TUE, a new energy-efficiency metric applied at ORNL's Jaguar

Michael K Patterson[1,5], Stephen W Poole[2,5], Chung-Hsing Hsu[2,5], Don Maxwell[2], William Tschudi[3,5], Henry Coles[3,5], David J Martinez[4,5], Natalie Bates[5]

[1]Intel Architecture Group, Intel Corporation, Dupont, Washington, USA
michael.k.patterson@intel.com
[2]Oak Ridge National Laboratory, Oak Ridge, Tennessee, USA
[3]Lawrence Berkeley National Laboratory, Berkeley, California, USA
[4]Sandia National Laboratories, Albuquerque, New Mexico, USA
[5]Energy Efficient HPC Working Group, Anderson Island, Washington, USA

Abstract. The metric, Power Usage Effectiveness (PUE), has been successful in improving energy efficiency of data centers, but it is not perfect. One challenge is that PUE does not account for the power distribution and cooling losses inside IT equipment. This is particularly problematic in the HPC (high performance computing) space where system suppliers are moving cooling and power subsystems into or out of the cluster. This paper proposes two new metrics: ITUE (IT-power usage effectiveness), similar to PUE but "inside" the IT and TUE (total-power usage effectiveness), which combines the two for a total efficiency picture. We conclude with a demonstration of the method, and a case study of measurements at ORNL's Jaguar system. TUE provides a ratio of total energy, (internal and external support energy uses) and the specific energy used in the HPC. TUE can also be a means for comparing HPC site to HPC site.

Keywords: HPC, energy-efficiency, metrics, data center

1 Introduction

This Whitepaper is a collaborative effort of the Metrics team of the Energy Efficient HPC Working Group (EEHPC WG). It reviews successes and issues with Power Usage Effectiveness (PUE) and explores some of the gaps in the metric. It disassembles the metric, applies the same simple logic to the IT, and then to the whole; including the IT and Infrastructure. This methodology is shown to produce two new metrics, with the higher level metric being a combination of PUE and IT-power usage effectiveness (ITUE) yielding total-power usage effectiveness (TUE). These new metrics can be used to understand the entire energy use from the utility to the silicon. It can model the entire energy stack and allow exploration of how trade-offs in the infrastructure or the IT can help change the total efficiency. Previously that total efficiency could neither be measured nor trended without these proposed metrics.

adfa, p. 1, 2011.

2 Background

Power Usage Effectiveness (PUE), introduced in a paper by Malone and Belady [1], provides a simple metric that is used to give comparative results between data centers or of a single data center over time. The metric provides a simple way to understand the energy consumed by the infrastructure for a given IT load. In 2007 the Uptime Institute reported the average enterprise data center PUE was around 2.5. [2] This meant that the data center used 2.5X the energy needed to run the IT equipment by itself. The extra energy was used for cooling, lighting, maintaining standby power generation, and power conversion losses.

The Green Grid has written a number of White Papers since the original work [3,4]. Most recently The Green Grid, DOE, EPA, ASHRAE and others produced a white paper that represents a consensus definition including how and where to measure the metric [5]. PUE is defined as:

$$PUE = \frac{Total\ Data\ Center\ Annual\ Energy}{Total\ IT\ Annual\ Energy} \quad (1)$$

The metric has progressed in maturity and its widespread use has been responsible for the energy efficiency focus and resulting progress in energy efficiency of data center infrastructure since its definition. Admittedly it is at a very high level, a fine-grained evaluation of each term and components of the terms can be found in [6].

3 The Challenge

PUE, while very successful in driving energy efficiency of the infrastructure for data centers, is not perfect. Its advantages are its simplicity, both the math and the concept. However, it is not the be-all and end-all metric for data centers. That metric would entail computational performance and energy: a ``miles per gallon" metric for data centers. A much improved metric would be a Data Center Productivity Index (DCPI). This would be the useful work divided by the total facility energy (DCPI=Useful Work/Total Facility Energy). Useful work is difficult to define since there are many diverse computational tasks, so today there is no definition of such a metric. An exception might exist in the HPC world where more common benchmarks and applications tend to exist. One such benchmark is LINPACK metric [5]. It is not an application but simply solves a dense system of linear equations. This benchmark generally represents only a small fraction of actual applications, but it is commonly run in most HPC systems, and is used for Top500 [6] and Green500 rankings. As stated it is a poor indicator of all but a very few workload types and therefore not really an indicator of an individual clusters productivity, but it is widely run as a benchmark. The EEHPC WG is concurrently working on how to measure energy consistently and appropriately for HPC benchmarks. With definition of a Productivity

Index type of metric still well in the future, there are still other metrics which can be defined. There are two specific issues with PUE that need to be understood when using it. This paper proposes a methodology to address one of them.

One issue with PUE comes from its focus on the infrastructure. Consider a given data center with a known PUE. Assume that data center then goes through a refresh and upgrades their old IT equipment with new more efficient equipment. The new will likely provide more compute capability and possibly use less energy (IT manufacturers continue to reduce energy use at part-loads and idle, providing an overall reduction in IT energy). One interesting result here is that, if the infrastructure is left alone and runs just as it did before the new equipment was brought in, the PUE will go up. While this may be troubling to some, it actually is a non-issue. First, anytime new IT equipment is brought in, the infrastructure should be reviewed for needed changes and efficiency opportunities, this is more of an operational issue than a metric issue. The fact that the PUE went up is an indication that the infrastructure energy use did not scale with the IT load. Second, and more importantly, PUE is an infrastructure measure, to trend changes in the infrastructure over time, not to trend changes in the IT equipment. Changing the IT equipment is changing the baseline.

The second issue, the subject of this paper, is that of shifting cooling or power conversion loss. By definition, everything outside the IT is infrastructure, and everything inside is IT. As in the paragraph above, if the IT load and IT equipment remains fixed and you are only tracking your own data center energy efficiency over time, PUE can be used to guide facility operational or infrastructure efficiency improvements. The difficulty comes when infrastructure loads are moved from inside to outside the box (or vice versa). To illustrate this point consider three data centers with identical workloads and numbers of servers. Consider a data center (data center (a)) using free cooling, moving outdoor air into the building with building level fans. Then the IT level fans (considered as part of the "IT energy" in PUE) will move that cool air through the IT equipment. Now consider the neighboring data center (b). It has a different configuration with no building fans and using only the fans in the IT equipment to move the air (ramping up existing IT fans or using larger fans in the IT equipment). In this case the infrastructure load goes down, and the IT load goes up. The PUE will drop in this case. At the third data center (c) fans were removed from the IT equipment altogether and only the building fans provide air movement. Data center (c) will have the lowest IT load and a higher infrastructure load. Because of this, it will have the worst PUE of the three. In order of PUE, (b) is likely the lowest, then (a), then (c). Can we conclude that (b) is the best design and that (c) is the worst? Not at all. In fact PUE should not be used for this type of conclusion. The only valid way to determine the most energy efficient design would be to measure total energy, and we can do this because we started with identical output as an assumption. With the reality that all data centers are different in the number of servers and workloads, how would one compare the increasingly common case of infrastructure (cooling or power conversion or both) moving across the IT boundary?

4 Metric Proposal

ITUE is proposed as a possible solution. ITUE is intended to be a "PUE-type" metric for the IT equipment rather than for the data center. PUE is total energy divided by IT energy, analogously, ITUE is defined as total IT energy divided by computational energy.

$$ITUE = \frac{Total\ Energy\ into\ the\ IT\ Equipment}{Total\ Energy\ into\ the\ Compute\ Components} \quad (2)$$

As PUE identifies the infrastructure burden on the IT equipment, ITUE would identify the same for computing. Data centers have cooling devices, UPS, and PDUs, and other systems supporting the IT equipment. Similarly, IT equipment has internal fans, power supplies, and voltage regulators (VRs), etc.. The compute components can be defined as the CPU, memory, and storage, etc. The math and structure of PUE and ITUE are the same.

Now, these two metrics can be combined.

$$TUE = ITUE \times PUE \quad (3)$$

TUE is the total energy into the data center divided by the total energy to the computational components inside the IT equipment. Figure 1 illustrates the differences between PUE, ITUE, and TUE. Note that in equation 4, "IT" represents the IT equipment or everything inside the server or cluster. In equation 5 however, "IT" represents only the compute components (CPU, memory, fabric) but not cooling, power supplies or voltage regulators. Those (cooling, power supplies, and voltage regulators) are part of "IT" in equation 4. The definition of "a" through "i" in Equations 4 - 6 come from Figure 1.

$$PUE = \frac{Cooling + PowerDistribution + IT_{equip}}{IT} = \frac{a+b}{d} \quad (4)$$

$$ITUE = \frac{Cooling + Pwr\ Dist + Misc + IT_{comp}}{IT} = \frac{g}{i} \quad (5)$$

$$TUE = ITUE \times PUE = \frac{a+b}{i} \quad (6)$$

Additionally, the IT equipment list in PUE would necessarily include the network switches, I/O subsystem, and storage. ITUE and TUE should also be extended to cover the full spectrum of the IT equipment in the data center. The graphics and coverage in this paper are compute centric primarily for simplicities sake and not to exclude anything in the IT suite of equipment.

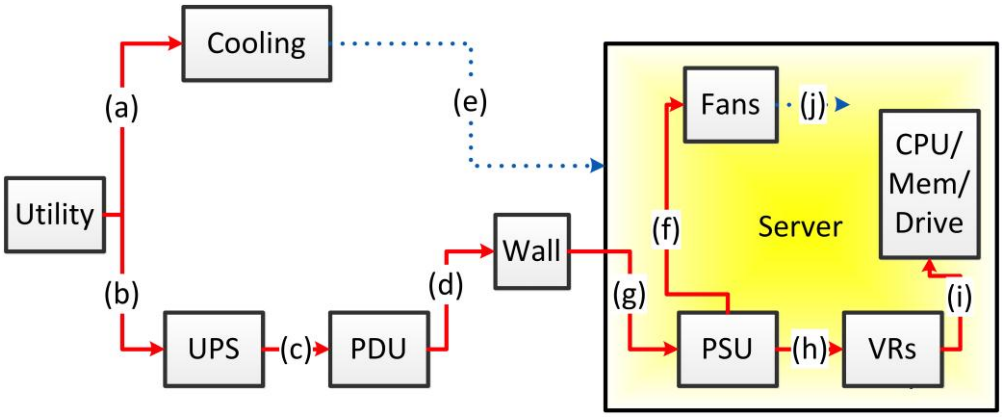

Fig. 1. Schematic of the combined Data Center and IT Equipment

Now that we have defined TUE as a function of the well understood PUE and the new ITUE we can apply it. Recall the comparisons of data centers (a), (b), and (c). The first (a) had fans in both the room and in the IT equipment. The second (b) had fans only in the IT. And the third (c) only had fans in the room, not in the IT equipment. While we cannot yet determine which uses the least energy, it is easy to see that our PUE fan energy accounting problem (where $PUE_b << PUE_a << PUE_c$) can be resolved. The mathematics of TUE do not favor one over the other as all the fans are in the numerator in all three cases. We would expect that all three TUEs to be much closer to each other than their respective PUEs, but more importantly, we can now use TUE to measure all three and to determine which data center and IT combination is actually the most energy efficient.

Another possible methodology for developing greater use and understanding of the metric could be the analogous historical development of PUE. [5] describes a good / better / best scheme; the simplest way to get a number is to use the readout of the UPS output power as the IT inlet power. For a data center with no better way to get IT inlet, this is at least "good". "Better" is using the PDU output as IT input, with "best" being direct measure of the IT energy. Similarly, a good / better / best approach to ITUE may help in its eventual adoption. The "good" may be as simple as the energy leaving the PSU minus the fan energy for the denominator, with the PSU "in" (wall socket energy) as the numerator. "Best" would be direct measurement of component level energy consumption.

This good-better-best approach certainly applies to measurement of the value of "i" in Fig 1. While many manufacturers now measure these values and they have become critical to node and system level power management, the energy use at the component level is at best "available with a little work" depending on the suppliers manageability

interface. Over time it will become more readily available for two reasons. First, if it is asked for by a growing community looking at ITUE and TUE, the eco-system will respond. This has happened already for the measurements needed for PUE. Second, as we proceed towards hard power limits in the exascale timeframe, the ability of the HPC applications to become energy aware can only happen with this data more fully exposed.

PUE's strict definition is the total ***annual*** energy divided by the IT ***annual*** energy. This is done to ensure any seasonal impacts are included in the number. Measuring PUE during the winter at a data center with extensive free cooling could skew that value significantly. Similarly, TUE and ITUE are defined as ***annual*** values as well. It may be beneficial and informative for an individual site to calculate the min or max values of these to help characterize their system (e.g. winter vs summer PUE), but for all three metrics they should only be reported as annual numbers.

The true value of ITUE is likely in the discussions around more advanced and more integrated infrastructure solutions. Difficulties with the simple concept of PUE come about when the line between infrastructure and IT are not clear. For example, many large supercomputers come with an integrated cooling system. Some components of these systems would generally be part of a data center room infrastructure in a more standard situation, but in these large specialized systems the standard IT servers, storage, and network are also not as easily split between infrastructure and IT. TUE and ITUE used with PUE, can be useful in being able to compare different data centers.

5 Demonstration of the Metrics

Consider a data center with a PUE of 1.6. For this example assume that the data center infrastructure efficiency is independent of the specifics of the IT and its particular efficiency. Compare this to a similar second data center, each having the same number of servers. The output of each data center will be assumed equal. Data center (a) (PUE = 1.6) has servers with standard or low first-cost components, particularly the fans, power supplies, and voltage regulators. The new data center (b) servers have high efficiency power supplies and fans. The physical infrastructure of data center (b) is identical to (a). From earlier discussions we know that the PUE of data center (b) would be lower, which has been a valid criticism of the PUE metric.

A detailed platform design model [7] shows that data center (a) uses servers with a power draw of 330 W, while data center (b) servers draw only 266 W as shown in Table 1. Recall that PUE_a = 1.6 = Power+Cooling+IT/IT. If we assume 10,000 servers in the space, the IT load is 3.30 MW, and the data center infrastructure uses 1.98 MW. The new data center's PUE (with identical infrastructure) with an IT load of 2.66 MW

is PUE=1.74. (The new data center could have some turn-down efficiency, but we assume not for the method's demonstrations). From this perspective, it looks like high efficiency components are a bad idea because the PUE is worse.

Table 1. – Server power use by platform and component

	a) Low Eff (W)	*b) High Eff (W)*
Total Platform	330	266
PSU	58	18
VRs	56	38
Fan	18	12
Processor, Memory, Other	198	198

The platforms were analyzed using the model of [7], and fan power, PSU losses, and board level conversion losses were identified. All other loads are considered compute power or "IT"; including the processors, memory, storage, network cards, etc…So for the low efficiency servers in data center (a) we have:

$$ITUE_a = \frac{18 + 58 + 56 + 198}{198} = 1.67 \quad (8)$$

And for the high efficiency servers in (b)

$$ITUE_b = \frac{12 + 18 + 38 + 198}{198} = 1.34 \quad (9)$$

The efficient platform has the lower $ITUE_a$ of 1.34. It carries a 34% "overhead" for power and cooling losses versus 67% for the lower efficiency version ($ITUE_b$=1.67).

From here, a higher level comparison of the two data centers using TUE can be made. Table 2 shows that for total efficiency, data center (b) with the high efficiency IT equipment is more efficient. TUE_b at 2.33 is better than TUE_a at 2.67, even though PUE originally had indicated the opposite. Additionally, with our earlier premise that output from both data centers is the same, the efficiency of the two is related directly to the one with the lower energy, and as expected TUE_b (higher efficiency) correlates with the lower total site power number of 4.64 MW. If the infrastructure can scale with the IT load, the actual PUE can similarly be used to calculate TUE.

Table 2.) Power and efficiency numbers of example data centers

	a) Low Eff	*b) High Eff*
Total Platform	3.31 MW	2.67 MW
Infrastructure	1.99 MW	1.99 MW
Total Site Power	5.3 MW	4.66 MW

PUE	1.6	1.74
ITUE	1.67	1.34
TUE	2.67	2.33

6 Case Study using ITUE

In this section we apply these concepts to the Jaguar system at Oak Ridge National Laboratory.

6.1 The Jaguar Supercomputer

The Jaguar system [8] consists of 200 Cray XT5 cabinets. Each cabinet contains three backplanes, a blower for air cooling, a power supply unit, and twenty-four blades. There are 4,672 compute blades and 128 service blades in Jaguar. A compute blade consists of four compute nodes, each having two six-core 2.6 GHz AMD Opteron processors. Two 4 GB DDR2 memory modules are connected to each processor. A compute blade also has a mezzanine card to support Cray's SeaStar2+ interconnect between nodes. A service blade consists of two nodes, a mezzanine card, and two PCI risers connecting to an external file system.

Jaguar uses both air and liquid to cool the system. Jaguar's liquid-cooling system uses both water and refrigerant R-134a. Cool air is blown vertically through a cabinet from bottom to top by a single axial turbofan. As the heat reaches the top of the cabinet, it boils the refrigerant which absorbs the heat through a change of phase from liquid to gas. The gas is converted back to liquid by the chilled-water heat exchanger inside a pumping unit where the water absorbs the heat and dissipates it externally. There are 48 external liquid-cooling units (denoted as XDPs) used for Jaguar.

6.2 Energy Efficiency Analysis

The site distributes 13.8kV power to the Computer Science Building (CSB) in which Jaguar is located. Transformers at the CSB convert the power to 480 Vac, and switch-boards (MSB) feed the power to Jaguar cabinets. The switchboards also provide 480 Vac connections to 48 XDPs. Inside a cabinet, the power supply unit (PSU) converts the 480 Vac power into 52 Vdc and deliver it to the blades. Each blade has an inter-mediate bus converter (IBC) that converts the 52 Vdc power into 12 Vdc. This power then traverses the blade and reaches the point of load (POL) next to the compute components (such as processors, memory modules, and mezzanine cards). The POL further converts the 12 Vdc power into 1.3 Vdc for the processors, and 1.8 Vdc for the memory.

Figure 2 depicts the Jaguar power delivery network inside of a cabinet. Orange boxes represent compute components. Brown boxes indicate where the electrical power can be monitored. For Jaguar, there are two locations where we can monitor the power: One is at the output of the switchboard, and the other is at the output of the power supply unit. Apparently, the power monitoring capabilities of the Cray XT5 are limited. Power can only be monitored at the cabinet level --- not at the blade level. For January 2011, the average aggregate output power from the switchboards and from the cabinet power supply units are 5,259.56 kW and 4,209.90 kW, respectively.

To calculate Jaguar's ITUE for January 2011, the efficiency ratings of IBC and POL are needed. In fact, the best way is to be able to monitor the power draw at the outputs of IBCs and POLs. Unfortunately, Jaguar does not provide this monitoring capability. The next best way is to get the efficiency ratings from vendors. Vendors often have this data but consider them proprietary. As a result, we examine HPC systems from other vendors' public information. This is because the vendors are more likely to use similar state-of-the-art packaging technologies for their systems. Following the methodologies around the JUGENE supercomputer [9] and the K supercomputer [10], we determine that the IBCs and POLs in Jaguar have the combined efficiency of 84%. The Cray blowers were estimated to carry a 7% penalty as reported in [11].

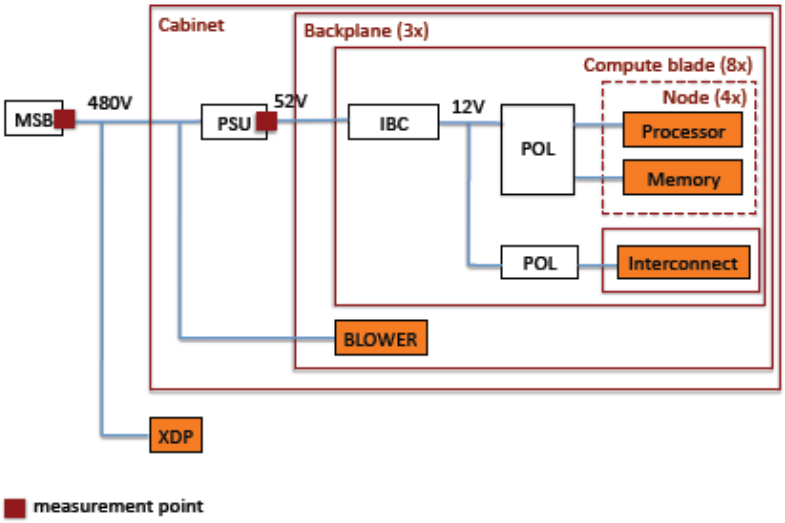

Fig. 2. Power monitoring capabilities for Jaguar

We calculate the Jaguar ITUE as follows. Since the average power attributed for compute is 4,209.90 kW x 84% = 3,536.32 kW, the metric value can be calculated as 5259.56 / 3536.32 = 1.49. That is, for every kW supplied for computing, there is additional 0.49 kW supplied for cooling and lost in power distribution. Using our estimate of ITUE = 1.49 and the PUE of the CSB as 1.25, the Jaguar TUE = $ITUE_J$ x PUE_{CSB} = 1.86.

7 Challenges and Future Work

Similar issues still exist in ITUE and TUE as do in PUE. These simply need to be understood and dealt with. For example, when more efficient IT equipment is installed in a data center and nothing else is done, PUE will go up (the denominator went down more than the numerator). Similarly if new lower power CPUs or DIMMs were installed in a server, ITUE (and TUE) will go up (again, the denominator went down more than the numerator).

A complicating factor of PUE and ITUE is temperature. The temperature at which the data center, as well as the IT components, operates at can significantly affect both values pushing one up and perhaps the other one down. The ability for the data center operator to pick the right temperatures is advantageous in the pursuit of overall highest efficiency (minimize TUE). Because of this, PUE or ITUE methods should not specify a temperature. However, the temperatures must be consistent. Measuring PUE at a given data center configuration with a certain temperature, then measuring ITUE at a different configuration and IT inlet temperature would render the TUE value invalid. Reporting temperatures during which PUE, ITUE, and TUE were measured would be beneficial in others understanding of the overall thermal management strategy of the data center and IT equipment.

Another issue is defining more precisely what is considered to be a compute load versus support or infrastructure loads. Certainly CPU, memory, memory controller, MIC or GPU processors are all compute. Fans, pumps, PSUs, VRs are all infrastructure. But what of disk drives? Solid state disk drives would seem to be compute, but much of a standard disk drive is spinning the disk. For consistency we suggest all storage be considered compute. Status lights are infrastructure. Baseboard or motherboard controllers are infrastructure. Using the data center level analog, the baseboard controller would be the same as the building control system.

Long term, being able to measure ITUE, at least in large scale HPC systems may be a useful capability to build into the equipment, but for now the development of the concept will give us a tool with which to extend the PUE concept to the IT equipment and then to the combined infrastructure and IT installation.

A good estimate of Jaguar's TUE (1.86) and ITUE (1.49) is now published. Jaguar has been decommissioned and replaced by Titan. Work to define these values for that system are ongoing The intention is to continue this line of work, add further refinements and begin to do comparisons with other HPC sites to be able to measure the true efficiency of the site and cluster together.

8 Conclusions

The Energy Efficient High Performance Working Group has proposed two new metrics to improve the tracking and comparison of energy efficiency in data centers. PUE has been as successful as it has because of its simplicity. ITUE has been developed as a direct analog of PUE; PUE for the server. While this value is of interest the true richness comes when multiplied by PUE to get TUE for the data center. This metric surpasses the value of PUE as it now includes the IT support inefficiencies that PUE left out.

ITUE and TUE and their measurement capability will take time to develop (as did PUE), but their use can drive greater efficiency and clearer comparisons in the data center.

9 Acknowledgements

The work we have done for the power analysis of Jaguar is supported by the Extreme Scale Systems Center at Oak Ridge National Laboratory funded by the United States Department of Defense.

The work was performed at the Oak Ridge National Laboratory, which is managed by UT-Battelle, LLC under Contract No. De-AC05-00OR22725.

Accordingly, the U.S. Government retains a non-exclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for U.S. Government purposes.

The EEHPC WG is an open collaborative group focusing on Energy Efficient High Performance Computing, partially supported the Federal Energy Management Program.

10 References

[1] Malone, C., Belady, C., “Metrics to Characterize Data Center & IT Equipment Energy Use”, Proceedings of 2006 Digital Power Forum, Richardson, TX, 2006

[2] Uptime Institute, “Uptime Institute 2007 Data Center Industry Survey”, www.uptimeinstitute.com, 2007,

[3] The Green Grid, White Paper #6 “Green Grid Data Center Power Effieicny Metrics: PUE and DCiE, Dec, 2008, http://www.thegreengrid.org/~/media/WhitePapers/White_Paper_6_-_PUE_and_DCiE_Eff_Metrics_30_December_2008.pdf?lang=en
[4] The Green Grid, “White Paper #49-PUE: A Comprehensive Examination of the Metric” October 2, 2012, http://www.thegreengrid.org/en/Global/Content/white-papers/WP49-PUEAComprehensiveExaminationoftheMetric
[5] EPA, DOE, TGG, ASHRAE, et.al. “Recommendations for Measuring and Reporting Overall Data Center Efficiency”, May 17, 2011 EPA Website: http://www.energystar.gov/ia/partners/prod_development/downloads/Data_Center_Metrics_Task_Force_Recommendations_V2.pdf?7438-21e8,
[6] http://top500.org/project/linpack/ Top500 [Internet]. Top 500 Supercomputing Sites; (c2000-2012) [cited 2012 October 31]. Available from http://top500.org/project/linpack
[7] Intel, Power Joint Engineering Team System Power Calculator, January 2013, unpublished
[8] A.S. Bland, W. Joubert, R.A. Kendall, D.B. Kothe, J.H. Rogers, and G.M. Shipman. Jaguar: The world’s most powerful computer system – an update. In Cray Users Group, May 2010.
[9] M. Hennecke, W. Frings, W. Homberg, A. Zitz, M. Knobloch, and H. Böttiger. Measuring power consumption on IBM Blue Gene/P. In International Conference on Energy-Aware High Performance Computing, September 2011.
[10] H. Maeda, H. Kubo, H. Shimamori, A. Tamura, and J. Wei. System packaging technologies for the K computer. Fujitsu Scientific and Technical Journal, 48(3):286–294, July 2012.
[11] J. Rogers, B. Hoehn, and D. Kelley. Deploying large scale XT systems at ORNL. In Cray Users Group, May 2009.