Power Use of Disk Subsystems in Supercomputers

Matthew L. Curry
Sandia National Laboratories
P.O. Box 5800
Albuquerque, NM 87185-1319
mlcurry@sandia.gov

H. Lee Ward
Sandia National Laboratories
P.O. Box 5800
Albuquerque, NM 87185-1319
lee@sandia.gov

Gary Grider
Los Alamos National
Laboratory
MS B260
Los Alamos, NM 87544
ggrider@lanl.gov

Jill Gemmill
Clemson University
324 Fluor Daniel
Clemson, SC 29631
gemmill@clemson.edu

Jay Harris
Clemson University
340 Computer Court
Anderson, SC 29625
jayh@clemson.edu

David Martinez
Sandia National Laboratories
P.O. Box 5800
Albuquerque, NM 87185-0823
davmart@sandia.gov

ABSTRACT

Exascale will present many challenges to the HPC community, but the primary problem will likely be power consumption. Current petascale systems already use a significant fraction of the power that an exascale system will be allotted. In this paper, we show measurements for real I/O power use in three large systems. We show that I/O power use is proportionally fairly low per machine, between 4.4 and 5.5% of the total consumption. We use these measurements to motivate a burst-buffer checkpointing solution for power-efficient I/O at exascale. We estimated this solution to use approximately 6.6% of the exascale machine power budget, which is on par with today's systems.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: Miscellaneous

General Terms

Measurement

Keywords

Power efficiency, high performance computing, data storage

1. INTRODUCTION

Many predictions about exascale machines, including those of the Exascale Computing Study [20], indicate that power is a major obstacle for developing future large platforms. In response, many groups in the high performance computing (HPC) community are undergoing deep investigations to discover inefficiencies. This certainly includes the storage community, where several power-related discussions have taken place. While concern abounds, there is little publicly released data to frame discussions on the power use of data storage infrastructure for large computers.

This paper provides compute and storage power use from three institutions: Clemson University, Los Alamos National Laboratory, and Sandia National Laboratories. The systems described are capability and capacity systems, with power measurements being taken while running normally. While the machines vary from each other substantially, there is one commonality: All disk storage infrastructure combined used only a small fraction of the power consumed by the entire machine. We also show through vendor predictions and the power data obtained that, by using new power saving techniques, an exascale storage system should consume a similarly small proportion of power.

2. MOTIVATION

There is no question that HPC requires a significant amount of power. One reason is that application data is too large to fit into the smallest and fastest memories, necessitating the use of a memory hierarchy. A typical system would include registers, several levels of cache, main memory (DRAM), and disk. Worse, the disks could be located in another machine that is accessed through a network, as is the case in the most popular massively parallel computer architectures [6, 17].

Traditionally, the main concern is that increasing latency imposed by the memory hierarchy influences the amount of time needed to access data. In the exascale era, the amount of power used is just as important, and is also directly related to where the data is within the memory hierarchy. It makes intuitive sense that, the deeper the data is in the hierarchy, the more power is necessary to retrieve the data. However, the power required to move between layers can be significantly different.

Moving data between DRAM and off-chip cache requires approximately 0.001 and 0.1 nanojoules of energy per byte [11]. Interestingly, this is much less than the amount of energy to move the same amount of data between off-chip cache and on-chip cache, which is approximately between 0.1 and 1 nanojoules per byte [11]. However, both of these figures are drastically less than the amount required for to move data

PDSW'11, November 13, 2011, Seattle, Washington, USA.
Copyright 2011 ACM 978-1-4503-1103-8/11/11 ...$10.00.

from a disk: 100-1000 nanojoules per byte [11]. Secondary storage appears to be ripe for power optimization.

The true power consumption of I/O, however, depends on several different factors like the application's behavior, the architecture and configuration of the I/O servers within the system, and administrative policies of the system. For example, an application whose primary operation is double-precision matrix multiplication will use proportionally less power for I/O than an application that performs a large map-reduce job. To understand real-world power load, one must measure real-world systems. Each site was asked to measure the compute and I/O power use for their machines individually in order to provide an accurate characterization of their machines' power use. It is worth noting that this survey only included compute, storage, and networking infrastructure. No sites included other consumers of power, like cooling.

3. SITE SURVEYS

3.1 Los Alamos National Laboratory

Los Alamos Laboratory and Sandia National Laboratories, through the Alliance for Computing at Extreme Scale (ACES), procured, operate, and use Cielo, the primary capability computing platform for both labs [2]. Cielo has achieved 1.11 PF using LINPACK [7], which makes it the sixth most powerful supercomputer on the June 2011 Top500 list [25]. The compute section of Cielo, a Cray XE6, has 8,944 dual-socket nodes, populated with 2.4 Ghz eight-core AMD Magny-Cours processors, with 32GB of DDR3 memory per node [24]. Cielo is connected to a 10 PB Panasas storage system that manages hardware-accelerated RAID 6 arrays with eight disks each.

There are other machines in the same building that use an enterprise storage model; i.e., there is another centralized 10 PB high-speed parallel file system that is available to several machines including Roadrunner, a heterogenous platform that contains IBM Cell processors, and 800 TF of capacity compute clusters. The SAN is PaScalBB, a high-bandwidth 10 Gb ethernet fabric [3]. Together, all of the clusters and storage form the secure environment at Los Alamos National Laboratory, comprising approximately 3.5 PF of compute and 20 PB of storage. The secure environment uses approximately 16.5 MW of power total.

Figure 1 depicts the power use of Cielo individually, along with the power use of the entire secure computing environment at Los Alamos. The power use of storage was measured by reading the power draw directly from the power distribution units several times throughout the course of a normal working day, when the utilization was high. Cielo uses, on average, 4.3% of its power for disks, SAN, and I/O servers, while the entire secure computing environment in aggregate uses 2.4% for those components. The storage infrastructure is clearly a minor consumer of power compared to compute nodes.

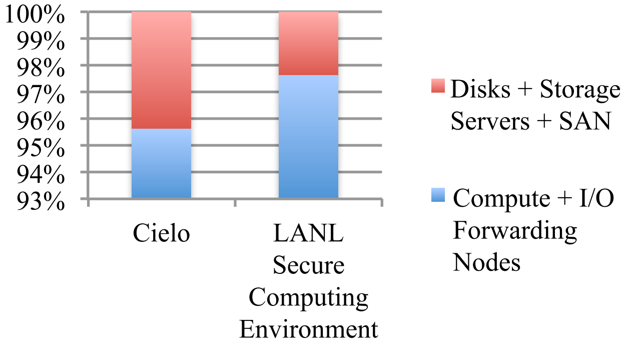

Figure 1: Power use for Los Alamos Laboratory supercomputers and storage infrastructure for Cielo individually, and for the entire secure computing environment that includes Roadrunner, Cielo, and capacity machines

Make/Model	CPU	RAM	Qty
Sun X6250	Intel Xeon L5420	32 GB	430
IBM dx340	Intel Xeon E5410	16 GB	340
Dell PE1950	Intel Xeon E5410	12 GB	258
Dell PE1950	Intel Xeon E5345	12 GB	257
Sun X2200	AMD Opteron 2356	16 GB	256
HP DL165	AMD Opteron 6176	48 GB	70
Sun X4150	Intel E5410	16 GB	10
HP DL580	Intel Xeon X7542	512 GB	6
Sun X4600	AMD Opteron 8220	256 GB	1
HP DL980	Intel Xeon X6560	2 TB	1

Table 1: Types of compute nodes in Palmetto

3.2 Clemson University

Palmetto is Clemson University's largest compute cluster. It represents a successful instantiation of the condominium model of supercomputing, where users contribute funds for hardware to secure time on the platform [1, 5]. This model creates a varied mix of hardware that can participate in computations. Palmetto is used to run a large number of applications, including molecular dynamics, econometrics, network simulation, biophysics, genomics, and combustion codes. It achieved 92 TF on the LINPACK benchmark, securing it the 96th slot on the June 2011 Top500 list [7, 25].

Table 1 shows a list of all machines within the cluster, sorted by quantity. Nodes are connected to a Myrinet fabric. Once or twice per year, Clemson will make a bulk hardware purchase on behalf of those buying into Palmetto. The vast majority of compute capability in the cluster (1611 nodes, or 14008 cores) were purchased in this manner. A small number of large-RAM nodes were purchased for specific applications, and are not included in the general availability pool of resources. Palmetto's storage hardware includes 32 Dell R510 servers with Dell PERC H700 controllers. Each controller runs a RAID 5 array with five 1 TB disks, with a total usable space of 256 TB. An OrangeFS scratch file system [19] of 115 TB occupies this space, with the remainder unused. A further 360 TB of data storage is provided by three DDN arrays: Two DDN 6620s with sixty 2 TB disks apiece that host home and project directories; and one currently unused DDN 9550 with five trays, each containing forty-eight 500 GB disks [14]. The DDN 6620 units are configured with Sun QFS file systems [18]. Two Sun F5100 servers with eighty 22 GB flash disks each are used for QFS metadata.

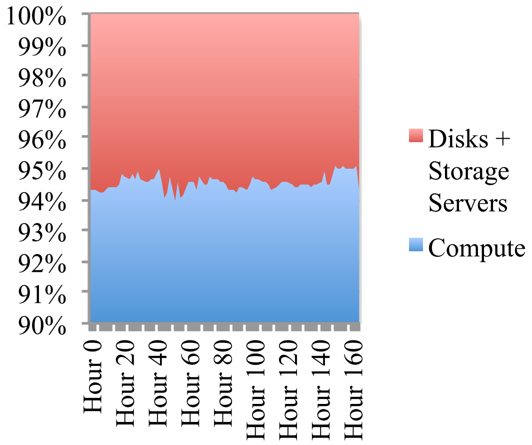

Figure 2: Power use for Palmetto and its storage infrastructure

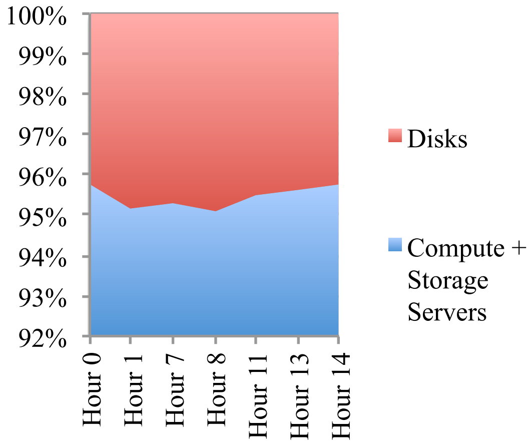

Figure 3: Predicted power use for Red Sky

In measuring the power use of Palmetto, each power distribution unit was sampled every two hours. During this period, June 20th, 2011, through June 27th, 2011, Palmetto's utilization averaged 54% [4]. Figure 2 shows that Palmetto's storage power use ranges between 4.9% and 6.1% of the total machine's consumption throughout the testing period, or 20 to 24 kW total. Meanwhile, total compute power draw ranged from 359 to 427 kW. During higher utilization periods, it is likely that the compute power draw is higher. This would reduce the proportion of power used by storage.

3.3 Sandia National Laboratories

Red Sky is the premier capacity compute platform at Sandia National Laboratories. When combined with Red Mesa, a similar machine hosted at Sandia for the National Renewable Energy Laboratory [16], it ranked 16th on the June 2011 Top500 list, having achieved 433.5 TF on the LINPACK benchmark [7, 25]. One of the machine's notable features is that it is the first Infiniband cluster with a 3-D torus topology [9]. It is composed of Sun X6275 blades, each containing two Intel X5570 processors and 12 GB of RAM. The storage nodes are nearly identical to the compute nodes, save connection to the storage, and are in the same racks. Red Sky hosts 3 PB of raw storage organized as Linux software RAID 6 arrays. These provide the storage for several Lustre file systems.

Red Sky is designed to be split between two modes of operation simultaneously: Classified and unclassified. The unclassified partition of the machine, at the time of this writing, was configured to be five-sixths of the nodes in Red Sky, or about 54% of the nodes in Red Sky and Red Mesa combined [21]. This portion of the machine includes a 750 TB Lustre file system for scratch, and two other 18 TB Lustre file systems for home and project directories. We project the results based on this portion of the machine, as it is intended to be a reasonable partition that can be used independently.

Rather than measure the power use of the whole machine, we measured three racks (one compute, two storage) and extrapolated it to the size of Red Sky's unclassified partition, which contains 40 racks of compute and six racks of storage. Several times throughout a 14-hour period, we read the power use directly from the compute rack's power distribution unit, and measured the draw of the disk cabinets from the circuit panel board using a Fluke 33 clamp meter [15]. The compute rack consumed between 23.2 and 27.2 kW, while the two storage racks together consumed a steady 16 kW throughout the test period. Figure 3 shows power use as extrapolated to the full size of Red Sky's unclassified partition. We predict 928-1088 kW are consumed by compute and I/O nodes, while 48 kW are consumed by disks. This implies that 4.2-4.8% of the machine's power is used for disks.

4. DISCUSSION

The data strongly shows that storage consumes much less power than compute. Figure 4 shows the power use data for each machine, normalized by the LINPACK score for each platform. The proportion of power consumed by storage is remarkably similar among the machines sampled, between 0.17 and 0.31 kW/TF. In comparison, computation uses up to 6.7 kW/TF. As mentioned previously, Palmetto's data shows low power consumption per TF because it was lightly loaded. An interesting feature of Red Sky and Palmetto is the steadiness of storage power consumption. Red Sky showed no variance throughout the testing period, while only 8.3% of Palmetto's 84 samples varied from the mode.

4.1 Extrapolation to Exascale

While many options are being contemplated, it is clear that an exascale machine will be quite different than today's machines. Currently, the supercomputers at Los Alamos National Laboratory in aggregate provide approximately 3.5 PF of peak compute power with 16.5 MW of power consumption. However, the first exaflop machine will be limited to approximately 20 MW [23]. This implies a radical transformation in the compute hardware used to construct the machine. Does disk infrastructure have to undergo a similar transition? The power use of current infrastructure can guide our analysis.

At 30 kW/PB, Cielo stands as a machine that has extremely efficient power use for I/O and infrastructure. How-

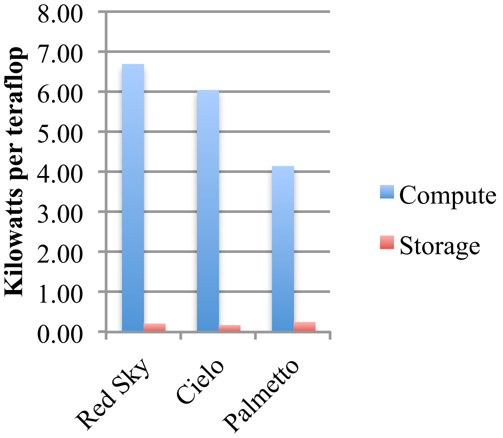

Figure 4: Power consumed per teraflop on LINPACK

ever, without a change of strategy, an exascale storage system may consume too much power. To understand why, we analyze a common workload for large scale machines, checkpointing. One rule of thumb is that an application should spend less than 5% of its execution time writing checkpoints to storage. With today's disks, this necessitates that the system purchaser buy enough disks to provide necessary bandwidth. The storage space obtained from the purchase is often far in excess of what is otherwise required in a scratch file system.

This assumption will also be true in the future. In 2018, when the first exascale system will appear, there will be at least 32 PB of RAM within the system [23]. Disks are projected hold 29.52 TB, and have a bandwidth of 384.2 MB/s [8]. The system is expected to require between 320 PB to 1 EB of storage space [23]. Disregarding fault tolerance measures like RAID, this can be satisfied with 10,847 to 33,898 disks. However, in order to accept a checkpoint every hour, which may be necessary with an exascale machine [8], the system will need to provide at least 106.7 TB/s of bandwidth. This requires over 277,633 disks! If we hold power per disk constant to our findings, this system would require about 13 MW of power, or 65% of the exascale system's power budget, *without* extra capacity for fault tolerance.

Grider has described and analyzed a different alternative involving a flash memory burst buffer [8]. Briefly, a small flash-based store can be used to quickly accept checkpoints from compute nodes within five minutes. The flash can be sized to contain three checkpoints, which are slowly transferred to the disk-based system during the 55 minutes between checkpoints. This uses only 9.7 TB/s of disk bandwidth, which can be provided by approximately 25,247 disks. Once again assuming the same power per disk as today, the disk-based storage system would require 1.2 MW, or 6% of the machine's power budget.

For the flash portion of the system, we must extrapolate from properties of contemporary flash memory. The Intel 320 flash drives up to and including the 160 GB models can sustain 1 MB/s of streaming write bandwidth for every gigabyte of capacity [10]. This would imply that, even today, the necessary 90 PB flash partition would deliver 90 TB/s of bandwidth, which is close to the target performance of the burst buffer system. Flash also currently has a lower power consumption per unit of storage under full utilization [10, 13], and consumes less power per device under real-world workloads [22], than hard disks. One can assume that adding 10% capacity in flash adds up to 10% additional power consumption, resulting in a system that uses about 6.6% of 20 MW. Grider demonstrates the efficacy of this approach with projected data from manufacturers [8], lending further weight to these predictions. Unfortunately, we cannot share that proprietary data directly in this paper.

5. CONCLUSION

Exascale is both a daunting problem and a lofty goal. Many who have analyzed the challenges state plainly that power is one of the biggest problems facing the operators of an exascale machine. Forward-looking members of the HPC community are searching for every opportunity to make significant reductions in power use for computer components by altering architecture, strategy, and design. This discussion is still ongoing, including within the storage community.

We have shown through empirical measurements of power use at three HPC sites that storage is not a large user of power within these machines. Inefficiencies in the power feed systems of the data center is often a larger consumer of power [12]. Further, power-reducing strategies like flash burst buffers can divorce the storage system from the capacity-bandwidth relationship of the past, allowing a modest system to have both high bandwidth and capacity for bursty, write-mostly workloads like checkpointing. While the storage community should continue to investigate novel ways to reduce power use within supercomputers, the majority of power use reduction efforts would be more fruitful if spent on subsystems that use significantly more power.

6. ACKNOWLEDGEMENTS

We would like to acknowledge Edward Duffy and Randall Martin of Clemson University for their responsiveness to our requests for information about Palmetto's hardware and configuration. Further, we acknowledge Steve Monk of Sandia National Laboratories for providing information on Red Sky's architecture.

This work was supported by the United States Department of Energy under Contract DE-AC04-94AL85000.

7. REFERENCES

[1] P. Agnihotri, V. K. Agarwala, J. J. Nucciarone, K. M. Morooney, and C. Das. The Penn State computing condominium scheduling system. In *Proceedings of the 1998 ACM/IEEE conference on Supercomputing (CDROM)*, Supercomputing '98, pages 1–23, Washington, DC, USA, 1998. IEEE Computer Society.

[2] J. Ang, D. Doerfler, S. Dosanjh, S. Hemmert, K. Koch, J. Morrison, and M. Vigil. The alliance for computing at the extreme scale. In *Proceedings of the Cray Users Group 2010*, 2010.

[3] H.-B. Chen, P. Fields, and A. Torrez. An intelligent parallel and scalable server I/O networking environment for high performance cluster computing systems. In *Proceedings of the International*

Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA), 2008.

[4] Clemson University. Cluster statistics for Palmetto. http://cluster-usage.clemson.edu/.

[5] Clemson University. The Palmetto cluster condominium program. http://citi.clemson.edu/condoprogram. Accessed on September 13, 2011.

[6] Cray, Inc. Cray XE6. http://www.cray.com/Assets/PDF/products/xe/CrayXE6Brochure.pdf, 2010.

[7] J. J. Dongarra, P. Luszczek, and A. Petitet. The LINPACK benchmark: Past, present, and future, August 2003.

[8] G. Grider. Exa-scale FSIO. HEC-FSIO workshop presentation. http://institute.lanl.gov/hec-fsio/workshops/2010/presentations/day1/Grider-HECFSIO-2010-ExascaleEconomics.pdf, August 2010.

[9] S. Holinka. Red Sky at night, Sandia's new computing might. *Sandia Lab News*, 61(24), December 2009. http://www.sandia.gov/LabNews/ln12-18-09/labnews12-18-09.pdf.

[10] Intel Corporation. Product brief: Intel solid-state drive 320 series. http://www.intel.com/content/dam/doc/product-brief/ssd-320-brief.pdf.

[11] B. Jacob, S. Ng, and D. Wang. *Memory Systems: Cache, DRAM, Disk*. Morgan Kaufmann, 2008.

[12] D.-H. Kim, T. Yu, H. Kim, H. Mok, and K.-S. Park. 300V DC feed system for Internet data center. In *Proceedings of the IEEE Eighth International Conference on Power Electronics and ECCE Asia (ICPE ECCE 2011)*, pages 2352 –2358, June 2011.

[13] S. T. LLC. Data sheet: Barracuda LP. http://www.seagate.com/docs/pdf/datasheet/disc/ds_barracuda_lp.pdf.

[14] R. Martin. Personal communication, September 2011.

[15] D. Martinez. Personal communication, September 2011.

[16] National Renewable Energy Laboratory. Sandia's new supercomputer helps energize NREL's research. http://www.nrel.gov/news/press/2009/775.html, December 2009. Accessed on September 15, 2011.

[17] NR Adiga, et al. An overview of the BlueGene/L supercomputer. In *Proceedings of the 2002 ACM/IEEE conference on Supercomputing*. IEEE Computer Society Press, 2002.

[18] Oracle Corporation. Sun QFS software. http://www.oracle.com/us/products/servers-storage/storage/storage-software/031712.htm. Accessed on September 15, 2011.

[19] Orange file system. http://www.orangefs.org/. Accessed on September 15, 2011.

[20] Peter Kogge, et al. Exascale computing study: Technology challenges in achieving exascale systems. Technical Report TR-2008-13, University of Notre Dame, September 2008.

[21] Red Sky hardware environment. Sandia internal resource. Archived copy available upon request.

[22] E. Seo, S. Park, and B. Urgaonkar. An empirical analysis of the energy efficiency of flash-based SSDs. In *Proceedings of the First Workshop on Power-Aware Computing and Systems (HOTPOWER 2008), co-located with OSDI 2008*, December 2008.

[23] R. Stevens and A. White. A DOE laboratory plan for providing exascale applications and technologies for critical DOE mission needs. http://computing.ornl.gov/workshops/scidac2010/presentations/r_stevens.pdf, July 2010. SciDAC Workshop.

[24] B. Tomlinson, J. Cerutti, and R. A. Ballance. Cielo computational environment usage model. Technical Report LA-UR 10-07492, Los Alamos National Laboratory, June 2011. http://www.lanl.gov/orgs/hpc/cielo/docs/CieloUsageModel.pdf.

[25] Top500 list. http://www.top500.org.