem or 15 compute nodes.
- For Level 2, the compute node subsystem must be measured and all other subsystems participating in the workload must be measured and, if not measured, their contribution must be estimated.

  Measure the largest of at least 10 kW of power, or $\frac{1}{8}$ of the compute-node subsystem, or 15 compute nodes.

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

To estimate the power consumption of a subsystem when measurement is not possible, use an upper bound derived from the maximum specified power consumption of all hardware components. The submission must include the relevant manufacturer specifications and formulas used for power estimation.

Include additional subsystems if needed.

Refer to Section 3.6 Aspect 3: Subsystems Included in Instrumented Power for more information about included subsystems.

Refer to Section 3.5 Aspect 2: Machine Fraction Instrumented for information about measuring a subset of the compute subsystem and extrapolating.

[ ] **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**

Reporting information about the cooling system temperature is optional. It is requested to provide a description of the cooling system as well as where and how the temperature was measured.

Refer to Section 3.8 Environmental Factors for more information.

3 Reporting Power Values (Detailed Information)

This section contains:

- The exact requirements for each quality level of each aspect.
- A description of what must be included with a power measurement submission. It also describes some optional information that submitters may decide to include.
- 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 Terminology and Specifications

This section defines meter accuracy requirements. First, Subsection 3.1.1 defines the general accuracy requirements. The following section describe the different types of measurements and the required sampling rates. Measuring devices must also meet the Level requirements as defined in Sections 3.2 and 3.3.

3.1.1 Accuracy Requirements

For Level 3, it is required to use any revenue grade meter, any meter accepted by SPEC power, or any meter documented to have an accuracy of 1% or better. Level 2 requires a meter with a minimum documented accuracy of 2%. Level 1 requires a meter with a minimum documented accuracy of 5%.

The required accuracy as percentage refers to the relative error specified for the power meter. If the power meter has multiple measurement intervals with different accuracies, all measurement intervals used during the benchmark run must meet the above requirements. Error in this context refers to the sum of statistical and systematic error. It is acceptable if the error is documented as a hard bound, as the 1σ standard deviation, or as the full width at half maximum (FWHM).

All requirements and operating conditions that are required for the power meter to achieve its documented accuracy must be met. For Levels 1 and 2, it is acceptable to assume that the AC net power is a sine wave function.

If the site uses multiple independent power meters for a Level 2 or a Level 3 measurement, and if all employed power meters measure an identical fraction of the system, the requirements are relaxed in the following way.

- Each individual power meter must have at least a documented accuracy of 3%.
- The documented error of the individual meters divided by the square root of the number of employed meters must meet the above requirements of 2% for Level 2 and 1% for Level 3.

In this case, the second requirement ensures that the standard deviation of the measurement of the entire run, using error propagation, and assuming a Gaussian distribution, and assuming that the errors of the power meters are uncorrelated, meets the requirements. For example, assume 4 power meters for a Level 3 measurement, each measuring 25% of the total system. In that case, each meter must have a documented accuracy of $1\% \cdot \sqrt{4} = 2\%$ to get a final accuracy of 1%. In other words, you can use a single meter with 1% accuracy, 4 meters with 2% accuracy, or 9 meters with 3% accuracy to reach 1% for Level 3.

Revenue grade meters are defined by ANSI C12.20. For meters accepted by the SPEC power, 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`

3.1.2 Sampling and Power Readings

For Levels 1 and 2, the power meter must internally sample the instantaneous power at a constant frequency of at least once per second. Usually the net frequency varies by a small percentage around 50 Hz or 60 Hz. It is allowed that the measurement sample rate is not strictly constant, but it can be adjusted such that each sample interval covers a certain constant number of full sine waves. In that case, the intervals must not exceed one second by more than one sine wave, and the power readings must be weighted by the duration of the sample interval when building the final power average.

(We do not specify exactly how such a power sample must be measured in order to give the users greater flexibility in the choice of the power meter equipment. Usually such a power sample, which is taken internally at least every second, will again consist of multiple internal measurements. I.e. the power meter will internally measure voltage and current much more frequently than once per second. It will either integrate their product or fit sine waves to them, determine the power factor, and multiply the wave amplitudes by the power factor. In this way, it generates one power sample.)

Sampling in an AC context requires a measurement stage that determines the true power delivered at that point (denoted a *power sample*) 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 once per second a sample is added to the buffer used to calculate the average. 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 3.6 Aspect 3: Subsystems Included in Instrumented Power for details.

The power meter is not required to report every power sample it takes. Instead, the power meter may use an internal buffer and aggregate several samples over time and then only report one average power reading at a certain constant frequency. In the following, we refer to such a reported power measurement of the power meter as a *power reading*. The intervals between the power readings must be no longer than 10% of the duration of the core phase. If the power meter reports every sample, then a power sample and a power reading are the same.

The power meter must use all power samples taken during an interval to compute the average power reading. Consider that one sample per second is the minimum acceptable sampling rate. Usually, the power meter will take many samples per second and report power readings on about a per-second basis.

3.1.3 Power-Averaged and Total Energy Measurements

This subsection describes the difference of power-averaged measurements as required for Level 1 and Level 2 as compared to total energy measurements as required for Level 3. Levels 1 and 2 specify power measurements. Level 3 specifies an energy measurement, but reports a power value.

Power-Averaged Measurements (Levels 1 and 2)

The reported power values for Levels 1 and 2 are power-averaged measurements over a certain time period (e.g. the core phase or the full run). 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 specified interval. This interval is short compared to the full time of the power-averaged measurement. The power meter device numerically averages of all the instantaneous power measurements samples during that interval. This constitutes one power reading reported by the device covering that interval. Finally, the power-averaged measurement for the required time period is the average of all these power readings obtained from the device in that period. All power readings reported by the device for that period must be used except for one case: Exclude those readings for intervals that do not lie completely in the requested time period. (Usually the reading intervals and the measured time period (e.g. the core phase) are not synchronized.) We require that the internal device sampling rate must be at least 1 Hz and constant. The intervals between the power readings must be constant and no longer than 10% of the duration of the core phase. This ensures that the parts of the core phase that are not measured at the beginning and end are not significant. As noted in the previous subsection, the sampling rate may vary slightly to accomodate for variations in the net frequency. In that case, the averaging process must weight the power readings from the power meter according to the respective interval lengths.

Consider Level 1, which requires only one reported power value. This reported power value may consist of several power readings taken at a constant frequency with interval length shorter than 10% of the core phase. Each such reading itself must be the average of instantaneous power samples measured at least once per second and averaged over the reading interval. For Level 1, the time period that is measured must be the core phase of the run, which must be at least one minute long (see 3.3.1 for definition of core phase). For example, if the device reports one power measurement every 5 seconds (each being the average of at least 5 power samples) and the core phase takes 10 minutes, Level 1 reports one power-averaged value that averages over all the 120 measurements obtained during the core phase. (Usually the power reading intervals and the core phase are not synchronized, such that the first reading does not lie completely in the core phase. In that case the average is only over the 119 readings completely in the core phase.)

Level 2 also requires that power readings be taken at a constant frequency and the device must sample internally at least once per second. It requires the submission of the power-averaged measurement for the core phase (as Level 1 does), for the full run, and for a series of power-averaged measurements of equal length at regular intervals during the full run. The measurement interval must be short enough so that at least 10 measurements are reported during the core phase of the workload. This series of measurements in total must cover the full run. Alternatively, instead of this series of power-averaged measurements, it is acceptable to submit the full set of power readings from the device that are used to calculate the average power during the core phase and during the full run.

The measurement for the core phase for Level 2 is identical to the measurement in Level 1. Assume again one measurement every 5 seconds (consisting of 5 samples each), a core phase of 10 minutes and a full run of 15 minutes. In that case, the submission includes one power-averaged measurement over 120 power readings for the core phase, one power-averaged measurement over 180 readings for the full run, and for instance a series of 30 power-averaged measurements over 6 power readings each, once every 30 seconds.

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

Total Energy Measurements (Level 3)

Level 3 specifies a total energy measurement that, when divided by the measured time, also reports power. An integrated measurement is a continuous sum of energy measurements. The measuring device samples voltage and current many times per second and integrates those samples to determine the next total energy reading. Typically, there are hundreds of measurements per second. The power meter should sample voltage and current at a minimum of 120 Hz for a DC measurement and at a minimum of 5 kHz for an AC measurement.

Level 3 reports an average power value for the core phase, an average power value for the whole run, at least 10 energy values at regular intervals within the core phase, and the elapsed times between the first and last energy readings in both the core phase and the full run. These last and first measurements in the core phase must be timed such that no more than a total of 10 seconds (5 seconds each at begin and end) of the core phase are not covered by the total energy measurement. The final average power value for the core phase is the difference between the first and last energy readings divided by the elapsed time between first and last reading. The final value for the full run must be obtained analogously.

The difference between the initial and final energy readings are the total energy measurement in Joules. The reported values after the energy has been divided by the elapsed time are watts as for Levels 1 and 2.

3.2 Aspect and Quality Levels

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

3.3 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.

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, voltage and current sampled at 5 kHz for AC / 120 Hz for DC
1b: Timing	At equal intervals across the entire core phase of the run, which must be at least one minute	At equal intervals across the full run	At equal intervals across the full run
1c: Measurements	Core phase average power	• Core phase average power • Full run average power • 10 average power measurements in the core phase • Idle power	• Core phase total energy • Full run total energy • 10 total energy measurements in the core phase • Idle power
2: Machine fraction	The entire system; at least 40 kW; or the largest of $\frac{1}{10}$ of the compute subsystem, 2 kW, or 15 compute nodes	The greater of $\frac{1}{8}$ of the compute-node subsystem or 10 kW or 15 compute nodes - alternatively the entire system	The whole of all included subsystems
3: Subsystems	Compute-nodes measured, interconnect measured or estimated	Compute-nodes measured, all participating subsystems measured or estimated	All participating subsystem must be measured
4: Location of measurement	Upstream of power conversion **OR** Conversion loss modeled with manufacturer data	Upstream of power conversion **OR** Conversion loss modeled with off-line measurements of a single power supply	Upstream of power conversion **OR** Conversion loss measured simultaneously
4b: Meter accuracy	5%	2% (see Section 3.1)	1% (see Section 3.1)

- 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 measurement values at a regular sampling rate in the core phase of the run.

Levels 2 and 3 require reported measurement values at regular intervals for the following reasons:

- 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.
- With multiple periodic measurements, more measurement values are included before and after the benchmark run. This ensures that a uniform standard of *beginning* and *end* of the power measurement can be applied to all the power measurements from different sensors.

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

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 and 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.

3.3.1 Core Phase

All submissions must include the average power within the core phase of the run.

The core phase is usually considered to be the section of the workload that undergoes parallel execution. It typically does not include the parallel job launch and teardown and is required to run for at least one minute.

In order to correlate the power measurement and the performance measurement to obtain the power efficiency, power and performance must be measured over the exact same time period. Therefore, the core phase is defined as the time period that is used for the performance calculation. Sometimes, the exact same time is impossible to measure due to power meter restrictions. Therefore, Level 1 and Level 2 allow a small part at the beginning and at the end which is not measured (see Section 3.1.3).

For example, the core phase of the Linpack workload is the portion of the code that actually solves the matrix. It is the numerically intensive solver phase of the calculation. In case of the standard HPL implementation of the Linpack benchmark as provided by netlib.org (`http://www.netlib.org/benchmark/hpl/`) this is the time spent in the routing `HPL_pdgesv`. Note that HPL as of version 2.1 contains a timer routine that prints start and end time of the core phase to facilitate power measurements, which looks like:

```
HPL_pdgesv() start time Wed Jun 30 21:49:08 2015
HPL_pdgesv() end time Wed Jun 30 22:23:15 2015 .
```

3.3.2 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.

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 high enough sampling rates to not be subject to these sampling artifacts.

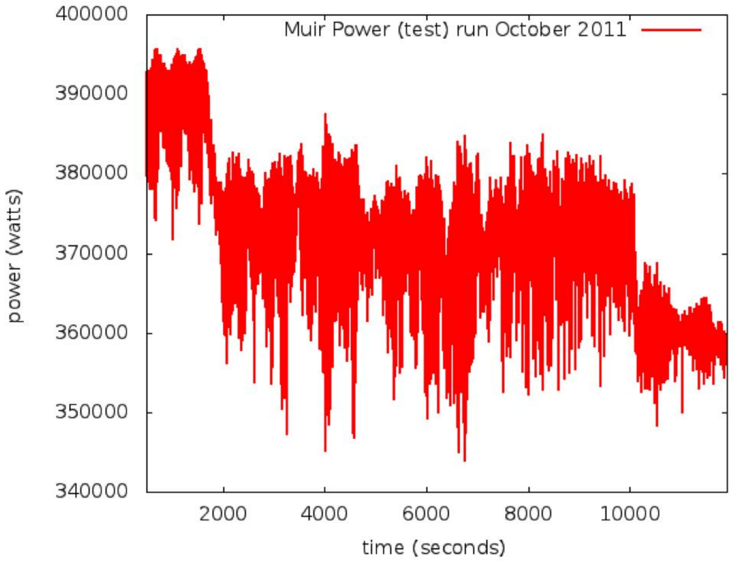


Figure 3.1: Power Profile HPL Run

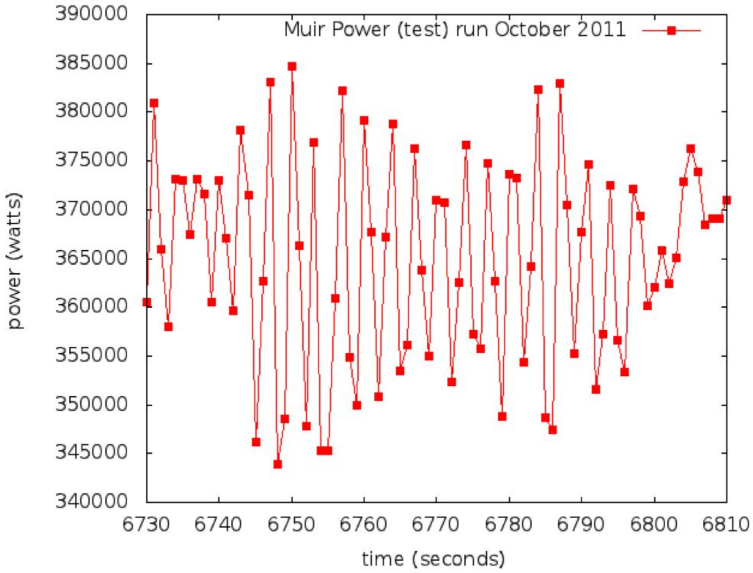


Figure 3.2: Spread of Power Measurements

L3 requires a power-meter device with a higher inherent measurement granularity as compared to L1 and L2. L1 and L2 allow 1-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.

3.3.3 Level 1

Level 1 submissions include a measurement of the average power during the core phase. 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 the core phase of the run which must be at least one minute long. Exclude power readings for time intervals that do not lie completely in the core phase.

3.3.4 Level 2

Level 2 submissions include a measurement of the average power during the core phase of the run, the average power during the full run, and a series of regular power-averaged measurements. 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.

The submitted value for the core phase must be the average of all power readings taken during the core phase of the run. In addition, the average of all power readings during the full run must be submitted. In both cases, exclude power readings for time intervals that do not lie completely in the respective time period of core phase and full run. The core phase is required to run for at least one minute.

On top of these full measurements, a series of power-averaged measurements of equal length at regular intervals must be submitted for the full run. These intervals must be short enough so that at least 10 measurements are reported during the core phase of the workload. This series of measurements in total must cover the full run.

Figure 3.3 illustrates the series of power measurements at regular intervals of Aspect 1 of a 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.

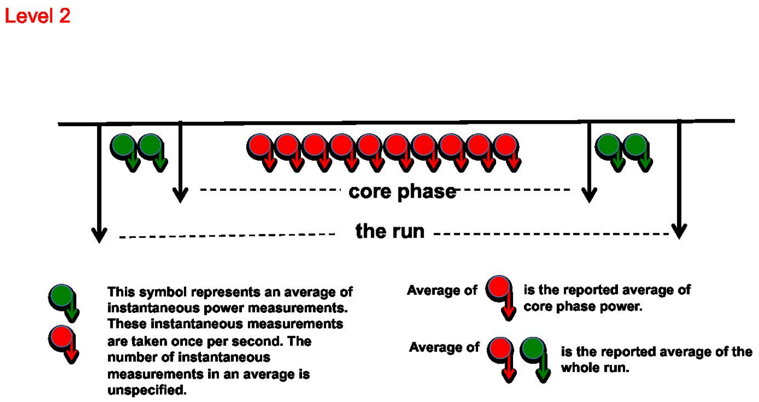

Level 2
core phase
the run
This symbol represents an average of instantaneous power measurements. These instantaneous measurements are taken once per second. The number of instantaneous measurements in an average is unspecified.
Average of
is the reported average of core phase power.
Average of
is the reported average of the whole run.

Figure 3.3: Aspect 1 of Level 2 Power Measurements

3.3.5 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 at high rate and integrate those samples to determine the next total energy consumed reading. The sampling rate must be at least 120 Hz for DC, and at least 5 kHz for AC. Sampling at a greater rate is permitted.

The intervals at which total energy readings are reported must be short enough so that at least 10 reported readings fall within the core phase of the workload. Note that each reported energy reading is the integral over many internal power samples.

The measured energy for the core phase is the last measured total energy within the core phase minus the first measured total energy within the core phase. The final power is calculated by dividing this energy by the elapsed time between these first and last energy readings. These last and first measurements in the core phase must be timed such that no more than a total of ten seconds (five each at the beginning and end) of the core phase are not covered by the total energy measurement. The procedure for the full run is analogous.

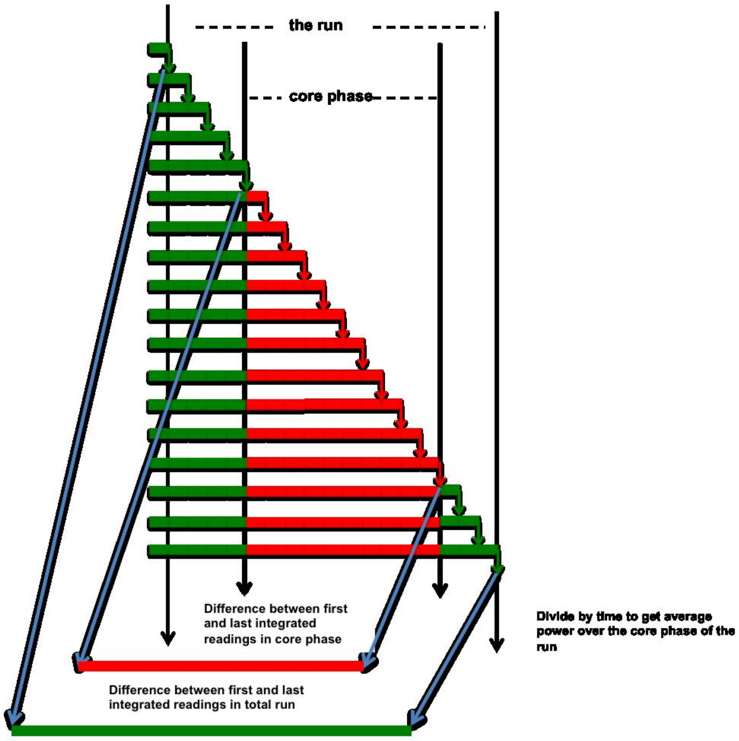

the run
core phase
Difference between first and last integrated readings in core phase
Difference between first and last integrated readings in total run
Divide by time to get average power over the core phase of the run

Figure 3.4: Aspect 1 of Level 3 Power Measurements

Figure 3.4 illustrates Aspect 1 of 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.

3.4 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 3.5. Keep this graph for reference, but do not provide it as part of the submission.

3.5 Aspect 2: Machine Fraction Instrumented

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

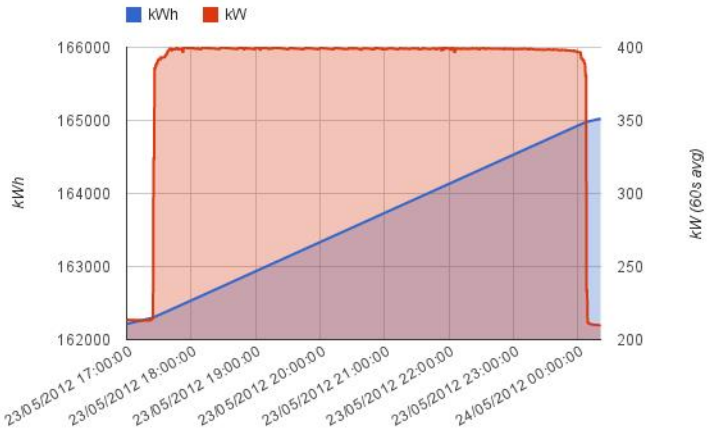


Figure 3.5: Power and Energy During the Workload
(used with permission from Université Laval, Calcul Québec, Compute Canada)

Level 3 requires that the entire machine be measured. Level 2 requires a higher fraction than Level 1.

Levels 1 and 2 do not measure the power of the entire system, but require only a measurement of a fraction of the system and estimate the total power consumption by extrapolation. When calculating the average power of the full machine for Levels 1 and 2, the measured power must be divided by this fraction to extrapolate to 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 obtain the 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 with respect to the compute-node subsystem for each quality level are as follows.

- L1: Any of
  - The entire machine
  - At least 40 kW
  - Whichever is largest of: a minimum of 2 kW of power, $\frac{1}{10}$ of the system, or 15 nodes
- L2: The largest of $\frac{1}{8}$ of the compute-node subsystem or at least 10 kW or 15 compute nodes. Alternatively, the entire system if smaller than 15 nodes.

- L3: The power use of the whole machine must be measured.

For measurements of other subsystems (network, storage, etc.), measure at least $\frac{1}{10}$ for Level 1, $\frac{1}{8}$ for Level 2, and everything for Level 3.

The fraction of a subsystem that is measured must be chosen such, that the entire subsystem is a real multiple of the fraction. For instance, a cluster consisting of blade servers with 8 nodes per chassis requires a measurement of the entire chassis, hence a multiple of 8 nodes. Even if individual node power could be measured, that would not be valid as it would not include the common chassis.

The requirements for Level 1 and Level 2 include the measurement of at least 15 nodes. For a cluster consisting of individual servers, a node is simply defined as one server. There are architectures with multiple individual servers in one chassis, e.g. 2U twin or twin^2 servers with 2 and 4 servers per chassis and blade servers with many more nodes in one chassis. Usually, all the nodes in one chassis share redundant power supplies. Hence, in that case you have to measure the entire chassis, including all the power supplies of a chassis. The measurement of a chassis is then considered a measurement of as many nodes as there are nodes in the chassis.

For a measurement of only a fraction of a subsystem, this fraction must be chosen randomly. In this context, random refers to the entity that is measured. For instance, if you measure nodes then choose random nodes, if you measure entire racks, choose random racks. It is not acceptable to screen the system for components with lower power consumption and then measure only that fraction, nor is it acceptable to replace components in the chosen fraction with others that have lower power consumption.

3.6 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 measurement. The explicitly measured value must be used as is.

For Level 1, both the compute-node subsystem and the interconnect must be reported. The compute-node subsystem power must be measured. The interconnect subsystem participating in the workload must also be measured or, if not measured, the contribution must be estimated. Include everything that you need to operate the interconnect network that is not part of the compute subsystem.

For Level 2, measure the compute node subsystem. Other subsystems participating in the workload must be measured or estimated.

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. 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 1 and 2, estimations of other subsystems besides the compute subsystem must be performed by substituting the measurement by an upper bound derived from the maximum specified power consumption of all hardware components.

Measurements may include infrastructure that is shared. It is allowed but not required to exclude subsystems that don't participate in the workload (e.g. storage subsystems). However, if these 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 subsystem is not part of the rack or cabinet being measured and it does not participate in the workload, it need not be measured. For some systems, it may be impossible not to include a power contribution from certain subsystems that are not used by the application. In this case, provide a list of the measured subsystems, but do not subtract an estimated value for the subsystems that are not used.

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.

In other words, for Level 2 and 3 you have to include everything that you cannot turn of during the benchmark run. It only allowed to exclude subsystems (or parts thereof) from the measurement in the first place, but their contribution must not be subtracted afterwards. For Level 1 this applies as well, but only for the network subsystem.

3.6.1 Heterogeneous Systems

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. Each heterogeneous set must consist of identical nodes.

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. Measure at least $\frac{1}{10}$ of each heterogeneous set for Level 1, and at least $\frac{1}{8}$ of each heterogeneous set for Level 2. Then, extrapolate the total power per each heterogeneous set as you would do for a non heterogeneous cluster. Finally, accumulate the contributions of all heterogeneous sets.

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

3.7 Aspect 4: Instrumentation Location 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.

Measurements of power or energy are typically made at multiple locations in parallel across the computer system. For example, such locations can be at the entrance to each separate rack, or at the output connector 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 3.7.1 Adjusting for Power Loss.

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

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

   OR

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

   OR

C) At a location just inside of the computer system that is electrically equivalent to location B) above. This includes the following.
   - At any location within a passive PDU, at the input to the PDU, at the output connector(s) of the PDU, or anywhere in between
   - At the input connector to the first power-modifying component (for example, the Blue Gene bulk power module, Cray AC/DC converters, and possibly the input connector to one or more crate power supplies)

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

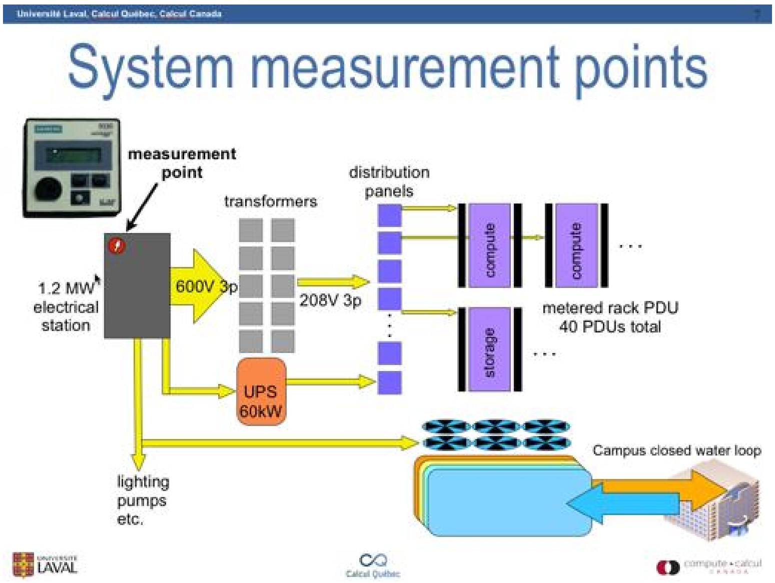
Refer to Section 3.1 for the specifications of the required accuracy of the power meter devices.

3.7.1 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 location (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.



Université Laval, Calcul Québec, Calcul Canada

System measurement points

Figure 3.6: Example of a Power Measurement Schematic
(used with permission from Université Laval, Calcul Québec, Compute Canada)

3.7.2 Data Center Schematic

Figure 3.6 is an example of a simple power measurement schematic. This example shows only one power measurement location.

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.

3.8 Environmental Factors

Reporting information about the cooling system temperature is optional. It is requested to provide a description of the cooling system as well as where and how the temperature was measured. Reporting temperature allows for better comparison of reported results, as there is a indirect correlation between temperature and power.

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 Management

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.

This document has been created in a github repository and the Change Management process leverages the features of github. This repository is public and anyone can access the document. The repository is located at `https://github.com/EEHPCWG/PowerMeasurementMethodology` Anyone can submit issues against the document. Editing the document is restricted to specific individuals from The Green500 and the Energy Efficient High Performance Computing Working Group. Github provides full history tracking for the entire life of the document. A released version of the document will be posted on the Green500 website.

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 Computing Working Group.

6 Definitions

Core Phase

The core phase is defined as the time period that is used for the performance calculation in the benchmark.

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)

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

Power-Averaged Measurement

A power-averaged measuremend over a certain time is the numerical average of all power readings of a power meter reported during that time. The power meter internally samples the instantaneous power used by a system at some fine time resolution (say, once per second). The power meter either reports every such power sample as a power reading or averages all power samples taken during a certain interval and then reports only one single power reading as measurement for that interval.

Sampling and Power Readings

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.

The power meter may either report every power sample that is measured. In that case, power sample and power reading are the same. Or the power meter may internally average all power samples over a certain interval and then report only one power reading covering that entire interval.

Workload

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