A Statistical Study for File System Meta Data On High Performance Computing Sites

Submitted in partial fulfillment of the requirements for

the degree of

Master of Science

in

Information Networking

Yifan Wang

B.S., Information Engineering, Southeast University

Carnegie Mellon University
Pittsburgh, PA

May, 2012

Acknowledgements

I would like to express my gratitude to my advisor Prof. Garth Gibson for working closely with me through the entire study, offering precious advice and guidance and helping connect with industrial and academic HPC sites. I would like to thank my academic adviser Prof. Greg Ganger for offering valuable advice on the paper and defense. I wish to acknowledge Meghan W. McClelland from Los Alamos National Lab (LANL) for providing me current data and timely feedback on the report. I also wish to thank Machael Stroucken for helping collect data from Parallel Data Lab (PDL) clusters and offering comments. This paper and relevant research is supported by CMU / Los Alamos National Lab Institute for Reliable High Performance Information Technology.

Abstract

High performance parallel file systems are critical to the performance of super computers, are specialized to provide different computing services and are changing rapidly in both hardware and software, whose unusual access pattern has drawn great research interest. Yet little knowledge of how file systems evolve and how the way people use file systems change is known, even though significant effort and money has been put into upgrading storage device and designing new file systems. In this paper, we report on the statistics of supercomputing file systems from Parallel Data Lab (PDL) and Los Alamos National Lab (LANL) and compare the current data against their statistics 4 years ago to discover changes in technology and usage pattern and to observe new interesting characters.

Keywords: file system, storage, HPC

Table of Contents

List of Tables

List of Figures

1. Introduction

As a continuous work of a project started from 2006 at CMU, our goal is to continue studying static file tree attributes and to compare these data with previous record to analyze the impact of software and hardware evolution. In the past, people have collected data from working file systems and study how files change in term of change in file size, file age and other attributes. Yet, very few researches are focused on file systems used in high performance computing sites. There is neither research reporting how file systems change in super computers nor papers explaining the reasons lying behind the change of file system meta-data. However, this is very important for file system designers and users since they want to make the most from expensive storage technology.

In this paper, we collected data from Parallel Data Lab (PDL) and Los Alamos National Lab (LANL) in the year 2011 and 2012. In addition to comparing the recent data with each other, we also compare the data with a 2008 report. In order to support a concrete comparison, we only compare the file systems used for similar applications. In total, we collected data from four PDL machines and twelve LANL machines, which are used for home directories, cluster machines, administration use and archive. There are two major challenges in the whole project. The first is that it takes a very long time for file system administrators to run our data collection script and this disrupts the normal usage of file system. Another difficulty is that our tool assumes that the target file system supports POSIX command, which might not be the case in real situation. Some file system administrator reports us with file system meta-data in their own format and we need to make modifications in format in order to incorporate those into our report.

We have also modified our tool fsstats to collect file system meta-data. In addition to the old attributes that were also collected in 2008 version of fsstats, at this time we collect data of new properties including name space property and

full path name length. We do not present the analysis for symbolic link and hard ink in order to be focused on more interesting properties.

Section 2 introduces how to read the graphs presented in the paper. Section 3 introduces the new version of fsstats. Section 4 presented some of the recent researches. Section 5 presents the hardware and software property of the file system we study. Section 6 shows how file systems evolve from 2008 to 2011 by comparing data generated from file systems used for similar purposes. Section 7 gives a more comprehensive analysis into the recent file systems property. Section 8 provides a conclusion for our research and section 9 explains what our next step of research is.

2. Reading Graphs

Our graphs show cumulative distributive functions (CDF); that is, the part of samples (like file system file size histogram) whose property of interest is smaller than a given number (like file size). According to our observation, lots of data are gathered together in two parts: under 10% and from 90% to 100%. Researchers also pay special attention to the data in that part. In order to highlight that area, for most cases, we plot the y axis in three sections in the figure. Part 1 is the log scale of base 10, to represent 0.001% to 10%. Part 2 is the linear space to represent 0% to 1% percent. Part three is the log scale space from 90% to 99.999%. Some of the X-axis is in log scale and some others are in linear scale, and this is specified in the footnotes. We did not curve fit our data.

2.1 Legends

We use the following legends to represent lines in out graph.

PDL-1-08, PDL-2-08: volumes were used by PDL, data were collected in 2008.

LANL-SCR1-08, LANL-SCR2-08, LANL-SCR3-08: scratch space in LANL. Data were collected in 2008.

LANL-LNFS, LANL-GNFS: NFS file system from LANL. Data were collected in 2011.

LANL-ARCH: Archive system from LANL. Data were collected in 2011.

LANL-PanFs-1~LANL-PanFS-9: PanFS file systems from LANL. Data were collected in 2011 and 2012.

3. Data Collection Methodology

We have made some modifications to the 2008 version of fsstats script to collect more attributes of interest. The general procedure remains the same. The script traverses through the whole directory tree and use POSIX command lstat to extract meta-data from a file. Then instead of storing the exact number, meta-data are put into histograms. For instance, the meta-data for a file with size 5 KB will be collected into the bucket of [4KB, 8KB] of the file size histogram. This method is used to keep privacy for fsstats user as opposed to keeping exact data for all files. We have incorporated three new histograms in addition to the previous histograms.

Full Path Name Length Histogram: This histogram bins data based on full path name length. The full path name length is measured in characters and character '/' is also counted into the length. We think this is important for two reasons. Firstly, in 2008 we only studied the name length for regular files while neglecting the name length for directories. Now, we can include the directory name length into consideration. Secondly, full path name length is important for file system designers using big table. This can offer a solid background for them to design the hash function.

File Count In Name Space: This describes the name space property. We always hope to know, what people's habit of using file system is. We want to answer the questions like whether people tend to put files in shallow depth of the file tree or they tend to hide most files deep in file trees. We want to know, how people manage their name space. Do they tend to maintain a flat name space or do they tend to develop a very deep file tree? Here we study how

many files are placed within a certain depth of the name space. For instance, people might store 1000 files at depth 1 and 10000 files at depth 2. Considering people will not use a too complex directory tree structure, we collect the number of files linearly on the depth from depth 1 to depth 32. Afterwards, we use log scale on the depth. For example, we collect number of files in depth 33 to 64 together.

File Size In Name Space: This describes another aspect of name space property. We want to answer the question whether people tend to store big files shallow in name space or they tend to store them deep in name space. Similar to file count in name space histogram, the buckets are also divided half linear and half in log scale and 32 is the threshold between linear space and log space.

4. Related Work

File systems properties are widely studied under multiple occasions. For instance, Mayer and Bolosky studied directory depth, number of sub directories and number of files in their research in de-duplication technology in Microsoft [1].

There are also researches focusing on the studying file size and other attributes distributions [2] [3]. Within those papers, mathematic algorithms for curve fitting are largely used in order to discover a general rule. General workloads are also studied and benchmark results are presented for some file systems. [8][9][10]

There are researches focusing on comparing the temporal change of meta-data. Bolosky presented his research in Microsoft, which compares data over five years [4]. It also gives us a description of file system attributes distribution on personal computers.

In 2008, PDL CMU released a study on file system meta-data for HPC sites [5], which is the first study on super computers.

Label	Date	Type	File System	Tot. Size (TB)	Tot. Space (TB)	# files (M)	Max Size (GB)	Avg Size (MB)	# dirs (K)	Max Dir ents	Avg dir ents	Max name bytes
PDL-Home	9/11	Home	WAFL	1.18	1.21	14	30.7	.09	1,269	571,665	12	143
PDL-OsDist	9/11	OS	WAFL	.015	.016	.5	.467	.05	54.1	54,067	12	110
PDL-Cirrus	9/11	FS	WAFL	.81	.767	.218	48.8	3.78	30	5,625	14	80
PDL-VM	9/11	VM	WAFL	.05	.054	0	17	1632	0.014	10	3	24
PDL-1-08	4/09	Project	WAFL	3.93	3.63	11.3	23.4	.37	821	56,960	15	255
PDL-2-08	4/09	Project	WAFL	1.28	1.09	8.11	23.4	.17	694	89,517	14	255
LANL-LNFS	7/11	Cluster	NFS	.02		.272	.465	.064	352	342	8	7
LANL-GNFS	7/11	Cluster	NFS	1.41		5.69	32.16	.33	815	565	8	8
LANL-ARCH	7/11	Archive	GPFS	52.8		106	820	21	47,356	384	2	9
LANL-SCR1-08	4/10	Scratch	PanFS	9.2	10.7	1.52	134	6	120	14,420	14	90
LANL-SCR2-08	4/10	Scratch	PanFS	25	26	3.3	978	8.2	241	50,000	15	73
LANL-SCR3-08	4/10	Scratch	PanFS	26	29	2.58	998	10.9	374	45,002	8	65
LANL-PanFS-1	1/12	Compute	PanFS	24.9	28.66	4.4	285	5.9	4,640	62,430	73	69
LANL-PanFS-2	1/12	Compute	PanFS	286.34	327.9	8	297.4	37.2	9,330	203,040	45	83
LANL-PanFS-3	1/12	Compute	PanFS	25.58	28.3	1.9	143.8	14.3	2,025	144,427	13	80
LANL-PanFS-4	1/12	Compute	PanFS	14.02	16.2	2.2	332.2	6.6	3,968	490,399	7	95
LANL-PanFS-5	1/12	Compute	PanFS	313.11	361.9	32.1	3,659	10.2	32,614	415,310	77	87
LANL-PanFS-6	11/11	Compute	PanFS	192.79	219.4	15.3	1,276	13.2	15,438	142,940	107	211
LANL-PanFS-7	11/11	Compute	PanFS	382.6	439.8	43.2	2,669	9.3	43,606	359,023	120	133

Label	Date	Type	File System	Tot. Size (TB)	Tot. Space (TB)	# files (M)	Max Size (GB)	Avg Size (MB)	# dirs (K)	Max Dir ents	Avg dir ents	Max name bytes
LANL-PanFS-8	11/11	Compute	PanFS	43.97	48.7	4.6	85.4	10	493801	296,183	16	133
LANL-PanFS-9	11/11	Compute	PanFS	28.35	31.6	5	254.4	5.3	7,651	721,281	10	123

Table 1: This table summarizes the data generated from 22 file systems. Here we report when data was collected, what is the file system used for, what is the file system, what is the total file size, what is total disk space consumed, how many files are there in the system, what is the biggest files size, what is the average file size, how many directories are there within that file system, what is the maximum entry number in that directory, what is the average number of entries in the directory and what is the maximum name length of files in that file system. For PDL-VM, the entry for total file number is 0 million. This is because it is a very small file system and contains only 34 files. For LANL-LNFS, LANL-GNFS, and LANL-ARCH, total space consumed is not available now.

5. File System Studied

Data for 2008 PDL and LANL machines can be found in the previous report [5]. Specified file system statistics in 2011 are listed in table 1.

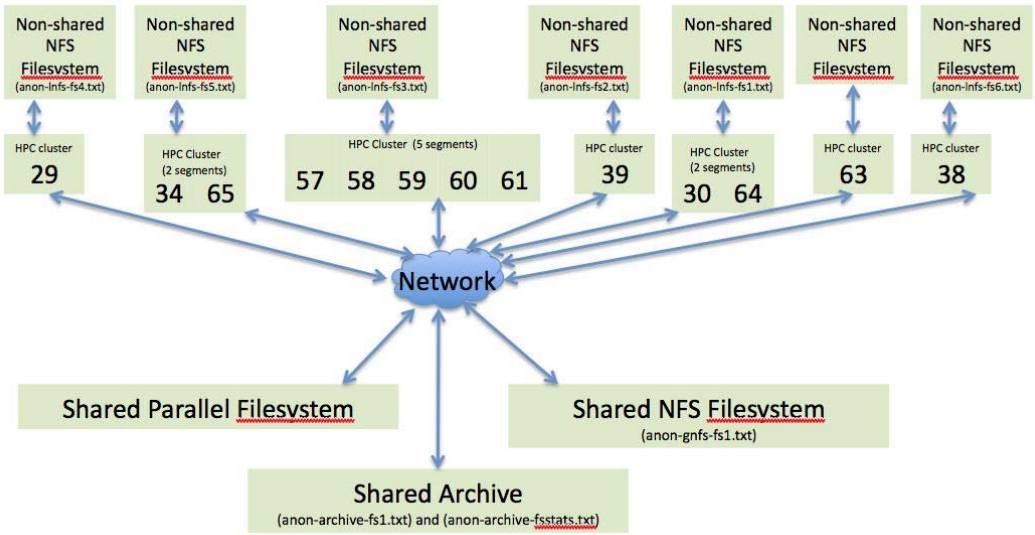

Figure 1: Description of file systems structure from LANL. Numbers are the machine IDs used by the LANL to indicate the source of data. For instance, data in NFS File System 1 are generated from No. 30 and No. 64 HPC cluster machines. There are four major divisions of LANL machines, NFS non-shared file systems, NFS-shared file system, archive file system and shared parallel file system. Thanks to LANL for providing the data

5.1 2011 PDL file systems

2011 PDL data are generated from two volumes. Both volumes use WAFL file system from NetApp. PDL-Home and PDL-OsDist are generated from a volume with name ‘Flow’. For hardware, it uses FAS 3050 from NetApp. ‘Flow’ volume has 28disks; each of the disks is 400 GB, 7200 RPM SATA. The total 28 disks are divided into two groups and each of them uses 12+2 RAID-DP. PDL-Home is used to store home directories and it can be accessed by all PDL researchers. PDL-OsDist is used to store operating system distributions and other machines may boot from it. It can only be modified by administrator. The other volume is called ‘Valeve’ and it includes PDL-Cirrus and PDL-VM. ‘Valve’ volume has 40 disks of 7200 RPMSATA 136 GB. PDL-Cirrus is used to store open source software for Cirrus. It also has some files used for administration only. PDL-VM is used to store virtual machine image for PDL people to use. It can only be accessed by administrator.

5.2 LANL NFS machines and LANL Archive

LANL-LNFS is used for local cluster only and it includes six non-sharing cluster file systems. It is used as home directory for cluster machines. LANL-GNFS is used to be shared by all LANL users and hence it is called global file system. LANL-LNFS and LANL-GNFS are using NFS file system. Both of them are used as home directories. LANL Archive is used as archive directory for LANL and it runs GPFS. For LANL-ARCH, there is a tape recording system using HPSS as backend. As shown in Figure 1, LANL-LNFS is a summary for all the non-shared NFS files systems; LANL-ARCH is the shared archive; and LANL-GNFS is the shared NFS file system.

5.3 LANL PanFS machines

LANL provides us with 9 machines running PanFS file system [11]. Some of them are using RAID 5 with 64 KB

stripe size and some others are using RAID 10. There are usually three groups of users for the PanFS machines. Group 1 use this machine to hold code snippets and hence they will generate lots of small files on the machine. Group 2 users use this for testing and will generate some strange files on the machines. Group 3 store some big files on the file system. There is no clear division of which file system is used by which group. As shown in Figure 1, LANL-PanFS files systems are the shared file systems.

6. Results - Temporal Comparison

In this section, we will compare recently-collected data with the data generated in 2008. In order to provide a concrete comparison and analysis of data, we will study and compare the file systems used for similar purposes. Among the PDL machines, we will compare the PDL-Home, PDL-OsDist with PDL-1-08 and PDL-2-08. PDL-Home and PDL-OsDist are used to hold home directory and Operating System Distributions for PDL machine to boot from. On the other hand, PDL-1-08 and PDL-2-08 were also used to store home directory, Operating System distribution and video/audio files for I/O testing. Moreover, all file systems are WAFL provided by NetApp and there is no change in hardware.

Among LANL machines, nine LANL-PanFS file systems are compared against three LANL-SCR-08 file systems. All of them are used in cluster machines and can be accessed only by one local cluster, and they all use PanFS as the file system, running on RAID 5 with 64 KB stripe size or RAID 10. Within the LANL-Panfs-2011 machines, there exist small differences in the usage. File System 1-5 and 8-9 are more likely to be used to save big files for production testing. File System 6-7 are more likely to be used for holding codes, and thus, it is likely to be filled with small files. Detailed comparisons will be shown below.

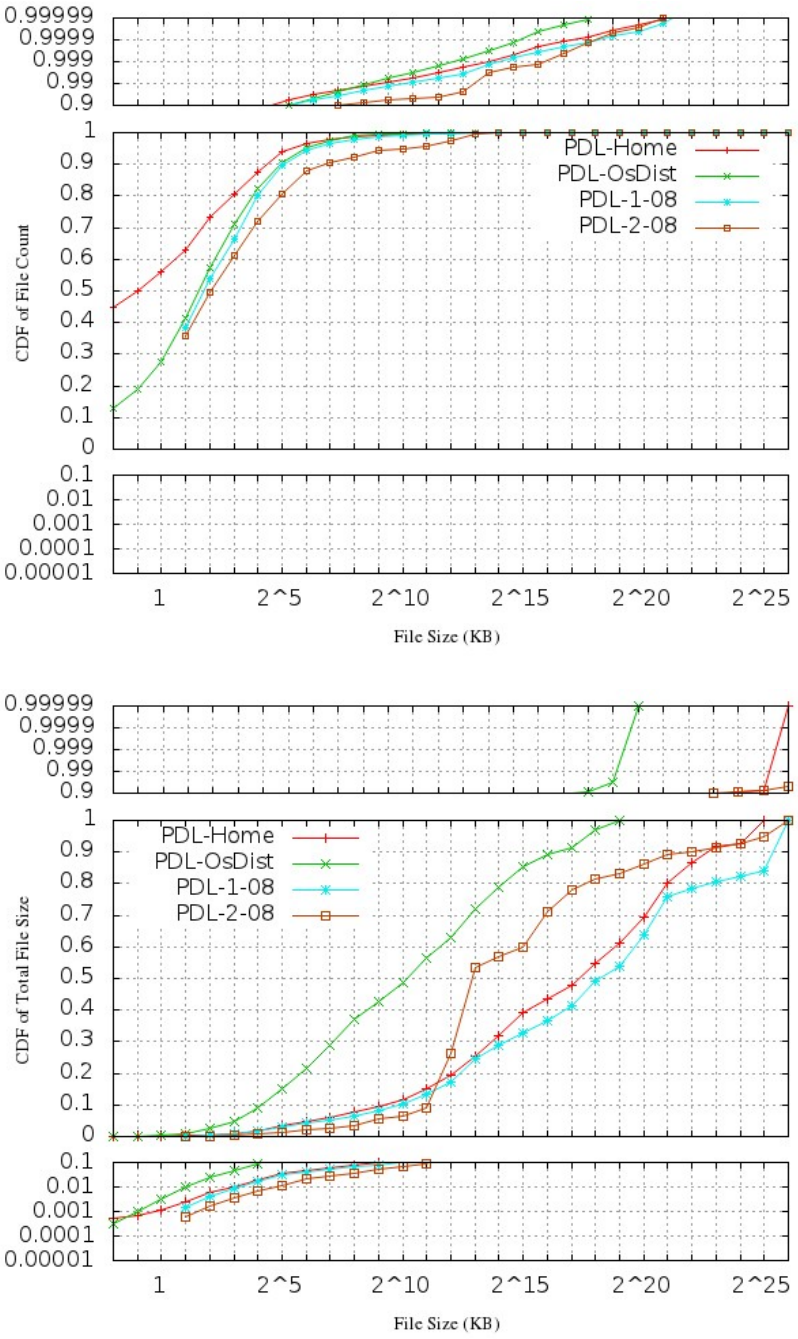

Figure 2: This figure describes the file size distributions on PDL machines. The upper graph shows a CDF of files of given size across PDL files systems. The lower graph shows a CDF of total file space in files of given size across PDL machines. The file systems studied are mainly used for storing home directories, operating system distributions and I/O testing files. The X axis is the last byte offset in kilo bytes. It is in log scale of base 2. The y axis is divided into three sections. It is log scale space from 0.001 percentile to 10 percentile in section 1. Upon part 1, there is linear space from 0 to 100 percentiles in section 2 and log scale space from 90 to 99.999 percentile in section 3. The legends are explained in section 2.

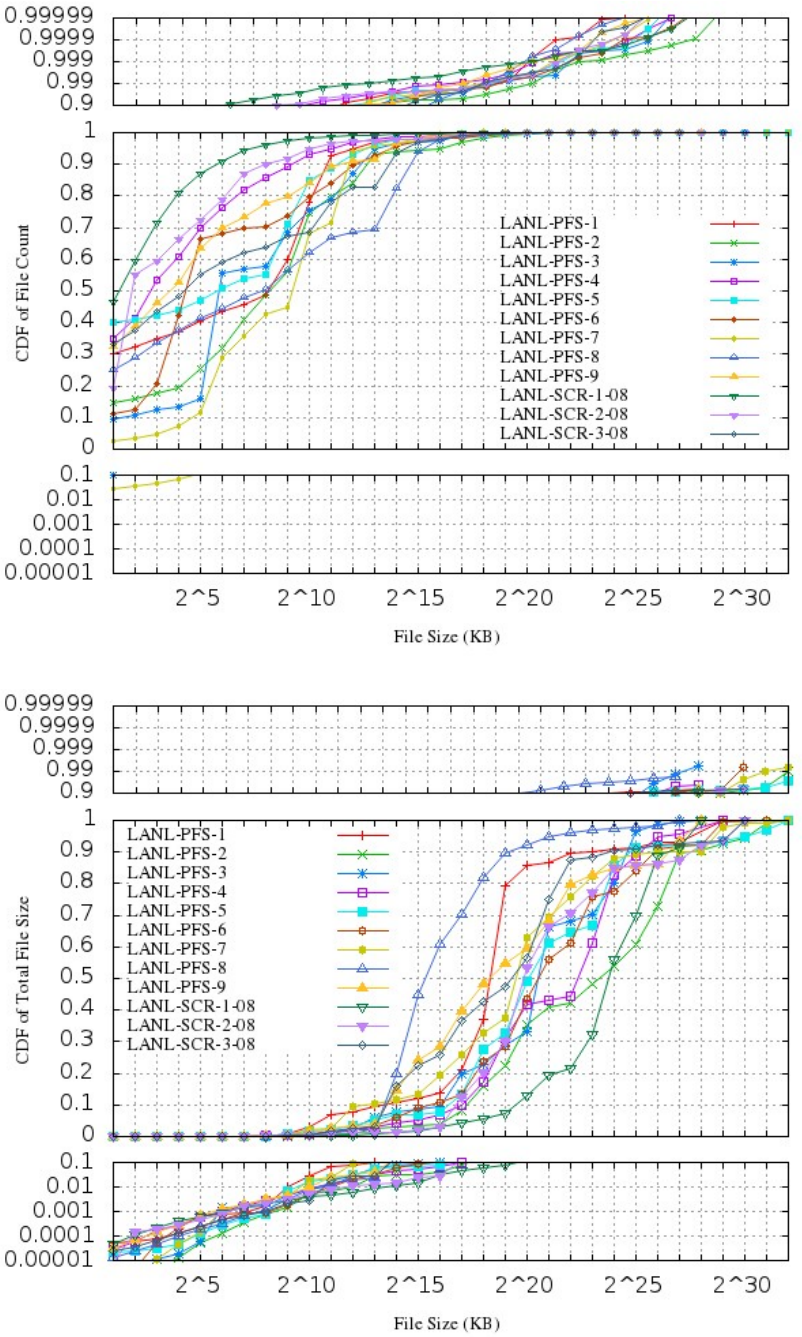

Figure 3: This figure describes the files size distribution on LANL machines. The upper graph shows a CDF of files of given size across LANL files systems. The lower graph shows a CDF of total file space in files of given size across LANL machines. The file systems studied are mainly used as cluster machines. The X axis is the last byte offset in kilo bytes. It is in log scale of base 2. The y axis is divided into three sections. It is log scale from 0.001 percentile to 10 percentile in section 1. Upon part 1, there is linear space from 0 to 100 percentiles in section 2 and log scale space from 90 to 99.999 percentile in section 3. The legends are explained in section 2.

6.1 File Size

For PDL machines, average file size ranges from 0.05 MB to 0.37 MB. Generally, the files on PDL machines 2011 are smaller than the files sampled on 2008. For home directory, average file size has dropped from 370 KB (PDL-1-08), to 90 KB (PDL-Home), as shown in Figure 2.

Moreover, the total file size in home directory also changes greatly; it drops from 3.93 TB (2008) to 1.2 TB (2011).

For LANL machines, file systems are growing larger. The average total size per file system has increased from 20 TB (2008) to 72 TB (2011), and average total number of files per file system has increased from 5.68 million (2008) to 48.4 million (2011). For each of the file system, the average file size has also increased. In 2008, average file size ranges from 6 MB to 10.9 MB, and in 2011, the average file size ranges from 5.9 MB to 37.2 MB. As shown in Figure 3, LANL-PanFS-2 has the biggest average file size, which is 37.2 MB. LANL-PanFS-7 has the biggest file, which is 2.7 TB. LANL-PanFS-7 is used for non-production testing and this might explain the existence of the big file. There is no clear change in the distribution over four years. When we compare Figure 2 and Figure 3, we can find that small files (smaller than 64 KB) take majority of total file number (greater than 90%). But disk is mainly occupied by big files.

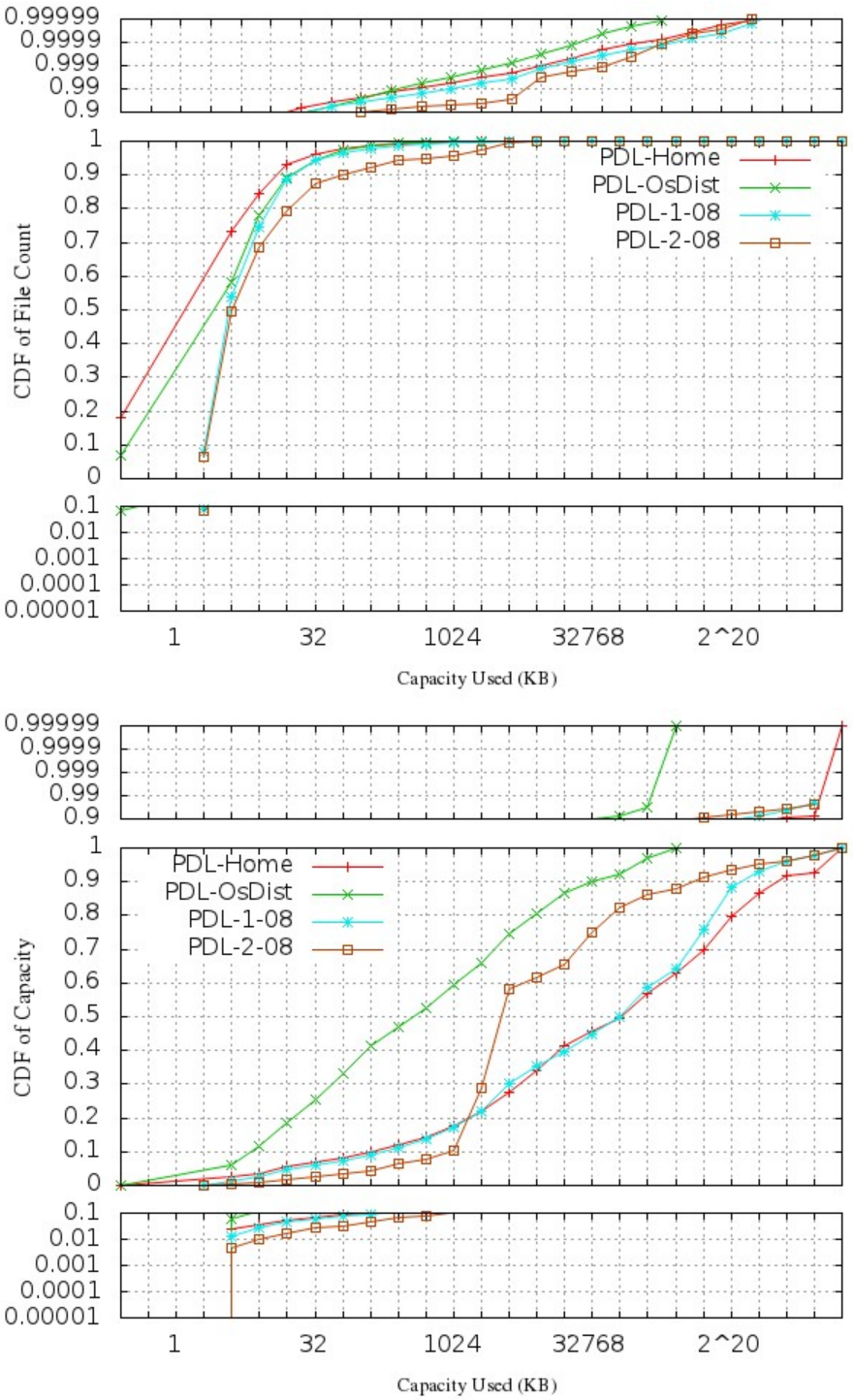

Figure 4: This figure describes the capacity distribution on PDL machines. The upper graph shows a CDF of files of given capacity across PDL files systems. The lower graph shows a CDF of total capacity in files of given size across PDL machines. The file systems studied are mainly used for storing home directories, operating system distributions and I/O testing files. The X axis is the consumed disk space in kilo bytes. It is in log scale of base 2. The y axis is divided into three sections. It is log scale from 0.001 percentile to 10 percentile in section 1. Upon

part 1, there is linear space from 0 to 100 percentiles in section 2 and log scale space from 90 to 99.999 percentile in section 3. The legends are explained in section 2. On the lower graph, some curves start from 4 KB for the sake that 4 KB is the smallest block size specified in that file system.

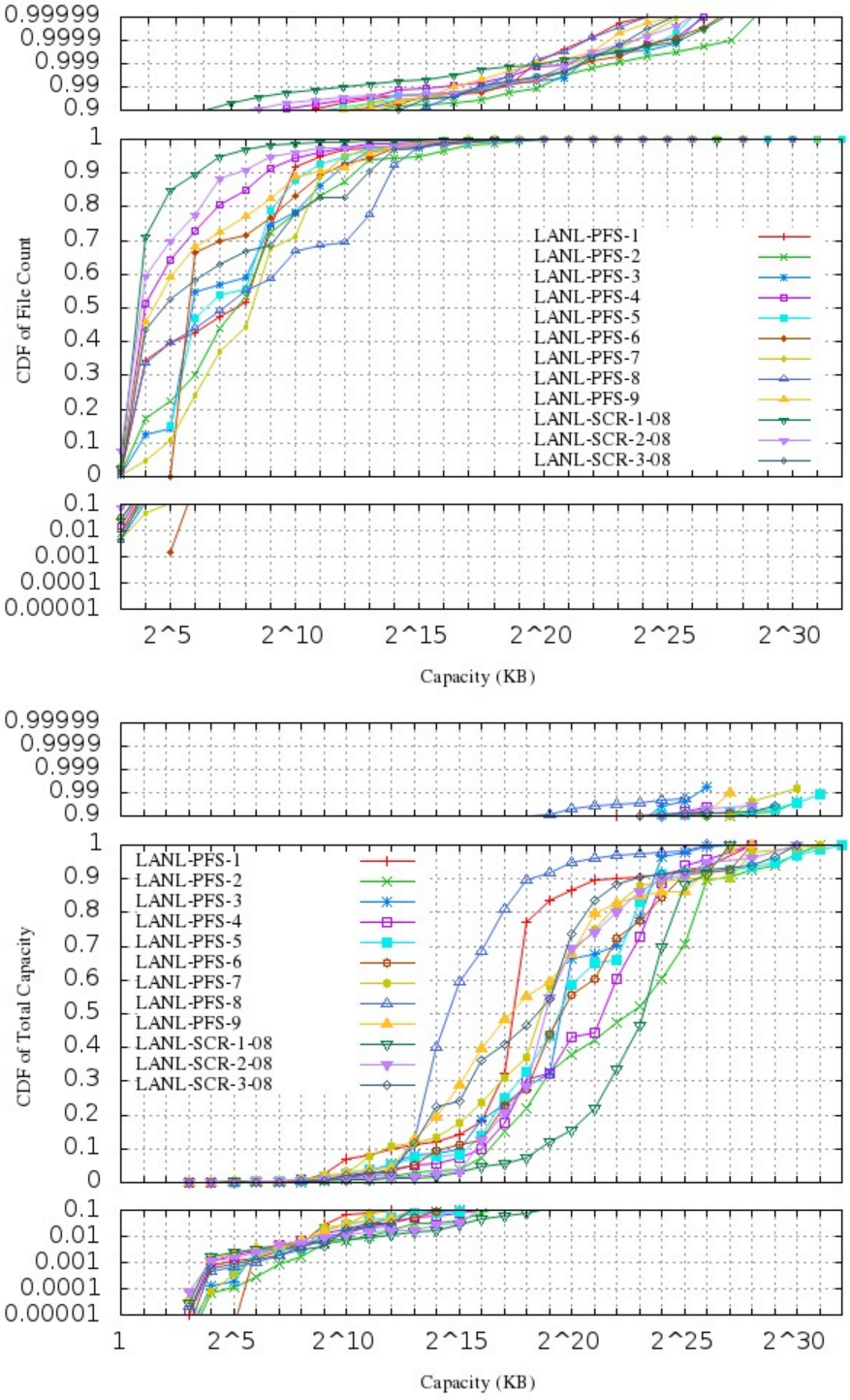

Figure 5: This figure describes the capacity distribution on LANL machines. The upper graph shows a CDF of

files of given capacity across LANL files systems. The lower graph shows a CDF of total capacity in files of given size across LANL machines. The file systems studied are mainly used as cluster machines. The X axis is the consumed disk space in kilo bytes. It is in log scale of base 2. The y axis is divided into three sections. It is log scale from 0.001 percentile to 10 percentile in section 1. Upon part 1, there is linear space from 0 to 100 percentiles in section 2 and log scale space from 90 to 99.999 percentile in section 3. The legends are explained in section 2. On the lower graph, some curves start from 8 KB for the sake that 8 KB is the smallest block size specified in that file system.

6.2 Capacity

As shown in Figure 4, in PDL machines, median of capacity ranges from 512 KB to 64 MB. We notice that curves for PDL-1-08 and PDL-2-08 start from 4 KB, curves for PDL-Home and PDL-OsDist start from 0.25 KB. This is because on PDL machines, the block size is 4 KB, so the distribution curve starts from 4 KB on PDL-1-08 and PDL-2-08. On PDL-Home and PDL-OsDist, some files take zero blocks and hence they are put into the bucket of [0, 0.25KB] in the histogram in our report. When comparing the four curves in Figure 4, we can find that PDL-OsDist has comparatively more files using small number of blocks. Its average capacity size is 512 KB and the second smallest average capacity is 2 MB.

Figure 5 presents the distributions for capacity on LANL machines. We can easily find that all curves start from 8 KB, which is the block size on LANL machines. The median ranges from 64 MB to 16 GB, which is significantly larger than PDL capacity. Similar to PDL-OsDist, LANL-PanFS-8 also has more files using small number of blocks, when compared with other machines. On the other hand, PDL-SCR1-08 has small number of large files (less than 10%) occupying more than 90% of total disk space.

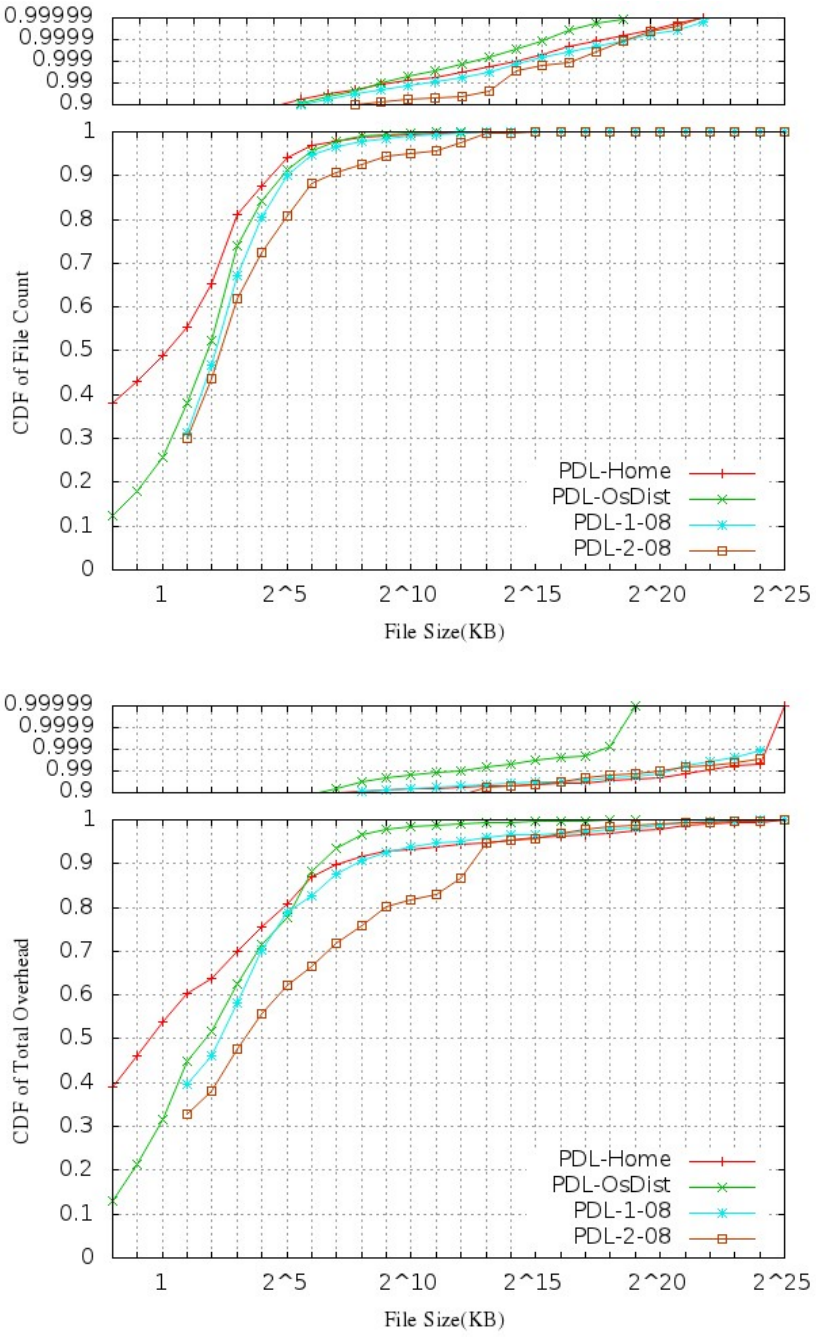

Figure 6: This figure describes the positive overhead distribution on PDL machines. The upper graph shows a CDF of files of given positive overhead across PDL files systems. The lower graph shows a CDF of total positive overhead in files of given overhead size across PDL machines. The file systems studied are mainly used for storing home directories, operating system distributions and I/O testing files. The X axis is the overhead size in kilo bytes. It is in log scale of base 2. The y axis is divided into two sections. There is linear space from 0 to 100 percentiles in section 1 and log scale space from 90 to 99.999 percentile in section 2. The legends are explained in section 2.

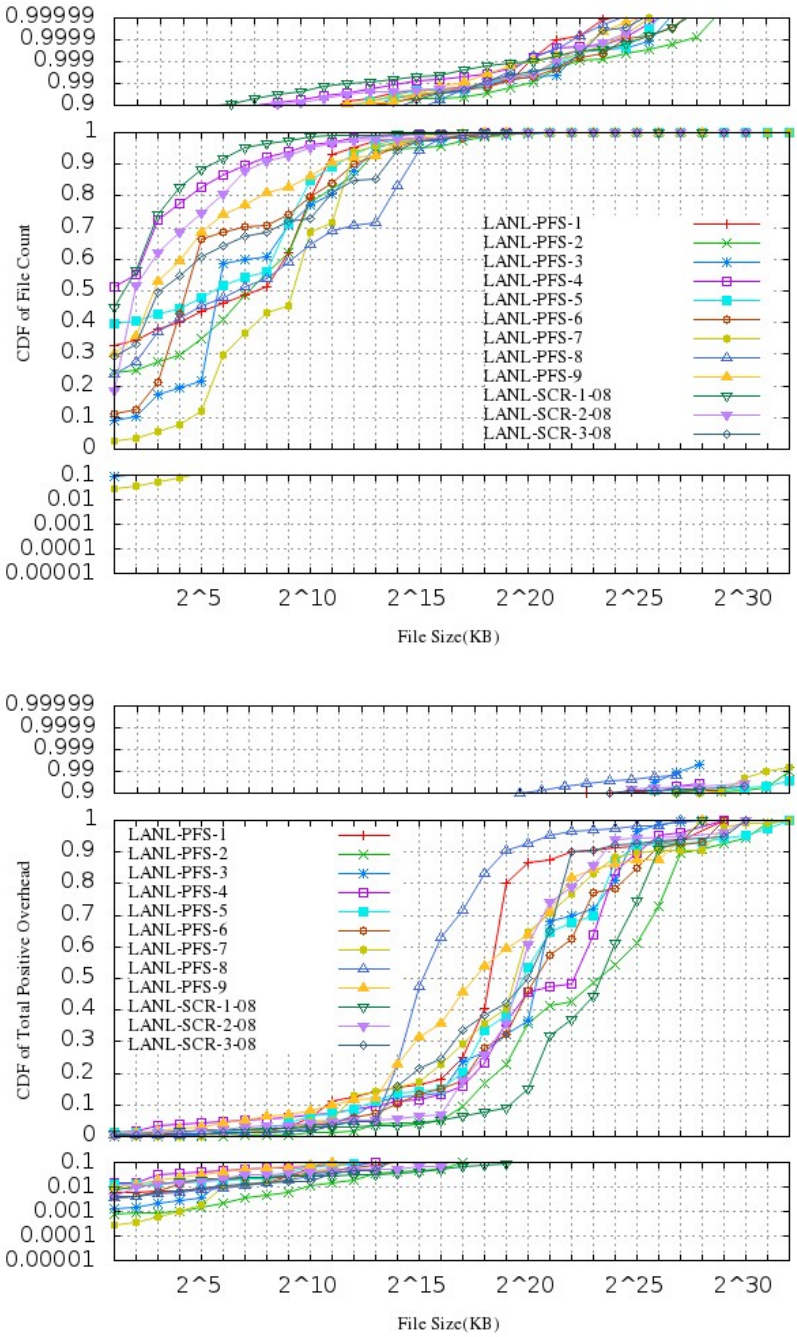

Figure 7: This figure describes the positive overhead distribution on LANL machines. The upper graph shows a CDF of files of given positive overhead across LANL files systems. The lower graph shows a CDF of total positive overhead in files of given overhead size across LANL machines. The file systems studied are mainly used as cluster machines. The X axis is the overhead size in kilo bytes. It is in log scale of base 2. The y axis is divided into three sections. It is log scale from 0.001 percentile to 10 percentile in section 1. Upon part 1, there is linear space from 0 to 100 percentiles in section 2 and log scale space from 90 to 99.999 percentile in section 3. The legends are explained in section 2.

6.3 Positive Overhead

The mean positive overhead on PDL machine ranges from 2.55 KB to 2.8 KB. The median for positive overhead lies within from 1 KB to 32 KB. The total positive overhead ranges from 1.29 GB to 35.3 GB, and PDL-Home has the most positive overhead. When we look at Figure 6, small files make a major contribution to the positive overhead both in number of files having positive overhead and in total positive overhead size. This can be explained for two reasons. First, the positive overhead histogram bucket starts from zero, which means that files with zero positive overhead are also counted into the distribution calculation. Secondly, internal fragmentation, due to fixed block size on file systems, is believed to be a major source for positive overhead. Therefore, positive overhead has limited relationship with original file size. As stated in 6.1 in this paper, files with small file size take a majority in total number of files and hence, they can make a major contribution to the total positive overhead. For instance, we can find that in Figure 6, files smaller than 32 KB take more than 80% in total file number and around 70% in the total positive overhead size. Moreover, PDL-Home has the most number of files (14 million) and it has also the biggest total positive overhead (35.3 GB). On the contrary, PDL-OsDist has the smallest number of files (0.5 million) and it also has the smallest total size of positive overhead, which is 1.29 GB.

In 2008, LANL machines have a positive overhead ranging from 725 KB to 1172 KB. After four years, we find that there is a significant amount of increase in positive overhead. In 2012, as we can find in Figure 7, positive overhead ranges from 521 KB to 4782 KB and more than 67% of file systems have a positive overhead larger than 1172 KB.

One thing comes to our attention is that on LANL-PanFS-5, a file with positive overhead as large as 0.5 TB is discovered. As general belief, this means that a large chunk of space is wasted and we will investigate this with LANL to discover the reason. LANL-PanFS-7 has 57.5 TB positive overhead in total which is around 16.7% of its total size.

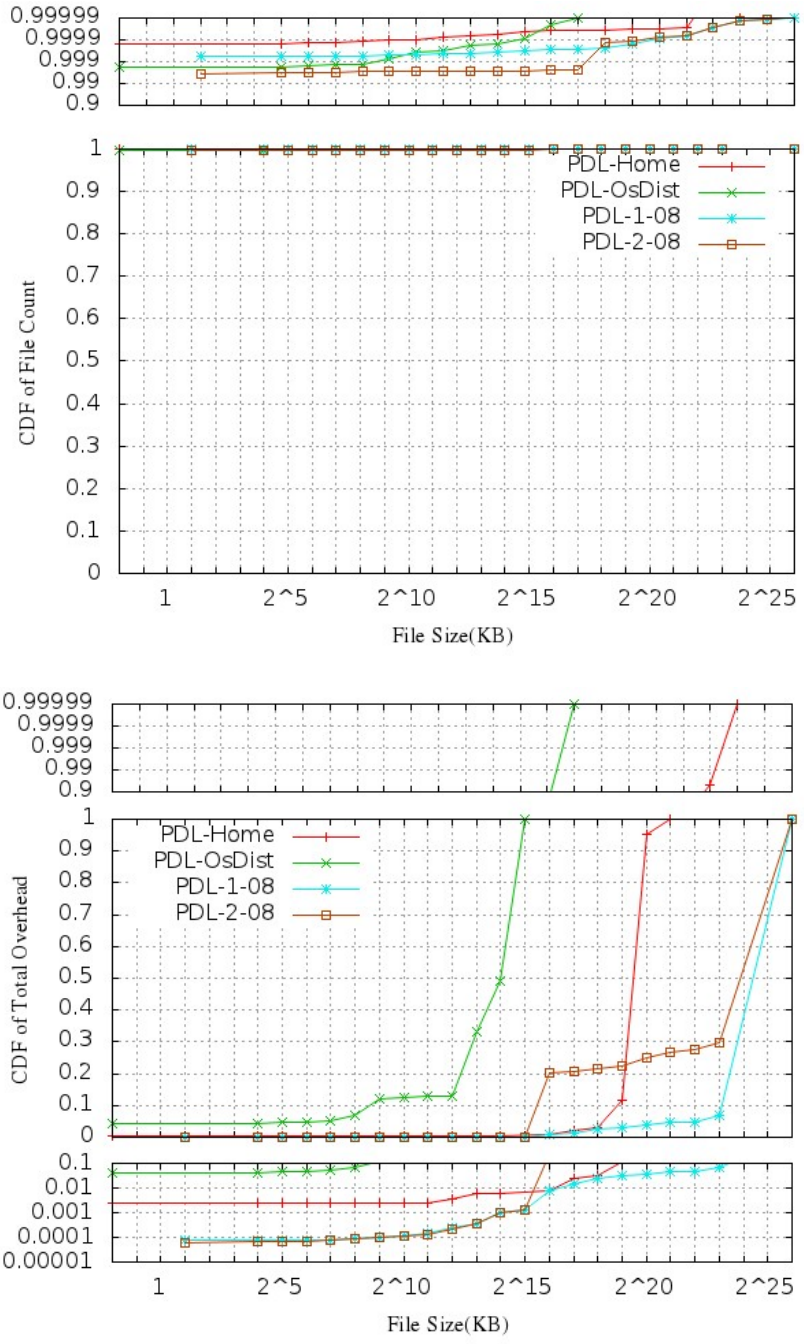

Figure 8: This figure describes the negative overhead distribution on PDL machines. The upper graph shows a CDF of files of given negative overhead across PDL files systems. The lower graph shows a CDF of total negative overhead in files of given overhead size across PDL machines. The file systems studied are mainly used for storing home directories, operating system distributions and I/O testing files. The X axis is the overhead size in kilo bytes. It is in log scale of base 2. On the bottom, the y axis is divided into three sections. It is log scale space from 0.001 percentile to 10 percentile in section 1. Upon part 1, there is linear space from 0 to 100 percentiles in section 2 and log scale space from 90 to 99.999 percentile in section 3. On the top, only

section 2 and section 3 of y axis are presented. The legends are explained in section 2 of the paper.

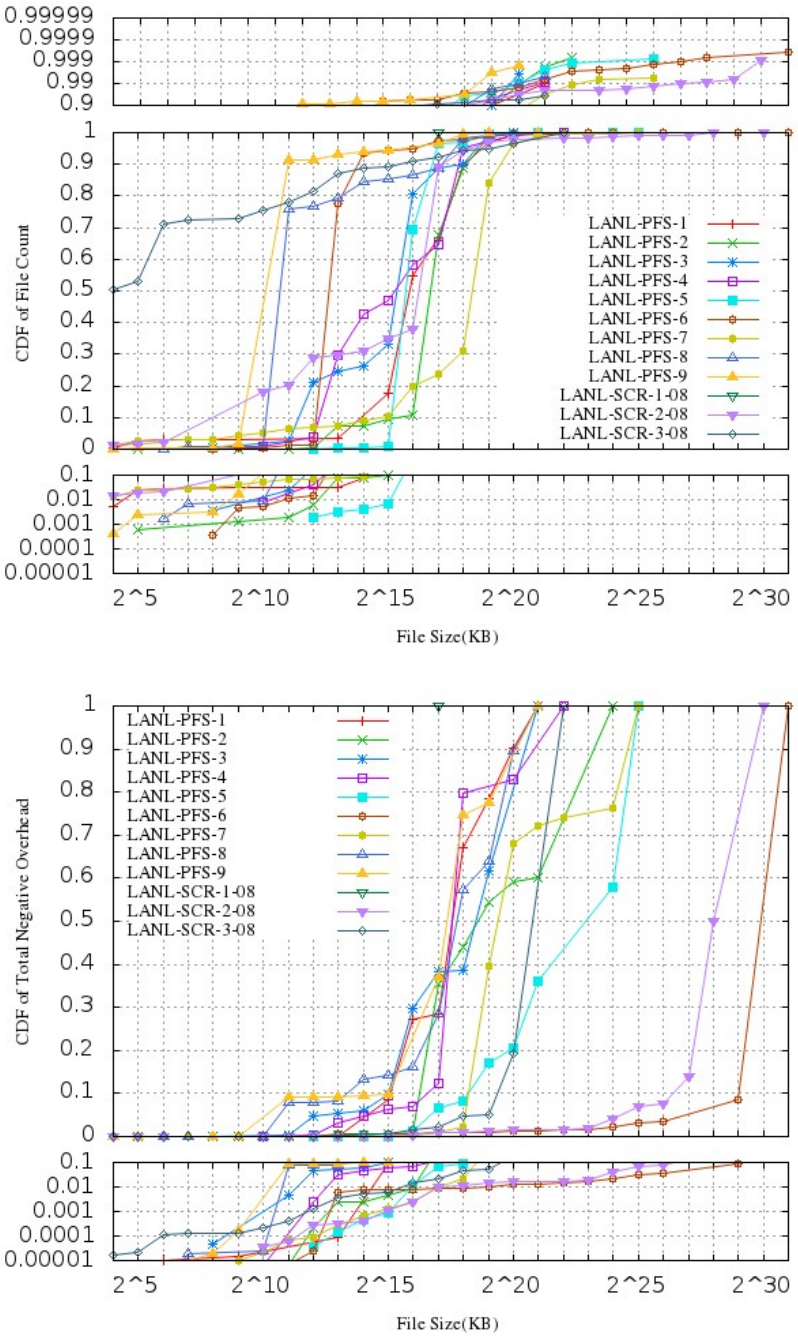

Figure 9: This figure describes the negative overhead distribution on LANL machines. The upper graph shows a CDF of files of given negative overhead across LANL files systems. The lower graph shows a CDF of total negative overhead in files of given overhead size across LANL machines. The file systems studied are mainly used as cluster machines. The X axis is the overhead size in kilo bytes. It is in log scale of base 2. On the top, the y axis is divided into three sections. It is log scale space from 0.001 percentile to 10 percentile in section 1.

Upon part 1, there is linear space from 0 to 100 percentiles in section 2 and log scale space from 90 to 99.999 percentile in section 3. On the bottom, only section 1 and section 2 of y axis is presented. The legends are explained in section 2.

6.4 Negative Overhead

Negative overhead means one file using disk space (represented by the number of blocks used) smaller than its last byte offset. Generally, three reasons might cause negative overhead: existence of log file, using lseek and existence of virtual machine image.

For PDL machines, negative overhead is becoming smaller. In 2011, average negative overhead ranges from 0.5 KB to 14 KB. As a comparison, in 2008, negative overhead ranges from 360 KB to 440 KB. This might be related to change of virtual machine image, since in PDL, virtual machine image is stored for studying and testing. The maximum size of negative overhead is also becoming smaller. In 2011, maximum negative overhead is 0.6 GB, found on PDL-Home. As a comparison, PDL-1-08 has maximum negative overhead of 35.7 GB. Reading Figure 7, we can find that 99% of files with negative overhead have overhead size less than 4KB. Moreover, there is a significant right shift in Figure 8 from 2011 curves to 2008 curves. This means that small files now contribute more significantly to the total negative overhead.

On LANL machines, some strange super big negative overhead exists. On LANL-PanFS-6, a file with 0.94 TB negative overhead is discovered. From a recent update from LANL, we know that the 0.94 TB negative overhead is made by one user. In 2008, big negative overhead (1.27 TB) is also discovered. All the 2008 and 2011 files with big negative overhead were found to be empty files with big last byte offset is created. There is no direct relationship between the work pattern and the creation of files with big negative overhead. Without Panfs-6 and PanFS-7, in 2011 LANL-PanFS file systems have negative overhead ranging from 0.69 GB to 50.7 GB. The average negative overhead ranges from 1.02 M to 256 MB, which is significantly larger than that on PDL machines.

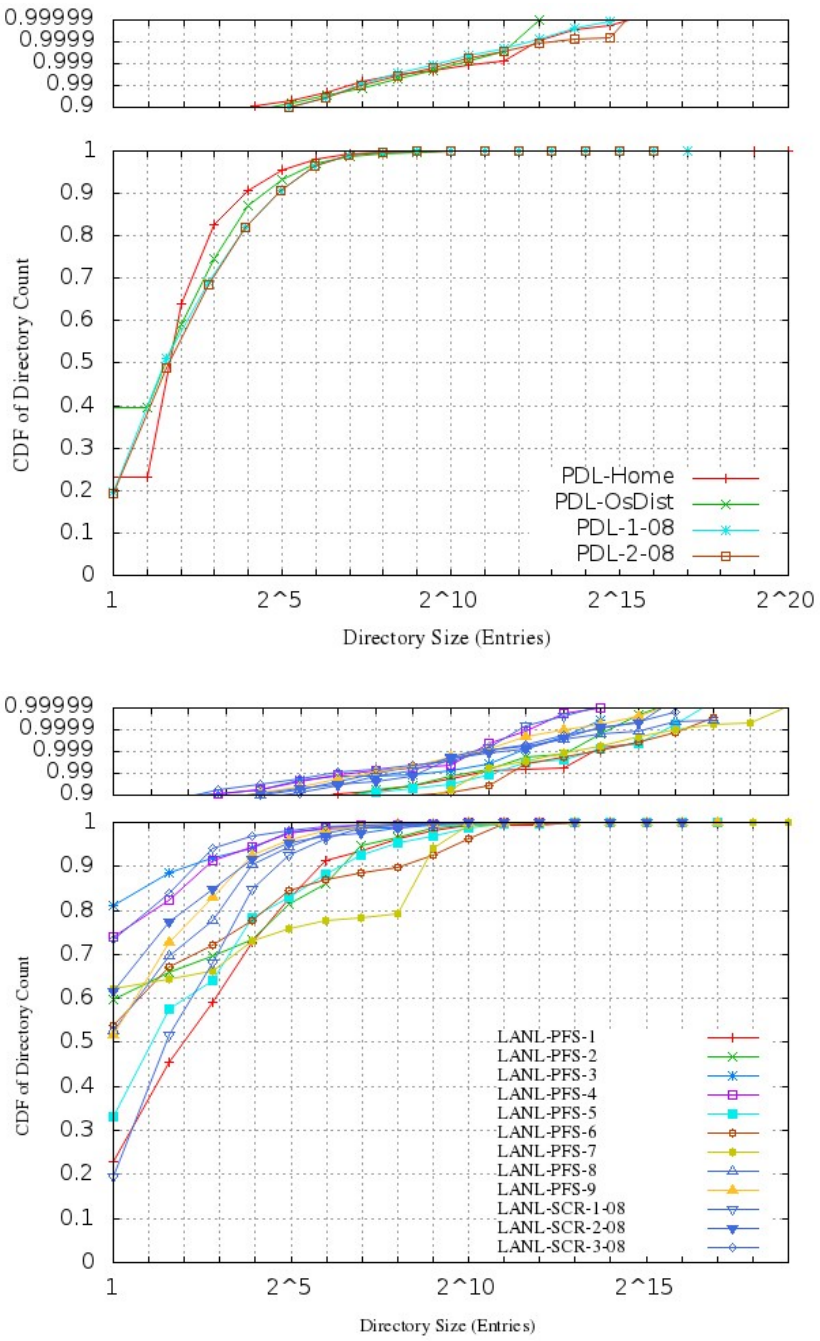

Figure 10: The figure describes directory size distribution on PDL machines and LANL machines. On the upper side, PDL machines directory size distributions are presented. On the lower graph, directory size distributions are presented for LANL machines. The directory size is measured by the number of entries within that directory. The X axis is the number of entries in the directory in log scale of 2. For both the graphs, the y axes are divided into 2 sections. There is linear space from 0 to 100 percentiles in section 1 and log scale space from 90 to 99.999 percentile in section 2.

6.5 Directory Size

In order to allow for a straightforward understanding into how big a directory is, we use the total number of entries lying in the directory as the size of the directory. Entry includes regular files, soft links and sub directories. Among PDL machines, medians of directory size range from 2 to 4 entries per directory, and over 90% of directories have less than 64 entries. The average directory size ranges from 12 entries per directory to 15 entries per directory. In 2011, PDL-Home has one directory with 571,665 entries under it and it is the biggest directory, which is much bigger than biggest directory size, 89,517, on PDL-2-08, 2008. In 2011, PDL-Home also has the largest total number of directories, which is 1269 K, compared to 821 K, the maximum number of 2008. When we compare home directories (PDL-Home and PDL-1-08), there is no significant change in the distribution, though the total amount of files and file size are quite different. This means that the way people organize their files has not changed significantly.

For LANL machines, firstly we can find that there is a significant increase in directory size from 2008 to 2011. In 2008, average entry number ranges from 8 to 15. In 2011, average ranges from 7 to 120. Actually, only LANL-PanFS-4 has average entry number smaller than 15. If we do not consider it, then the smallest average entry number is 16, even larger than the biggest average entry number in 2008. As can be seen in Figure 10, the medians of entry number range from one entry to four entries. The biggest directory we can find is on LANL-PanFS-4, which contains 490,399 directories, and this is much bigger than the biggest value, 50,000, in 2008.

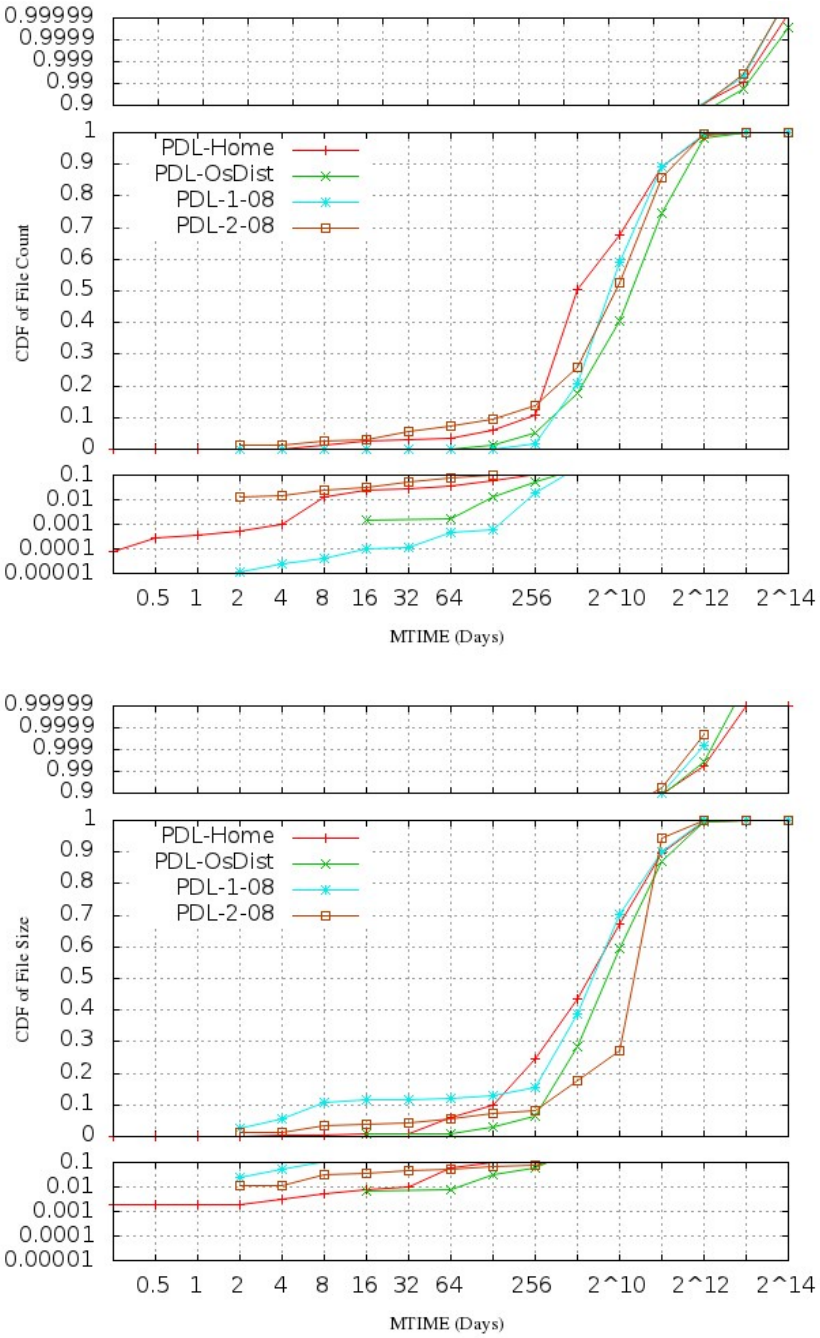

Figure 11: This figure describes the file age distribution on PDL machines. The upper graph shows a CDF of files of given age. The lower graph shows a CDF of file size of given file modification time. The file systems studied are mainly used for storing home directories, operating system distributions and I/O testing files. The X axis is the modification time in days. It is in log scale of base 2. The y axis is divided into three sections. It is log scale space from 0.001 percentile to 10 percentile in section 1. Upon part 1, there is linear space from 0 to 100 percentiles in section 2 and log scale space from 90 to 99.999 percentile in section 3. The legends are explained in section 2 of the paper.

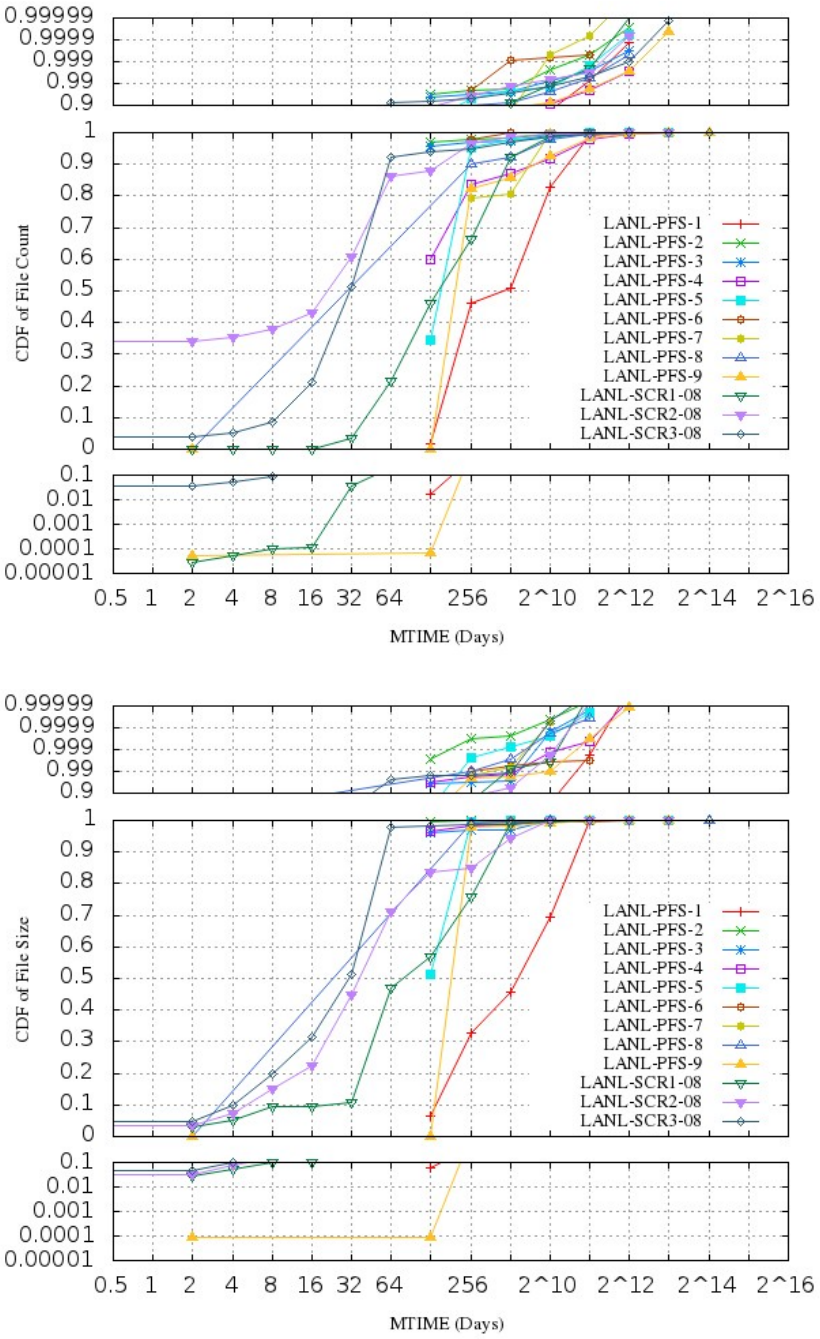

Figure 12: This figure describes the file age distribution on LANL machines. The upper graph shows a CDF of files of given age. The lower graph shows a CDF of file size of given file modification time. The file systems studied are mainly used as cluster machines. The X axis is the modification time in days. It is in log scale of base 2. The y axis is divided into three sections. It is log scale space from 0.001 percentile to 10 percentile in section 1. Upon part 1, there is linear space from 0 to 100 percentiles in section 2 and log scale space from 90 to 99.999 percentile in section 3. The legends are explained in section 2 of the paper.

6.6 Modification Time

Among PDL machines, average modification time ranges from 920 days to 1143 days. PDL-Home has the smallest average modification time, which is 920 days. PDL_OsDist has the biggest modification time, which is 1430 days. This may be related to its usage, holding operating system distributions for administration use. Median of modification time on PDL machines ranges from 256 days to 1024 days. It also comes to our notice that the home directory of 2011 also has lots of old files. As can be seen on Figure 11, on PDL-Home, no more than 30% of total storage is taken by files younger than one year old. Oldest file on PDL-Home can be as old as 15228 days. We assume that it is some old system file kept in the operating system. It also needs to be mentioned that the youngest file we discover on PDL-Home is -3516 days, which means it shows a date in the future. This is a sign of manual change, like using touch command.

On LANL machines as can be seen in Figure 12, the median of modification ranges from 16 days to 512 days. In 2008, average modification ranges from 65 days to 620 days. In 2011, average modification time ranges from 74 days to 333 days, which indicates that files on LANL-PanFS machines are much younger than files on PDL machines. We also find some very old files on LANL machines. The oldest file is 15428 days old, found on LANL-PanFS-1. As can be seen on Figure 12, LANL-PanFS-1 holds files much older than other machines and its curve is on the left-most position. Except LANL-PanFS-1, around 90% of files are younger than 256 days. 90% of total storage space is taken by files younger than 512 days old. We also notice that over 4 years, there is no significant change in file age on LANL machines, and no files have a negative age.

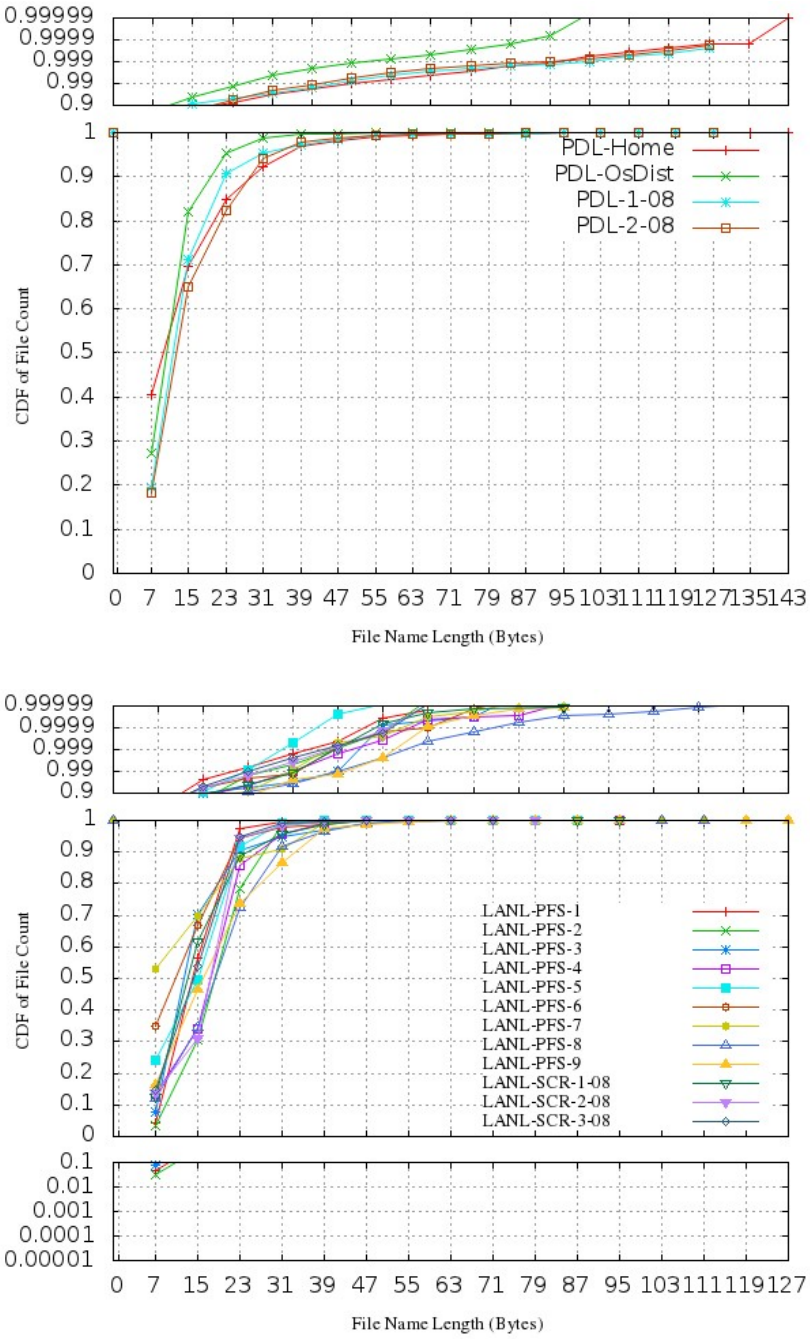

Figure 13: The figure describes file name length on PDL machines and LANL machines. On the bottom, file name length distributions are presented for LANL machines. The y axis is divided into three sections. It is log scale space from 0.001 percentile to 10 percentile in section 1. Upon part 1, there is linear space from 0 to 100 percentiles in section 2 and log scale space from 90 to 99.999 percentile in section 3. On the top, PDL machine file name length distributions are presented. Section 2 and Section 3 of y axis are presented. Both X axes are the length of the file name in the characters in linear scale. The legends are explained in section 2.

6.7 File Name Length

On PDL machines, average file name length ranges from 11 characters to 15 characters. The median ranges from 7 characters to 15 characters. The longest file name we found on PDL machine is 143 characters from PDL-Home. Over four years, there is no significant change in the file name distribution, which means that the naming pattern for PDL people remains largely the same.

LANL machines, as can be seen in Figure 13, have medians ranging from 7 characters to 15 characters. More than 90% of files have name length less than 31 characters. The longest name length we find is 127 characters on LANL-PanFS-9. In figure 13, all curves are quite close to each other and this indicates that there is no significant change in the distribution of file name length. We also need to note here that LANL generates the file name length data from un-anonymized directory tree. It uses the real file name instead of a fake one. Based on observations on PDL and LANL machines, we can know that people tend to give their files meaningful names, which are generally less than 32 characters.

7. Results – Peer Comparison

In this section, we will compare data generated from LANL and PDL. Similar to previous section, in order to provide concrete comparison, we will also compare the data based on similar usage.

For LANL machines, we will present the result generated from LANL-LNFS, LANL-GNFS and LANL-ARCH. LANL-LNFS is used by local cluster only, and it is not shared with any outside-cluster user. LANL-GNFS can be accessed by all people in LANL and hence its volume name is global NFS file system. Both of them are using NFS file system. Moreover, LANL-LNFS and LANL_GNFS are also used to hold home directories, which are very similar to PDL-Home. LANL-ARCH is used to hold to archive and it runs GPFS file system. It has a tape system using HPSS.

For PDL machines, besides PDL-Home and PDL-OsDist, which are same to previous section, we also introduce PDL-Cirrus and PDL-VM. Part of PDL-Cirrus is used to store open source software for Cirrus Operating System and Cirrus Operating System Distributions, and the other part is used for administration use. PDL-VM is used to store virtual machine image and very small amount of files are stored there.

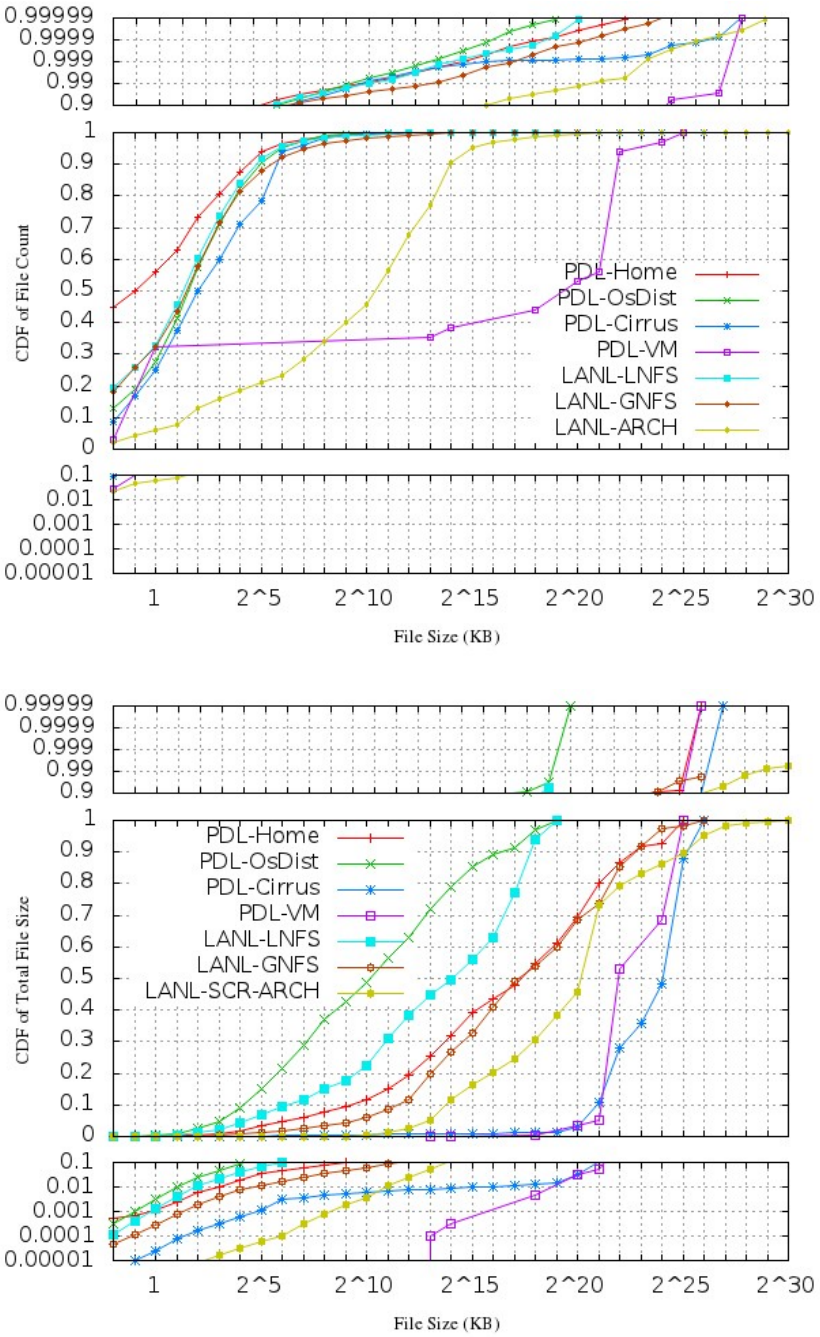

Figure 14: This figure describes the files size distribution on PDL and LANL machines. The upper graph shows a CDF of files of given size across PDL and LANL files systems. The lower graph shows a CDF of total file size in files of given size across PDL and LANL machines. The file systems studied are mainly used for storing home directories. The X axis is the last byte offset in kilo bytes. It is in log scale of base 2. The y axis is divided into three sections. It is log scale space from 0.001 percentile to 10 percentile in section 1. Upon part 1, there is linear space from 0 to 100 percentiles in section 2 and log scale space from 90 to 99.999 percentile in section 3. The legends are explained in section 2.

7.1 File Size

As can be seen in Figure 14, PDL-VM and LANL-ARCH are significantly different from other machines. For PDL-VM, this is because it has too small number of files (34 only). Therefore, its statistics can be influenced by particular files greatly. For LANL-ARCH, it is used as archive file system and therefore, the files stored within it are different from others. For LANL-ARCH, half of its files are smaller than 2 GB and average files size of it is 21 MB, which is significantly larger than other file systems. LANL-ARCH is also the biggest file system within the 8 file systems; its total size is 52.75 TB.

Without LANL-ARCH and PDL-VM, we can see that there is no significant difference between curves in Figure 14, which means that the usage is generally similar. It comes to our notice that PDL-Home is to the right of all the other curves, which indicates that small files take a larger percentage in it, compared with other machines. Within the remaining six machines, PDL-Cirrus has the biggest average file size, which is 3.78 MB. The medians of file size lie within 0.5 KB to 4 KB. We can also find that the files in LANL home directories (LANL-LNFS and LANL-GNFS) are bigger than that on PDL machines, since LANL home directory curves are to the right of PDL-Home.

7.2 Capacity

LANL does not provide us with the distribution of disk block usage per file. As a consequence, we cannot compare capacity, positive overhead and negative overhead between the home directories of LANL and PDL. However, we know that the smallest block size on LANL NFS machines are 256 KB, compared to 4 KB block size on PDL machines. We believe that this might be related to its storing bigger files. We also think that the big block size will have a big impact on positive and negative overhead.

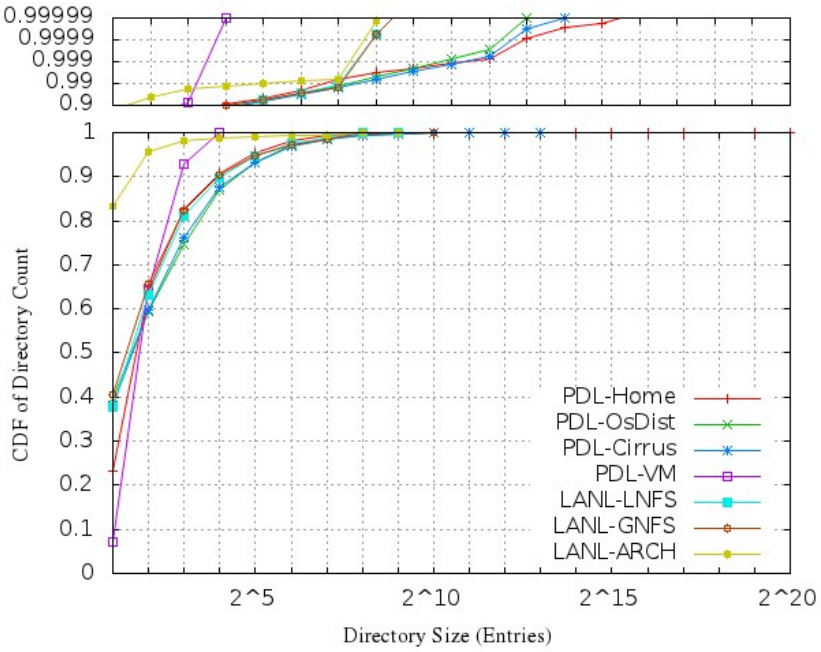

Figure 15: This figure describes the directory size distribution on PDL and LANL machines. The graph shows a CDF of files of given directory size across PDL and LANL files systems. The directory size is measured in the number of entries held within it. The file systems studied are mainly used for storing home directories. The X axis is the entry number. It is in log scale of base 2. The y axis is divided into two sections. There is linear space from 0 to 100 percentiles in section 1 and log scale space from 90 to 99.999 percentile in section 2. The legends are explained in section 2.

7.3 Directory Size

Here we use the number of entries in a directory as the metric to gauge the directory size. As can be seen in the graph, LANL-ARCH and PDL-VM have a different distributions from all the other file systems. LANL-ARCH has the smallest number of entries per directory. On average; one directory in LANL-ARCH contains only 2 entries. LANL-ARCH also has the most number of directories, 47356 K. PDL-VM has a very simple directory structure. It only has 20 regular files and 14 directories. Therefore, its distribution can be impacted by particular directories greatly.

Except LANL-ARCH and PDL-VM, let's take a closer look at figure 15. Directories of LANL NFS machines (LNFS and GNFS) have, on average, 8 entries within it. On LANL machine, the biggest directory contains 565 entries and it is located on LANL-GNFS. LANL-GNFS also has the most directories, which is 15 K. For PDL machine, its average directory size ranges from 12 entries to 14 entries per directory. PDL-Home has the most directories, which is 1269 K. PDL-Home also has the biggest directory, which has 571,665 entries under it. As a general trend, we can find that more than 90% of directories have less than 32 entries in it.

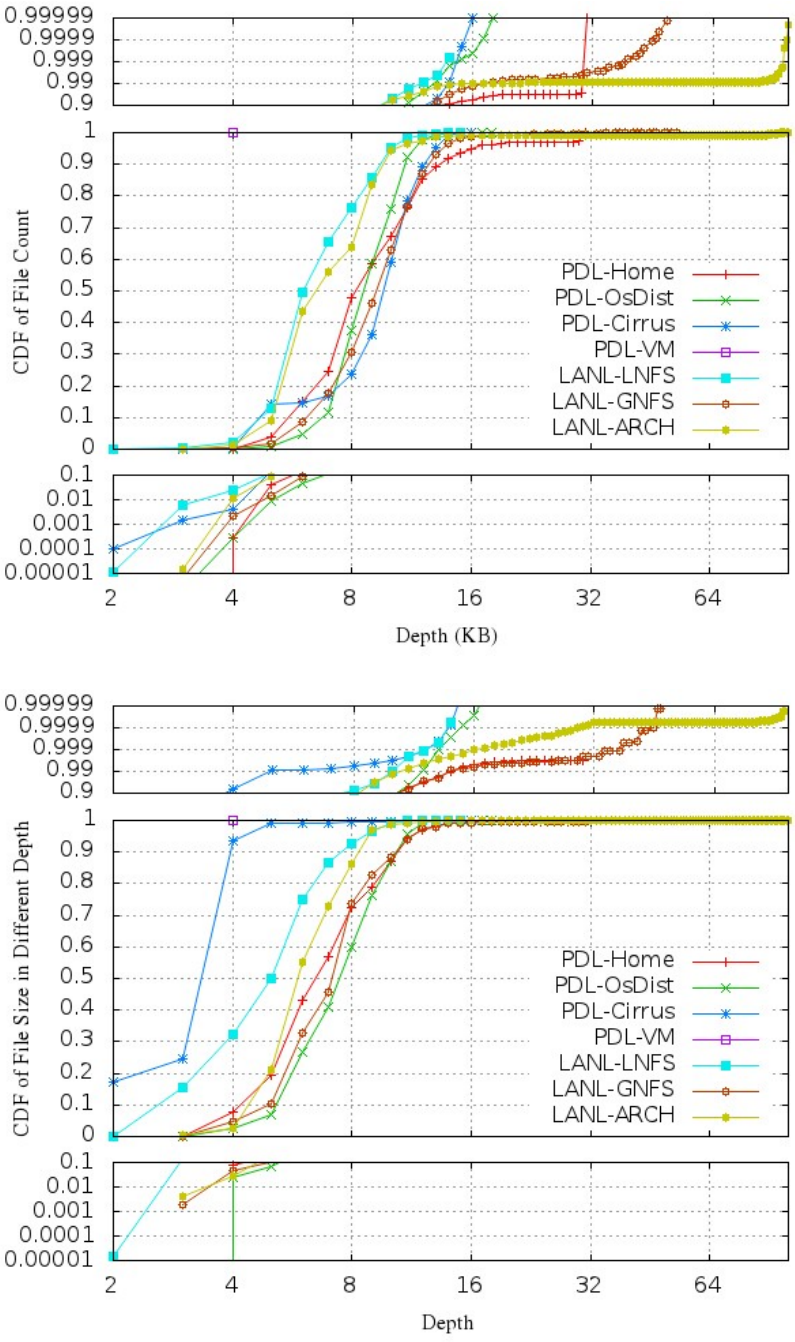

Figure 16: This figure describes the name space property on PDL and LANL file systems. On the top, the total number of files lying at a particular depth in the name space is studied. On the bottom, the total amount of storage space used at a particular depth in the name space is studied. The X axis is the depth in the name space. Before Depth 32, it is in linear scale and afterwards it is in log scale of base 2. The y axis is divided into three sections. It is log scale space from 0.001 percentile to 10 percentile in section 1. Upon part 1, there is linear space from 0 to 100 percentiles in section 2 and log scale space from 90 to 99.999 percentile in section 3. The legends are explained in section 2.

7.4 Name Space

In this part, we look into name space property of file systems. We mainly study the distribution of files and total storage space consumed at a certain level in the directory. We are mainly interested in how people use the file system and whether there is a change in the user habit between researchers in PDL and in LANL. We hope to know, whether people tend to store big files at shallower depth in namespace or deeper in name space, and whether people tend to store more files shallower or deeper.

In the CDF of File Size graph in Figure 16, PDL-VM is only one dot, this is because it has a very simple name space structure and all its regular files lie on Depth 4 in the directory tree. Looking at other file systems in CDF of File Size graph in Figure 16, we can find that for both PDL machines and LANL machines, more than 80% of disk space is consumed by files placed shallower than Depth 8 in directory tree. PDL-Cirrus tends to put heavy files shallower than the other machines and 99% percent of its total storage is placed shallower than Depth 4.

Now let us look at CDF of File Count graph in Figure 16. We can find that for both PDL and LANL, users tend to put files at Depth 7 to Depth 9. More than 70% of files are put within that depth. This matches the findings of Bolosky based on statistics generated from personal computers [4].

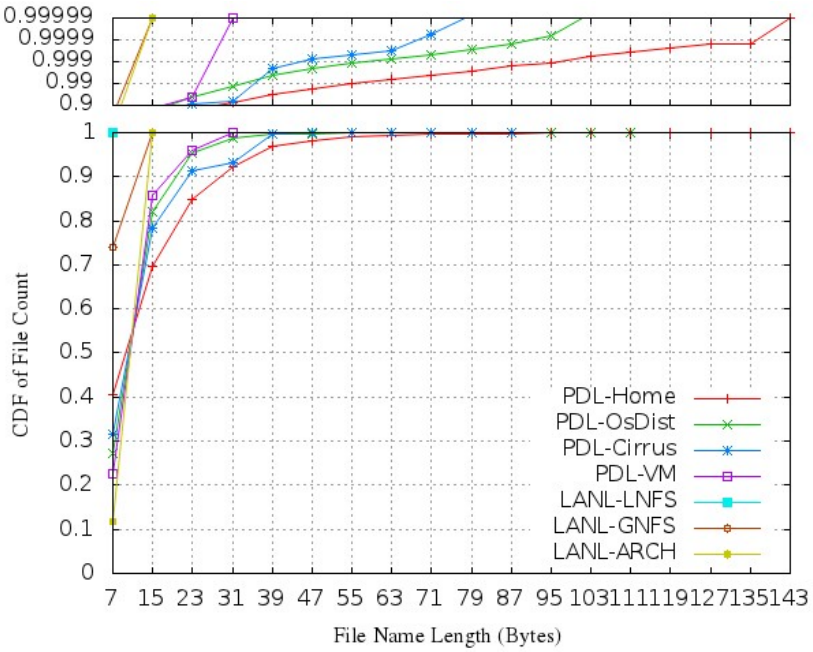

Figure 17: This figure describes the file name length distribution on PDL and LANL machines. The graph shows a CDF of files of given file name length across PDL and LANL files systems. The file name length is measured in characters. The file systems studied are mainly used for storing home directories. The X axis is the file name length in linear scale. The y axis is divided into two sections. There is linear space from 0 to 100 percentiles in section 1 and log scale space from 90 to 99.999 percentile in section 2. The legends are explained in section 2.

7.5 File Name Length.

For LANL NFS machines, average file name length ranges from 6 characters to 8 characters. On the other hand, PDL machines have an average file name length ranging from 10 characters to 14 characters. This might be related to the difference in naming mechanism. In LANL machines, all files are named in string of digits, counting from 0. For instance, there might exist file with full path name 0/1/2. Even under this naming mechanism, LANL-ARCH has file with longest name 15 characters. This is because archive systems hold so many files. As can be seen on Figure 17, on LANL-LNFS, the longest file name length is 7 characters and LANL-GNFS has maximum file name length 8 characters.

For PDL machines, maximum file name ranges range from 24 to 243 bytes, with longest file name appearing on PDL-Home. More than 90% of files haves name less than 31 characters.

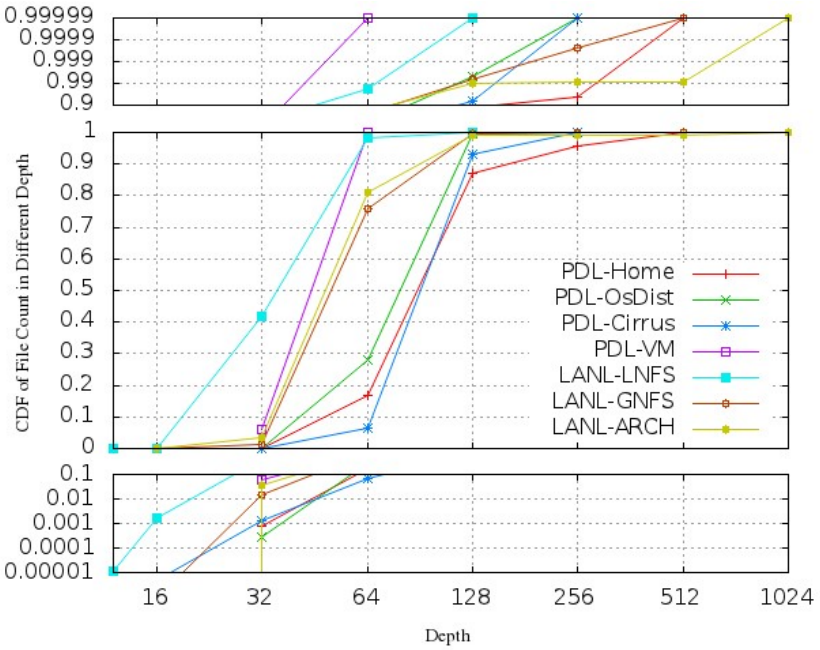

Figure 18: This figure describes the full path name length distribution on PDL and LANL machines. The graph shows a CDF of files of given full path name length across PDL and LANL files systems. The full path name length is measured in characters. The file systems studied are mainly used for storing home directories. The X axis is the full path name length in linear scale. The y axis is divided into three sections. It is log scale space from 0.001 percentile to 10 percentile in section 1. Upon part 1, there is linear space from 0 to 100 percentiles in section 2 and log scale space from 90 to 99.999 percentile in section 3. The legends are explained in section 2.

7.6 Full Path Name Length

In addition to study file name length, we also study the full path name length. This is because, in previous part, we only dig into the length of regular files while neglecting the name length of directories. Moreover, to design a file system using big table [7], the information of total path name is also important.

For LANL NFS machines, its average full path name length ranges from 37 characters to 56 characters. This is comparatively shorter than full path name length on PDL machines and this can be explained by that LANL also uses the same naming mechanism on directories. Its median full path name length ranges from 42 characters to 44 characters. Moreover, as can be seen in Figure 18, the distribution of full path name length for LANL home directories are close to each other and this can also be explained by using same naming mechanism.

For PDL machines average full path name length ranges from 41 bytes to 94 bytes. More than 70% of files have full path name length between 64 characters and 128 characters.

8. Conclusion

We collect data from 22 machines from PDL and LANL. Five of them were collected in 2008, thirteen of them are collected in 2011 and the rest four are collected in 2012. We collected statistics including file size, file capacity, file age (modification time, access time and change time), and file name length, file full path length, file system name space property, positive overhead and negative overhead.

We have offered a concrete comparison on file systems over four years. We divide and compare file systems based on their usage in order to offer a valid comparison. We compare PDL and LANL machines separately on all attributes of interest.

We also provide a comparison over the recent data. We compare the home directories from LANL and from PDL and make solid analysis on some special features of data representation.

Our primary contribution is to offer a valid comparison over time and between different HPC sites. A concrete analysis is also offered.

Our second contribution is to provide the new data to the public. All data will be presented on our websites www.pdsi-scidac.org/fsstats/.

9. Future Works

There are some areas concerning file system meta-data that is not touched upon by this paper. Our future research will focus on the following targets.

- Collect more data from other HPC sites in order to support a more concrete comparison over time. We have data generated from thirteen machines in 2008 and we hope to get more recent data from the same site. This will enable us to do a more concrete comparison.
- Study more deeply into the usage pattern. Instead of studying file in directories, we hope to study files in users' project. We want to dig into project property. This will rely on more sophisticated algorithm to tell projects from others.
- Study the impact on hardware impact. For LANL and PDL, there is very small change in the underlying hardware, and hence we cannot know if the new underlying hardware has an impact on file system meta-data or not. We want enlarge out data source and find the impact brought by hardware update.

References

[1] D. Meyer, W. Bolosky, “A study of practical deduplication,” in Proceedings of the FAST ’11 Conference on File and Storage Technologies, San Jose, CA, USA, 2011. [Online]. Available: www.usenix.org/event/fast11/tech/full_papers/Meyer.pdf . [Accessed: 15 Aug. 2011].

[2] M. Mitzenmacher, “Dynamic models for file sizes and double pareto distributions,” Internet Mathematics, vol. 1, no. 3, pp. 305-333, 2003. [Online]. Available: www.eecs.harvard.edu/~michaelm/postscripts/im2005a.pdf. [Accessed: 17 July. 2011].

[3] K. Evans and G. Kuenning, “A study of irregularities in file-size distributions,” in Proceedings of the International Symposium on Performance Evaluation of Computer and Telecommunication Systems, San Diego, CA, USA, July, 2002. [Online]. Available: www.cs.hmc.edu/~geoff/papers/filesize02.ps. [Accessed: 17 March. 2012].

[4] N. Agrawal, W. Bolosky, J. Douceur, J. Lorch, “A five-year study of file-system metadata,” in Proceedings of the FAST ’07 Conference on File and Storage Technologies, San Jose, CA, USA, 2007. [Online]. Available: research.microsoft.com/pubs/72896/fast07-final.pdf. [Accessed: 17 March. 2012].

[5] S. Dayal, “Characterizing HEC storage systems at rest,” Parallel Data Lab, Carnegie Mellon University, Pittsburgh, PA, USA, 2008. [Online]. Available: www.pdl.cmu.edu/PDL-FTP/PDSI/CMU-PDL-08-109.pdf. [Accessed: 17 March. 2012].

[6] A. Downey, “The structural cause of file size distributions,” in Proceedings of 9th International Workshop on Modeling, Analysis, and Simulation of Computer and Telecommunication System, York, UK, 2001. [Online]. Available: allendowney.com/research/filesize/Downey01Structural.ps.gz. [Accessed: 17 March. 2012].

[7] F. Chang, J. Dean, S. Ghemawat, and etc., “Big Table: A distributed storage system for structured data,” in Proceedings of the OSDI’06, 7th USENIX Symposium on Operating Systems Design and Implementation, Seattle, WA, USA, 2006. [Online]. Available: research.google.com/archive/bigtable-osdi06.pdf. [Accessed: 17 March. 2012].

[8] A. Tanenbaum, J. Herder, and H. Bos, “File size distribution on UNIX systems – then and now,” in ACM SIGOPS Operating Systems Review, vol. 40, issue. 1, 2006. [Online]. Available: dl.acm.org/ft_gateway.cfm?id=1113364&type=pdf. [Accessed: 17 March. 2012].

[9] D. Roselli, J. Lorch, and T. Anderson, “A comparison of file system workloads,” in Proc. of 2000 USENIX Annual Technical Conference, San Diego, CA, USA, 2004. [Online]. Available: www.usenix.org/event/usenix2000/general/full_papers/.../roselli.pdf. [Accessed: 17 March. 2012].

[10] Open Systems Group, Standard Performance Evaluation Corporation, “Network File System Benchmark,” [Online]. Available: http://www.spec.org/benchmarks.html#nfs, 2008. [Accessed: 27 Apr. 2012].

[11] B. Welch, M. Unangst, Z. Abbasi, G. Gibson, and etc., “Scalable performance of the Panasas parallel file system,” in Proc. of the FAST ’08 Conference on File and Storage Technologies,San Jose, CA, USA, February, 2008. [Online]. Available: http://repository.cmu.edu/cgi/viewcontent.cgi?article=1733&context=compsci. Accessed: 17 March. 2012].