Characterizing Energy Efficiency in I/O System for Scientific Applications

Javier Panadero, Sandra Méndez, Dolores Rexachs and Emilio Luque
Computer Architecture and Operating Systems Department (CAOS)
Universitat Autònoma de Barcelona
Barcelona, Spain
javier.panadero@campus.uab.es
{sandra.mendez, dolores.rexachs, emilio.luque}@uab.es

***Abstract*—The increasing complexity of scientific applications and the increase of scalability in high performance computing systems demand a more powerful Input/Output system. This requirement is present in both performance and power consumption. For this reason, performance, power consumption, energy and energy efficiency have become critical measures in Input/Output systems. Nowadays, when a High Performance Computing center buys a system of storage not only does it take into account the price, but also the cost of its useful life cycle as well as the energy cost. This paper proposes a methodology to characterize the energy efficiency of the Input/Output system, considering the Input/Output system at a device level, I/O library and file system. The methodology provides a wide range of I/O system parameters that have an impact on the energy efficiency. Furthermore, we evaluate the impact of Input/Output intensive applications in energy efficiency.**

Keywords-Energy Efficiency; Consumption; I/O System; HPC; Methodology.

I. Introduction

Energy efficiency has become an extremely important consideration for the storage system, due to several factors, among which the most important is the scalability of the system, because we will not be able to expand the system if we have consumed all the available energy [1]. Another important factor to consider is the cost of the Input/Output (I/O) system, due to the Kilowatt/hour rate imposed by the electricity company. It is because of this reason that nowadays, when a High Performance Computing center buys a system of storage, not only does it take into account the price, but also the cost of its useful life cycle.

These days, we can find similar rankings to the Top500 [2] such as the Green500 [3], where we can obtain a list of the supercomputers with the highest energy efficiency in computing. The Green500 updates its ranking 3 times a year in order to increase awareness about power consumption and energy. Furthermore, they promote alternative total costs and ensure that supercomputers do not only simulate the climate change but they do not help to its degradation. For I/O systems, it is not easy to find analysis for comparing energy efficiency.

In considering power consumption and energy efficiency, the scientific community has an extremely important challenge to overcome [4]. When we take into account both the energy efficiency and a determinate performance of the I/O system, it is important to have common sense for the configuration of the I/O system. First of all, we are going to consider, how to do an energy diagnosis; What can characterize of the I/O system; How can analyze power consumption and energy efficiency; What metric should we use. All these questions are necessary to be able to plan and to propose improvements of energy efficiency in the configurations of the I/O system. In this paper, we offer a methodology for characterizing energy efficiency in the I/O system. The proposed methodology relates application phases to power consumption phases throughout the execution time. This methodology considers the I/O system at device level, I/O library and global file system. On the other hand, it extracts information about throughput, power consumption, energy and energy efficiency in a system with different device access patterns. Considering this information, we analyze the impact of energy efficiency for the different configurations of the I/O system. The methodology allows us to characterize I/O scientific applications such as EarthScience, NuclearPhysics, CombustionPhysics, etc. This methodology serves as a starting point to be able to decide on the dimensioning of the I/O system that improves its energy efficiency. We also analyze the impact of the executed benchmarks in the characterization of the I/O system.

In this paper, we use a Watts UP pro ES digital power meter to take measurements and analysis for the I/O system. This meter provides a sampling each second.

This paper is organized as follows: in the next section we briefly review related work. In Section III, we introduce our methodology for characterizing energy efficiency in the I/O system. Then, in Section IV, we expose and analyze the experimental results. Finally, in Section V, we present the conclusions and future work.

II. Related Work

The work that we introduce in this paper is related to the analysis and characterization of the I/O system.

Ge [4] proposes a methodology to profile the performance, energy and energy efficiency considering the parallel I/O access patterns and the CPU frequency. This study differs from our work since we do not just consider the I/O access patterns.

 ISBN: 978-1-61208-172-4

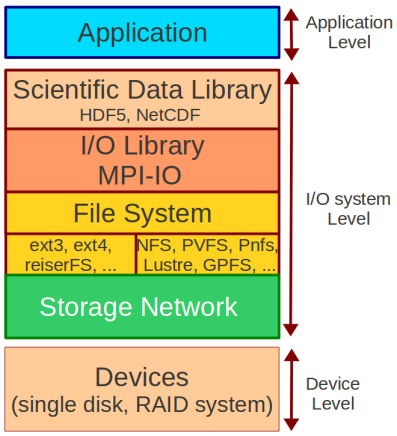


Fig. 1. I/O system

Where?	Benchmark	System characterization	Devices characterization
What?	·Access Patterns ·Type of operations (write and read) ·Size request ·Number of processes ·I/O library ·Phases of the benchmark	Bandwidth and power on: ·I/O library ·Filesystem Local and Global (parallel, distributed and network) ·Connection Storage (NAS, DAS, SAN, NASD) ·Buffer/Cache ·Interconnection Network	·Bandwidth and power on: •Devices •RAID level •Disk Cache ·Energy control ·Rotation Disk
How?	·Benchmarks Configuration ·Benchmark Documentation ·Trace of benchmark	·Filesystem Benchmark (Iozone, Bonnie++) ·I/O library benchmark (IOR, PIOBench)	·Disk benchmark (Iozone, Bonnie++) ·Tools system linux (hdparm, vmstat, iostat)
Measurements?		·Performance: Bandwidth (MB/sec) ·Power: watts (W) ·Energy: Joules (J) ·Energy efficiency (MB/J)	·Performance: Bandwidth (MB/sec) ·Power: watts (W) ·Energy: Joules (J) ·Energy efficiency (MB/J)

Fig. 2. Characterization of the I/O system

On the other hand, Hylick [5] proposes an analysis of power consumption at device level, considering the dependence of the locality of the data and the block size of access. Our study differs from this one since we do not consider just the I/O at the device level but also the I/O at the system level.

Another study proposed by Sehgal [6] analyzes the energy and energy efficiency by considering several Linux local file systems modifying the default options. Our work differs from his study because we do not consider modifying the default options but instead we consider different file system levels (local, distributed and parallel).

In terms of massive storage, Dong [7] proposes an analysis about power characteristics of read/write operations compared with the power efficiency of different RAID levels. Our work considers RAID levels as a part of the characterization of device level.

III. Proposed Methodology

In this section, we detail our methodology for the characterization of the energy efficiency in terms of performance, power, energy and energy efficiency.

We characterize at device level and at I/O system level. Fig. 1 illustrates the I/O system; we characterize the elements that include the device level and the I/O system level. The methodology is divided into four parts:

- Metrics selection used during the characterization,
- Characterization at device level,
- Characterization at I/O system level and
- Characterization of the benchmark parameters.

Fig. 2 illustrates these parts and, in addition, the information obtained in each part of the characterization.

A. Metrics used in the methodology

In this section, we detail the metrics used in the methodology for: performance, power, energy and energy efficiency.

The performance of the I/O operations is normally quantified using throughput (number of megabytes transferred per second) and/or IOPs (Input/Output Operations Per Seconds). The throughput (MB/sec) is obtained directly by the I/O benchmarks or Linux I/O tools and IOPs are obtained indirectly analyzing the bandwidth and the duration time of I/O operations.

We use the watt (W) as metrics for power and the Joule (J) for the energy, both included in the international unit system. The power meter obtains Watts directly and energy is derived indirectly from the total execution time of the benchmark per average consumption power of the benchmark execution.

There is not a standard measurement for energy efficiency. Due to the interrelation between performance and energy, Liu [8] proposes two new metrics for energy efficiency: IOPS/Watt and MBPS/Kilowatt. For this study we have chosen MBPS/watt as efficiency metrics. We use the equation mentioned above introducing the energy. For this reason, we use the equation (1) and we obtain MB/J as the final equation for the energy efficiency.

$$1\ Joule = 1\ Watt * 1\ Second \quad (1)$$

B. Characterization at the device level

This phase consists of the characterization of the devices and RAID system using I/O benchmark as Iozone [9] or/and Bonnie++ [10]. These benchmarks generate and measure a wide variety of file disk operations. During the characterization, we consider the access patterns (sequential, random, stride), the request size of operations and the type of access (block or character). For these operations, we characterize the bandwidth, the power consumption, the energy and the energy efficiency. We also consider the device's different states of power consumption.

C. Characterization of the I/O system level

This phase consists of the characterization of the I/O library, the file system (local, distributed and parallel), storage connection (NAS, SAN, DAS, NASD) and the system buffer cache. We obtain the bandwidth, the power consumption, the energy and the energy efficiency. In order to characterize the I/O system level, one could use the IOzone or/and Bonnie++

file system benchmarks. IOR [11] or/and PIO-bench [12] could be used as I/O library benchmarks. These benchmarks leverage the scalability of MPI to accurately calculate the throughput of a given number of client machines. As we have already illustrated in figure 1, this characterization is linked with the characterization at the device level because the data follows a process until it is written or read in the final device (single device or RAID system). For this reason, at the same time we carried it out at the I/O system level and at the device level.

D. Benchmark parameters characterization

To characterize the different levels, we use I/O benchmarks. These benchmarks have many configuration parameters, the access patterns (sequential, random, stride), the request size of operations, the type of access (block or character), the number of processors, the type of I/O library, amongst others.

The objective is to tune the specific parameters of the benchmarks, according to the characterization that we are doing.

E. Characterization of the I/O system

We carried out our characterization on two different systems. The first system characterized was a Pentium 4 single core, with a 512 MB RAM memory and a single device Seagate Barracuda ATA ST340016A. It also has a capacity of 80GB and a cache disk of 2MB. Fig. 3 illustrates the power consumption specifications of each state and also the transitions of the device. The local file system used was Ext4 with a DAS store network. The I/O library used was MPICH. The second system characterized was the cluster Aohyper. This cluster has 8 dual Core nodes AMD Athlon(tm) 64 X2, 2 GB RAM memory, 150 GB local disk. The local file system used was Ext4. Also has a 1 NFS server with RAID 1 (2 disks) with 230GB capacity and RAID5 (3 disks) with stripe=256KB and 917GB capacity, both with write-cache enabled (write back). The networks used were two Gigabit Ethernet network, one for communication and one for data.

We utilized the Watts UP pro digital power meter to measure. It was connected to the output AC power source of the computer. This meter provides a sampling each real-time second.

These two systems have been characterized. Now, we present how we characterized the I/O system. Although we expose the workstation characterization, this characterization is extensible for the cluster characterization or another system. We describe the workstation characterization, however it could be extended for the cluster characterization or any other system.

1) The device characterization: What we did, first of all, was to characterize the effects of power-saving using state controls. In order to do that, we used the benchmark IOR, which was executed twice with an interval of 60 seconds between each execution. Fig. 4 illustrates the result of the execution with power-saving using state controls and without power-saving. Fig. 5 shows the power consumption and the energy required for the two executions. The result of execution

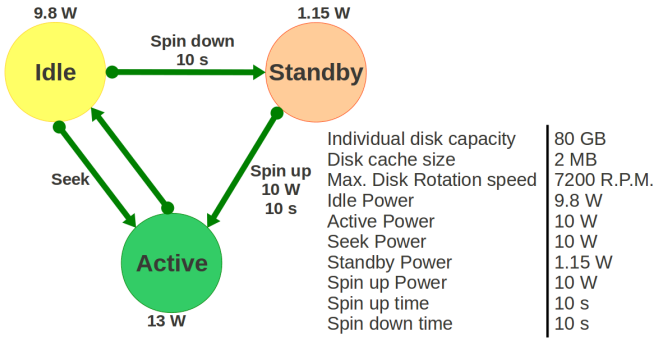


Fig. 3. Specifications of power consumption

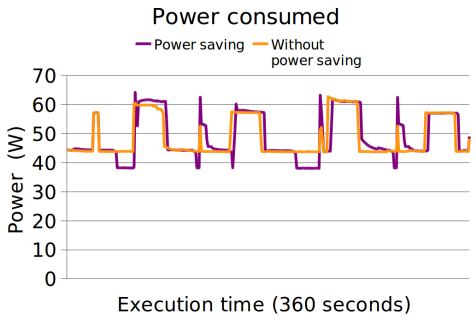


Fig. 4. Execution both power saving and without power saving

is better without power-saving states than with power-saving states. This is due to several factors such as: short Standby periods, the cost of spin-up transactions (time and power) and the peaks obtained during the spin-up transactions. Applications with short standby periods do not take advantage of power saving states.

After that, we characterized different access patterns (sequential and random). In order to do that, we used the benchmark IOzone with different requests sizes and a file size of 1GB. Fig. 6 illustrates the result of the characterization for sequential access patterns. Finally, Fig. 7 illustrates the characterization for random access patterns. We observe the following trends if we execute the sequential characterization and the set of request sizes tested in read operations. In the

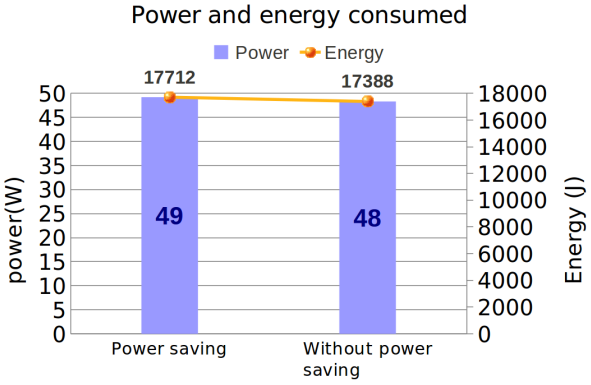
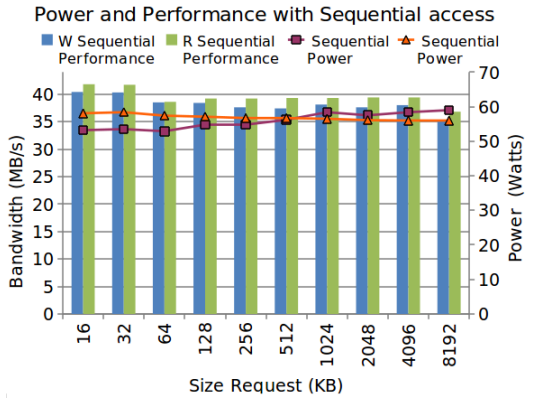
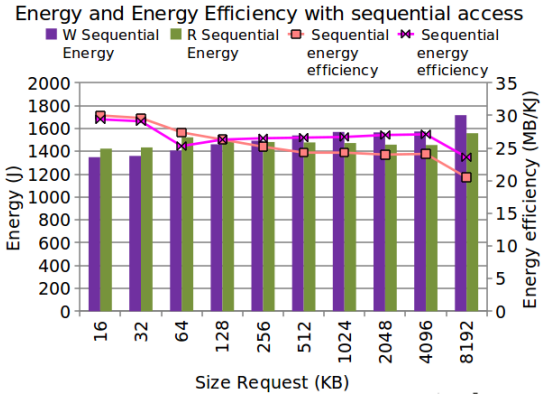


Fig. 5. Power and energy for execution with power saving

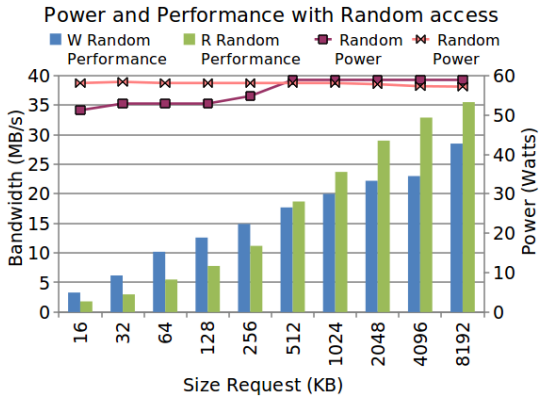


(a) Bandwidth and Power

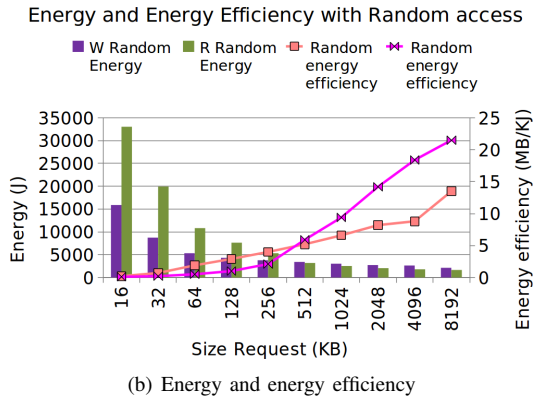


(b) Energy and energy efficiency

Fig. 6. Characterization for sequential access patterns



(a) Bandwidth and Power



(b) Energy and energy efficiency

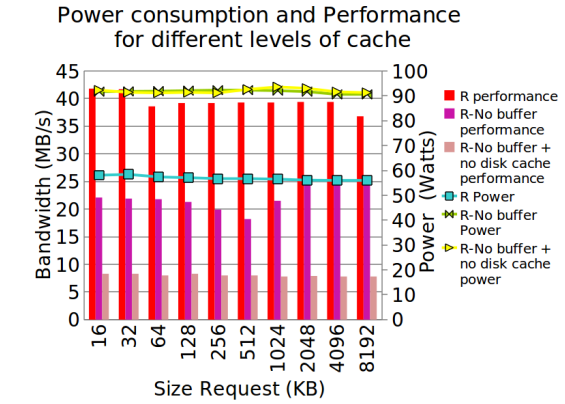
Fig. 7. Characterization for random access patterns

case of a very small-sized request (16 KB - 32 KB), we obtain a larger bandwidth, more power, less energy and an increase in energy efficiency. On the other hand, in the case of a request size of 8192KB, we obtain the worst bandwidth and the lowest power of the whole set of request size. The result is larger energy consumption and the worst energy efficiency. In write operations, in terms of bandwidth, energy and energy efficiency show the same trends. However, in terms of power, by increasing the request size, the power increases. Despite the small power variation that we observe, it has an influence in the metrics of energy and energy efficiency. We observe the following trends if we execute the random characterization and the set of request size tested in read and write operations. In the case of a larger request size, we obtain a larger bandwidth, less energy and more energy efficiency. Because of the data locality, as we increase the request size, data transferring time is longer than data seeking time. Power consumption has two different trends depending on the operation type. In write operations, if we increase the request size we obtain more power consumption. On the other hand, in read operations, if we increase the request size the lowest power consumption is obtained.

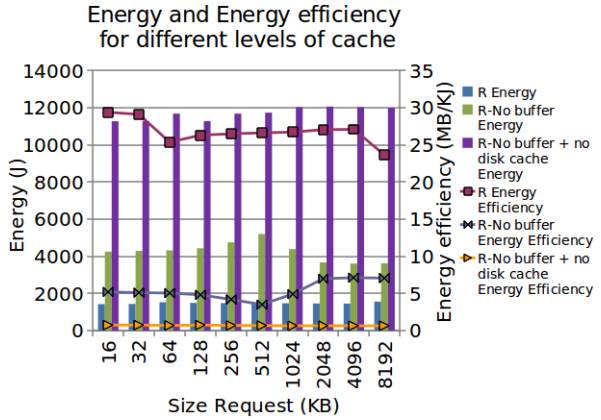
2) I/O system characterization: At I/O system level, we characterized the influence of different cache levels on the system. It is worth mentioning that we have included the cache disk as part of the system cache hierarchy. We call the buffer memory of the operating system buffer cache. For that reason, we used the benchmark IOzone with different request sizes and size file of 1GB. Fig. 8 illustrates the result of execution with different levels of cache disabled. In this figure, we observe the following trends: bandwidth and energy efficiency decrease whereas power and energy increase as we disable levels of cache. There are no substantial differences between the power consumption without buffer cache disabled and disk cache enabled; and there is no difference either between power consumption without buffer cache disabled and disk cache disabled. On the contrary, there are differences in energy and energy efficiency for those configurations, because of the influence of the bandwidth.

Moreover, we also characterized the influence of the I/O library. To achieve these study objectives, we characterized the MPICH library. In order to do that, we used the benchmark IOR for 1 core, with different request sizes and a file size of 1GB. The goal was to observe the influence in energy efficiency with the insertion of the new layer in the I/O stack. In Fig. 9, we observe the following trends: bandwidth and energy efficiency decrease whereas energy increases for a larger request size. Power consumption in write operations increases until a request size of 512 KB and then, it begins to decrease. In read operations, in the case of request size (16KB – 512 KB), we obtain the same trend in relation to power. It

 ISBN: 978-1-61208-172-4

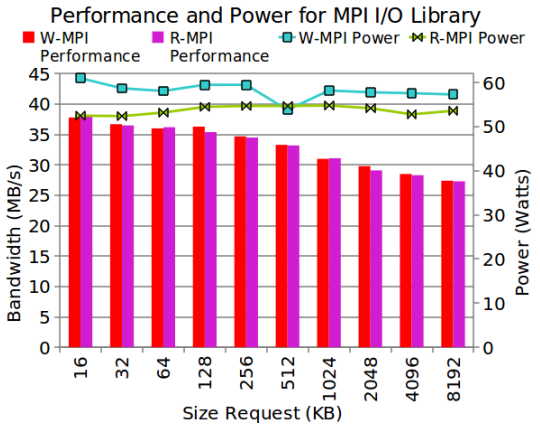


(a) Bandwidth and Power

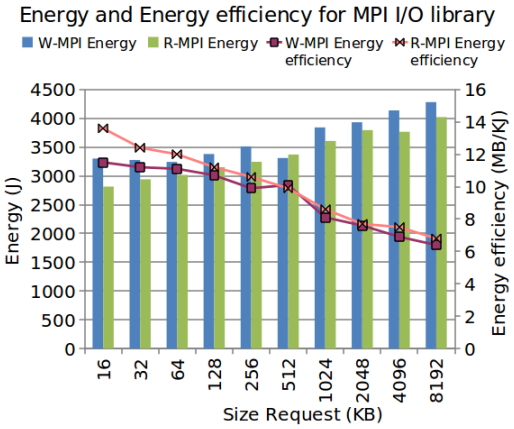


(b) Energy and energy efficiency

Fig. 8. Execution with different levels of cache disabled



(a) Bandwidth and Power



(b) Energy and energy efficiency

Fig. 9. Characterizating the influence of the I/O library

begins to decrease from a 512 KB request size on. It also influences energy efficiency.

IV. Experimentation

In order to evaluate the methodology for the characterization, we selected the Block Tridiagonal (BT) application of NAS Parallel Benchmark (NPB) suite. We executed this application in the systems that were previously characterized. On the workstation, the application was executed in its class B and subtype full. That configuration writes a file size of 1,5 GB. Whereas for the execution on the cluster, we executed the application in its class C and subtype full for 16 processors. That configuration writes a file size of 6,5 GB.

Fig. 10 shows the trace of the application for the workstation. The violet color represents write operations and the green color represents read operations. We observe 3 different phases; involves the computing and discontinuing write operations of 128 KB request size. The second phase is I/O intensive in write operations, whereas the last phase is I/O intensive in read operations.

Fig. 11 illustrates power consumption during the application execution. After an initial time when the state of the device was idle, we executed the application. We observed 2 distinct phases; in the first phase, we observed more consumption than in the second phase. This is due to application compute

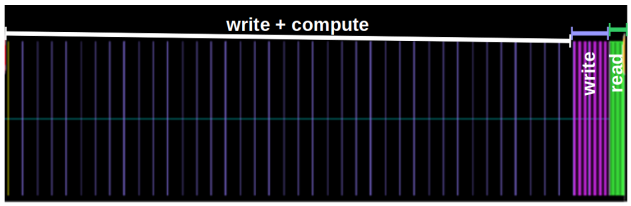


Fig. 10. Trace of the application NAS BT class B subtype FULL

and write discontinuous operations that took place during the first phase. Whereas in the second phase only intensive I/O operations without compute were made. During the first phase, we observed peaks of consumption, which are caused by the addition in write operations of power to compute power.

Fig. 12 shows the trace of the application for the cluster. The application was executed with two different RAID configura-

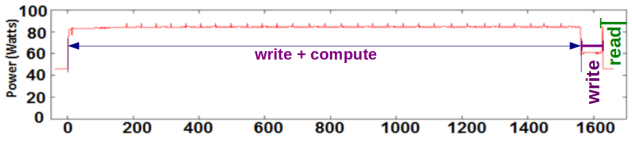


Fig. 11. Power of the execution application BT class C subtype FULL

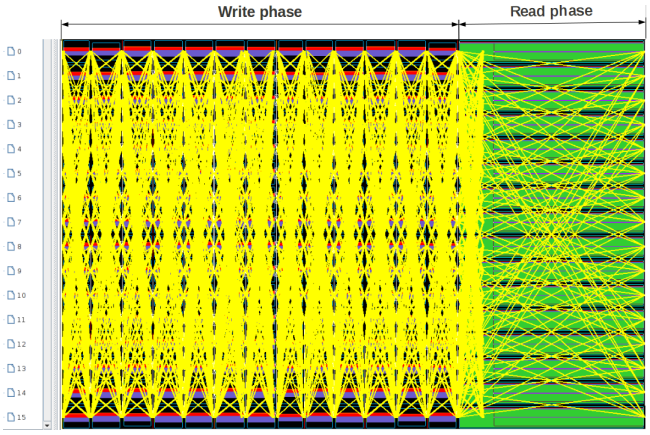
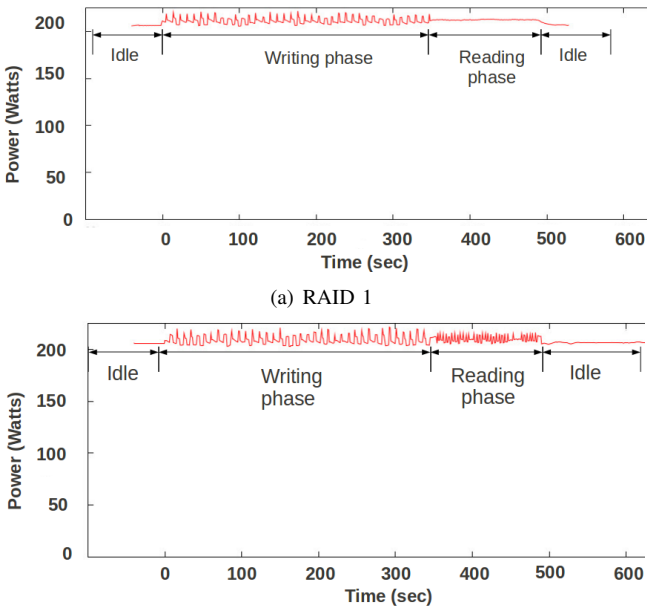


Fig. 12. I/O Phases of NAS BT-IO by 16 processes FULL subtype



(a) RAID 1

(b) RAID 5

Fig. 13. Power consumed for NAS BT-IO by 16 processes

tions (RAID 1 and RAID 5). The yellow color represents write operations and the green color represents read operations. We observe 2 different phases; the first phase is I/O intensive in write operations of 128 KB request size, whereas the second phase is I/O intensive in read operations.

Fig. 13(a) illustrates the power consumption during the application execution on the cluster for a RAID 1 configuration, whereas the Fig. 13(b) illustrates the power consumption during the application execution on the cluster for a RAID 5 configuration.

After an initial time where the state of the devices was idle, we observed 2 distinct phases for the two configurations. The first configuration phase is the power consumption for I/O intensive write operations; whereas the second phase is the power consumption for I/O intensive read operations. We observed that because the I/O system is apart from the compute

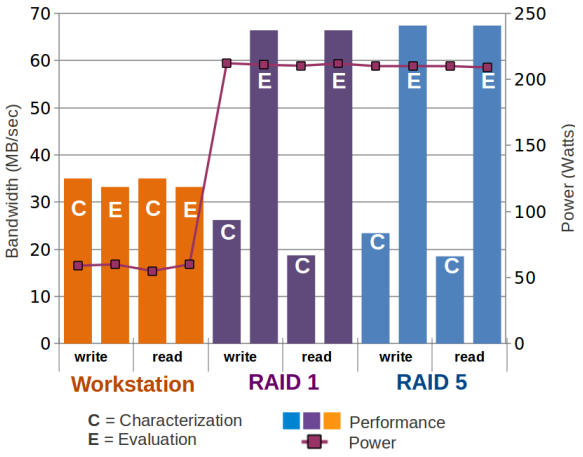


Fig. 14. Real and references values obtained

TABLE I
PERCENTAGE (%) DEVIATION OBTAINED

Operation type	write	read
WorkStation	3%	7%
RAID 1	0.6%	0.8%
RAID 5	0%	0.5%

nodes, the power consumption values obtained are only of I/O operations. Because of this, the intensive computation does not have in the results.

Fig. 14 illustrates real values obtained during the application execution and reference values obtained during the characterization for all systems characterized. We show the performance and the power consumption. In terms of workstation, we selected real values of performance and power consumption obtained during the execution of the application's second phase because it is the I/O intensive phase. We selected the reference values results of a 128Kb size request obtained in Fig. 9, because the BT application is a parallel MPI application that uses a 128 KB request size in I/O operations.

In terms of cluster values, because the I/O system is apart from the compute nodes, we selected the real values obtained in each phase. We selected the reference values from the cluster characterization in the same way as we did before for the workstation architecture. We observe close results in power consumption in all characterized systems. However, in bandwidth, we observe significant differences for cluster configurations. These differences are because the application did not manage to stress the system.

Table I, shows the power consumption deviation of reference values with real values. We observe close results between real and reference values.

V. CONCLUSION

In this paper, we propose a new methodology to characterize the energy efficiency in the I/O system. The methodology takes into account performance and energy. Moreover, it extracts information about the bandwidth, the power consumption, the energy and the energy efficiency from different I/O benchmarks. We evaluated the methodology with real applications

 ISBN: 978-1-61208-172-4

and we observed that the reference values of characterization were close to the real values obtained with the application's execution. We also observed that in intensive operations of the I/O system, power consumption changed to a small extent. However, that change did modify energy and energy efficiency.

This paper is just a small part of our research and will serve to find new ways of investigating. Our final goal will be to propose a methodology for dimensioning the I/O system in terms of energy and energy efficiency. That methodology will be able to characterize, analyze and evaluate the I/O system for dimensioning. We are also looking for a new method to identify the significant phases in terms of power consumption at device level considering the writing and reading blocks. In order to carry this out, we are also working in new metrics for energy efficiency.

ACKNOWLEDGMENT

This research has been supported by MICINN-Spain under contract TIN2007-64974.

REFERENCES

[1] T. Minartz, J. Kunkel, and T. Ludwig, "Simulation of power consumption of energy efficient cluster hardware," *Computer Science - Research and Development*, pp. 165–175, 2010. [Online]. Available: http://www.springerlink.com/content/r21r2376730p7161/fulltext.pdf

[2] T. 500, "Top 500 supercomputing list," Tech. Rep., [Retrieved: September, 2011]. [Online]. Available: http://www.top500.org

[3] G. 500, "Green computing list," Tech. Rep., [Retrieved: September, 2011]. [Online]. Available: http://www.green500.org

[4] R. Ge, X. Feng, S. Subramanya, and X. he Sun, "Characterizing energy efficiency of i/o intensive parallel applications on power-aware clusters," in *Parallel Distributed Processing, Workshops and Phd Forum (IPDPSW), 2010 IEEE International Symposium on*, april 2010, pp. 1 –8.

[5] A. Hylick, R. Sohan, A. Rice, and B. Jones, "An analysis of hard drive energy consumption," in *Modeling, Analysis and Simulation of Computers and Telecommunication Systems, 2008. MASCOTS 2008. IEEE International Symposium on*, sept. 2008, pp. 1 –10.

[6] P. Sehgal, V. Tarasov, and E. Zadok, "Optimizing energy and performance for server-class file system workloads," *Trans. Storage*, vol. 6, pp. 10:1–10:31, September 2010. [Online]. Available: http://doi.acm.org/10.1145/1837915.1837918

[7] Y. Dong, J. Chen, and T. Tang, "Power measurements and analyses of massive object storage system," in *Computer and Information Technology (CIT), 2010 IEEE 10th International Conference on*, 29 2010-july 1 2010, pp. 1317 –1322.

[8] Z. Liu, F. Wu, X. Qin, C. Xie, J. Zhou, and J. Wang, "Tracer: A trace replay tool to evaluate energy-efficiency of mass storage systems," in *Proceedings of the 2010 IEEE International Conference on Cluster Computing*, ser. CLUSTER '10, 2010, pp. 68–77, [Retrieved: June, 2011]. [Online]. Available: http://dx.doi.org/10.1109/CLUSTER.2010.40

[9] W. D. Norcott, "Iozone filesystem benchmark," Tech. Rep., [Retrieved: September, 2011]. [Online]. Available: http://www.iozone.org/

[10] R. Coker, "Bonnie++ filesystem benchmark," Tech. Rep., [Retrieved: September, 2011]. [Online]. Available: http://www.coker.com.au/bonnie++/

[11] . S. J. Shan, Hongzhang, "Using ior to analyze the i/o performance for hpc platforms," LBNL Paper LBNL-62647, Tech. Rep., [Retrieved: September, 2011]. [Online]. Available: www.osti.gov/bridge/servlets/purl/923356-15FxGK/

[12] F. Shorter, *Design and analysis of a performance evaluation standard for parallel file systems*, [Retrieved: September, 2011]. [Online]. Available: http://books.google.com/books?id=g7sEOAAACAAJ

 ISBN: 978-1-61208-172-4