Node Variability in Large-Scale Power Measurements: Perspectives from the Green500, Top500 and EEHPCWG *

Thomas Scogland
Green500 List and Lawrence Livermore National Laboratory
scogland1@llnl.gov

Jonathan Azose
University of Washington, Seattle, WA, USA
jonazose@uw.edu

David Rohr
Frankfurt Institute for Advanced Studies, Goethe University, Frankfurt
drohr@cern.ch

Suzanne Rivoire
Sonoma State University, Rohnert Park, CA, USA
rivoire@sonoma.edu

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

Daniel Hackenberg
TU Dresden
daniel.hackenberg@tu-dresden.de

ABSTRACT

The last decade has seen power consumption move from an afterthought to the foremost design constraint of new supercomputers. Measuring the power of a supercomputer can be a daunting proposition, and as a result, many published measurements are extrapolated. This paper explores the validity of these extrapolations in the context of inter-node power variability and power variations over time within a run. We characterize power variability across nodes in systems at eight supercomputer centers across the globe. This characterization shows that the current requirement for measurements submitted to the Green500 and others is insufficient, allowing variations of up to 20% due to measurement timing and a further 10-15% due to insufficient sample sizes. This paper proposes new power and energy measurement requirements for supercomputers, some of which have been accepted for use by the Green500 and Top500, to ensure consistent accuracy.

1. INTRODUCTION

With power requirements for supercomputers soaring and becoming the main limiters of system performance, quantifying, comparing and analyzing the efficiency of supercomputers has never been more important nor more difficult. The extreme and constantly expanding scale of systems, both in number of components and in power draw, makes taking full-system measurements infeasible for many sites. As a result, most reported power and energy measurements of supercomputers, such as those published by the Green500 [7] and the Top500 [22], are based on performance measured for a full run across the entire system, but power measured only on subsets of a system and potentially a subset of the runtime. The accuracy and comparability of the resulting measurements is highly dependent on the measurement methodology and the degree to which the subset of the system behaves, on average, in the same way as the full system.

A common set of requirements has been defined by the Energy-Efficient HPC Working Group (EE HPC WG) power measurement methodology [5] that is used by the Green500 and Top500 lists. The methodology defines three measurement levels providing different accuracies, with Level 1 the lowest and Level 3 the highest. In an ideal world, all measurements would use Level 3, but its high level of accuracy also requires a precise, comprehensive and sometimes prohibitively expensive measurement infrastructure. Instead, the most achievable measurement level, Level 1, is the most commonly used methodology. Of the 267 submitted measurements on the November 2014 Green500 list, 233 submissions used power estimates based on derived numbers rather than measurement, 28 used Level 1, and only 6 used a higher measurement level. With the vast majority of actual measurements using Level 1, its accuracy and comparability are extremely important to the value of the data collected by both of these lists as well as any others that use the methodology.

During the development of the methodology, an inter-node variance of up to 5% was expected, and had been shown to be normal variation for computers at the time [11]. Recently however, different measurements of the same system using the Level 1 methodology have been found to vary by more than 20% [16, 4]! This variability has significant ramifications for Green500 rankings. For instance, the advantage of the current 1st ranked system over the current 3rd ranked

*This material is based upon work supported by the U.S. Department of Energy's Lawrence Livermore National Laboratory; Office of Advanced Scientific Computing Research (LLNL-CONF-669882); the SIMOPEK project, which has received funding from the German Federal Ministry for Education and Research under grant number 01IH13007A, at the Leibniz Supercomputing Centre (LRZ) with support of the State of Bavaria, Germany, and the Gauss Centre for Supercomputing (GCS); Calcul Québec at Université Laval and Compute Canada, which is funded by the Canada Foundation for Innovation (CFI) and the Gouvernement du Québec; Oak Ridge Leadership Computing Facility at the Oak Ridge National Laboratory, which is managed by UT Battelle, LLC for the U.S. DOE (under the contract No. DE-AC05-00OR22725).

SC '15, November 15-20, 2015, Austin, TX, USA

 ISBN 978-1-4503-3723-6/15/11…$15.00

DOI: http://dx.doi.org/10.1145/2807591.2807653

system is less than 20%. At SC '13 and SC '14 BoF sessions, there were presentations that showed variations of 10% and 20%, respectively, due to usage of different measurement intervals. These findings demonstrate the significance of the improvements to the measuring methodology suggested in this paper.

There are two primary sources of variation in the measurements: variation in the power consumed over time, and variation of the power consumed by individual components in the subset measured. Both have grown in relative size lately. The variation over time is a side effect of modern, fast, heterogeneous systems, which do not achieve a constantly high utilization of their peak compute capability. This is addressed in Section 3.

Variability of components has numerous causes. One is variability in the manufacturing process, where the amount and location of imperfections in the substrate or circuit paths themselves result in different amounts of leakage, and thus different efficiency levels. A recent trend is that hardware vendors adapt their hardware better, both to these fluctuations in the manufacturing process and to operating conditions such as temperature, which leads to secondary causes of variability. With the emergence of heterogeneous clusters, understanding variability becomes even more complicated. Section 5 sheds light on two examples: "identical" GPU boards that the same clock speed but different programmed voltage IDs; and automatic fan speed regulation, which affects power consumption. The turbo modes of modern CPUs and temperature are other reasons. A detailed analysis of the sources of component variability should be performed in future work.

In this paper, we analyze the sources of measurement variability and determine new guidelines for achievable and accurate measurement of supercomputer power. To ensure fairness and accuracy in the published measurements, these new guidelines on minimum node count and measurement period are being integrated into the submission requirements for the Green500 and Top500 lists.

With the suggested methodology, sites can determine how many components or nodes must be measured in order to characterize system-level power with reasonable accuracy. Even sites with system-level power measurement capabilities tend not to reserve a full machine for post-acceptance benchmarking, and instead run on a subset of the system and extrapolate. Other use cases of system-level power characterizations include architectural trending, system modeling (design, selection, upgrade, tuning, analysis), procurement, operational improvements and power capping. Our guidelines also serve as instructions for extrapolating Total Cost of Ownership (TCO) from smaller test systems during procurement or initial testing phases. In particular, the observed variations of 20% in power consumption lead directly to a possible 20% increase in electricity costs, which nowadays are a significant contribution to the TCO. Hence, a higher accuracy is desirable.

We start by presenting background on the currently recommended measurement methodology and related work in Section 2. We present results and analysis of power variability in a workload across time in Section 3 and analysis of inter-node power variability in Section 4. In Section 4.2, we analyze the node power consumption variations of several large-scale systems and derive a recommended node subset size to perform reasonably accurate full-system power consumption extrapolations. GPUs have been becoming more and more important in the HPC sector in recent years. Section 5 presents case studies of GPU results on the L-CSC [16] and Titan clusters in order to examine the validity of our approach for systems with GPUs. Lastly we present our final suggestions and conclusions in Section 6.

2. BACKGROUND

This section provides related work along with a description of the current EE HPC WG power measurement methodology.

2.1 Related Work

For energy-efficiency benchmarks to provide fair comparisons between systems and to illuminate architectural trends, they need to provide standard methods of measuring system power. At the single-node level, the Standard Performance Evaluation Corporation (SPEC) benchmarks provide a precise and widely used set of rules for measuring system power during the execution of a workload. These methods were initially developed for the server-oriented SPECpower benchmark [20] but have since been added to the SPEC OMP2012 [15] and SPEC ACCEL [13] suites. The SPEC methodology provides reliably accurate power and performance measurements, but at the cost of requiring the use of their software and a pre-vetted set of high-quality meters. This approach gives terrific accuracy where it is practical, but does not scale well to supercomputing, where several distributed meters are often required to measure even a significant subset of a system.

By contrast, power measurement methodologies appropriate for large-scale systems running a specific workload are still evolving. Fully instrumenting all nodes is still impractical at most facilities [11, 19] or can lack accuracy [9]. Hsu describes the different possible levels of instrumentation [12].

Several HPC benchmarks currently include power measurements (optional or required). The Green500 [7], which uses High-Performance Linpack (HPL) as its workload, uses energy efficiency in the form of FLOPS/Watt as the metric of comparison to rank the fastest supercomputers in the world by their energy efficiency. Its floating-point performance-oriented counterpart, the Top500 [22], also accepts power measurements, representing how efficient the supercomputer is when running at full performance. The Green Graph 500 [8] and Graph500 are analogous benchmarks with graph analysis as the workload of interest. Half of the Green500 power results are actually based on vendor specifications and extrapolation rather than physical measurements [19]. These derived numbers still fill in gaps in the list today for the sites that choose not to measure their power consumption. The Green500 rules needed to evolve in order to make the measurement of systems achievable for sites with low to average instrumentation, leading Subramaniam [21] to recommend extrapolation from measuring a subset of the system, an approach also favored by Kamil [14].

Even when all nodes of the system are identical, the well-documented manufacturing variation across processors [1, 17] means that a subset of nodes may not accurately represent the system as a whole. However, very few studies have rigorously examined the loss in accuracy from extrapolating power measurements from a subset of nodes [11].

Other approaches to this question come from prior work in large-scale power modeling, which has examined the ques-

tion of how large a system subset must be used to train a power model. Fan was able to extrapolate coarse-grained power models from a single node for lightly used servers, but only after adding a large constant offset to account for variation in idle power and for networking components [6]. Davis [3], in the only study to apply statistical methods to this question, demonstrated that node-to-node variation can significantly affect the accuracy of power models for four small clusters running data-intensive workloads. They propose using a very conservative Chernoff-Hoeffding bound to select the subset size, and they note that their workloads are not homogeneous, with substantial differences in nodes' average power. For regular workloads (i. e. balanced equally across all nodes), such as HPL, we find that a much less conservative bound is sufficient to produce highly accurate estimates of full-system power consumption.

2.2 Measurement Methodology

The EE HPC WG power measurement methodology [5] has been adopted by the Green500 and Top500 as a common baseline for measurement requirements. It defines three levels of measurement quality, each of increasing accuracy but also increasing difficulty to use at large scale. Each level defines different requirements for each of four aspects of the measurement:

1. Measurement duration and granularity
2. How much of the system is measured
3. Which subsystems must be included in the measurement
4. Where in the power hierarchy the measurements may be taken

In the following paragraphs, we give a brief outline of some aspects of the current methodology, which Table 1 summarizes. One goal of the Level 1 specification was to encourage more people to measure the efficiency of their systems and participate in rankings such as the Green500. In order to make it approachable, the requirements for the measurement equipment had to be low. Level 1 measures the performance of only the core phase of a benchmark: that is, the time period in which the actual computation of the benchmark happens. It does not include setup and tear-down time. Within the core phase of the Linpack benchmark, traditional HPC systems have shown a quite flat power consumption over time (see next section). There are usually some variations at the very beginning (for instance because of warming up of hardware components) and at the very end (where, in the case of Linpack, the remaining matrix size becomes small). Therefore, Level 1 requires only a measurement during 20% of the core phase. This 20% period must be within the middle 80% of the core phase, to exclude the first 10% and the last 10% of the time where the power profile is not flat.

A measurement of the entire facility power usually includes other components such as storage, other compute clusters, and infrastructure. As such, it cannot be used to get an accurate power measurement of an isolated supercomputer. It can be challenging to measure the power consumption of an entire supercomputer, which can drain several megawatts of power, on its own. Therefore, Level 1 requires a power measurement of only $\frac{1}{64}$ of the compute nodes participating in the benchmark and allows this measurement to be scaled up linearly to approximate the full-system power. In order to achieve a certain level of accuracy for low-power nodes, the methodology requires at least 2 kW of power to be measured. To simplify the measurement, Level 1 does not require the inclusion of secondary infrastructure, like the network or the infrastructure nodes.

It is not clear whether these requirements for the measurement will be suitable in the future. In order to enable measurements of higher accuracy, the current methodology specifies two additional levels with higher requirements. The main differences are that Level 2 and 3 both require a measurement during the entire core phase of the run, they require a measurement of a larger fraction of compute nodes ($\frac{1}{8}$ for Level 2 and all nodes for Level 3), and they require all required infrastructure components (everything that cannot be switched off for the benchmark run) to be measured or estimated.

Scogland [19] presents data at the different levels for a few real-world supercomputers, and shows that the Level 1 and Level 2 methodologies can significantly overstate a system's energy efficiency. However, because the three levels of measurement differ in more ways than the size of the system subset being measured, it's not clear from that work how much of the difference between levels has to do with the question of extrapolation alone. We have since found that the both the measurement phase and the machine fraction, as well as subset selection, play key roles in measurement accuracy. Each of these will be discussed in greater detail in the following sections.

Table 1: Summary of the EE HPC WG methodology's requirements by quality level

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 middle 80% of the core phase of the run	Ten equally spaced power averaged measurements spanning the full run	Continual measurement across the full run
2: Machine fraction	The greater of 1/64 of the compute subsystem or 2 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 modeled with manufacturer-supplied data	Upstream of power conversion or modeled with off-line measurements	Upstream of power conversion or conversion loss measured simultaneously

	HPL runtime	Core phase power (kW)	First 20%	Last 20%
Colosse	7 hours	398.7	398.1	398.2
Sequoia	28 hours	11,503.3	11,628.7	11,244.2
Piz Daint	1.5 hours	833.4	873.8	698.4
L-CSC	1.5 hours	59.1	63.9	46.8

Table 2: Runtime and average power in kilowatts for different segments of each HPL test run

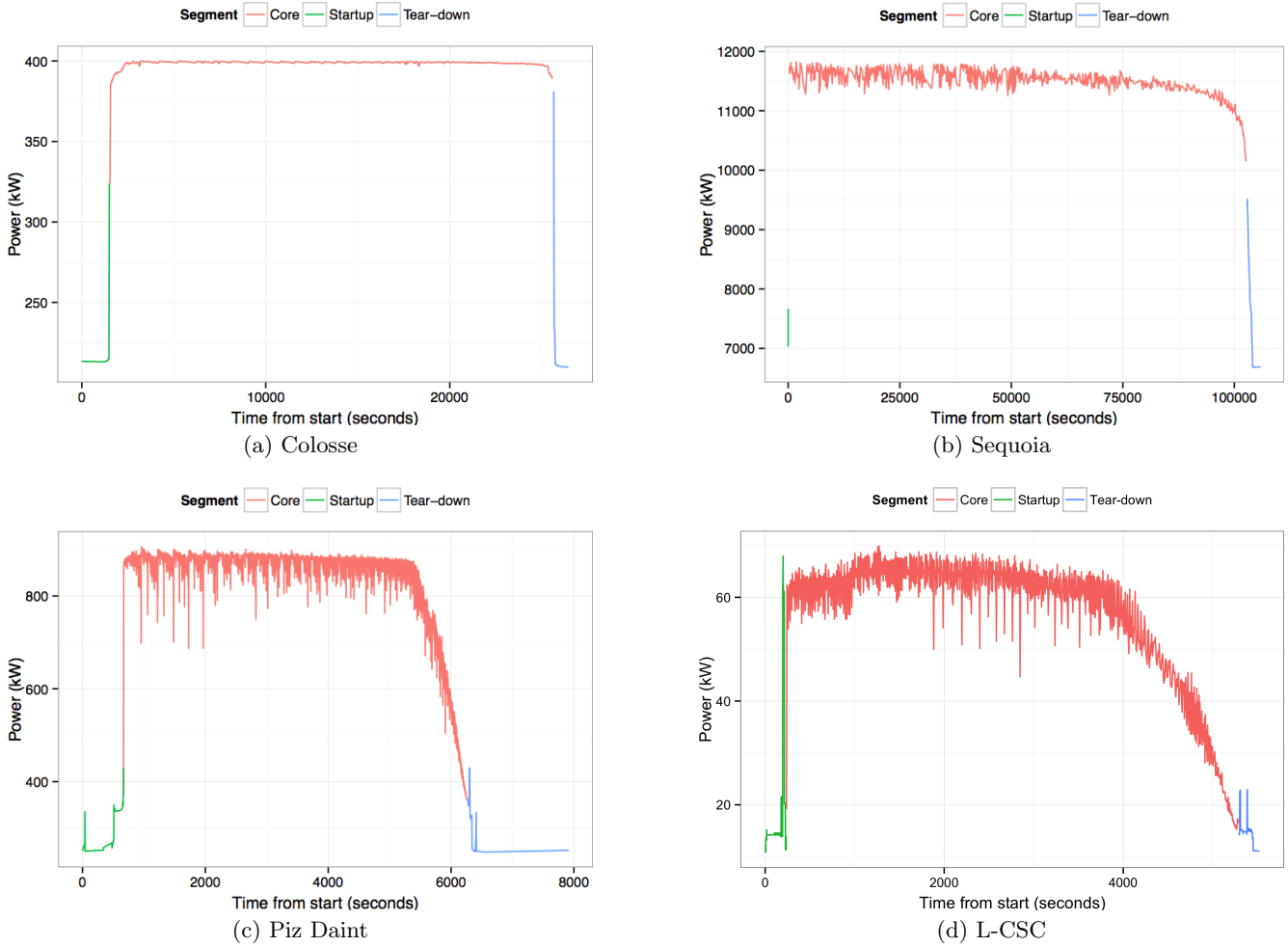
3. POWER VARIABILITY OVER TIME

The first source of measurement variability we will discuss comes from the portion of a workload run that is sampled. In previous years, the Green500 list and others have advocated measuring as little as 20% of the middle 80% of the core phase of the run, partly in order to support measurements on runs of great length or with minimal equipment. Figure 1 shows the average power over time for HPL on four systems, along with a table of segment averages in Table 2.

On the most "traditional" of these designs, the Colosse system, the HPL run is over 7 hours long and yields a very flat power curve, with the set-up and tear-down time almost completely lost in the overall runtime. The absolute numbers bear this out: the average measurements for the entire core phase, the first 20% of the run, and the last 20% of the run are all within 0.25% of one another. No matter what segment of the core phase is measured, the extrapolated power is a close match for the actual average over the full run.

Sequoia – actually Sequoia-25, a temporary combination of the Sequoia and Vulcan systems at LLNL – is by far the largest system in the group, consisting of nearly 2 million cores. Its runtime is typical of HPL runs for large-scale CPU systems, in that it runs for over 28 hours in total. The pattern appears more jagged, but still largely flat. The numbers show a slightly larger variation than Colosse, with a difference of approximately 3.5% between the average power of the first 20% and the last 20% of the core phase. Based on these results, a subset approach is reasonable assuming that the workload is balanced and consistent as with those tested.

Moving on to the heterogeneous CPU/GPU systems tells a different story, however. Piz Daint, of the Swiss National



(a) Colosse (b) Sequoia (c) Piz Daint (d) L-CSC

Figure 1: System average power over time for Linpack (Data taken from public data published by the Green500 list on green500.org)

Supercomputing Centre, is more indicative of an average heterogeneous system run. As one of the longer GPU HPL runs listed by the Green500, the Piz Daint run is still only about 1.5 hours long, and visibly more jagged and sloped than either of the CPU-only systems discussed above. In fact, the difference between an average over the first 20% and last 20% of the core phase is more than 20%! In-core HPL, which is the version favored for NVIDIA GPU-based systems, uses only the GPU memory to store the matrices, and thus necessarily runs in less overall time than the CPU systems, which tend to fill main memory. Taken to the extreme, some runs have been as short as five minutes on systems large enough to qualify for the Top500. Equally extreme is the data from the current #1 system on the Green500, the L-CSC cluster, which shows a first 20% average power of 63.9 kW and a last 20% of core phase average of only 46.8 kW. Based on the current requirements, a measurement for this machine could also vary by over 20%.

The fact that the power of an HPL run tails off as the matrix shrinks near the end of the run, especially on GPU systems, has been exploited to achieve better-looking results before. The TSUBAME-KFC cluster yielded a 10.9% reduction in its power consumption measurement for the Green500 in November 2013 by selecting an "optimal" time interval [4]. Rohr et al. [16] have shown that the L-CSC cluster would have been able to submit a result with 23.9% improved power efficiency to the Green500 in November 2014 by tweaking the time interval.

Another potential cause of power variation over time is dynamic voltage and frequency scaling (DVFS), which many computers use to improve their operational efficiency. The L-CSC cluster could reach a 22% improvement in energy efficiency in the Linpack benchmark through DVFS [16]. In fact, the current methodology specification explicitly allows DVFS for this reason. However, this leads to an obvious problem when the power measurement does not cover the entire core phase. The power consumption will usually be lowest during the period where DVFS selects the lowest processor voltages. By placing the power measurement interval in this period, the power measurement could completely avoid the period where the processor runs at higher frequencies and drains more power.

Given this issue, effectively a way to "game the system," as well as the lack of generalizability to workloads with more complex patterns, we recommend ensuring that all power measurements are taken for the full period of the core phase of a given workload. This means the power measurement should cover exactly the time period that is used to measure the performance, and preferably include a number of measurements before and after as well. While this may preclude measurements with some equipment options or of some runs of sufficient length, the variability caused by allowing a shorter period makes the results unreliable.

4. INTER-NODE POWER VARIABILITY

In cases where it is infeasible to measure power consumption across a complete supercomputer, a practical alternative is to estimate power consumption by measuring a random sample of nodes. How many to measure, however, is not entirely straightforward.

For example, the current Level 1 methodology requires that at least 1/64 of a supercomputer's nodes be measured. This methodology can produce power estimates with dramatically lower accuracy for small supercomputers compared to large supercomputers. Supercomputers can range from only a few hundred nodes to tens of thousands, depending on the configuration. Those discussed later in this paper run from 210 to $\sim$ 19000 for example, a gap of nearly two orders of magnitude. For a hypothetical supercomputer with 210 nodes and a true value of $\sigma/\mu = 2\%$ (where μ is the measured average power consumption, and σ is the standard deviation of the measurement), the Green500 methodology would require at least 4 nodes to be measured. Based on 4 nodes, we would be able to say with 95% certainty that our estimate of the total power usage is within 3.2% of the true total. In contrast, for a supercomputer with 18,688 nodes and $\sigma/\mu = 2\%$, the Green500 methodology would recommend that at least 292 nodes be measured. From a sample of 292 nodes, we would be 95% certain that the estimated power usage is within 0.2% of the true total. That is, although both supercomputers have the same relative variability, this methodology produces a sample that is an order of magnitude less accurate on the smaller supercomputer.

4.1 Experimental Results

To gain insight into the node variability of current systems, we conduct a study across six large-scale systems located at Technische Universität Dresden, Calcul Québec, the French Alternative Energies and Atomic Energy Commission (CEA), the Leibniz Supercomputing Center (LRZ), and Oak Ridge National Laboratory (ORNL) see Table 3. Each system ran a balanced workload high in floating-point computation, and power was measured on each node individually.

A histogram of the per-node power results is presented in Figure 2. While the range and counts differ for each system, the distributions remain relatively similar, even across the GPUs in the ORNL results. All systems show power distributions that are roughly unimodal with few outliers, suggesting that it may be appropriate to model these distributions as Gaussian. Given the assumption that the per-node power consumption is approximately normally distributed,

Table 3: Test systems

	CPUs per node	RAM per node	components measured	workload
Calcul Québec	2x Intel X5560	24 GiB	480x2 nodes	HPL
CEA (Fat)	4x Intel X7560	16x4 GiB	316 nodes	HPL
CEA (Thin)	2x Intel E5-2680	16x4 GiB	640 nodes	HPL
LRZ	2x Intel E5-2680	32 GiB	512 nodes	MPrime [18])
ORNL	1x AMD 6274	32 GiB	GPUs in 1000 nodes	Rodinia CFD [2])
TU Dresden	2x Intel E5-2690	8x4 GiB	210 nodes	FIRESTARTER [10]

we can construct a statistical basis for the number of nodes that must be sampled to produce a given confidence in the final result. This assumption is unlikely to be met in the general case, where computational load may be distributed unevenly between nodes. However, for the case of balanced workloads as in these benchmarks, we find that the assumption of a normal distribution is appropriate. We therefore proceed upon the assumption of approximate normality, with the caveat that this methodology will not be appropriate in scenarios where the distribution of per-node power consumption contains many outliers or is heavily skewed.

Based on our sample data, we provide a formula for the necessary number of nodes in a sample in order to obtain a selected level of accuracy. Consider a supercomputer with N nodes with a true per-node mean power consumption of μ. Suppose we select a subset of n nodes at random, measuring time-averaged power consumption on each of these n nodes. Denoting these measurements by $X_1, \ldots, X_n$, a reasonable estimate of the true mean μ over all nodes is the mean of the measurements of the sub-sample, or $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$. From this sample, we will be able to make a statement that with $(1-\alpha)\times 100\%$ certainty, the difference between $\hat{\mu}$ and μ is at most $\lambda \cdot \mu$. What values of α and λ are "reasonable" depends a great deal on perspective, but a common baseline is to use confidence level $(1-\alpha) = 95\%$, and accuracy $\lambda = 1\%$, which would allow a statement with 95% certainty that an estimate of the per-node power consumption is off by no more than 1%.

So long as the per-node power usages are approximately normally distributed and n is small relative to N, a given confidence interval for the mean can be calculated based on $\hat{\sigma}$, the sample standard deviation, and $t_{n-1,1-\alpha/2}$, or the $1-\alpha/2$ quantile from a t distribution with $n-1$ degrees of freedom:

$$\mathrm{CI} = \hat{\mu} \pm \frac{t_{n-1,1-\alpha/2} \cdot \hat{\sigma}}{\sqrt{n}}, \tag{1}$$

For large values of n (e.g. $n \geq 20$), we can approximate quantiles from a t_{n-1} distribution with quantiles from a standard normal distribution, where $z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile from a standard normal distribution. So for large n an approximate confidence interval is given by:

$$\mathrm{CI} \approx \hat{\mu} \pm \frac{z_{1-\alpha/2} \cdot \hat{\sigma}}{\sqrt{n}}, \tag{2}$$

Translated into a restriction on sample size, we want to choose n so that the confidence interval half-width is no more than $\lambda \cdot \mu$, where μ is approximated with the estimated sample mean $\hat{\mu}$.

$$\frac{z_{1-\alpha/2} \cdot \hat{\sigma}}{\sqrt{n}} \leq \lambda \cdot \mu. \tag{3}$$

Solving for n then yields the general formula:

$$n \geq \left(\frac{z_{1-\alpha/2}}{\lambda} \cdot \frac{\hat{\sigma}}{\hat{\mu}} \right)^2. \tag{4}$$

This can be further refined to reduce the requirement that the sample size n be small relative to the total number of nodes N. In this case we can introduce a finite population correction into the initial formula for the confidence interval. Carrying this finite population correction through the same logic yields a two-step procedure for providing a sample size recommendation:

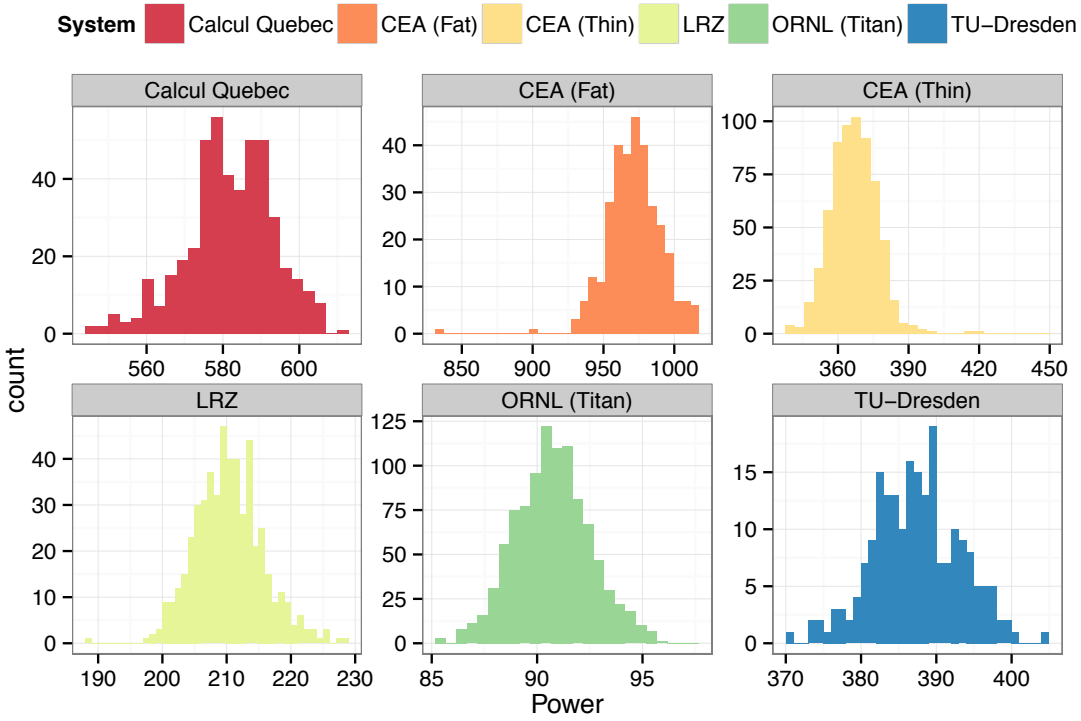


Figure 2: Histograms of whole-node power under load across systems

Table 4: Comparison of four supercomputers. Columns in the table: N, the total number of nodes (or blades in the case of Calcul Québec); $\hat{\mu}$, an estimate of the mean per-node (or per-blade) power usage in Watts; $\hat{\sigma}$, an estimate of the standard deviation of the per-node power usages; $\hat{\sigma}/\hat{\mu}$, the ratio of those two estimates.

	Nodes/Blades (N)	Sample mean ($\hat{\mu}$)	Std. deviation ($\hat{\sigma}$)	$\hat{\sigma}/\hat{\mu}$
Calcul Québec	480	581.93	11.66	2.00%
CEA (Fat)	360	971.74	19.81	2.04%
CEA (Thin)	5040	366.84	10.41	2.84%
LRZ	9216	209.88	5.31	2.53%
Titan	18688	90.74	1.81	1.99%
TU-Dresden	210	386.86	5.85	1.51%

$$n_0 = \left(\frac{z_{1-\alpha/2}}{\lambda} \cdot \frac{\hat{\sigma}}{\hat{\mu}}\right)^2 \qquad n = \frac{n_0 N}{n_0 + N - 1}. \tag{5}$$

First we compute n_0, the minimum required sample size if N were infinite, and then adjust this downward based on the true value of N.

4.2 Generalizing the Requirements

The sample size recommendations in Equation 5 rely on known values for five parameters: N (the total number of nodes), α (a function of desired confidence), λ (the desired accuracy), $\hat{\sigma}$ (sample standard deviation), and $\hat{\mu}$ (sample mean power consumption). While N is known in advance and α and λ may be selected based on the desired accuracy, $\hat{\sigma}$ and $\hat{\mu}$ are unknown without sampling. This leads to a paradoxical situation in which a sample must be taken in order to determine an appropriate sample size. Depending on the context for wanting to determine a sample size, it may be reasonable to simply take a small initial sample (e. g. of $n = 10$ nodes) to obtain estimates of μ and σ in order to determine an appropriate size for a final sample. As a requirement for the EE HPC WG methodology, requiring a sample to determine sample size is impractical both to implement and to check. An alternative is to estimate the ratio $\hat{\sigma}/\hat{\mu}$ by examination of similar systems.

Table 4 provides details of estimates of $\hat{\sigma}$ and $\hat{\mu}$ on the four test systems discussed in Section 3. Although the absolute mean power usage differs substantially between systems, all four have estimated values of σ/μ that fall approximately within the range $1.5\% - 3\%$. The GPU-based system L-CSC presented in Section 5), as a second GPU system apart from the Titan cluster, shows an even lower variability $\hat{\sigma}/\hat{\mu}$. This affirms our assumption that $1.5\% - 3\%$ holds valid for GPU-based systems as well as traditional CPU systems.

Equation 5 shows that, for cases where N is far greater than n, the adjustment factor n/n_0 is close to 1. In other words, the required sample size depends mostly on λ and $\hat{\sigma}/\hat{\mu}$ and only marginally on the total population size. Hence, we can obtain a sample size n for a reasonably large cluster N, and then use that sample size for all clusters. Consequently, that increases the achieved accuracy for smaller systems slightly.

Finally, we can translate our recommendations into a table of sample sizes for various values of the parameters. Table 5 shows recommended sample sizes for different values of λ and of σ/μ, fixing α at a conventional 0.05, for a 95% confidence interval, and N at a conservative value of 10,000. Given the values we have seen in practice, a σ/μ of 0.028 is the highest we found, and the standard variance of power measurement equipment of 1-1.5%, a measurement of 16 nodes or more may be reasonable. For purposes that require tighter bounds, as many as 370 may be needed, or as few as four, but until attempting to reach accuracy below 1% the required values remain reasonable.

These sample size recommendations rely on an assumption that the distribution of per-node power consumption in each system is normally distributed. In fact, visual inspection of the distributions in Figure 2 reveals the presence of outliers in several of the systems that are of a larger magnitude than we would typically see arising in truly normal data. We should check for all the available data that any violations of normality are small enough that the sample size determination procedure is still valid. In generalizing these recommendations, we will need to assume that any other tests being run are on sufficiently similar systems with similar workloads so that normality is not badly violated.

We performed bootstrap re-sampling to confirm that violations of true normality in each of these systems did not impact the calibration of confidence intervals. For each of our five data sets, we performed a simulation study wherein we simulated new data from the observed empirical distribution and repeatedly estimated the mean from subsamples of the simulated data. Specifically, we repeated the following procedure 100,000 times for a range of sample sizes n:

1. Simulate a complete supercomputer of N nodes by re-sampling with replacement from the collection of nodes observed in the real data.
2. Generate a sample of n nodes by sampling without replacement from the full simulated supercomputer.
3. Using the formula in Equation 1, obtain a mean estimate along with 80%, 95%, and 99% confidence intervals from the sample.
4. Check whether the confidence intervals contain the true mean power usage for the full N nodes.

Table 5: Table of recommended sample sizes for a system with $N = 10000$ nodes

		σ/μ		
		0.02	0.03	0.05
λ	0.5%	62	137	370
	1%	16	35	96
	1.5%	7	16	43
	2%	4	9	24

Results of these simulations for the LRZ data are plotted in Figure 3. If the normality assumption is approximately right and/or the sample size n is sufficiently large, this should be a well calibrated procedure. That is, an 80% confidence interval from the above procedure should contain the true mean power usage (of the simulated complete data) 80% of the time. These simulations show good calibration even as low as $n = 5$.

Based on these simulations, for any sample of size $n \geq 3$, violations of the normality assumption don't cause miscalibration of 80%, 95%, or 99% confidence intervals. Simulation studies on the other systems reveal that the normality assumption is appropriate for all systems we have tested, with good calibration as low as $n = 5$ on all systems.

Although the normality assumption is appropriate for small n, a separate issue is that in producing recommended sample sizes, we propose to approximate the t-quantile $t_{n-1,1-\alpha/2}$ with the normal quantile $z_{1-\alpha/2}$. This approximation causes slight under-coverage at small values of n. For example, for samples of size $n = 15$, approximating the t quantile with a normal quantile will produce 95% confidence intervals which are roughly 9% too narrow.

5. AN OUTLOOK TO SOURCES OF NODE VARIABILITY ON A GPU CLUSTER

The subset sample size analysis in the previous section is a statistical method to cope with node variability. In addition to coping with it, investigating and eliminating the sources of node variability can improve the accuracy of power measurements and also the energy efficiency of a system in some cases. Hence, in addition to the mere fact of variability between nodes, we are also interested in the sources of that variability. Section 3 presented some CPU systems and one GPU system. This section presents a case study of a multi-GPU system very different from the one in Section 3 to gain insight into the causes of variability. We did not include the results presented here in the earlier sample size analysis, due to a necessarily small sample size for results on this system.

All experiments in this section are derived from an OpenCL version of the Linpack benchmark on the **L-CSC** (Lattice Computer for Scientific Computing) cluster [16]: A multi-GPU cluster installed at GSI research facility featuring four

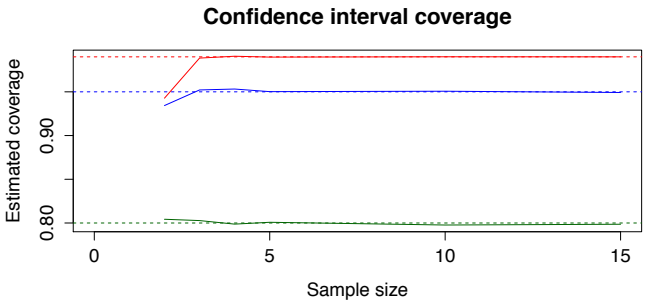


Figure 3: Coverage of 80% (green), 95% (blue), and 99% (red) confidence intervals in simulation studies based on the pilot sample of 516 nodes of the LRZ supercomputer. Solid lines show the simulated coverage while dashed lines show the target coverage. Each data point on this plot is calculated from 100,000 simulations of samples from a simulated full supercomputer.

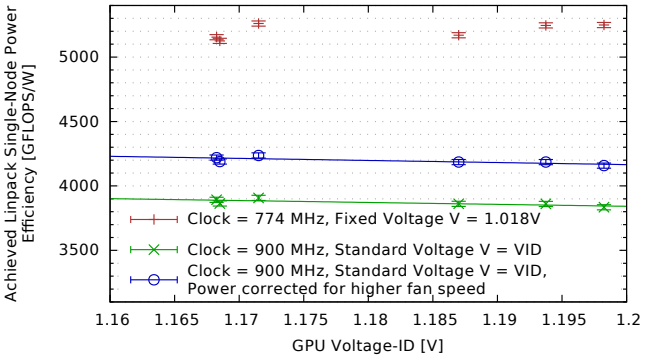


Figure 4: Power efficiency of individual nodes achieved in single-node Linpack on Lattice-CSC cluster.

AMD FirePro S9150 GPUs per node. With four GPUs per node, the GPUs are the main contributor to both performance and power consumption. This makes the L-CSC system quite distinct from the CPU-based supercomputers discussed in the previous sections and also from the Titan GPU cluster, which features only 1 CPU and 1 GPU per node.

Power efficiency depends to a great deal on the frequency and voltage selected for the GPUs. Best efficiency is achieved with the minimum voltage that ensures stable operation at a given frequency. Every GPU ASIC behaves a bit differently, so vendors program Voltage-IDs (**VIDs**) into the ASIC, which define sufficient voltage for a given frequency for that specific ASIC. For consistency, we ensure that all four GPUs in a node have the same VID value.

L-CSC used dynamic frequency and voltage scaling to obtain the best power efficiency for its Green500 submission [16]. For the Green500 run, the GPUs were not operating at the voltage defined by their VIDs. Instead, a comprehensive search of the frequency and voltage space found that the most efficient frequency for Linpack on L-CSC is 774 MHz, and the lowest stable voltage at this frequency is 1.018 V.

The other major source of variability noticed at the node level is the system fans, which vary by more than 100W depending on the current temperature and load. By default, system fans are regulated automatically, and cause larger variances in power efficiency than the actual CPU/GPU variability. To control for this variability, in all tests we have fixed the fans of all nodes to the lowest speed that maintains the thermal limits.

Figure 4 shows the power efficiency achieved in single node Linpack on several nodes of L-CSC, classified by the GPUs' VIDs on the x-axis. The figure presents two measurements: one with fixed ASIC settings of 774 MHz and 1.018 V (ignoring the VID) and one with default settings of 900 MHz, where the voltage is defined by the VID. For the 900 MHz settings we used faster fan settings to remain in thermal limits. We have measured the difference in fan power consumption, and we add a third dataset to the figure, which is derived from the 900 MHz test but corrected for the higher power consumption due to the higher fan speed. Since the offset due to fan speed is constant, both curves have the same slope which shows the variability of the GPUs: by trend, the GPUs with higher VID, and thus higher voltage, drain more power and are less efficient. However, it is obvious that the effect of the fans is much larger.

We draw the following conclusions from the figure:

- The standard deviation of the power efficiency of the most efficient configuration is 1.2 % and is thus smaller than the deviations for the CPU systems and the Titan GPU system in Table 4). Considering Titan and L-CSC, we don't see any evidence that we should expect GPU systems to have more node-to-node variability than CPU systems.
- Among all the presented systems, L-CSC has the lowest variability among the nodes. Although we did not analyze this in sufficient detail, there is strong indication that this is due to the additional measures we have taken for L-CSC, to mitigate the sources of node variability., i. e. fixing voltage and fan speed.
- Surprisingly, the efficiency in the most efficient configuration with identical voltage is unrelated to the VID. We had expected that even though we had fixed the voltage to the same level, the VID would still have an influence, as it classifies the quality of an individual ASIC. This does not seem to be the case.
- The power variability due to the different fan speeds is many times more significant than the variability of the GPUs themselves.
- Running at the default settings, the nodes with higher VID are slightly less efficient than those with lower VID. The differences are small but there is a clear trend. This is in complete concordance with our expectations, as the GPUs with higher voltage drain more power.
- It is possible to screen processors (CPUs and GPUs) via software for the ones with the lowest VIDs. In this way, if the voltage is not fixed, by measuring only nodes with low VID, it is possible to obtain a favorably biased efficiency result.

This case study shows that the suggested sample size from the previous section yields an accurate measurement for the GPU cluster L-CSC. This affirms that our approach is valid for GPU clusters as Titan and L-CSC as well. However, we want to examine more GPU clusters in the future in order to confirm this claim.

On top of the sample size suggestions, the case study yields two additional suggestions how sources of node variability can be mitigated.

- The fans of all nodes should be pinned to the same speed. This has a larger influence than processor variability.
- If possible, one should scan for processors with middle VIDs and use such processors for the measurement.

We will have to examine the applicability of these two recommendations in real-world situations. Depending on the possibilities to manage fan speeds and to read out VIDs, the suggestions can be hard or impossible to follow. In addition, we will examine further sources of variability and confirm these conclusions by repeating the case study on different systems.

6. CONCLUSIONS AND NEXT STEPS

Power consumption measurements for large-scale HPC systems present a significant technical challenge. In this paper we have evaluated the current requirements imposed on extrapolating full-system measurements from reference subsets of the nodes and the runtime of a workload. The pivotal questions concern the required length of the measurement and the required size of the reference subset.

By analyzing the power profiles of four supercomputers running Linpack, we find that the variation in power consumption in different phases of the run has greatly increased with recent system designs. What originally was a source of up to 3% error has now been shown to be higher than 20% in multiple cases, which we consider unacceptable. We conclude that the only sensible alternative is to require the measurement of the entire core phase of the workload under test.

To evaluate inter-node power variability we present benchmark results from balanced, floating-point intensive workloads run on six large-scale systems with at least 200 nodes. The power distribution has proved to be near-normal for all systems tested. Based on this finding, we then used a statistical approach to determine the subset size (node count) that is required to do an extrapolation that yields the required accuracy (e.g., 1.5%) with a desired certainty (e.g., 95%).

Our approach requires knowledge of the quotient of power measurement standard deviation and per-node power consumption, which we extrapolate to be approximately within the range 0.015–0.025 based on our analysis. Assuming a conventional 95% confidence interval, we find a measurement of at least 11 nodes to be reasonable even for very large systems. Our overall recommendation is to require that 16 nodes be measured, or 10% of nodes, whichever is larger. This matches the requirement to reach our desired confidence interval with one level greater overall variability than we're currently seeing in practice. We also recommend that all submissions include an assessment of their measurement accuracy.

Our recommendation is based on data from different types of HPC systems and covers the majority of current HPC system types. The dataset includes Intel Xeon E5 (78.1% share, 392 systems in the Top500 Nov. 2014) and one of the largest NVidia K20x accelerated systems (20% share of the accelerated systems, Top500 Nov. 2014). Our data from the L-CSC cluster suggests that systems accelerated by AMD FirePro GPUs show similar behavior. Still, it would be interesting to include more systems in the analysis in the future, especially those with other accelerator or processor designs.

Our methods and analysis will remain valid for new large-scale systems as long as the application under test is regular. The specific percentage and count may shift if the level of variability increases significantly in the exascale timeframe, but our methods would show this and provide new baseline requirements.

While we focus our experiments on homogeneous CPU clusters, we found strong indications that our evaluation holds true for GPU-accelerated systems as well. Finally, we have presented an outlook on how the measurement methodology could be refined in the future to take into account sources of node variability in order to improve the accuracy.

Our recommendations with respect to the measurement duration and the number of nodes to measure have been adopted by the power measurement methodology of the Energy Efficient High Performance Computing Working Group, which is in force for the Green500 and Top500 lists. These recommendations will be adopted into the submission requirements for each list in the late 2015 time frame. The suggestions that emerged from our case study on the sources of component variability (fan speed and voltage) are still too

loose to lead to a meaningful specification. These should be analyzed in detail and refined in future work, but we assume that they can improve power measurement accuracy in the future.

Acknowledgments.

We would like to thank all the people and sites who contributed power measurement data including: the GSI Helmholtz Center for Heavy Ion Research, Francis Belot of CEA, and Florent Parent of Calcul Québec.

7. ADDITIONAL AUTHORS

Additional authors: Torsten Wilde (Leibniz Supercomputing Ctr., `Torsten.Wilde@lrz.de`), James H. Rogers (Oak Ridge National Laboratory, `jrogers@ornl.gov`), Devesh Tiwari (Oak Ridge National Laboratory, `tiwari@ornl.gov`). Devesh Tiwari couldn't be listed as an author in the metadata of the paper because of the limit on the number of papers a PC member can co-author.

8. REFERENCES

[1] B. Balaji, J. McCullough, R. K. Gupta, and Y. Agarwal. Accurate characterization of the variability in power consumption in modern mobile processors. In Workshop on Power-Aware Computing Systems, HotPower '12, 2012.

[2] S. Che, M. Boyer, J. Meng, D. Tarjan, J. W. Sheaffer, S.-H. Lee, and K. Skadron. Rodinia: A benchmark suite for heterogeneous computing. In IEEE International Symposium on Workload Characterization (IISWC), pages 44–54, 2009.

[3] J. D. Davis, S. Rivoire, M. Goldszmidt, and E. K. Ardestani. Accounting for variability in large-scale cluster power models. In 2nd Exascale Evaluation and Research Techniques Workshop (EXERT), 2011.

[4] T. Endo, A. Nukada, and S. Matsuoka. TSUBAME-KFC: Ultra green supercomputing testbed. Presented at International Conference for High Performance Computing, Networking, Storage and Analysis (SuperComputing, SC13), 2013.

[5] Energy Efficient High Performance Computing Working Group (EE-HPC-WG). Energy efficient high performance computing power measurement methodology, version 1.2rc2. http://www.green500.org/sites/default/files/eehpcwg/EEHPCWG_PowerMeasurementMethodology.pdf.

[6] X. Fan, W. Weber, and L. A. Barroso. Power provisioning for a warehouse-sized computer. In 34th International Symposium on Computer Architecture (ISCA), June 2007.

[7] Green500. http://www.green500.org.

[8] The Green Graph 500. http://green.graph500.org/.

[9] D. Hackenberg, T. Ilsche, R. Schöne, D. Molka, M. Schmidt, and W. E. Nagel. Power measurement techniques on standard compute nodes: A quantitative comparison. In IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS), 2013.

[10] D. Hackenberg, R. Oldenburg, D. Molka, and R. Schöne. Introducing FIRESTARTER: A processor stress test utility. In International Green Computing Conference (IGCC), 2013.

[11] D. Hackenberg, R. Schöne, D. Molka, M. S. Müller, and A. Knüpfer. Quantifying power consumption variations of HPC systems using SPEC MPI benchmarks. Computer Science - R&D, 25(3-4):155–163, 2010.

[12] C.-H. Hsu and S. W. Poole. Power measurement for high performance computing: State of the art. In International Workshop on Power Measurement and Profiling (PMP), 2011.

[13] G. Juckeland, W. Brantley, S. Chandrasekaran, B. Chapman, S. Che, M. Colgrove, H. Feng, A. Grund, R. Henschel, W.-M. Hwu, H. Li, M. S. Müller, M. Perminov, P. Shelepugin, K. Skadron, J. Stratton, A. Titov, K. Wang, M. van Waveren, B. Whitney, S. Wienke, R. Xu, and K. Kumaran. SPEC ACCEL – a standard application suite for measuring hardware accelerator performance. In 5th International Workshop on Performance Modeling, Benchmarking and Simulation of High Performance Computer Systems, 2014.

[14] S. Kamil, J. Shalf, and E. Strohmaier. Power efficiency in high performance computing. In IEEE International Symposium on Parallel and Distributed Processing, pages 1–8, April 2008.

[15] 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 B. M. Chapman, F. Massaioli, M. S. Müller, and M. Rorro, editors, OpenMP in a Heterogeneous World, volume 7312 of Lecture Notes in Computer Science, pages 223–236. Springer Berlin Heidelberg, 2012.

[16] D. Rohr, M. Bach, G. Nešković, V. Lindenstruth, C. Pinke, and O. Philipsen. Lattice-CSC: Optimizing and building an efficient supercomputer for Lattice-QCD and to achieve first place in Green500. In Proceedings of the International Supercomputing Conference, 2015.

[17] B. Rountree, D. H. Ahn, B. R. de Supinski, D. K. Lowenthal, and M. Schulz. Beyond DVFS: A first look at performance under a hardware-enforced power bound. In Workshop on High-Performance, Power-Aware Computing (HPPAC), 2012.

[18] J. Russell and R. Cohn. Prime95. 2012.

[19] T. R. Scogland, C. P. Steffen, T. Wilde, F. Parent, S. Coghlan, N. Bates, W. Feng, and E. Strohmaier. A power-measurement methodology for large-scale, high-performance computing. In 5th ACM/SPEC International Conference on Performance Engineering, ICPE '14, pages 149–159. ACM, 2014.

[20] SPEC Power and Performance Committee. SPEC power and performance benchmark methodology. Technical report, Standard Performance Evaluation Corporation, 2010.

[21] B. Subramaniam and W.-c. Feng. Understanding power measurement implications in the Green500 list. In IEEE/ACM International Conference on Green Computing and Communications (GreenCom), pages 245–251, Dec 2010.

[22] Top 500 supercomputing sites.

http://www.top500.org.