International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:3, No:12, 2009

Analysis of Long-Term File System Activities on Cluster Systems

Hyeyoung Cho, Sungho Kim, and Sik Lee

***Abstract*—**I/O workload is a critical and important factor to analyze I/O pattern and to maximize file system performance. However to measure I/O workload on running distributed parallel file system is non-trivial due to collection overhead and large volume of data. In this paper, we measured and analyzed file system activities on two large-scale cluster systems which had TFlops level high performance computation resources. By comparing file system activities of 2009 with those of 2006, we analyzed the change of I/O workloads by the development of system performance and high-speed network technology.

***Keywords*—**I/O workload, Lustre, GPFS, Cluster File System

I. Introduction

AS cluster systems are becoming more popular in various areas and users are increasing, I/O workload analysis is required to use cluster systems more efficiently. Through I/O workload analyses, we can predict I/O access patterns of users. System performance can be increased by optimization which considers user I/O access patterns[1].

Over the past few decades a large number of studies have been made on analysis of I/O workloads. The analysis result of file system I/O workloads can be used for managing file systems and designing new systems. To maximize performance with limited resource, file system managers trace file system, analyze I/O workload and optimize file system based on the I/O workload data. In addition, I/O workload data is one of the most important factors for the design of new file system.

Even though I/O workload is significant for this variety of purpose, it is difficult to get the I/O trace data on the fly system of real world, because of collection overhead and large volume of data. Especially, for data-intensive high-end applications I/O workload analysis on a live distributed parallel file system such as PVFS[2], Luster[3], and GPFS[4], is non-trivial due to the following reasons. First, the distributed file system physically organizes a number of different systems. Second, the file system trace should not affect its performance. [1]

As the first step for analysis of I/O workload on Distributed and Parallel File System, we designed and implemented a parallel file system logging method for high performance computing using shared memory-based multi-layer scheme[5]. It minimizes overhead with reduced logging operation response time and provides efficient post-processing scheme through shared memory. In large-scale distributed environment, separated logging server can collect sequential logs from multiple clients in a cluster through packet communication. As the next step, we monitored and measured file system on distributed environment to understand over all file system activities and target files which were traced and analyzed.

In this paper, we measured and analyzed file system activities on two large-scale cluster systems in distributed environment at 2006[6] and 2009. The contributions of our study are:

- We analyzed recent file system activities by monitoring two running distributed and parallel file systems at 2006 and 2009.
- We measured long-term file system activities on running two large-scale cluster systems for more than 6 months.
- We found the change of file system activities through improving system performance and high-speed network technology. For that, we monitored and analyzed file system activities of 2006 and those of 2009.

The rest of the paper is organized as follows. We reviewed and discussed related works in Section 2. In Section 3, we explained our monitoring systems and user characteristics. We showed the results of file system activities about files, directories and the user spaces in Section 4, 5 and 6 respectively. In Section 7, a short conclusion and future works were given.

II. Related works

Previously, there were a number of interesting and noteworthy studies on I/O workload analysis. Table 1 shows previous studies of I/O workloads. Satyanarayanan[7] analyzed file system access patterns on CM-5 system, a national supercomputer. Jone[8] and Timothy[9] monitored UNIX file system to collect information that would be useful in designing and managing a file system. However most of studies had been made more than 10 years ago.

In addition, target file systems of most preceding studies were local file systems[10,11,12]. Even though the target file system was a share file system, it was a local file system that was connected by NFS[8,9,11]. I/O workload researches on distributed and parallel file systems were not many. Nils et. al.[12] observed iPCS/860 at NASA's Ames Research Center and on Thinking Machines CM-5 at the National Center for supercomputing Application. They monitored Intel's Concurrent File system (CFS) and Scalable File System(SFS). Phyllis[13] observed a parallel file system, Intel's PFS. However the objective of those researches was to analyze I/O workload of a specific application, not I/O workload of a full file system[14].

Hyeyoung Cho, Sungho Kim and Sik Lee are with the Supercomputing Center, Korea Institute of Science and Technology Information(KISTI), 335 Gwahangno, Yuseong-gu, Daejeon, Korea 305-806.(e-mail: chohy@kisti.re.kr, sungho@kisti.re.kr, siklee@kisti.re.kr)

TABLE I PREVIOUS STUDIES OF I/O WORKLOADS

Year	Study	System/File System/Protocol	All FS	Applications	FS Types	Characteristics
1981	M. Satyanarayanan[7]	TOPS-10/CM-5	O	-	PFS	A research Server at a University
1995	John K. Ousterhout et al.[8]	BSD	O	-	LC, NFS	University Servers for staffs, electrical engineering students
1995	Phyllis E et al.[13]	Intel Paragon XP/S, Intel's PFS	-	O	PFS	Scientific applications
1996	Nils Nieuwejaar et al.[12]	Intel PCS/860,CFS, CM-5, SFS	O	-	PFS	Research Servers
1997	Evgenia Smirni et al.[14]	Intel PFS	-	O	PFS	Scientific applications
1998	Timothy J. Gibson[9]	UNIX-based FS/NFS	O	-	LC, NFS	University file systems
2001	Allen B. Downey[10]	-	O	-	LC	local file systems, web servers, web clients,
2002	Drew Roselli et al.[9]	FS(VxFS, NTFS)	O	-	LC	Research Server, Web Sever, PC
2003	Daniel Ellard et al.[11]	NFS	O	-	NFS	University Servers
2009	Hyeyoung(our research)	Lustre, GFS	O	-	PDFS	Cluster System for Scientific applications

*LC: Local File System, PDFS: Parallel & Distributed File System, CFS: Cluster File System

According to the review of previous studies, new efforts about I/O workload analyses are needed by the following reasons:

First, most of studies had been made more than 10 years ago. There have been significant improvements in network bandwidth and computing performance during past few decades. File system designers and managers want to know file system I/O workload information on recent technology environment, such as with high-speed network and over hundreds TFlops computing power.

Second, I/O workload researches on distributed and parallel file systems were not many. Target file systems of most previous studies were local file system. Even though the target file system was a share file system, it was a local file system connected by NFS. Although there were several studies of I/O workload analysis on distributed and parallel file systems, the objective of those researches was to analyze I/O workload of a specific application, not I/O workload of a full file system.

III. SYSTEM ENVIRONMENT

A. Monitoring Systems

For collecting file system information in cluster systems, we monitored two systems, Hamel and Tachyon. Hamel cluster system is a part of 3th supercomputer on KISTI(Korea Institute Science and Technology Information). It was serviced from January, 2005 to September, 2008 to users. Hamel system has 512nodes(512CPUs) and the theoretical peak performance is 2.867TFlops.

Tachyon cluster system is a part of 4th supercomputer on KISTI. It was built on 2008 and has been servicing from August, 2008 to users. It has more than 3008 CPUs of AMD Opteron and is constructed 188 nodes. Each node has main memory of 32GB. It is composed by SUN Blade 6048s and the theoretical peak performance reaches 24TFlops. Tachyon system was ranked at No. 130 in the 32rd edition of the TOP500 list in June 2008[15].

Hamel and Tachyon system are the large-scale systems which have thousands of logins per month. Figure 1 shows login statistics of two systems. The averages of login numbers per month on Hamel and Tachyon are 3,275 and 5,266 respectively.

Table 2 shows storage list of trace systems. We examined user's home and scratch storages in two cluster systems. Home storage has user's configuration files, source data and execution files, etc. The purpose of scratch storages is to provide a large amount of disk space with very high speed. Users use the scratch storages for I/O intensive applications.

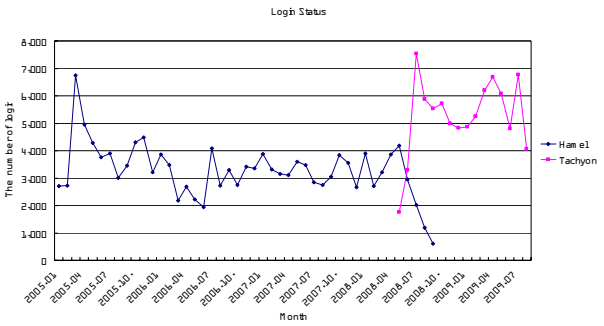

Fig. 1. Login Statistics

TABLE II LIST OF TRACE SYSTEMS

System Name	Target Storage	Total Volume	Average Space Usage	FS Types	System Service Periods	Date of Traces	Description
HAMEL	/home2	3.2TB	29.4%	GPFS	2005.01~ 2008.09	2006.06-2007.12	home directory
HAMEL	/ytmp	2.1TB	76.6%	GPFS	2005.01~ 2008.09	2006.06-2007.12	scratch directory
TACHYON	/home01	11TB	14%	Lutre	2008.08 ~ present	2009.1-2009.9	home directory
TACHYON	/work01	31TB	44%	Lustre	2008.08 ~ present	2009.1-2009.9	scratch directory
TACHYON	/work02	54TB	43%	Lustre	2008.08 ~ present	2009.1-2009.9	scratch directory

Through the comparing file system behaviors of the similar users between 2006 and 2007, we analyzed changes of file

system activities by growth of system performance and high-speed network technology.

B. Users Characteristic

Hamel and Tachyon were designed and have been serviced as public research resources to universities, government research institutes, government agencies, industries etc. Those systems have been used for a variety of research areas, such as, physics, chemistry, environment, biology and so on.

Figure 2 shows the distribution of CPU utilization on the two systems. The monitoring period of Hamel system included all servicing period, from January, 2006 to September, 2008. That of Tachyon was from August, 2008 to July, 2009.

The users of two systems are similar researchers who are examining about applied science topics. In the case of Hamel, the percentages of use were 43.5% for Physics, 22% for Machinery, 12.2% for Chemistry, 9.6% for Electrical/Electronics and 3.6% for Atmosphere/Environment. In the case of Tachyon, the percentages of use were 32.5% for Physics, 26.1% for Chemistry, 21.2% for Machinery, 8.3% for Electrical/Electronics and 8.3% for Atmosphere/Environment.

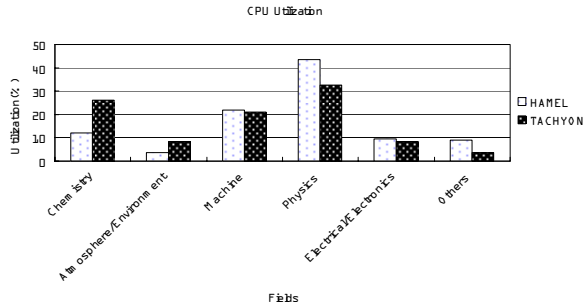

Fig. 2. CPU utilization of research areas

IV. Files

A. File Types

This section describes our findings regarding file types. We observed the file type distribution of two systems, each on 2006 and 2009. Figure 3 shows cumulative distribution functions(CDFs) of file types by count of files at home storage. The percentages of file count in 2006 were less than 1 % for configuration files and symbolic links files, 8.48% for directories, 14.90% for execution files and 75.59% for other data files. The percentages of file count in 2009 were 2 % for configuration files and symbolic links, 15.26% for directories, 7.21% for execution files and 75.44% for other data files. One interesting discovery is that the percentage of directories on home storage is increased by 8.86% between 2006 and 2009. Figure 4 describes CDFs of file types by size at home storage. There was no significant difference at the data between 2006 and 2009.

Figure 5 and 6 illustrate CDFs of file type by file count and size at scratch storage. We observed that the percentage of directories was increased by 1.51% between 2006 and 2009, like home directory. Also the percentage of execution files was increased by 1.49%.

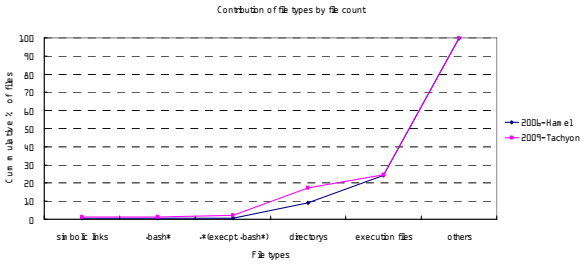

Fig. 3. CDFs of file types by file count at home

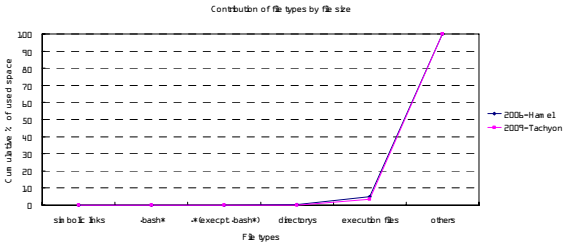

Fig. 4. CDFs of file types by file size at home

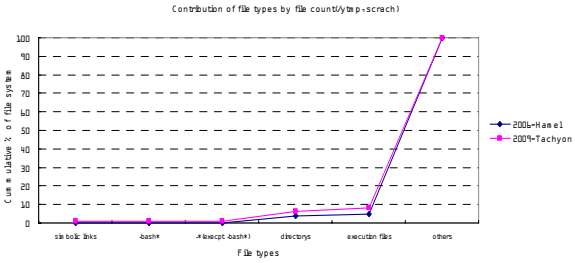

Fig. 5. CDFs of file types by file count at scratch

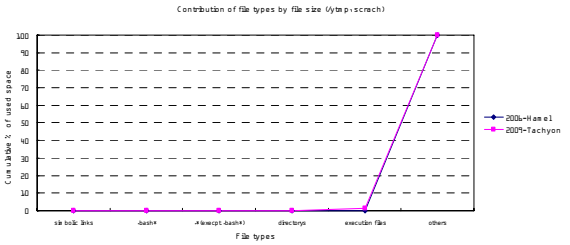

Fig. 6. CDFs of file types by file size at scratch

B. File Size

This section describes our results regarding file size. Figure 7 shows Histograms of file size distribution by file count on scratch storages. We observed that over 55% of files were smaller than 1 megabyte and 30% of files were between 3 megabytes and 64 megabytes in Hamel. In the case of Tachyon, over 83% of files were smaller than 1 megabyte and 6% of files were between 4 megabytes and 64 megabytes.

Figure 8 shows Histograms of file size distribution by file capacity. We found that 25.43% of files were between 64

megabytes and 100 megabytes and 11.05% of files were between 4 megabytes and 64 megabytes in Hamel. Especially over 40% of files were 400 megabytes and 500 megabytes. The reason including the high percentage of files between 400 megabytes and 500 megabytes is that users used Hamel system as a backup server for multimedia data. In the case of Tachyon, 18% of files are between 4 megabytes and 64 megabytes and 13 % of files are between 100 megabytes and 200 megabytes. We observed that 1 gigabyte or more files specially were increased in terms of capacity. The reason was that system managers made big size of files for data backup.

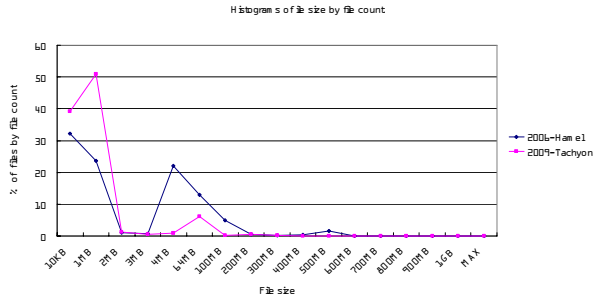

Fig. 7. Histograms of file size distribution by file count

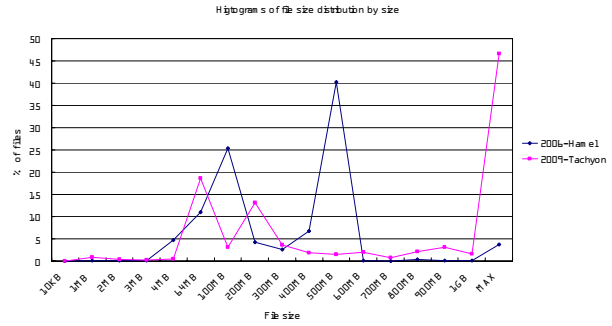

Fig. 8. Histograms of file size distribution by file size

C. File Age

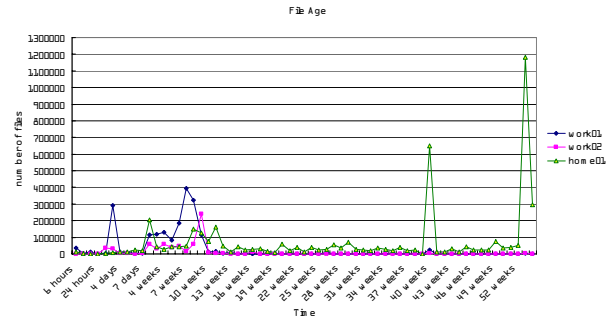

Fig. 9. File Age

This section describes our findings regarding the file age. Figure 9 illustrates the file ages of home and scratch storages. The averages of file age were 49 days for work01, 65 days for work02 and 483days for home. The result of scratch was reflected the scratch management policies such as the guarantee period of data.

V. Directory

A. Directory Proportion

This section describes our detections regarding directory. Table 3 shows the total number of file and directory of trace systems. We found that the home directory has a high proportion of directories compared to the scratch directory.

TABLE III TOTAL NUMBER OF DIRECTORIES AND FILES

System Name	Target Storage	Description	# of directories	# of files	files : directories
HAMEL	/home2	home directory	822	3,070	21.12%
HAMEL	/ytmp	scratch directory	58,364	329,544	15.05%
TACHYON	/home01	home directory	197,900	3,312,898	5.64%
TACHYON	/work01	scratch directory	121,062	7,350,507	1.62%
TACHYON	/work02	scratch directory	32,384	1,045,243	3.01%

B. Contained file count per directory

Figure 10 plots CDFs of directories by contained file count. The percentage of directories which contained 10 or less files is 73.35% for home directory and 77.53% for scratch directories. The average of contained file count per directory is 16.85 for home directory and 46.15 for scratch directories. The maximum number of contained file count per a directory is 20,748 for home and 153,497 for scratch storages.

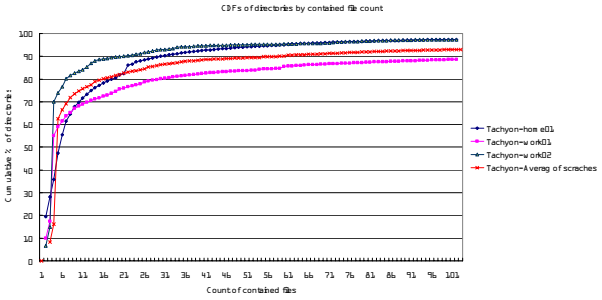

Fig. 10. CDFs of directories by contained file count

C. Directory Depth

This section describes our findings regarding the depth of directory. The depth of directory means the level for root directory in the namespace tree. For example, In the case of home storage, the depth of /home directory is 0. In the case of /home/hycho directory, the depth of directory is 1.

Figure 11 plots histograms of directories by directory depth. From Figure 11, we observe that the 90% of directories has 9 or less for home, 11 or less for work01 and 7 or less for work02 in Tachyon system. The 90% of directories has 7 or less for home and 6 or less for ytmp in Hamel. Most of the directories have less than 10 directory depth. We detected that the averages of directory depths were increased by 1.16 in home directory and by 3.22 in scratch directory between 2006 and 2009.

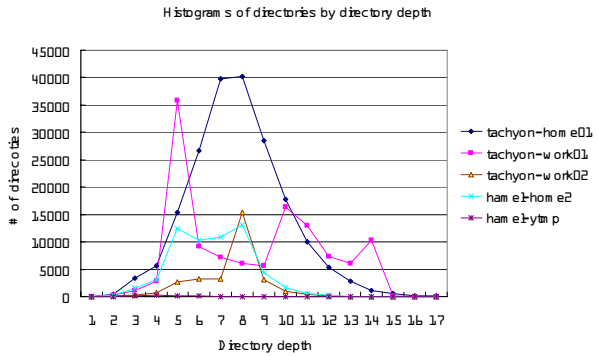

Fig. 11. Histograms of directories by directory depth

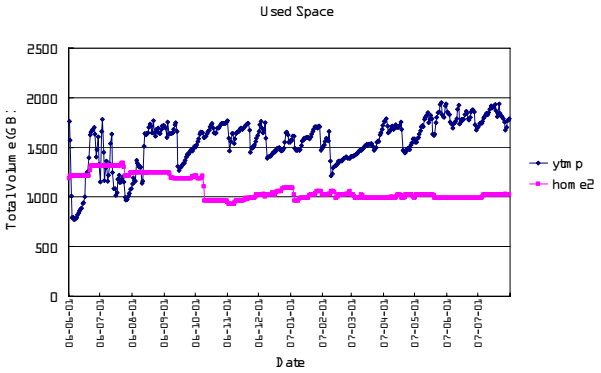

Fig. 13. Changes of used Space on Hamel

TABLE IV STATICSTICS OF DIRECTORY DEPTH

System	*Tachyon-home01*	*Tachyon-work01*	*Tachyon-work02*	*Hamel-home2*	*Hamel_ytmp*
MAX	*47*	*14*	*15*	*14*	*8*
Average	*6.81*	*7.41*	*6.49*	*5.65*	*3.73*

VI. USER SPACE

For building and managing a cluster system, it is important to predict changes of data capacity. This section describes our results regarding changes of user space. Figure 12 and 13 illustrate chances of used space on Tachyon and Hamel. We notice that home directory did not have a lot of chances in capacity because user's home directory contains configuration files and source files. While scratch directory has a large amount of changes in capacity compared with home directory. The averages of incremental volume per day were 6.39 gigabytes for home and 197.02 gigabytes for scratch storage. The averages of incremental volume per day were increased by 4.35 times for home and 17.42 times for scratch between 2006 and 2009.

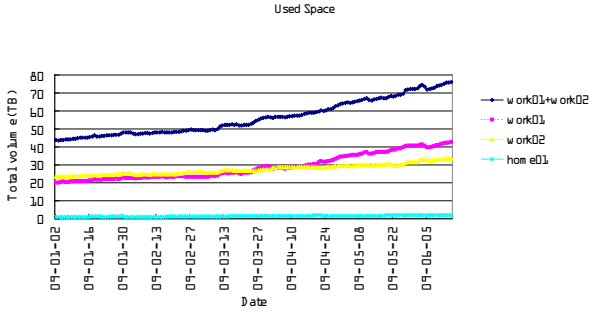

Fig. 12. Chances of used Space on Tachyon

VII. CONCLUSIONS AND FUTURE WORKS

In this paper, we measured and analyzed long-term file system activities on two large-scale realistic cluster systems in the distributed environment. Through comparing file system activities of 2006 with those of 2009, we found important similarities and differences including file types, file size, directory and user space with the growth of system performance and high-speed network technology. All systems have still small size of files. The file size for data backup is increased from 400~500 megabytes to 1 gigabyte or more. From the facts of increasing percentage of directories and the average of directory depths, we detected that users are managing files more structurally using directories recently. The information about directory depth can be considered to design file system directory structure, such as the format of the indirect maps.

Home storage is managing more systemically using directories compared with scratch storage. While home directory did not have a lot of chances in capacity, scratch directory had a lot of changes. The average of incremental volume per day was 6.39 gigabytes for home and 197.02 gigabytes for scratch storage. The average of incremental volume per day was increased by 4.35 times for home and 17.42 times for scratch between 2006 and 2009. This information can be reflected to design a next generation cluster system. This depth understanding of file system activities is useful to provide keen insight into the design and analysis of file system for performance gains, such as the file system capacity and cache policy of file system.

In the future, we are planning to analyze both I/O workload of overall file system and that of high erformance applications. In this research, we monitored and measured file system activities on distributed environment to understand over all file system activities. From the results we recognized some important characteristics of our files. We can classify our files using the results, such as types, permanence, file age, etc. In the next step, we will do modeling I/O workload based on the our file system characteristics. And we will analyze I/O workload of high performance computing system and application based the file system modeling. Through the analysis of file system, we will recognize the characteristic of both user I/O pattern and overall

system. In addition, the results of analysis will be used as a feedback for file system management and optimization.

REFERENCES

[1] John K. Ousterhout, Hervg Da Costa, David Harrison, John A. Kunze, Mike Kupfer, and James G. Thompson, "A Trace-Driven Analysis of the UNIX 4.2 BSD File System," ACM SIGOPS Operating Systems Review archive, Volume 19, Issue 5, pp. 15~24, 1985.
[2] PVFS web size, http://www.pvfs.org
[3] Lustre web site, http://wiki.lustre.org
[4] GPFS Wikipedia, http://en.wikipedia.org/wiki/GPFS
[5] Hyeyoung Cho, Sungho Kim and SangDong Lee, "Design and Implementation of Shared Memory based Parallel File System Logging Method for High Performance Computing," Volume 45, 2008.
[6] Hyeyoung Cho, Kwangho Cha and Sungho Kim, "Analysis of File System Workloads on Hamel Cluster System," 2006 Autumn Conference, Korea Information Processing Society, 2006.
[7] M. Satyanarayanan, "A Study of File Sizes and Functional Lifetimes," In Proceedings of the 8th Symposium on Operating Systems Principles, pp. 96-108, 1981.
[8] John K. Ousterhout, Hervg Da Costa, David Harrison, John A. Kunze, Mike Kupfer, and James G. Thompson, "A Trace-Driven Analysis of the UNIX 4.2 BSD File System," ACM SIGOPS Operating Systems Review archive, Volume 19, Issue 5, pp. 15~24, 1985.
[9] Timothy J. Gibson and Ethan L. Miller, "Long-Term File Activity Patterns in a UNIX Workstation Environment," in the Proceedings of the 15th IEEE Symposium on Mass Storage Systems, pp. 355-272, 1998.
[10] Allen B. Downey, "The structural cause of file size distributions," ACM SIGMETRICS Performance Evaluation Review, Volume 29, pp. 328 – 329, 2001.
[11] Drew Roselli, Jacob R. Lorch,, "A comparison of file system workloads," USNIX, 2002.
[12] Nils Nieuwejaar , David Kotz , Apratim Purakayastha , Carla Schlatter Ellis , Michael L. Best, "File-Access Characteristics of Parallel Scientific Workloads," IEEE Transactions on Parallel and Distributed Systems, v.7 n.10, pp.1075-1089, October 1996.
[13] Phyllis E. CrandallRuth A. AydtAndrew A. ChienDaniel A. Reed, "Input/Output characteristics of scalable parallel applications," in the Proceedings of the ACM/IEEE Supercomputing conference, 1995.
[14] Evgenia Smirni and Daniel A. Reed, "Workload characterization of input/output intensive parallel applications," In the Proceedings of the Conference on Computer Performance Evaluation Modeling Techniques and Tools for computer performance evaluation, Volume 1245, LNCS, pp 169-180, June 1997.
[15] Top500 Supercomputing Website, http://www.top500.org

Hyeyoung Cho received the M.E. degree in computer engineering from the Information and Communications University, Daejeon, Korea in 2004. She is currently a researcher of the Supercomputing Center of Korea Institute of Science and Technology Information. Her interests include cluster system, distributed and parallel file system and high-performance computing. She is a member of the Korea Information Science Society.

Sungho Kim received the Ph.D. degree in aerospace engineering from Korea Advanced Institute of Science and Technology, Daejeon, Korea in 1999. He is currently a chief researcher of the Supercomputing Center of Korea Institute of Science and Technology Information, Daejeon, Korea. He performed many national projects related to cluster computer architecture, system software and grid technology. He is now one of the key members to design 4th supercomputer of KISTI Supercomputing Center and other related projects. His research interests include cluster system and cloud computing.

Sik Lee received the Ph.D. degree in chemistry from Pohang University of Science and Technology, Korea (1996) and was researcher at Massachusetts Institute of Technology, Cambridge University, and the University of Pennsylvania. He is currently the Department Head of Application and Support at the Supercomputing Center of Korea Institute of Science and Technology Information, Daejeon, Korea. He performed many national projects on high performance computing applications. His research interests include molecular modeling, bioinformatics and high-performance computing.