Efficiency Evaluation of the Input/Output System on Computer Clusters*

Sandra Méndez, Dolores Rexachs and Emilio Luque

Computer Architecture and Operating System Department (CAOS)
Universitat Autònoma de Barcelona, Barcelona, Spain

{sandra.mendez,dolores.rexachs,emilio.luque}@uab.es

Abstract. The increase of the complexity of scientific applications that use high performance computing require more efficient Input/Output (I/O) systems. In order to efficiently use the I/O is necessary to know its performance capacity to determine if it fulfills applications I/O requirements. This paper proposes the efficency evaluation of the I/O system on computer clusters. This evaluation is useful to study how different I/O system will affect the application performance. This approach encompasses the characterization of the computer cluster at three different levels: devices, I/O system and application. We select different system and we evaluate the impact on performance by considering both the application and the I/O architecture. During I/O configuration analysis we identify configurable factors that impact the performance of I/O system. Furthermore, we extract information in order to determine the used percentage of I/O system by an application on a given computer cluster.

Keywords: Parallel I/O System, I/O Architecture, I/O Configuration, I/O Path Level, I/O inefficiency.

1 Introduction

The increase of processing units, the advance in speed and compute power, and the increasing complexity of scientific applications that use high performance computing require more efficient I/O systems. Due to the historical "gap" between the computing performance and I/O performance, in many cases, the I/O system becomes the bottleneck of the parallel systems. The efficient use of the I/O system and the identification of I/O factors that influence the performance can help to hide this "gap". To efficiently use the I/O system it is first necessary to know its performance capacity to determine if it fulfills the application I/O requirements.

There are several papers on performance evaluation of I/O system. Roth [1] presented event tracing for characterizing the I/O demands of applications on the Jaguar Cray XT of supercomputer. Fahey [2] experimented in the I/O system of the Cray XT, and the analysis was focused in the LUSTRE filesystem.

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

Laros [3] made a performance evaluation of I/O configuration. Previous papers do not consider directly the I/O characteristics of applications.

We propose the efficiency evaluation of the I/O system by analyzing each level on the I/O path. Furthermore, we taking into account the application I/O requirements and the I/O architecture configuration. The proposed methodology has three phases: characterization, the analysis of I/O system, and the efficiency evaluation. In the application's characterization phase, we extract the I/O requirements of the application. In the I/O system characterization we obtained the bandwidth and IOPs (I/O operations per second) at filesystem level, interconnection network, I/O library and I/O devices. Furthermore, we identify configurable or selectable factors that have an impact the I/O system performance. We search this factors in the filesystem level, I/O node connection, placement and state of buffer/cache, data redundancy and service redudancy. We collect metrics of the application execution on I/O configurations In the evaluation phase, the efficiency is determined analyzing the difference between measured values and characterized values.

The rest of this article is organized as follows: Section 2 introduces our proposed methodology. In Section 3 we review the experimental validation of this proposal. Finally, in the Section 4, we present conclusions and future work.

2 Proposed Methodology

The I/O in the computer cluster occurs on a hierarchal I/O path. We see I/O system as show in Fig. 2(a). The application carries out the I/O operations in this hierarchical I/O path. The I/O path levels are: I/O library (high and low level), filesystem (local and global), network (I/O o shared with computing), and I/O devices. Although, the placement of the filesystem and interconnection network can vary by depending of the I/O system configuration. The application also can use I/O libraries of high (NetCDF, HDF5) or low level (MPI-IO). In order to evaluate the I/O system performance is necessary to know its capacity of storage and throughput. The storage depend of amount, type and capacity of devices. The throughput depend of IOPs (Input/Output operations per second) and the latency. Moreover, this capacity is diferent in each I/O system level. Furthermore, the performance depend on the connection of the I/O node, the management of I/O devices, placement of I/O node into network topology, buffer/cache state and placement, and availability data and service. In order to determine whether an application uses I/O system capacity is necessary know its I/O behavior and requirements. The methodology is shown in Fig. 1. This is used to evaluate the efficiency of I/O system and identify the possible points of inefficiency. The efficiency is based in the used performance percentage by the application on each I/O path level. Also, when the cluster have different selectable or configurable parameters, the methodology is used to analyze which I/O configuration is the more appropiate for an application.

2.1 Characterization

This is applied to obtain the capacity and performance of I/O system. We also obtain I/O requirements and behavior of the application. Here we explain the system characterization and the scientific application charaterization.

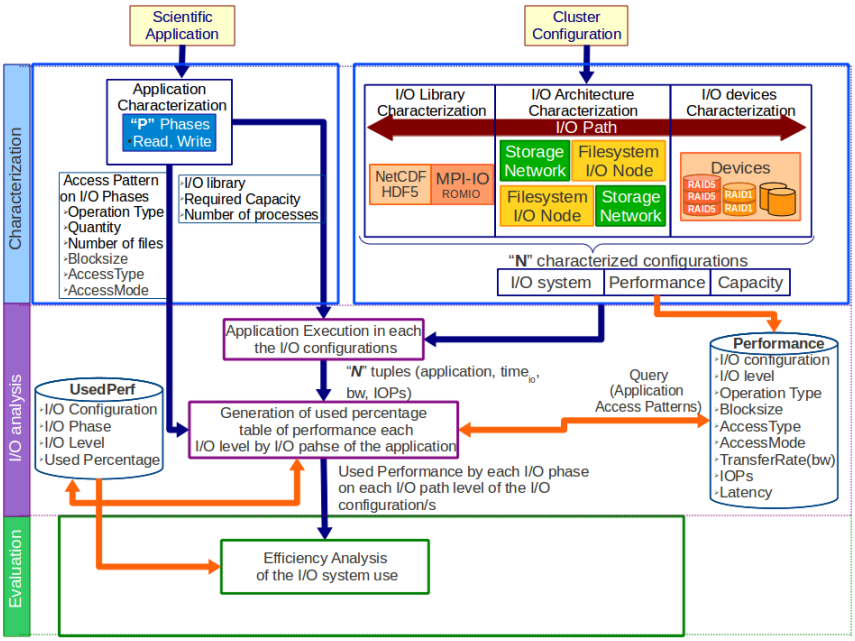


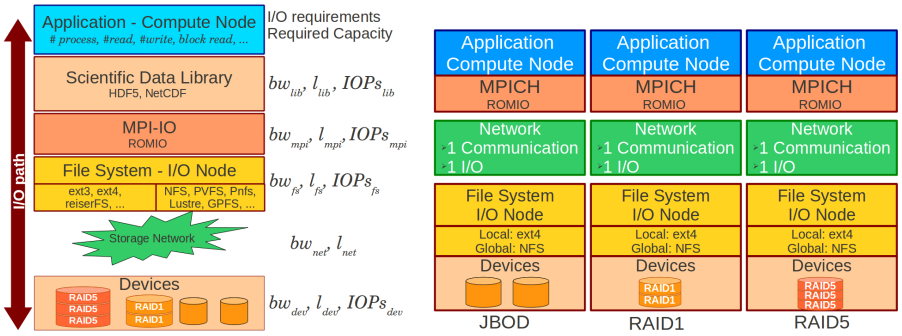
Fig. 1. Methodology for Efficiency Evaluation on I/O System

I/O System and Devices Parallel system is characterized at I/O library level, I/O Node (global filesystem and interconnection system) and devices (local filesystem). We characterize the bandwidth (bw), latency (l) and ($IOPs$) for each level, as shown in Fig. 2(a). Fig. 3(a) shows "what" and "how" we obtain this information for the I/O system and Devices. Furthermore, we obtain characterized configurations in each I/O path level. The data structure of I/O system performance for local and global filesystem, and I/O library following is shown:

- Operation Type (enumerate {0 (read), 1 (write)})
- Block size (double (MBytes))
- Access Type (enumerate {0 (Local), 1 (Global)})
- Accesses Mode (enumerate {0 (Sequential), 1 (Strided), 2 (Random)})
- transfer Rate (double (MBytes/second))
- Latency (double (microsecond))
- IOPs (integer)

To evaluate global filesystem and local filesystem, IOzone [4] and/or bonnie++ [5] benchmarks can be used. Parallel filesystem can be evaluated with the IOR benchmark [6]. The b_eff_io [7] or IOR benchmarks can be used to evaluate the I/O library. To explain this phase we present the characterization for the I/O system of cluster Aohyper.

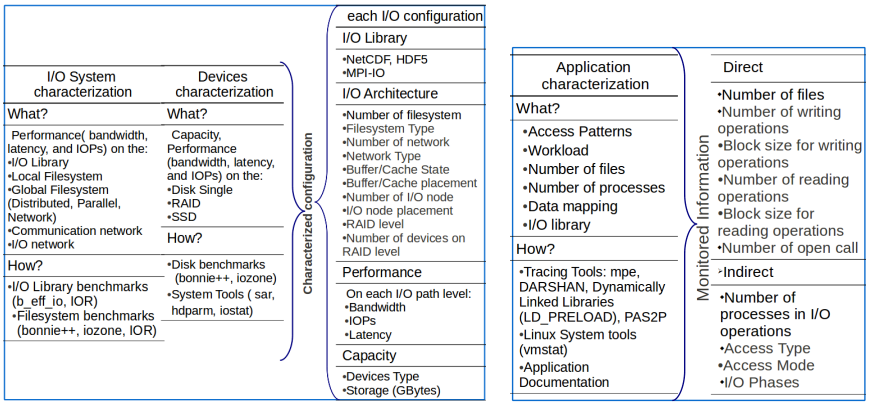
Cluster Aohyper has the following characteristics: 8 nodes AMD Athlon(tm) 64 X2 Dual Core Processor 3800+, 2GB RAM memory, 150GB local disk. Local filesystem is linux ext4 and global filesystem is NFS. The NFS server has a



(a) I/O system Characterization (b) I/O configurations of the Aohyper cluster

Fig. 2. Characterization of I/O System

RAID 1 (2 disks) with 230GB capacity and RAID 5 (5 disks) with stripe=256KB and 917GB capacity, both with write-cache enabled (write back); two Gigabit Ethernet network, one for communication and the other for data. NFS server is

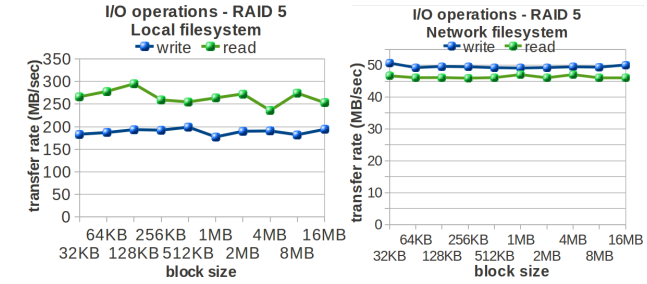


(a) I/O system and Devices (b) Application

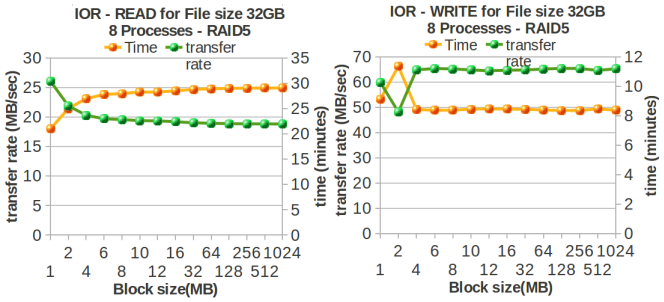
Fig. 3. Characterization Phase

an I/O node for shared accesses. Also, there are eight I/O-compute nodes for local accesses and the data sharing must be done by the user. Cluster Aohyper, at device level, has three I/O configurations (Fig. 2(b)). JBOD configuration are single disks without redundancy. RAID 1 configuration has a disk with its mirror disk and RAID 5 has five active disks. The parallel system and storage devices characterization were done with IOzone. Due to space, we show only the characterization of the RAID 5 configuration. Fig. 4 shows results for network filesystem, local filesystem, and I/O lybrary for RAID 5. The experiments were performed at block level with a file size which doubles the main memory size and block size was changed from 32KB to 16MB. The IOR benchmark was used to analyze the I/O library. It was configured for 32GB size of file on RAID configurations and 12 GB on JBOD, from 1MB to 1024MB block size and transfer block size of 256KB. It was launched with 8 processes.

Scientific Application We extract the type, quantity and operations size of I/O at library level. Fig. 3(b) shows "what", "how", and the monitored infor-



(a) Local filesystem and Network filesystem



(b) I/O Library

Fig. 4. Characterization of RAID 5 Configuration

mation of the application. This information is used in the evaluation phase to determine whether application performance is limited by the application characteristics or by the I/O system. To evaluate the application characterization at process level, an extension of PAS2P [8] tracing tool was developed. We incorporate the I/O primitives of MPI-2 standard to PAS2P. These are detected when the application is executed. To do we used dynamic link with LD_PRELOAD. With the characterization, we propose, to identify the significant phases with an access pattern and their weights. Due that scientific applications show a repetitive behavior, P phases will exist in the application.

To explain the methodology, the characterization is applied to Block Tridiagonal(BT) application of NAS Parallel Benchmark suite (NPB)[9]. The BTIO benchmark performs large collective MPI-IO writes, and reads of a nested strided datatype, and it is an important test of the performance a system can provide for noncontiguous workloads. After every five time steps the entire solution field, consisting of five double-precision words per mesh point, must be written to one or more files. After all time steps are finished, all data belonging to a single time step must be stored in the same file, and must be sorted by vector component, x-coordinate, y-coordinate, and z-coordinate, respectively.

NAS BT-IO full subtype has 40 phases to write and 1 phase to read. Writing operation is done each 120 message sent with their respective Wait and Wait_All. The reading phase consists of 40 reading operations done after all writing procedures are finished. This is done for each MPI process. Simple subtype has the same phases but each writing phase does 6,561 writing operations. The reading

phase consists of 262,440 reading operations. The characterization done for the class C of NAS BT-IO in full and simple subtypes is shown in TABLE 1.

Table 1. NAS BT-IO Characterization - Class C - 16 and 64 processes

Parameters	**full 16p**	**simple 16p**	**full 64p**	**simple 64p**
$numFiles$	1	1	1	1
$numIO_{read}$	640	2,073,600 and 2,125,440	2560	8398080
$numIO_{write}$	640	2,073,600 and 2,125,440	2560	8398080
bk_{read}	10 MB	1.56KB and 1.6KB	2.54 MB	800 bytes and 840 bytes
bk_{write}	10 MB	1.56KB and 1.6KB	2.54 MB	800 bytes and 840 bytes
$numIO_{open}$	32	32	128	128
$accessType$	Global	Global	Global	Global
$accessMode$	Sequential	Sequential	Sequential	Sequential
$numProcesos$	16	16	64	64

2.2 Input/Output Analysis

The I/O configurations of cluster computer are composed of I/O library, I/O architecture and I/O devices. The I/O parameters configurables or selectable are shown in Fig. 3(a), they are labeled as "each I/O configurations". The selection of I/O configuration depends of I/O requeriments of the application and the user requirements. In order to select the configurations, we considered the I/O library, number of processes and the capacity required by the application. The RAID level will depend on what the user is willing to pay. For this article we have selected three configurations: JBOD, RAID 1 and RAID 5.

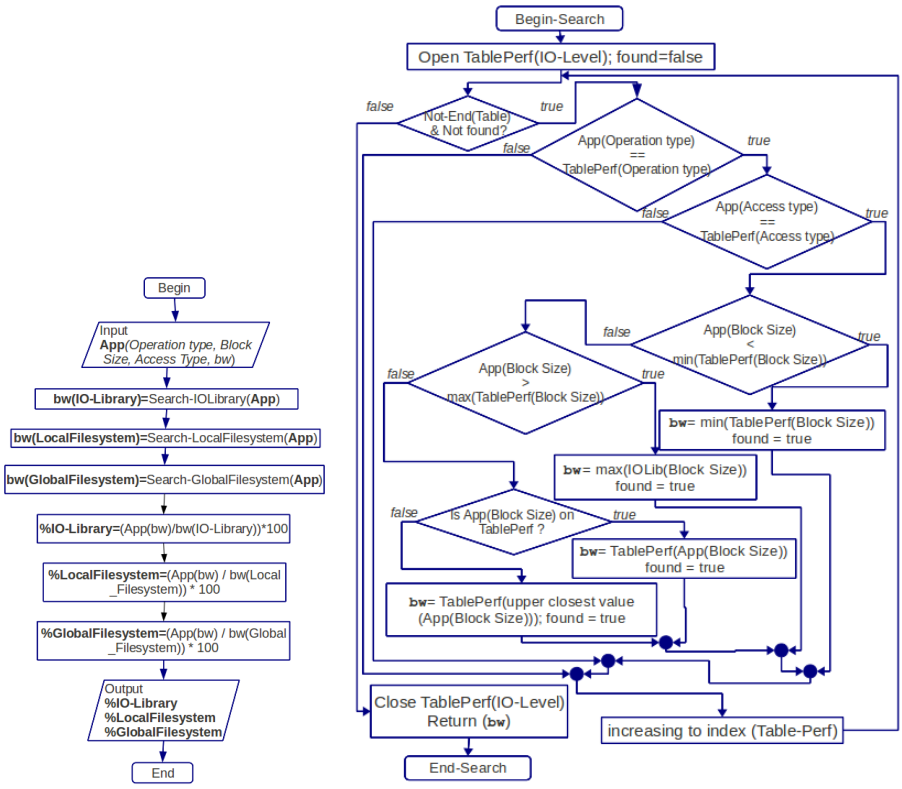
We extracted I/O behavior of application in the 1st phase, now we evaluate the application in selected configurations to view its behavior. The metrics for the application are: execution time, I/O time (time to do reading and writing operations), I/O operations per second (IOPs), latency of I/O operations and throughput (number of megabytes transferred per second). A file is generated with the used percentage by the application on the each I/O configurations, "P" I/O phases and I/O path levels (denoted by UsedPerf in Fig. 1 on the I/O analysis phase). The processes of generation of used percentage is presented in the Fig. 5(a). The algorithm to search the transfer rate on each I/O level is shown in the Fig. 5(b); and it is applied in each searching stage of Fig. 5(a).

2.3 Evaluation

We evaluate the use efficiency of I/O system based in the characterized values and measured values. The efficiency evaluation of the I/O system uses a file with the characterized values for the each I/O configurations of the computer cluster (denoted by Performance in Fig. 1).

Following with our example, we analyze NAS BT-IO on the Aohyper cluster. Fig. 6 shows the execution time, the I/O time and throughput for NAS BT-IO class C using 16 processes executed on the three configurations. The evaluation is for full (with collectives I/O) and simple (without collectives) subtypes. The used percentage of I/O system is shown in TABLE 2. The full subtype is an efficient

implementation for NAS BT-IO and we observe for the class C that the capacity of I/O system is exploited. But, for the simple subtype this I/O system is used only at about the 30% of performance on reading operations and less than 15% on writing operations. NAS BT-IO simple subtype carries out 4,199,040 writes and 4,199,040 reads with block sizes of 1,600 and 1,640 bytes (TABLE 1). This has a high penalization in the I/O time impacting on the execution time (Fig. 6). For this application in the full subtype the I/O is not factor bounding because the capacity of I/O system is sufficient for I/O requirements. The simple subtype does not achieve exploit of the I/O system capacity due to its access pattern.



(a) Generation Algorithm of used percentage

(b) Searching algorithm

Fig. 5. Generation used percentage

3 Experimentation

In order to test the methodology, an evaluation of NAS BT-IO for 16 and 64 processes in a different cluster was done, this cluster is called cluster A. Cluster A is composed of 32 compute nodes: 2 x Dual-Core Intel (R) Xeon (R) 3.00GHz,

Table 2. Percentage (%) of I/O system use for NAS BT-IO on I/O phases

I/O configuration	I/O Lib write	NFS write	Local FS write	I/O Lib read	NFS read	Local FS read	SUBTYPE
JBOD	101.47	117.70	78.00	309.74	127.93	60.00	FULL
RAID1	140.24	120.20	54.04	310.00	128.04	43.63	FULL
RAID5	88.60	115.18	29.69	303.11	125.20	22.76	FULL
JBOD	25.06	26.06	15.33	54.29	28.96	18.61	SIMPLE
RAID1	27.75	30.65	13.37	5448	31.98	12.68	SIMPLE
RAID5	24.60	29.52	8.07	56.77	31.40	5.55	SIMPLE

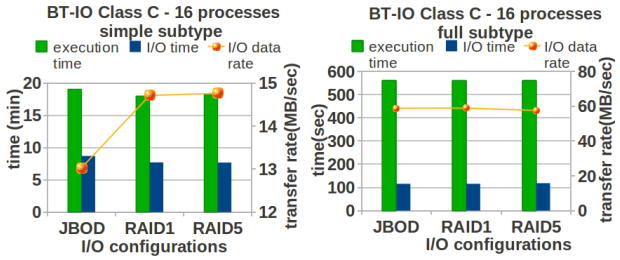
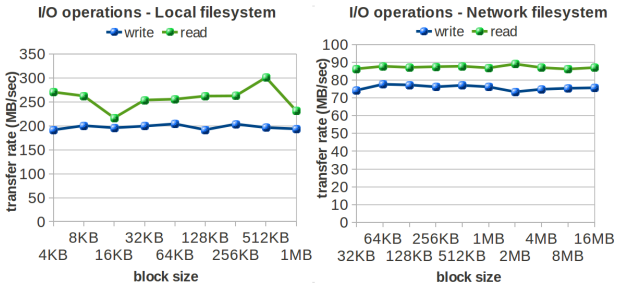


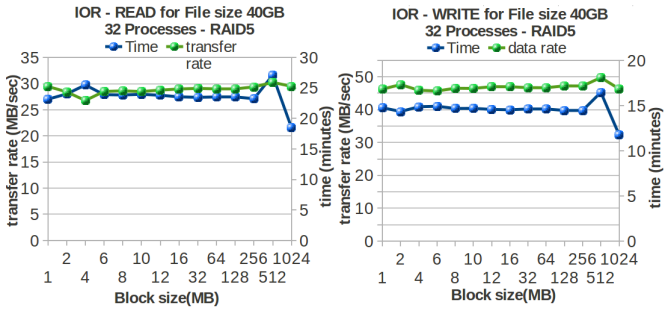
Fig. 6. NAS BT-IO Class C 16 Processes

12 GB of RAM, and 160 GB SATA disk Dual Gigabit Ethernet. A front-end node as NFS server: Dual-Core Intel (R) Xeon (R) 2.66GHz, 8 GB of RAM, 5 of 1.8 TB RAID and Dual Gigabit Ethernet. Cluster A has an I/O node that provides service to shared files by NFS and storage with RAID 5 level. Furthermore, there are thirty-two I/O nodes for local and independent accesses.

Due to the I/O characteristics of the cluster A, where there are no different I/O configurations, we used the methodology to efficiency evaluate of the I/O system for NAS BT-IO. Characterization of I/O system on cluster A is presented in Fig. 7. We evaluate the local and network filesystem with IOzone. Due to this cluster is restricted, the characterization in local file system was done by system administrators. IOR benchmark to evaluate the I/O library was done with 40 GB filesize, block size from 1 MB to 1024 MB, and 256 KB transfer block. The characterization for 16 and 64 processes is shown in TABLE 1. NAS BT-IO was executed for 16 and 64 to evaluate the use of the I/O system on cluster A. Fig. 8 shows the execution time, the I/O time and throughput for NAS BT-IO full and simple subtypes. TABLE 3 shows the used percentage on I/O library, NFS and Local filesystem. The full subtype is an efficient implementation that achieves more than 100% of the characterized performance on the I/O library for 16 and 64 processes. Although, with a greater number of processes, the I/O system influences on the run time of the application. NAS BT-IO full subtype is limited in the A cluster by computing and/or communication. NAS BT-IO full subtype does not achieve 50% of NFS characterized values and the I/O time is increased with larger number of processes, due to communication among processes and the I/O operations. NAS BT-IO simple subtype is limited by I/O for this A cluster I/O configuration. The I/O time is upper to 90% of run time. For this system



(a) Local filesystem and Network filesystem



(b) I/O Library

Fig. 7. Characterization of A Cluster Configuration

the I/O network and communication are bounding the application performance.

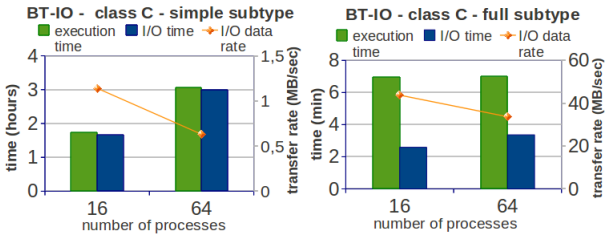


Fig. 8. NAS BT-IO Clase C - 16 and 64 processes

4 Conclusion

A methodology for efficiency evaluate of I/O system on computer clusters was shown. Such methodology encompasses the characterization of the I/O system at three different levels: devices, I/O system and application. We analyzed and evaluated different systems and we calculated the use by the application (% of use) of the I/O system on different I/O path levels. The methodology was applied in two different clusters for NAS BT-IO benchmark. The performance of both I/O systems were evaluated using benchmarks, and we characterized the application. Also, we show the use of the I/O systems done by NAS BT-IO, that has been evaluated on each I/O path level of the I/O configurations.

As future work, we are defining an I/O model of the application to support the evaluation, design and selection of configurations. This model is based on the characteristics of the application and I/O system, and it is being developed to determine which configuration of I/O meets the performance requirements of the user, taking into account the application I/O behavior in a given system. We will extract the functional behavior of the application, and we will define the I/O performance for the application given the functionality of application at I/O level. In order to test other configurations, we are analyzing the simulation framework SIMCAN [10] and planning to use such tool to model I/O architectures.

Table 3. Percentage (%) of I/O system use for NAS BT-IO on I/O phases

I/O configuration	I/O Lib write	NFS write	Local FS write	I/O Lib read	NFS read	Local FS read	SUBTYPE
16	70.74	43.39	16.27	112.21	36.16	13.56	FULL
64	80.26	49.76	18.66	128,69	41.47	15.55	FULL
16	2.45	1.58	0.57	3.86	1.28	0.45	SIMPLE
64	0.67	0.43	0.16	1.05	0.35	0.12	SIMPLE

References

1. P. C. Roth, "Characterizing the i/o behavior of scientific applications on the cray xt," in *PDSW '07: Procs of the 2nd int. workshop on Petascale data storage.* USA: ACM, 2007, pp. 50–55.
2. M. Fahey, J. Larkin, and J. Adams, "I/o performance on a massively parallel cray xt3/xt4," in *Parallel and Distributed Procs, 2008. IPDPS 2008. IEEE Int. Symp. on*, 14-18 2008, pp. 1 –12.
3. J. H. Laros *et al.*, "Red storm io performance analysis," in *CLUSTER '07: Procs of the 2007 IEEE Int. Conf. on Cluster Computing.* USA: IEEE Computer Society, 2007, pp. 50–57.
4. W. D. Norcott, "Iozone filesystem benchmark," Tech. Rep., 2006. [Online]. Available: http://www.iozone.org/
5. R. Coker, "Bonnie++ filesystem benchmark," Tech. Rep., 2001. [Online]. Available: http://www.coker.com.au/bonnie++/
6. . S. J. Shan, Hongzhang, "Using ior to analyze the i/o performance for hpc platforms," LBNL Paper LBNL-62647, Tech. Rep., 2007. [Online]. Available: www.osti.gov/bridge/servlets/purl/923356-15FxGK/
7. R. Rabenseifner and A. E. Koniges, "Effective file-i/o bandwidth benchmark," in *Euro-Par '00: Procs from the 6th Int. Euro-Par Conference on Parallel Procs.* London, UK: Springer-Verlag, 2000, pp. 1273–1283.
8. A. Wong, D. Rexachs, and E. Luque, "Extraction of parallel application signatures for performance prediction," in *HPCC, 2010 12th IEEE Int. Conf. on*, sept. 2010, pp. 223 –230.
9. P. Wong and R. F. V. D. Wijngaart, "Nas parallel benchmarks i/o version 2.4," Computer Sciences Corporation, NASA Advanced Supercomputing (NAS) Division, Tech. Rep., 2003.
10. A. Núnez, *et al.*, "Simcan: a simulator framework for computer architectures and storage networks," in *Simutools '08: Procs of the 1st Int. Conf. on Simulation tools and techniques for communications, networks and systems & workshops.* Belgium: ICST, 2008, pp. 1–8.