A Power-Measurement Methodology for Large-Scale, High-Performance Computing

Thomas R. W. Scogland
Virginia Tech
tom.scogland@vt.edu

Craig P Steffen
University of Illinois
csteffen@ncsa.illinois.edu

Torsten Wilde
Leibniz Supercomputing Ctr
Torsten.Wilde@lrz.de

Florent Parent
Calcul Québec
florent.parent@calculquebec.ca

Susan Coghlan
Argonne National Laboratory
smc@anl.gov

Natalie Bates
Energy Efficient HPC Working Group
natalie.jean.bates@gmail.com

Wu-chun Feng
Virginia Tech
feng@cs.vt.edu

Erich Strohmaier
Lawrence Berkeley National Laboratory
EStrohmaier@lbl.gov

ABSTRACT

Improvement in the energy efficiency of supercomputers can be accelerated by improving the quality and comparability of efficiency measurements. The ability to generate accurate measurements at extreme scale are just now emerging. The realization of system-level measurement capabilities can be accelerated with a commonly adopted and high quality measurement methodology for use while running a workload, typically a benchmark.

This paper describes a methodology that has been developed collaboratively through the Energy Efficient HPC Working Group to support architectural analysis and comparative measurements for rankings, such as the Top500 and Green500. To support measurements with varying amounts of effort and equipment required we present three distinct levels of measurement, which provide increasing levels of accuracy. Level 1 is similar to the Green500 run rules today, a single average power measurement extrapolated from a subset of a machine. Level 2 is more comprehensive, but still widely achievable. Level 3 is the most rigorous of the three methodologies but is only possible at a few sites. However, the Level 3 methodology generates a high quality result that exposes details that the other methodologies may miss. In addition, we present case studies from the Leibniz Supercomputing Centre (LRZ), Argonne National Laboratory (ANL) and Calcul Québec Université Laval that explore the benefits and difficulties of gathering high quality, system-level measurements on large-scale machines.

Categories and Subject Descriptors

C.4 [**Performance of Systems**]: Measurement Techniques

ICPE'14, March 22–26, 2014, Dublin, Ireland.

ACM 978-1-4503-2733-6/14/03 ...$15.00.
http://dx.doi.org/10.1145/2568088.2576795.

General Terms

Measurement, Performance, Energy, Power, Methodology

Keywords

Power-measurement methodology, Green500, Top500, high-performance computing, datacenter

1. INTRODUCTION

The energy efficiency of large-scale, high-performance computing (HPC) systems has become a key factor in design, procurement, and funding decisions. While many benchmarks exist for evaluating the computational performance of supercomputers, there remains a lack of standard methods for the accurate evaluation of energy efficiency at scale. In early 2011, the Energy Efficient HPC Working Group (EE HPC WG) [5] undertook a survey of power submissions to the Green500 [8] and Top500 [16] lists. The survey demonstrates that there is a wide variation in the quality of the measurements [6]. Some of the power submissions were very comprehensive and reflected a high level of quality. A number of the submissions were based on sampling and extrapolation and reflected a moderate level of quality. Even so, nearly half of the Green500 list power numbers were not based on measured power; rather they were derived from vendor specifications.[1] The survey identified the following methodology complexities and issues:

- Fuzzy lines of demarcation between the computer system and the data center infrastructure, e.g., fans, power supplies, liquid cooling
- Shared resources, e.g., storage and networking
- Limitations on data center and system instrumentation for system level power measurement

This paper describes the results of a collaborative effort led by the EE HPC WG with the Green500 [8], the Top500 [25] and The Green Grid [24] to address the complexities and

[1]Because submissions for power use in the Top500 were more sparse, the Green500 list was used as the larger sample set.

issues identified in the survey. The output of this collaborative effort is an improved power and energy measurement methodology for use with any system-level HPC benchmark. There are increasingly rigorous levels of measurement described by the methodology.

Few HPC systems today possess the instrumentation required to measure their power in its entirety without including other unrelated systems or subsystems in their datacenters. Further, adding large-scale or facility-level measurement equipment to existing systems and facilities is a difficult proposition. To accommodate systems that cannot feasibly be instrumented with the equipment necessary to produce the highest quality measurement, the methodology specifies three levels of measurement of increasing quality and complexity. Level 1 is similar to the Rev0.9 Green500 run rules, a single average power measurement extrapolated from a subset of a machine. Level 2 is more comprehensive, but remains a power measurement based on a subset of the overall system. Level 3 is the most rigorous of the three but only possible at a few sites. However, it offers a verifiably accurate full system measurement at any scale.

A broad community-based process was followed for developing the improved power measurement methodology [3]. The EE HPC WG has almost 400 members with 50% from government agencies, 30% from industry, and 20% from academic institutions. There are members from 20 different countries, mostly the United States and Europe. This methodology was generated and went through review with multiple opportunities for participation from the entire EE HPC WG. It also went through a review and approval from the Green500, the Top500 and The Green Grid. The methodology has gone through two testing phases with feedback from alpha testing resulting in modifications for the revision that went though beta testing. Both test phases included reporting on the test results to the broader community at major supercomputing conferences (ISC12 and SC12). The Green500 implemented this improved methodology as of its June 2013 List.

The rest of the paper is laid out as follows. Section 2 presents related work in measurement methodologies. Section 3 describes the methodology along with each of the levels. Sections 4 though 6 describe the experiences of the Leibniz Supercomputing Centre (LRZ), Argonne National Laboratory (ANL), and Calcul Québec Université Laval as illustrative case studies for the methodology. We present our conclusions in Section 7. Finally, in Section 9, we recognize as additional authors the invaluable contributions of the many people who participated in this collaborative and largely volunteer effort.

2. RELATED WORK

Benchmarking efforts have three inter-related and yet distinct elements: workload, metrics, and methodology. This paper focuses on the methodology, specifically the methodology used to measure system power while running an HPC workload.

There are several benchmarking efforts that attempt to characterize HPC architectural trends and that include a power measurement (some required and others optional). The most widely recognized benchmarking efforts are the Top500 [25] and Green500 [8], both of which use High Performance Linpack (HPL) [15] as the workload; additionally the Graph500 [1] has begun to gain traction with a workload that is focused on graph analysis and accepts power measurements as the Green Graph500 [2]. The Top500 accepts an optional power measurement, whereas it is the key thrust of the Green500. Since the inception of the Green500 [18], much work has been done to describe and evaluate a power-measurement methodology to use while running an HPC workload. Most of this work has been done for the Green500, but most recent and comprehensive is the exploration of the Green500's power measurement methodology limitations by Subramaniam and Feng [22].

The power-measurement methodology of the Standard Performance Evaluation Corporation (SPEC) [21] is likely the most cited methodology and framework for evaluating power and energy efficiency tied to a workload. This methodology was developed alongside the SPECpower benchmark, which evaluates the energy efficiency of a server, or set of servers, running the SPEC Java Business Benchmark (SPECjbb). Though it was developed for server workloads, the SPEC power measurement methodology and tools can be applied to others, as Hackenberg [10] demonstrates with his analysis of SPECMPI workloads run on a small cluster of nodes. The SPEC High Performance Group [19] has recently included the option and specification for power submissions as part of the SPEC OMP2012 benchmark suite [20] as an optional add-on and has since been analyzed by Muller [17]. While SPEC's methodology is precise and widely applied in multiple domains, it is not designed with supercomputer-scale systems in mind.

Whatever methodology it may be, the importance of having a common method goes beyond the benefit of an apples-to-apples comparison. One of the main purposes of benchmarking, particularly for the Green500 and Top500, is to analyze trends of the results over time or across designs. Analyses, such as Subramaniam and Feng's analysis of trends from the Green500 [23] and Hsu and Poole's analysis of trends in SPECpower [12], reveal information that would be at least obscured without a common methodology in each set of data. Yet more such trends might be found in far larger sets of data if a widespread common methodology can be established.

The intent of this effort is to push the envelope from prior related work with respect to power-measurement methodology and to do so along several dimensions, including the fraction of the machine that is instrumented, the time span for the measurement, and the machine boundary. In addition to improving the accuracy of the power-measurement methodology, we seek to accelerate the pace at which power-measurement capabilities evolve. For example, Hsu describes the evolving power measurement capabilities of HPC facilities and systems at seven levels (from the site all the way down to components) [11]. Our three-tiered quality levels are meant to raise the bar and accelerate the pace of evolution and adoption of high-fidelity, power-measurement capabilities.

In particular, this paper recommends a full-system, power measurement whereas Subramaniam [22] and Kamil [13] do not consider it a practical option and recommend extrapolating from a fractional system measurement. However, Laros [14] shows that there is value to full system measurements beyond improving the accuracy of the benchmarking effort. Benefits like those demonstrated by Laros partially motivated the full-system measurement encouraged by the EE HPC WG methodology.

3. MEASUREMENT METHODOLOGY

The methodology defined by the EE HPC WG defines three quality levels; essentially a *good*, *better*, or *best* rating system with Level 3 being the best. The quality ratings impose requirements on four different aspects of the power measurement:

1. The measurement itself, including the time span over which the measurement is taken, the time granularity, and the measurements reported
2. The fraction of the system that is instrumented
3. Subsystems that must be included in the measurement
4. Where in the power distribution network the measurements are taken

Increasingly stringent measurements are required in each of the four aspects for higher quality levels. Each level increases the measurement quality as well as its coverage of the machine infrastructure. For a measurement to qualify for a certain quality level, all aspects of the measurement must meet the requirements for that level, though they are allowed to exceed those requirements.

Level 1 is based on version 0.9 of the Green500 run rules [9]. We propose Level 3 as an enhanced energy measurement methodology that augments the ability to monitor and manage energy use. However, we understand that not all existing systems have the infrastructure to obtain Level 3 measurements. Hence, we define a Level 2 methodology as an intermediate step between Levels 1 and 3.

While each of the aspects listed above shifts for each level, there are several commonalities as well. All three levels require that you specify the power meter used. There are currently no requirements on the type or quality of power meter other than their sampling granularity. All levels also require that the power measurement is taken upstream of alternating current to direct current conversion, or accounted for by modeling or measuring the power lost during conversion.

The methodology distinguishes between the core phase of a workload and the entire workload. The core phase is the section of the workload that performs the main body of work evaluated or performed by the workload. The core phase does not include job launch and teardown time. While Level 2 and Level 3 require measurements across the entire run of an application, the span and frequency of measurements at all levels is defined in terms of the core phase. This decision was made in order to reasonably account for workloads with long setup and teardown times and short core phases that might otherwise focus the measurement on unimportant parts of the run.

Levels 2 and 3 require an idle power measurement. 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. As such, the idle measurement should be a near constant for a system given constant datacenter conditions, and the idle measurement need not be linked to a particular workload. The idle measurement may be made just before or after the workload is run, or independently so long as it is taken in the ready state. The idle measurement, while not a function of the workload being measured, serves as a baseline allowing the analysis of metrics such as static and dynamic power, energy proportionality, and others. These each offer important insights into the system as well as its interaction with the workload, revealing not only the energy consumed, but the amount consumed *because the workload is running*.

Table 1 summarizes the aspect and quality levels, with each defined in greater detail below.

3.1 Level 1

Level 1 requires at least one calculated power value. This value is the average of individual power measurements sampled at one-second intervals and taken over the required timespan. The required timespan is at least 20% of the workload's core phase or one minute, whichever is longer.

The requirement to sample at one-second intervals may be satisfied internally by the meter. All values reported by the meter need to be used in the calculation, though they may be aggregated at time scales larger than the sampling interval. For example, the meter may sample at one second intervals and report a value every minute, in that case measurements must be read once per minute and used in the calculation of the overall average power.

Level 1 requires that all the subsystems participating in the workload be listed, but that only the compute-node subsystem be measured. Level 1 requires that at least 1/64 of the compute-node subsystem or at least 1kW be measured, whichever is larger. The contribution from the remaining compute nodes is estimated by scaling up the sampled portion by the fraction that is monitored. If the compute node subsystem contains more than one type of compute nodes, at least one member from each type must be included in the measurement. The full system power should then be scaled proportionally for each type to determine the full system power.

In some circumstances it may be impossible to avoid a power contribution from other subsystems, such as integrated networking infrastructure in blade systems. In this case, subtracting an estimated value for the included subsystem is not allowed, but a list of what subsystems are included in this fashion may be included with the measurement.

3.2 Level 2

Level 2 requires two calculated average power values, one for the core phase of the workload and another for the entire workload. In addition, the complete set of measurements used to calculate these average values must be provided.

The complete set of measurements used to calculate the power values must be a series of equally spaced measurements taken during the run. These measurements must be included in the submission for verification purposes. The measurements must be spaced close enough so that at least 10 measurements are reported during the core phase of the workload, and a minimum of one each before and after the core phase. The reported average power for the core phase of the run is the numerical average of the measurements collected during the core phase. The reported average power for the whole run will be the numerical average of the power measurements for the whole run. As with Level 1, all reported measurements must be read, and all must be included in the average. Idle power must also be included, but may be taken as a separate event.

For Level 2, all subsystems participating in the workload must be measured or estimated. At least 1/8 of the compute-

Table 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 subsystems 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 off-line measurements of a single power supply	Upstream of power conversion **OR** Conversion loss measured simultaneously

node subsystem or at least 10 kW of power be measured, whichever is larger. The remaining part of the compute-node subsystem is extrapolated, and all types must be included, as with Level 1.

Other subsystems may be measured or estimated. If estimated, the submission must include the relevant manufacturer specifications and formulas used for power estimation.

3.3 Level 3

Level 3 submissions include the total energy over the course of the run, energy in the core phase, and the average power over those regions as computed from the energy. Each of these numbers is taken from a continuously integrated energy measurement as the total energy consumed to that point, each measurement will be the sum of the previous measurement and the energy consumed since it was taken, not as instantaneous power. In order to calculate the average power, the energy consumed in a given phase is computed by subtracting the total energy at the end of the phase from the energy at the start, and dividing by the time. 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. At least one measurement must fall before and one after the core phase. These must also be reported along with idle power. Unlike Levels 1 and 2, Level 3 need not be concerned about different types of compute nodes because Level 3 measures the entire system as a single unit. In addition to including the entire compute-node subsystem, all subsystems participating in the workload must be measured.

The measurements in the following sections of the paper are placed here as illustrations of the principles outlined in the power measurement specification. Normally, submissions would present one set of data targeted at the submission level desired by the submitter. These sections instead list multiple levels as illustrations of the constraints and ramifications of submitting at different quality levels with the intent of encouraging higher quality level measurements.

4. CASE STUDY FROM BADW-LRZ

The SuperMUC supercomputer at the Leibniz Supercomputing Center of the Bavarian Academy of Sciences and Humanities (BADW-LRZ) is one of the 10 fastest supercomputers in the world according to the June 2013 Top500 list. The data center housing SuperMUC was designed with an extensive monitoring infrastructure and provides many state-of-the-art measuring capabilities. This system is an ideal candidate to demonstrate the differences between the different levels described in the methodology.

The system consists of 18 islands, each of which are comprised of seven compute racks plus one network rack, and a total of 9,288 compute nodes. There are 2 power distribution units (PDUs) per rack; a PDU has 6 outlets and each outlet can power anywhere from 4 to 8 compute nodes.

SuperMUC is equipped with IBM 46M4004 Power Distribution Units, which are capable of sampling voltage, current and power at 120 Hz. Power values are averaged over 60 seconds and have a one minute readout interval. RMS current and voltage measurements have +/-5% accuracy over the entire range of supported voltages. For the SuperMUC tests, Level 1 and 2 power measurements are taken from the PDUs, sampled 120 times per second and recorded at ten equally spaced points once every 50 minutes.

Level 3 measurements used a Socomec Diris A40/A41 meter. This meter is a multi-function digital power and continuously integrating energy meter with a one-second internal measurement updating period and a 15-minute readout interval. The meter takes measurements up to the 63rd harmonic. The meter is International Electrotechnical Commission (IEC) 61557-12 certified. The energy accuracy is defined by IEC 62053-22 accuracy class index 0,5S. The power measurements have 0.5% accuracy. The meter is located after the transformers and after the dynamic UPS system and measures the power consumption for the entire room containing SuperMUC.

In order to provide easy comparison with widely available measurements, each case study is constructed around a run of the High Performance Linpack (HPL) [15] benchmark according to the version 1.0a EE HPC WG methodology

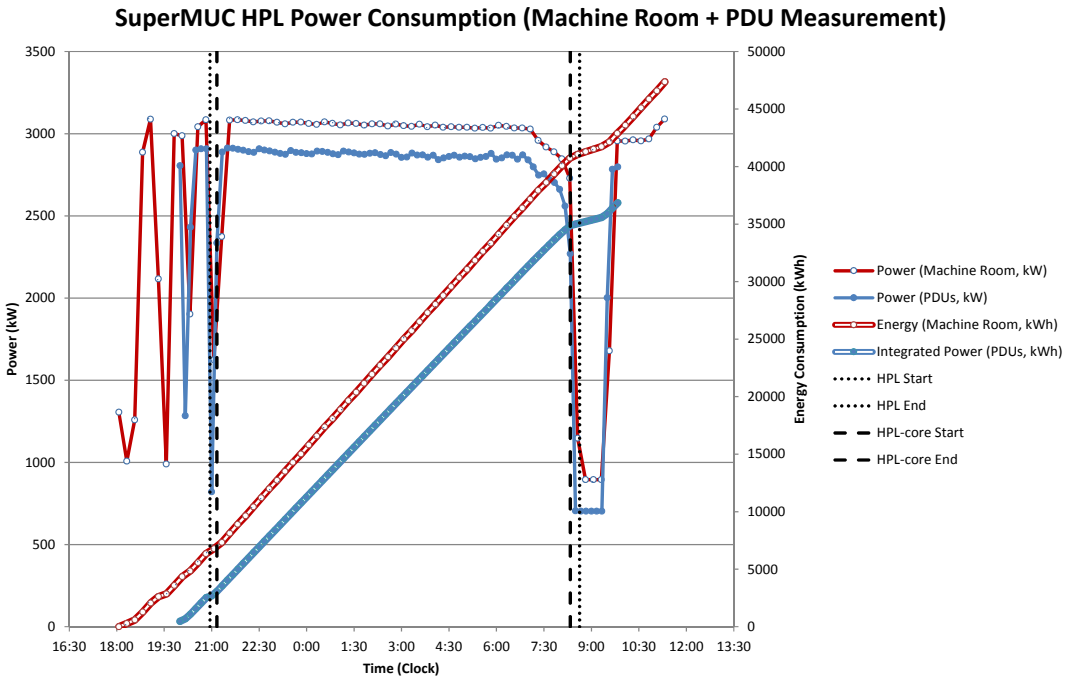

Figure 1: SuperMUC HPL power consumption ([machine room - Level 3] and [PDU - used for Level 1 and 2] measurement)

Table 2: SuperMUC: Level 1 results from one PDU outlet with 8 nodes

Average Power Location	Average Power Value
Average power for PDU outlet #6 with 8 nodes	2.126118kW
Average power per node	0.265765kW
Average power for the full machine during HPL core phase	2468.425kW

document [4].[2] The HPL run for SuperMUC is graphically represented in Figure 1. The run starts at 20:56 and ends at 8:37, for a duration of 701 minutes. The core phase, which is clearly visible in the power consumption of the benchmark as a long consistently high plateau of power consumption, starts at 21:10 and ends at 8:20 for a duration of 670 minutes. Given that length, the 20% of core phase runtime required for a level one measurement is 134 minutes.

4.1 Level 1

For Level 1, the reported value is the average over 140 minutes (23:20 - 01:40) which is just slightly more than 20% of the core phase. The power from one PDU outlet servicing 8 nodes is measured as 2126.118W, which is 265.77 watts/node. The extrapolated value for the entire machine (9288 compute nodes) is 2468.43 kW. Table 2 lists Level 1 power measurements.

[2]The current version of the methodology [3] requires the greater of 1/64th of the machine or 1kW for Level 1, and all measurements available during the run to be included at Level 2, when this study was conducted those requirements did not exist.

4.2 Level 2

Recall that Level 2 requires 1/8 of the system or 10kW, whichever is larger plus an additional system idle measurement. Fifteen power measurements were taken. The elapsed time of the Linpack run goes from 0 to 701 minutes with a measurement recorded every 50 minutes ending at 700min. The core phase begins at 14min and ends at 684min, placing 13 measurements within the core phase and satisfying the requirement for more than ten in that phase. The idle power is separately measured as 703kW for all of SuperMUC.

Level 2 also requires the measurement or estimation of all used subsystems (for example, networking). The two Ethernet switches in each compute rack of SuperMUC are automatically included when averaging over all PDU measurements for a rack. Additionally we must account for the power used by the InfiniBand and BNT-Ethernet switches which are located in a separate networking rack in each island. Recall that SuperMUC has 18 islands and hence 18 network racks, each of which has 10 PDU outlets; the system then has a total of 180 networking PDU outlets.

The power for one of these PDUs averaged over the full run is 415.15W and over the core phase is 416.08W. The value for the full system, all 180 networking racks, for the full run is 74.73kW, and the value for the core phase only is 74.89kW. The average power consumed by the IB switches does not depend significantly on the compute load of the system; it increases only by about 0.23%. This contrasts with the compute node behavior where the average power per node changes by about 10% between HPL core phase only and the full run.

In order to determine the skew introduced by choosing one or another of the minimum requirements for Level 2, either 1/8th of the system or greater than 10 kilowatts, each of the two is explored separately below.

4.2.1 Level 2 (1/8 System)

To measure 1/8 of the system, the power of 16 racks was measured. Of those fourteen contain 12 PDU outlets and 74 nodes, while the remaining two racks contain 72 nodes with the same number of PDU outlets. The total number of nodes measured come to 1,180 nodes, which is greater than 1/8 (1,161) of the total compute nodes (9288). The average power for a PDU is measured for both the full run, finding 1,670.04W, and for the core phase as 1848.36W.

To extrapolate to the entire compute subsystem, we multiply the average power for a PDU outlet by the number of PDU outlets measured, 192, divide by the number of nodes measured, 1,180, and multiply by the total number of compute nodes, or 9,288. Then, to extrapolate for the entire system, add the power measured for the network racks. Table 3 lists the details of the Level 2 power measurements.

4.2.2 Level 2 (>10 KW)

As compared to the 16 racks and 1,180 nodes required to conform to the 1/8th of the system requirement, the 10kW fraction of SuperMUC is miniscule. The power for just one rack with 2 PDUs, consisting of 6 outlets each, for a total of 12 PDU outlets and 74 nodes is measured. Since the fraction is so much smaller, a more efficient subset of the machine can be counted. The average power for a PDU outlet is lowered to 1,645.5W over the full run and 1,819.6 over the core phase only. Power consumption at the PDU outlet level is lowered by 24.54W and 28.76W respectively, changing nothing but the size of the machine fraction instrumented.

Extrapolating full-system power as before, the shifts seen in the PDU outlet-level measurements are magnified. At the full-system level, the smaller machine fraction consumes 2,553.1kW on average over the full run and 2,815.5 over the core phase. Counting the full system, the drop in power consumption shown by decreasing the fraction instrumented is 45.5 kilowatts over the full run or 52.75kW over the core phase. While the shift is less than 1/50 (i.e., 0.02) of the overall consumption, it is significant for large-scale systems. Therefore, the methodology requires the larger of the two to be used.

4.3 Level 3

Level 3 requires continuously integrated energy measurements of the complete system and as such, the data consist of accumulated energy reported every fifteen minutes. Figure 1 shows power and energy measurements vs. time. Over the full run the average power is computed to be 2,910.618kW, and 3,019.315kW over the core phase of the run. The Level 3 measurement is materially higher than either of the Level 2 measurements (357kW in the worst-case). Part of that difference comes from including portions of the system not included in the Level 2 measurement, for example the system data storage, and from really measuring all nodes. Another part comes from including infrastructure components such as the cooling system. This is in contrast to differences between the two different Level 2 measurements. There the main factor is the extrapolation of the final value from different system sizes.

4.4 Performance

When running on the entire SuperMUC system, all 9,288 nodes, the HPL benchmark reports an RMax of 2.582 Pflops.

Table 4: SuperMUC: Calculated megaflops/watt for the different quality levels

Quality Level	Mflops/Watt full run	Mflops/Watt core phase
L1	1055	1055
L2 (>10kW)	1011	917
L2 (>1/8)	994	900
L3	887	855

Table 4 lists the calculated efficiencies in Mflops/watt based on the power measurements gathered at each level.

Level 3 delivers the most accurate values, as it was obtained using a revenue-grade meter and measures the entire system including:

- Compute nodes
- Interconnect network
- GPFS mass storage systems
- Storage network
- Head/login and management nodes
- Internal warm water cooling system (machine room internal cooling such as water pumps, heat exchangers, etc.)
- PDU power distribution losses

Measurements for the lower levels were obtained using the PDUs, resulting in lower accuracy and additionally based only on the power of the compute and networking subsystems.

Since all requirements for each level were fulfilled, Table 4 illustrates the effect of attempting to compare results across disparate accuracy levels. As can be seen from the table the efficiency of the Level 1 core phase differs from that of Level 3 by 200Mflops/Watt or around 23% (1055 Mflops/Watt vs. 855 Mflops/Watt). This result is obtained without looking for the most energy efficient nodes for the Level 1 measurement. It is not hard to imagine that the difference could increase further if they were carefully selected for efficiency.

Level2 is closer to Level 3 (about 12% for the full run and about 5% for the HPL core run). But cherry picking would still be possible. Looking at the two possible requirements for a Level 2 the measurements show a difference of about 1.7%.

The comparison of the different measurement quality levels, that were all taken during the same run, shows that ranking different levels in one list would strongly favor lower level submissions. As a result of this, the Green500 has opted to require a level one measurement with every submission even if the submission also includes a higher level measurement to ensure comparability on the main list. We urge any other ranking organizations or procedures to perform separate rankings for each quality level, and discourage comparison across levels.

5. CASE STUDY FROM ARGONNE NATIONAL LABORATORY

The Argonne Leadership Computing Facility's (ALCF) new Mira supercomputer, sited at Argonne National Laboratory, debuted as the 3rd fastest supercomputer in the world (Top500, June-2012). Mira is a forty-eight rack IBM Blue Gene[3]/Q (BG/Q) with a peak compute performance of 10 PetaFlop/s (PF). The system has 49,152 compute nodes,

[3]Copyright 2013 by International Business Machine Corporation

Table 3: SuperMUC: Level 2 power measurements

	1/8th system		>10kW	
	Full run	HPL core phase	Full run	HPL core phase
Average power for one PDU outlet	1,670.039W	1,848.355W	1,645.500W	1,819.592W
Machine average power	2,523.875kW	2,793.357kW	2,478.391kW	2,740.601kW
Full machine + network	2,598.601kW	2,868.252kW	2,553.117kW	2,815.496kW

each with 16 cores and 16 gigabytes of memory, for a total of 786,432 cores and 786 terabytes of memory. The BG/Q is 93% water-cooled and 7% air-cooled, and, through innovative computer, network, and water-cooling designs, is one of the top energy efficient supercomputer architectures in the world.

The most challenging aspect of implementing the EE HPC WG power measurement methodology for Mira was inherently social/political and not technological. The ALCF BG/Q computers are located in the data center of the Theory and Computing Sciences (TCS) building. The TCS building is managed by a 3rd party. Because the building is not owned by ANL, the data from the building management system (BMS), which tracks energy usage over time, are not readily available to the tenants of the data center. Modifications to the system to add trending, gather data more frequently, etc., are infeasible for the same reason.

Because of these difficulties, ALCF was unable to measure Mira's HPL run at Level 3 for all aspects for an attempted submission in June 2012 to the Top500. Some Level 3 aspects were achieved. Table 5 shows the levels achieved for each aspect.

5.1 Measurement Specifications

As originally delivered, the IBM Blue Gene/Q provides several interesting power monitoring capabilities within the system, both at a coarse-grained level, with the Bulk Power Modules (BPMs) measuring one quarter of a rack each, and at a more fine-grained level, e.g. voltage domains within a node board (DCAs).[4] Figure 2 shows the locations of the monitors from the 480V distribution panels through to the fine-grained power monitor at the node board level.

Eaton Digitrip Optim 1050 meters measure the 480V power at the electrical distribution panels (Mira's racks are directly wired to these panels) mounted on the data center walls. The Eaton data are part of the BMS and are only readily accessible to the building management company. BPMs provide power measurements both upstream and downstream of the AC/DC power conversion. BPM data are gathered, along with other environmental data, into the overall control system database and are generally only accessible to the system administrator. The DCAs provide power measurements for the seven voltage domains on each node board and can be accessed by users from a running job. To determine overall power usage, the DCA power data would have to be adjusted for the power loss due to the AC/DC conversion using the BPM data.

The choice of measuring location depends not only on access but also on the planned usage for the data. For example, measurements at the distribution panels are impractical because measurements are taken every five minutes, one for each phase of their 3-phase AC input power supplies, resulting in only 30 data points per time step for the full Mira system. The data are also not kept beyond a few days, as the system tracking the data has very little storage space. The BPM measurements are primarily used to look at BPM efficiency and to monitor for potential problems with the BPMs. There are 36 BPMs in every rack (nine in a n+1 configuration for every 256 nodes), with four data points at each measurement (input and output current, input and output voltage), resulting in a total of 6,912 data points at each five-minute time step for Mira. The measurements are taken approximately every five minutes, for a total of 82,944 data points each hour. Users may request access to the BPM data, but the data are stored in a fire-walled database and are only provided as historical data. Therefore, if a user is interested in realtime power measurements, the user would use the DCA data, which are accessible from within a running job and provide much finer measurements both in time, measured every 560ms, and in space, 2 DCAs per every 32 nodes, for a total of 64 per rack. The DCA current and voltage are measured for each of seven voltage domains (described in Section 5.3) totaling 43,008 data points per time step, or 276,480,000 data points per hour. The DCA data would be appropriate for developing an application power signature, evaluating the impact of changes to an algorithm, or performing research into power management and reduction techniques for the next generation of computer systems.

Because data at the different locations are measured at different time scales and granularities, it can be challenging to compare the data between them. In addition, the panel measurements include other BG/Q racks, and, as such, cannot be accurately compared to the BPM and DCA data.

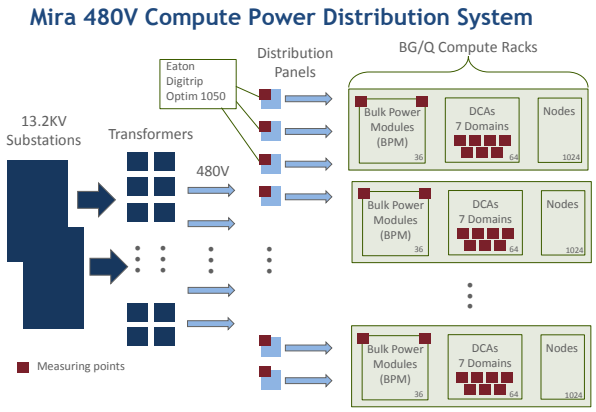

Figure 2: Power measurement locations for the Mira supercomputer

5.2 Mira Linpack Data

The Mira HPL run and evaluation is summarized in Figure 3, where the job starts at 23:21:30 and ends at 15:24:37

[4]See [26] for more information on the DCA power measurement capabilities

Table 5: Mira HPL aspect levels achieved

Aspect	Level Achieved	Notes
1a: Granularity	2	Bulk Power Modules (BPMs) sample instantaneously every 200us, but only measure average power and do not provide energy computed by continuously integrating power as required by Level 3
1b: Timing	2	195 point-in-time power averaged measurements were taken at equally spaced 5 minute intervals; Unable to measure integrated total energy values
1c: Measurements	2	Idle power measurement taken; more than 10 power measurements were taken within the core phase;
2: Machine fraction	3	All 48 racks (whole machine) were measured
3: Subsystems	3	Whole system
4: Point of measurement	3	Power measurements were taken both upstream and downstream of power conversion; power conversion measured simultaneously during the same run

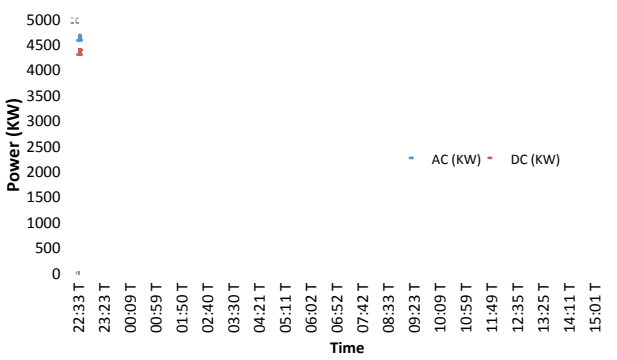

Figure 3: Power measurements during Mira Linpack run

Table 6: Mira Linpack data

Data Item	Value
Start Time	23:21:30
End Time	15:24:37
Performance (TFlops)	8,201
Mflops/Watt full run	1,824
Duration (s)	57,787.7
Idle power (kW)	1,549.10
Full run power (avg kW)	4,496.44

the next day for a duration of 963.1 minutes. BPM power measurements were automatically pulled by the control system and uploaded to the control system database every five minutes; these are plotted in Figure 3. Input power, specifically AC power at the entry to each BPM, is shown in blue, and output power (DC power at the exit of each BPM) is shown in red. The overall efficiency of the AC/DC power conversion across the entire time period measured was 93.3%.

Because the HPL code used was not modified to produce a timestamp at the start and end of the core phase [5], we do not know exactly when the core phase began and ended. From the Linpack output data, time was 57559.8 seconds; dgemm wall-time was 52784.8 seconds, and dtrsm wall-time was 342.1 seconds. The total time from the start of the job to the timestamp at the end of the job was 57787.7 seconds. Table 6 shows the Linpack measurements.

[5]Although HPL has since released this functionality, it was not readily available at the time of the test.

5.3 Reaching Level 3

Because the BPMs do not measure total energy, the Mira HPL power calculations were unable to reach Level 3. To address this issue, IBM, with a goal of achieving Level 3, has developed a firmware upgrade that can provide total energy measurements with sample rates over 2000 times per second at the DCAs. This modification was the direct result of IBM working with the EE HPC WG to determine the appropriate methodology.

The DCAs can be used by end users to measure CPU, memory, and network usage. The data gathered are downstream of the AC/DC conversion, and for a Level 3 full-system measurement these data would need to be adjusted for AC to DC conversion loss using measurements from the BPMs. Gathering DCA measurements requires modification of the code, and because the available DCA data at the time of the Mira HPL run was not Level 3 compliant, DCA measurements were not taken. At this time, installation of the firmware is not officially supported by IBM. With Mira in production, ALCF will not install the firmware until it becomes officially supported.

6. CASE STUDY FROM Calcul Québec Université Laval

Colosse is a Compute Canada cluster at Calcul Québec Université Laval. Installed in 2009 by Sun Microsystems, this cluster consists of 960 nodes (x6275 Sun blades with X5560 Nehalem processors) with 24 GB RAM, totaling 7,680 cores. The cluster also includes a Lustre filesystem offering one petabyte of storage (total of 28 OSS and MDS servers), and an Infiniband QDR fat-tree network built using two Sun M9 DSC648 core switches and 42 leaf switches.

Colosse is an air-cooled cluster installed in a three-level silo building. The site's cylindrical topology offers an outer cold air plenum in front of all racks, and the generated heat is captured in a central hot core cylinder. While the computer is air-cooled, the facility uses chilled water cooling at the infrastructure level. Chilled water input temperature is typically 5 °C, and the heated water is returned to the system at temperatures ranging from 25 to 28 °C. The cold air plenum temperature is maintained at 19 °C, at a relative humidity between 45% and 55%.

A 2 MW site transformer provides 3-phase 600 V. A Siemens 9330 power meter is connected on the 600 V power grid. This meter's accuracy specification complies with the IEC 687

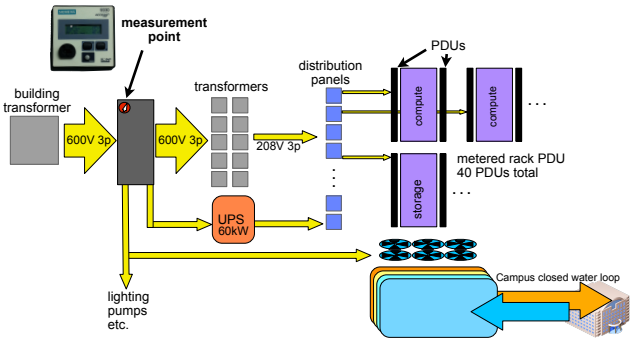

Figure 4: Power measurement points for the Colosse supercomputer

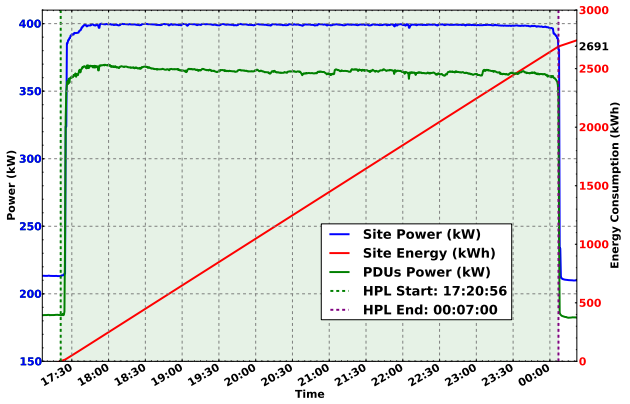

Figure 5: Power and energy measurements during the Colosse run

Class 0.5 specification and ANSI 12.20 Class 0.5. Ten 112 kVA transformers provide 3-phase 208 V to ten distribution panels.

The Colosse cluster uses about 30 % of the site's electrical and cooling capacity. It contains ten computer racks with 96 nodes per rack. Each rack is powered from two metered APC (Schneider Electric) AP7866 16.2 kW PDUs. The storage system, management servers, Infiniband and Ethernet switches are installed in ten computer racks. These racks are powered from two metered APC (Schneider Electric) AP7868 12.5 kW PDUs. Overall, the system has a total of 40 PDUs. The metering on these PDUs measures instantaneous current for each input phase and outlets. Continuously integrated total energy measurement, as required for a Level 3 measurement, is not available from these PDUs.

As an early adopter of the new measurement methodology, we set a goal to achieve a Level 3 measurement. The Siemens power meter was used since this meter provides the required energy measurement. This power meter is reachable through a TCP/IP connection, so we are able to adjust the measurement period according to the measurement requirements; the meter is capable of measuring up to 1,920 times per second. The integrated energy and power measurements are reported every 30 seconds.

Figure 4 shows the location of the Siemens power meter.

A drawback to using this power meter is that measurements include subsystems that are not required by the methodology (such as power transformers, UPS, site lightning, water pumps, fans).

Figure 5 and Table 7 show the Level 3 measurement results from May 2012. The energy measurement shows that 2691 kWh was consumed during the HPL run. By comparing the site power (measured from the Siemens power meter) and the PDU power (measured from instantaneous power measurements on the 40 PDUs), we note that the power transformers, water pumps, fans, and lighting consumed between 30 kW and 35 kW during the run.

The Colosse cluster has been in operation for over three years. Running HPL requires a significant maintenance window to get all systems working correctly. While this is usually not a problem for new systems under deployment (usually this is part of acceptance testing), it has an important impact for production systems. The power and energy requirements highlight the need for better instrumentation. New supercomputer deployments (or major renovations) should require power and energy sensors in the rack.

Table 7: Colosse power data

Data Item	Value
Start time	17:20:56
End time	00:07:40
Performance (TFlops)	77.89
Mflops/Watt full run	196
Duration (s)	24,404
Idle power (kW)	213.38
Full run power (avg, kW)	396.75

7. CONCLUSIONS

The energy consumption of larger HPC systems is an ever-growing concern. This paper described a more refined and rigorous power measurement methodology for HPC systems under workload runs. The methodology was developed by the Energy Efficient HPC Working Group in collaboration with Green500, Top500, and the Green Grid.

This paper considered three levels of quality of measurements: Level 1, which is similar to the Green500 version 0.9 run rules, and Levels 2 and 3, which include more comprehensive and integrated energy-usage measurements.

Case studies at LRZ, ANL, and Calcul Québec showed how this methodology can help these centers pinpoint the energy "hot spot" in a much clearer way than was previously possible. In particular, Level 3 measurements were indeed found to be both more precise and more accurate than Level 1 and Level 2 measurements. However, Level 3 measurements today require HPC-center, infrastructure-level instrumentation.

The challenges for attaining Level 3 measurements are not only technical, but also organizational and economic. Our case studies illustrated these challenges as well as demonstrated that measurements at different quality levels yielded different reuslts. Comparison of test results are best for data taken at the same quality level.

This is an important result. There are two main reasons for performance metrics. The first is to track a site's performance over time. The second is to be able to compare one site to another. The ability to do the latter is greatly enhanced by defining the levels.

Comparisons between sites at the same level are now meaningful. In the past, power measurements at different sites could be argued to have been different enough to not be

meaningful. For example, note LRZ's difference between L1 (1055 Mflops/Watt) and L3 (905 MFlops/Watt).

When trending an individual site over time, the worst thing to do would be to not measure the energy, claiming that one is waiting for Level 3 instrumentation to be added. The best approach is is to start with Level 1, which is reasonably achievable and then trend the cluster over time with a repeatable methodology.

The work of the EE HPC WG is also paying off in other aspects. Previously, the start and stop time for the HPL (HP Linpack) "core phase" was also not clear. The HPL code now includes a timestamp for the "core phase" start and stop time to better support the power measurement methodology. All sites are encouraged to use these timestamps.

Challenges also remain for larger installations regarding how to get to Level 3. At Argonne, IBM BG/Q responded to user desire for Level 3 measurements with a firmware upgrade, but it is not yet officially supported. The power meter installed at Calcul Québec for Level 3 measurements also had its issues in that it included power consumed by power transformers, water pumps, fans, and lighting, which consumed between 30 kW and 35 kW during the run, originally not part of Level 3. These examples point out the difficulties of getting to Level 3. These are best avoided by new cluster installations including the right measurement capability at the time of installation.

Future Work

Building more energy efficient HPC systems continue to be a major concern. In Europe, the EU COST Action IC0805 Open Network for High-performance Computing on Complex Environments and EU COST Action IC0804 Energy Efficiency of Large Scale Distributed Systems recently combined efforts proposing further work in this area [7].

The measurement quality level needs to be included when reporting power data for ranking sites such as the Green500 and Top500, because the different levels produce different results. The validity of comparing across sites is enhanced when reporting values and stating the Level used.

Level 3 measurement capabilities could be made a feature of HPC systems, as demonstrated by the IBM BG/Q. It is recommended that users ask for these kinds of capabilities from their vendors.

The EE HPC WG continues to develop and refine the methodologies, while at the same time exploring their applicability to measure the energy used in other performance metrics. In the near future, we are investigating the elimination or adjustment of Level 1 and increasing the specificity of Levels 2 and 3.

The focus on improving the ability to take precise and accurate power measurements is important for understanding architectural trends such as those provided for the Top500 and Green500. With these advanced measurement capabilities, larger HPC centers (such as LRZ) are more able to effectively drive energy efficiency programs.

8. ACKNOWLEDGMENTS

Portions of this research used resources of the Argonne Leadership Computing Facility at Argonne National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under contract DE-AC02-06CH11357.

Portions of this research used resources of Calcul Québec and Compute Canada, which is funded by the Canada Foundation for Innovation (CFI) and the Gouvernement du Québec.

Portions of this research have been carried out within the PRACE project, which has received funding from the European Community's Seventh Framework Program (FP7/2007-2013) under grant agreements no. RI-261557 and RI-283493, and at the Leibniz Supercomputing Centre (BADW-LRZ) with support of the State of Bavaria, Germany, and the Gauss Centre for Supercomputing (GCS), Germany.

9. ADDITIONAL AUTHORS

Anna Maria Bailey, LLNL
John Baron, SGI
Sergio Barrachina, UJI
Paul Coteus, IBM
Anne C. Elster, NTNU
Ladina Gilly, CSCS
Robin Goldstone, LLNL
Chung-Hsing Hsu, ORNL
Ted Kubaska, IEEE
James Laros, SNL
Yutong Lu, NUDT
David Martinez, SNL
Michael Patterson, Intel
Stephen Poole, ORNL
James H. Rogers, ORNL
Greg Rottman, ERDC
Mehdi Sheiklhalishahi, UNICAL
Daniel Tafani, LRZ
William Tschudi, LBNL

10. REFERENCES

[1] The Graph 500. `http://www.graph500.org/`.
[2] The Green Graph 500. `http://green.graph500.org/`.
[3] Energy Efficient High Performance Computing Power Measurement Methodology. `http://green500.org/sites/default/files/eehpcwg/EEHPCWG_PowerMeasurementMethodology.pdf`.
[4] Energy Efficient High Performance Computing Power Measurement Methodology version 1.0a. `http://green500.org/sites/default/files/eehpcwg/EEHPCWG_PowerMeasurementMethodology.pdf`.
[5] Energy Efficient HPC Working Group. `http://eehpcwg.lbl.gov/`.
[6] Energy Efficient HPC Working Group, Green500 and Top500 Power Submission Analysis. `http://eehpcwg.lbl.gov/documents`.
[7] EU COST Action IC0805: Open Network for High-Performance Computing on Complex Environments. `http://www.complexhpc.org`.
[8] Green500. `http://www.green500.org/`.
[9] Green500 Run Rules. `http://green500.org/docs/pubs/RunRules_Ver0.9.pdf`.
[10] D. Hackenberg and et al. Quantifying power consumption variations of hpc systems using spec mpi benchmarks. *Computer Science Research and Development*, 25:155–163, 2010.
[11] C. H. Hsu and S. Poole. Power measurement for high performance computing: State of the art. In *Proceedings of the International Green Computing Conference*, 2011.

[12] C. H. Hsu and S. Poole. Power signature analysis of the SPECpower_ssj2008 benchmark. In *Proceedings of the International Symposium of Performance Analysis of Systems and Software (ISPASS)*, 2011.

[13] S. Kamil, J. Shalf, and E. Strohmaier. Power Efficiency in High Performance Computing. In *Proceedings of the High Performance, Power Aware Computing Workshop (HPPAC)*, 2008.

[14] J. Laros. Topics on measuring real power usage on high performance computing platforms. In *Proceedings of the International Conference on Cluster Computing (CLUSTER)*, 2009.

[15] HPL A Portable Implementation of the High-Performance Linpack Benchmark for Distributed-Memory Computers. `www.netlib.org/benchmark/hpl/`.

[16] H. W. Meuer, E. Strohmaier, and H. D. Simon. Top500. `www.top500.org`.

[17] M. S. Müller, J. Baron, W. C. Brantley, H. Feng, D. Hackenberg, R. Henschel, G. Jost, D. Molka, C. Parrott, J. Robichaux, P. Shelepugin, M. van Waveren, B. Whitney, and K. Kumaran. SPEC OMP2012 — An Application Benchmark Suite for Parallel Systems Using OpenMP. In *Lecture Notes in Computer Science*, pages 223–236. Springer Berlin Heidelberg.

[18] S. Sharma, C. H. Hsu, and W.-c. Feng. Making a case for a green500 list. In *Proceedings of the High Performance Power Aware Computing Workshop*, 2006.

[19] SPEC High Performance Group. `http://www.spec.org`.

[20] SPEC OMP2012. `http://www.spec.org/omp2012/`.

[21] SPEC Power Committee. SPEC power and performance benchmark methodology. Technical report, Standard Performance Evaluation Corporation, June 2010. Tech. Rep. Version 2.0.

[22] B. Subramaniam and W. Feng. Understanding power measurement implications in the green500 list. In *Proceedings of the IEEE/ACM International Conference on Green Computing and Communications (GreenCom)*, 2010.

[23] B. Subramaniam, T. Scogland, and W. Feng. Trends in Energy-Efficient Computing: A Perspective from the Green500. In *roceedings of the 4th International Green Computing Conference, Arlington, VA*, June 2013. (to appear).

[24] The Green Grid. `http://www.thegreengrid.org/`.

[25] Top 500 Supercomputing Sites. `http://www.top500.org`.

[26] K. Yoshii, K. Iskra, R. Gupta, P. Beckman, V. Vishwanath, C. Yu, and S. Coghlan. Evaluating Power-Monitoring Capabilities on IBM Blue Gene/P and Blue Gene/Q. In *2012 IEEE International Conference on Cluster Computing (CLUSTER)*. IEEE.