The Energy Efficiency of the Jaguar Supercomputer

Chung-Hsing Hsu and Stephen W. Poole and Don Maxwell
Oak Ridge National Laboratory
Oak Ridge, TN 37831
{hsuc,spoole,maxwellde}@ornl.gov

December 7, 2012

1 Introduction

The growing energy cost has become a major concern for data centers. The metric, Power Usage Effectiveness (PUE), has been very successful in driving the energy efficiency of data centers. PUE is not perfect, however. It often does not account for the energy loss due to power distribution and cooling *inside* the IT equipment. In HPC space this is particularly problematic as the newer designs tend to move part of power and cooling subsystems into the IT hardware. This paper presents a preliminary analysis of the energy efficiency of the Jaguar supercomputer by taking the above issue into account. It also identifies the technical challenges associated with the analysis.

2 The Jaguar Supercomputer

The HPC system we study in this paper is the Jaguar supercomputer [1]. This system consists of 200 Cray XT5 cabinets in a configuration of eight rows by twenty-five columns with a footprint of roughly the size of a basketball court. Each cabinet contains three backplanes, a blower for air cooling, a power supply unit, a control system (CRMS), and twenty-four blades. Jaguar has two types of blades: compute blades and service blades. A compute blade consists of four nodes and a mezzanine card. A node has two six-core 2.6 GHz AMD Opteron 2435 processors. Each processor is connected to two 4 GB DDR2-800 memory modules. The mezzanine card supports the Cray SeaStar2+ three-dimensional torus interconnect. A service blade consists of two nodes, a mezzanine card, and two PCI risers to connect to a Lustre-based file system. There are 4,672 compute blades and 128 service blades in Jaguar.

Jaguar uses both air and liquid to cool the system. The ECOPhlex (short for PHase-change Liquid Exchange) cooling system, developed by Cray with Liebert, uses both water and refrigerant R-134a. Cool air flows vertically through the cabinet from bottom to top by a blower. As the heat reaches the top of the cabinet, it boils the refrigerant which absorbs the heat through a change of phase from a liquid to a gas. The gas is converted back to liquid by the chilled-water heat exchanger inside a Liebert XDP pumping unit where the water absorbs the heat and exhausts it externally. Each XDP unit has 240 kW cooling capacity, and there are 48 of them for Jaguar. In addition, Cray has been using a single axial turbofan in cabinets since the XT3. It is argued to be a lot more efficient than a large collection of less-powerful fans and the maintenance interval is longer, seven-and-a-half years versus a few months with the small fans.

3 Energy Efficiency Analysis

This section describes the energy efficiency analysis of Jaguar. We start with studying how the essential compute components (such as processors, memory, and interconnect) get their electrical power. At the very first power source, the Tennessee Valley Authority generates the needed electricity at one of its power plants, and supplies it to a power substation at ORNL via 161 kilovolts (kV) transmission lines. The substation converts the power into 13.8 kV and

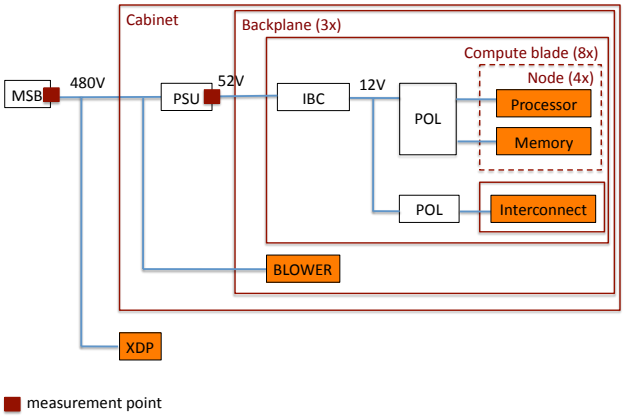


Figure 1: The power monitoring capabilities for Jaguar.

distributes it to the Computer Science Building (CSB). Transformers at the CSB convert the power down to 480 V, and switchboards (MSB) feed the power to Jaguar cabinets. Jaguar requires three full switchboards. The three switchboards also provide the 480 V 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 intermediate bus converter (IBC) that converts the 52 V power into 12 V. This power then traverses the blade and reaches the point of load (POL) next to the compute components. The POL converts the 12 V power into 1.3 V for the processors to use, and into 1.8 V for the memory to use.

Figure 1 depicts the power delivery network of Jaguar inside 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. Unfortunately, the power monitoring capabilities of Cray XT5 are limited. We cannot monitor the power at the blade level. We can only monitor the power at the cabinet level.

Based on Figure 1, we set up an array of equations to facilitate energy efficiency analysis. The energy efficiency metric we will use is called Information Technology Energy Usage (ITUE) [6], shown as Equation (1). ITUE is defined as the total energy supplying to IT hardware divided by the energy used directly for compute. In other words, ITUE is a "PUE" for the IT equipment. ITUE tries to capture support inefficiencies in the IT equipment such as fans, power supplies, and voltage regulators. Equation (2) states that the output power of the switchboards (P_{MSB}) accounts for the input power of both cabinets and XDP units. Equation (3) states that the input power of the cabinets (P_{CAB}) is the sum of the input power of the PSU and blower. Equation (4) states that the input power of the PSU (P_{PSU}) is partially used for compute and partially lost in three levels of conversion. We use Δ to denote conversion loss.

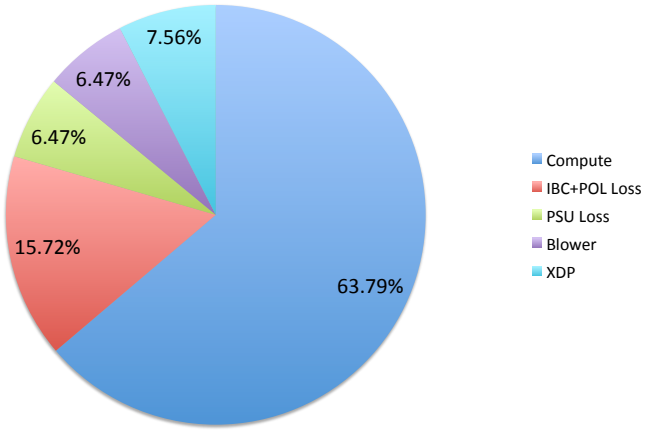
$$\text{ITUE} = P_{MSB}/P_{COMPUTE} \quad (1)$$

$$P_{MSB} = P_{CAB} + P_{XDP} \quad (2)$$

$$P_{CAB} = P_{PSU} + P_{BLOWER} \quad (3)$$

$$P_{PSU} = P_{COMPUTE} + \Delta_{PSU} + \Delta_{IBC} + \Delta_{POL} \quad (4)$$

For the entire month of January 2011, the average of the aggregate power demand from the output of switchboards is 5,295.50 kW. This observation is reflected in Equation (5). Similarly, the average of the aggregate power demand from the output of cabinet power supply units is 4,209.95 kW. This observation is reflected in Equation (6).



ITUE	P_{XDP}	P_{BLOWER}	Δ_{PSU}	$\Delta_{IBC} + \Delta_{POL}$	$P_{COMPUTE}$
1.57	400.21	342.67	342.67	832.20	3,377.75

Figure 2: The power breakdown of Jaguar at the switchboard level.

$$P_{MSB} = 5295.50\,\text{kW} \quad (5)$$

$$P_{PSU} - \Delta_{PSU} = 4209.95\,\text{kW} \quad (6)$$

Since the power monitoring capabilities of Jaguar is limited, we rely on the vendor data to help set up the final set of equations. Based on a pie chart of the power allocation within a cabinet in [5], we assume that the blower and rectifier loss were each about 7%, and the power train loss, the loss associated with all of the DC step down and deliveries, was about 17%. These assumptions are reflected in Equation (7–9).

$$P_{BLOWER} = 7\% \cdot P_{CAB} \quad (7)$$

$$\Delta_{PSU} = 7\% \cdot P_{CAB} \quad (8)$$

$$\Delta_{IBC} + \Delta_{POL} = 17\% \cdot P_{CAB} \quad (9)$$

Now that we have the complete array of equations for energy efficiency analysis, solving it is rather straightforward. Figure 2 depicts the power breakdown in a pie chart. For the ITUE metric, since the average power attributed directly for compute is 3,377.75 kW, the metric value can be calculated as 5295.50 / 3377.75 = 1.57. That is, for every joule contributed to compute, there is additional 0.57 joules consumed for power distribution and cooling. Ideally, we would like to see the ITUE value of 1.0 in an HPC system.

4 Discussion

This section discusses the technical challenges associated with the energy efficiency analysis of Jaguar. First, we try to understand the accuracy of the analysis results. Based on the method presented in [7], we can obtain the power of the blower in each cabinet. Specifically, the blower's power consumption is characterized as a function of the blower's frequency with the blower's frequencies being sampled at each cabinet. The frequency-power relationship of the blower is shown in Table 1, and linear interpolation is used to calculate between points. Using the method, we are

XT5 fan power						
Hz	40	45	50	55	60	65
kW	1.2	1.6	2.2	2.8	3.7	4.7

Table 1: The frequency-power relationship of the blower in a Cray XT5 cabinet.

able to derive P_{BLOWER} as 524.25 kW. In comparison with the value of 342.67 kW we derived in the analysis, the mismatch of the values is quite significant.

A similar suspicion arises when we try to understand the power demand of 48 XDPs. The analysis shows that the power demand P_{XDP} is 400.21 kW. We know that each XDP is rated at 5 horsepower (HP) and has 240 kW cooling capacity. Since 5 HP is equivalent to 0.746 * 5 = 3.73 kW, the power demand of 48 XDPs at the 100% loading level will be 3.73 kW * 48 / 85% = 210.64 kW assuming the motor efficiency of 85% in the XDPs. Since the total cooling capacity of 48 XDPs is 240 kW * 48 = 11,520 kW, it is impossible that the loading level of XDPs reached the 100%. The average loading level is probably around 5295.50 / 11520 = 46% or less. Hence, we expect the aggregate power load of XDPs to be half of 210.64 kW. Again, the mismatch of the values is significant.

The best way to address this technical challenge is to be able to monitor the power draw at the outputs of all the power conversion devices such as PSU, IBC and POL. Since it is inapplicable to Jaguar, we turn to another method which tries to figure out the efficiency rating of all these devices. This method has its own challenges, too. Vendors often have this data but consider them proprietary. In addition, the efficiency rating is actually a curve, not a single value. Depending on the load, the efficiency varies. In the following we present our best effort in using the method to evaluate the quality of the analysis results.

We know that Jaguar uses Cray high-efficiency (HE) cabinets in which Cray has invested in AC/DC power rectification with a 92%-plus efficiency rating [3]. Therefore, we assume the PSU in Jaguar has a 92% conversion efficiency. To get an idea for the efficiency ratings of IBC and POL in Jaguar, we examine HPC systems from other vendors that have this information in public. This is because the vendors are more likely to use similar state-of-the-art packaging technologies for theses systems. We examined the K computer [4] and an IBM Blue Gene/P (BGP) system at Forschungszentrum Jülich in Germany called JUGENE [2]. Both systems have PSUs with a 91% efficiency rating. Moreover, the combined efficiency of IBC and POL is between 84% and 86%. We assume that the IBC and POL in Jaguar has the combined efficiency of 84%. The following lists the new equations for the energy analysis of Jaguar.

$$P_{BLOWER} = 524.25 \text{ kW} \quad (10)$$

$$\Delta_{PSU} = (100\% - 92\%) \cdot P_{PSU} \quad (11)$$

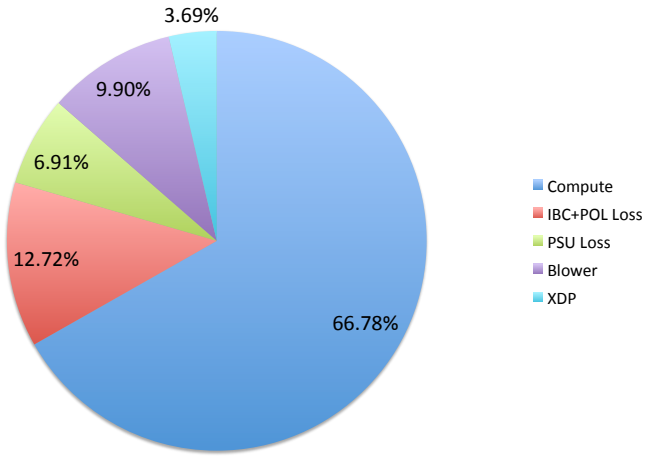
$$\Delta_{IBC} + \Delta_{POL} = (100\% - 84\%) \cdot (P_{PSU} - \Delta_{PSU}) \quad (12)$$

Using Equations (1–6) and (10–12), we can estimate the ITUE as 1.50. It is not too different from the value of 1.57 in the previous analysis. Figure 3 depicts the revised power breakdown in a pie chart.

As we have mentioned, the main technical challenge for calculating ITUE lies in the limited power monitoring capabilities of the system. We want to rely more on direct physical measurements and less on vendor's specification or indirect estimates. Fortunately, newer HPC systems such as IBM Blue Gene/Q (BGQ) and Cray XC30 allow for more power domains in a cabinet to be monitored. For example, one can monitor the input and output power of the PSU as well as seven power domains (such as processor and memory) on each node in a BGQ cabinet [8]. Similarly, all of the major power domains of Cray XC30 will report voltage and current on not only the output but also the input.

Finally, we attempt to compare Jaguar with other HPC systems in terms of the ITUE. Given the limited public information, we can only estimate the ITUE of JUGENE, a 72-cabinet BGP system. Figure 4 depicts the cabinet-level power delivery network of JUGENE. According to [2], we have the following array of equations for the first quarter of 2011. From these equation, we can calculate the ITUE as 1.36. In comparison with the ITUE of 1.5 for Jaguar, JUGENE has a lower ITUE value.

$$\text{ITUE} = P_{PSU}/P_{COMPUTE} \quad (13)$$



ITUE	P_{XDP}	P_{BLOWER}	Δ_{PSU}	$\Delta_{IBC} + \Delta_{POL}$	$P_{COMPUTE}$
1.50	195.22	524.25	366.08	673.59	3,536.36

Figure 3: The revised power breakdown of Jaguar at the switchboard level.

$$P_{PSU} = P_{COMPUTE} + P_{FANS} + \Delta_{PSU} + \Delta_{POL} \quad (14)$$

$$P_{PSU} - \Delta_{PSU} = 1650 \text{ kW} \quad (15)$$

$$P_{FANS} = 115.2 \text{ kW (estimated)} \quad (16)$$

$$\Delta_{PSU} = (100\% - 91\%) \cdot P_{PSU} \quad (17)$$

$$\Delta_{POL} = (100\% - 87\%) \cdot (P_{PSU} - \Delta_{PSU} - P_{FANS}) \quad (18)$$

There are several factors that lead JUGENE to have a smaller ITUE number. First, JUGENE uses hydro-air cooling. However, the calculation of its ITUE only accounts for the fan power. Second, JUGNE has one less level of DC/DC conversion, i.e., it does not have IBC, which makes its overall power conversion efficiency higher. Note that the newer BGQ system has IBC [8]. Third, the value of $P_{COMPUTE}$ is estimated rather than measured. While BGP allows the power at every POL to be monitored, the measured values are only partially reported in [2].

5 Conclusions

The growing energy cost has driven the popularity of the PUE metric for data centers. After all, one cannot improve if one cannot measure. In practice, however, PUE has only driven the invention of cooling technologies outside the IT equipment. As a result, ITUE has been proposed trying to address the energy loss due to power distribution and cooling inside the IT equipment. It is our opinion that the PUE values and the ITUE values in HPC space will not vary much. This is because HPC systems tend to use similar top-of-the-line system packaging technologies. While ITUE will be good for driving the second phase of invention, the ultimate challenge will be on how to improve the energy efficiency of compute.

Acknowledgment

This work was supported by the United States Department of Defense and used resources of the Extreme Scale Systems Center at Oak Ridge National Laboratory. 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

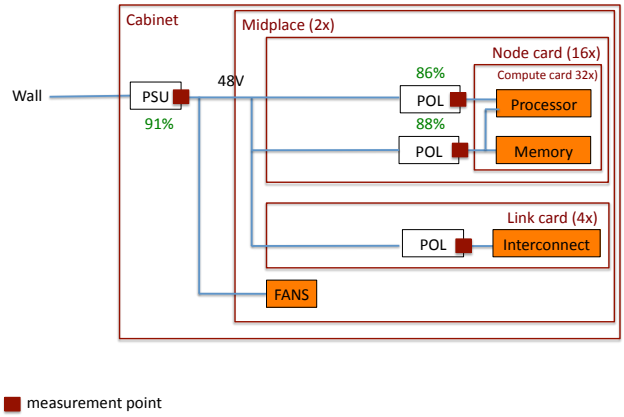


Figure 4: The power monitoring capabilities for JUGENE.

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.

References

[1] 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.

[2] 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.

[3] High-end HPC heats up: Cray and the path to exascale computing – IDC vendor spotlight. `http://www.cray.com/Products/XE/Resources.aspx`.

[4] 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.

[5] J. Rogers, B. Hoehn, and D. Kelley. Deploying large scale XT systems at ORNL. In *Cray Users Group*, May 2009.

[6] The Energy Efficient High Performance Working Group. Tue, a new energy-efficiency metric; moving pue inside the box, 2012.

[7] T. Wenning and M. MacDonald. High performance computing data center metering protocol. Resources on data center energy efficiency, Federal Energy Management Program, U.S. Department of Energy, September 2010.

[8] 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 *IEEE International Conference on Cluster Computing*, September 2012.