–Independent Work Report, Fall 2013 (Revised)–

File System Usage Patterns

Abstract

This paper studies the types of data that users store on their file systems and their patterns of interaction with that data. I compare the metrics gathered for this paper with those of previous studies to demonstrate the evolution of file system usage patterns and to predict future developments. In addition to the comparison, I summarize previous research on users' desires for file system functionality. Finally, I suggest a hybrid cloud-local file system that will improve the file system user experience. The data presented in this paper and previous ones demonstrates that the hypothesized system is more suited to users' present and future usage patterns compared to current file systems and that a prototype should be built for further testing.

Contents

1 Introduction **3**

2 Data Collection **5**

2.1 Sample Selection 5

2.2 Infrastructure 6

3 Research On Local Storage File Systems **8**

3.1 File Sizes 8

3.2 File Ages 11

3.3 File Types 13

3.4 File Count 15

3.5 File System Sizes 16

4 Previous Research on Cloud Storage **19**

4.1 Consumer Cloud Usage 19

4.2 Enterprise Cloud Usage 19

5 The Hybrid Cloud-Local File System **21**

5.1 File System Specifics 21

5.2 Do Users Want It? 22

6 Conclusion **25**

Introduction

Over previous decades, file system usage pattern research guided the development of file systems. Previous studies examined various aspects of file systems including rates of data storage and retrieval, the types of data stored, and the amount stored. [5, 22, p. 1, p. 93] Designers of industry standard file systems such as Microsoft's NTFS used this data to make feature inclusion and implementation decisions. [22, p.103-4] However, these papers are not the final word in file system research. They are old and user activity patterns may have changed since their publications. Additionally, the researchers limited their analyses to file systems which keep all data within the physical boundary of a single computer's case. File systems currently under development, particularly ones that utilize decentralized storage technologies, require guidance from research that does not have these limitations.

This paper addresses the age issue by combining previously existing data with a new study to analyze the development of file systems usage patterns over the past 15 years. Computers have gotten faster processors and larger hard drives. Internet speeds have increased about 50% per year over the last decade. [16] In response to these changes, users may be storing different types of media on their system. Or, they may be keeping all of their data on servers that they connect to over the Internet. The data in my study shows current tendencies for data storage.

This paper also offers a prediction of future file systems. I suggest an improved file system that utilizes both hard drive and Internet storage, a "hybrid cloud-local file system". Such a system will improve the user experience by optimizing the file access patterns most commonly found in file system usage data. For the rest of this paper, I shall refer to a collection of remote servers as "the cloud" and storage solutions built on these computers as "cloud-based storage".

This paper has four components:

1. The first section explains the data collection procedures. This will allow future researchers to collect data more easily and to reproduce my results.
2. The second section summarizes previous research on local storage file systems and compares

their results to my data. This will demonstrate the evolution and current state of file system usage patterns.

3. The third section reviews research on consumer sentiment toward cloud storage. This research will show users' opinions of non-local file system technologies.
4. The final section analyzes this and other papers' data in order to evaluate the validity of the hypothesized, hybrid cloud-local file system.

Data Collection

I collected the following data from the file systems of 24 subjects: total capacity, free space, and distribution of file sizes. Additionally, for each file type that consumed at least 0.1% of the used space on a file system, I collected the average size, total size, and average time since last access of files of that type.

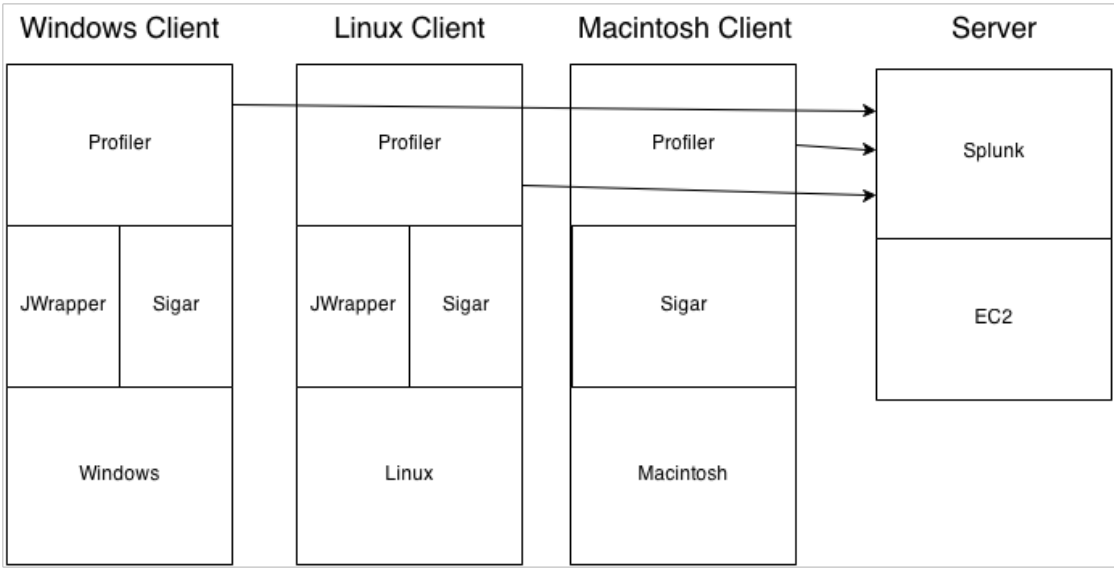
Sample Selection

I attempted to select a population that was representative of all computer users. Previous studies had focused on limited populations such as Microsoft employees. [1, p. 1] In order to get a sufficiently broad sample, I started by randomly selecting students from the entire undergraduate population of a university as of December 2013. The data that resulted from these subjects may be biased due to a low survey response rate of 2.93%. 11 students participated out of the 375 that were invited. It is plausible that only users with particular usage patterns would be willing to participate in the experiment of running a program that collects data from their computer and reports the results to a central server. While this low response rate is troubling, there are no superior methods for data collection. One cannot ethically collect information from subjects who are unwilling to participate. Nevertheless, future researchers should attempt to achieve a higher response rate as I was forced to select additional subjects using inferior, less random methods.

I directly recruited other users to participate in order to gather enough data to draw meaningful conclusions. I invited 13 friends and family members who were not part of the random sample. Their data may be biased as my network of acquaintances may not represent the average user. However, the recruited users come from a diverse set of age groups, from 20 to 70 years old, and occupations, such as student, engineer, and lawyer. The resulting data also appears to be representative of the entire universe of computer users. It includes all major operating systems, Windows, Linux, and Macintosh, and, as seen below, seems plausible when compared to previous papers. The 24 user sample is not optimal, but it is large enough and diverse enough to support meaningful conclusions about the average computer user.

Infrastructure

Figure 1: Data Collection Infrastructure



As described in the above picture, the experiment has two main technical components: a client and a server. The client runs on the users' computers and collects information about a user's file system. The server runs a program called Splunk which compiles the clients' data. In order to collect data from a sample that is representative of average computer users, as many people and computers must be able to run it. The profiler is easy to use so that less technical users can provide data. Additionally, the profiler runs on all major operating system, Windows, Linux, and Macintosh. The multiplatform requirement presents difficulties as the profiler must collect OS specific information, such as the file path format.

I made several language and library choices to accomplish this combination of ease-of-use, OS independence, and OS dependence. The profiler uses the Java programming language in order to be able to run on any computer. Additionally, it uses the JWrapper program and the Sigar libraries. JWrapper converts Java jar files into native executables for Windows, Linux, and Macintosh computers. [19] Since the Macintosh executable did not function properly, those users ran the jar using a bash file. Nontechnical users can run these executables and bash scripts with only a single click. The Sigar library provides Java with APIs for examining the file system. These libraries are

natively compiled for all targeted operating systems. They are then accessed through Java functions that provide OS specific information while allowing the programmer to write OS agnostic code. [21] Using these libraries, I wrote a profiler that easily downloads, executes, and reports back to a central server while running on any platform and requiring minimal user interaction.

I created a data collection server using the Splunk software running on an Amazon EC2 instance. EC2 instances are easy to maintain as Amazon handles the issues of server uptime and router configuration. The Splunk software automatically listens for incoming packets on a particular port and then provides a report of the data in those packets. I used this capability to track each client's report and then print out a file containing all of the data. The server and profiler pair provided the necessary infrastructure for my experiment.

Research On Local Storage File Systems

In this section, I will analyze current statistics on file sizes, file ages, file types, file counts, and file system sizes and the evolution of these values over the last 15 years. In order to do this, I will compare my data with that of four academic papers from the 1990s and 2000s. I will first examine *A Five-Year Study of File-System Metadata* (referred to as the Five-Year Study paper). This paper analyzes the computers of Microsoft employees from 2000 to 2004 including file size, age, and type. *File System Usage in Windows NT 4.0* (referred to as the NT paper) is a 1998 paper that measures the types and sizes of files stored in NTFS file systems as well as the file open, read, and write patterns of the Windows NT 4.0 operating system. *A Large-Scale Study of File-System Contents* (referred to as the Large Scale paper) is another 1998 study of Microsoft employees' computers. This paper investigates file and directory properties and how they differ depending on the occupation of the computer user. Finally, *A Study of Irregularities in File-Size Distributions* (referred to as the Irregularities paper) records the types of files used by Windows, Linux, and Macintosh users at Harvey Mudd College in 2001. The paper focuses on the effect of media files on file size distributions. [1]

File Sizes

This section will focus on the distribution of file sizes, the average file size, and the development of these values over the last 15 years. [2]

The two earliest papers find that most files are in the 1 to 10 KB range. The Large Scale paper shows a log-normal distribution of file sizes with a median of around 4 KB. 1.7% of all files are empty. [5, p. 3] The NT paper finds a similar distribution of file sizes, 40% of operations are to files smaller than 2KB. [22, p. 100]

[1] The data from these papers may provide a biased representation of the average computer user. It may overweight the usage patterns of highly technical users. The papers that use Microsoft employees may show more programming files than the average user. Additionally, the Irregularities paper largely uses students who go to an engineering school and stay on campus during the summer. [6] They are probably more technical than the average person.

[2] Please note that, as discussed above, conclusions based on my data may suffer due to my experiment's small sample size and imprecise values. Additionally, I use the NT paper's data even though it doesn't measure the same quantity as the other papers. It reports the sizes of files opened rather than those of all files stored on disk.

Figure 2: This Paper's Findings For File Size

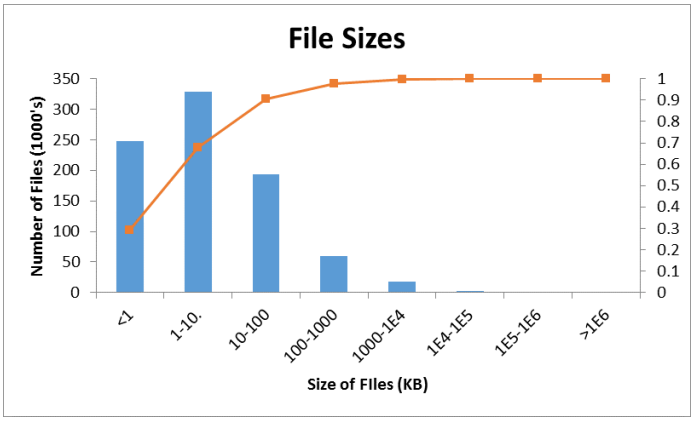


Figure 3: Previous Papers' Findings For File Size

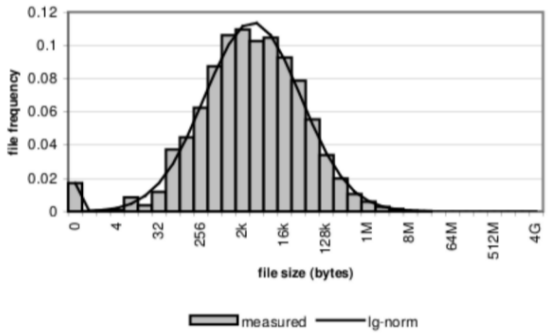


Figure 1: Histogram of Files by Size

Figure 4: Large Scale Paper, p. 3

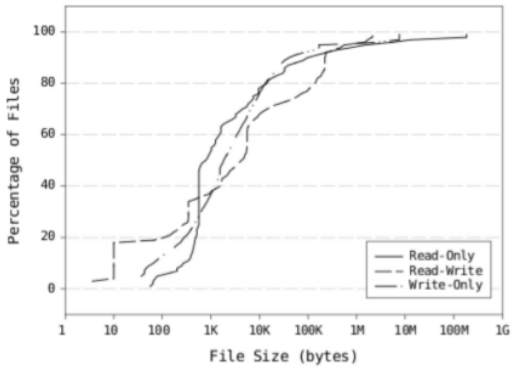


Figure 5: NT Paper, p. 8

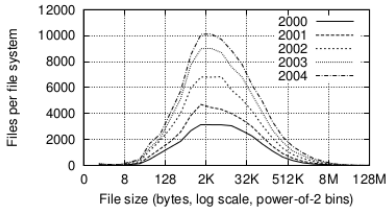


Figure 2: Histograms of files by size

Figure 6: Five-Year Paper, p. 4

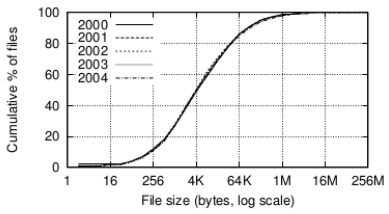


Figure 3: CDFs of files by size

Figure 7: Five-Year Paper, p. 4

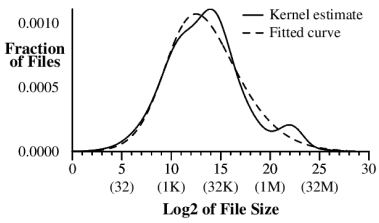


Figure 8: Irregularities Paper, p. 5

The two later papers find that files are larger but have a similar distribution compared to the previous papers. The Five-Year Study paper finds a log-normal distribution centered between 2KB and 32KB and that 1 to 1.5% of files are empty. Over the five years, the mean file size grows by roughly 15% per year from 108 KB in 2000 to 189 KB in 2004. [1, p.4] The Irregularities paper finds a log-normal distribution like that in the Five-Year paper. However, it reports a larger mean file size for Windows users, 720 KB. [6, p. 4-5]

My data follows that from the Five-Year Study. I find a mean file size of 316 KB and a roughly log-normal distribution in the above histogram. Additionally, the line in the above graph shows the CDF of this distribution. The CDF is included for easy comparison with the Five-Year Study paper. My CDF again shows a similar curve to previous work but with less granularity. Both CDF's show that roughly 20% of files are below 1KB in size and that most files are also below 1 MB in size. [1, p. 4]

There are two major shortcomings with my data, small sample size and imprecise values. I have fewer subjects than other papers as the Large Scale paper had over 4000 participants and the Five-Year Study had over 60000 compared to my 24. [5, 1, p. 2,p. 1] Additionally, my file size data is less precise than that of other papers. When recording the a user's file sizes, I grouped together files into bins starting in the range 0 to 1 KB and then grew the bins by a factor of 10 until 1 GB. I grouped together all files greater than 1 GB. This hides the distribution of files in dense regions such as below 1 KB and between 1 and 10 KB.

These errors affect my average file size and the log-normal distribution. My mean file size of 316 KB is half of the 600-700KB predicted by extending the Five-Year Study's 15% annual growth rate from 189 KB in 2004 to 2013. Since I have so few samples and since that paper's results most closely mirror my own, I accept the difference between its results and mine as a random deviation. This difference may be due to my small sample size and, if I had more subjects, my mean file size could increase. The distribution of my files is log normal if you accept the abrupt peak at 1-10 KB and the fact that the left tail is bunched together in one bin. Better data would divide this bin so that the left tail more smoothly trails off.

I conclude that file sizes are generally log-normal distributed around 6-100 KB with a significant number of empty files. The average file size is currently in the range of 300 to 700 KB. Finally, these values have increased over time and probably will continue to increase. With the exception of the Irregularities paper, each data set shows larger file sizes than the previous ones. The two later studies, the Five-Year and the Irregularities, differed in average file size by a factor of 7. This difference is most likely due to the different data samples. The Irregularities paper used Harvey

Mudd students while the Five-Year Study paper used Microsoft employees. Since my data more closely models that of the Five-Year study, I suggest that the Irregularities paper is an outlier and not representative of the general population.

File Ages

Figure 9: This Paper's Findings For File Ages

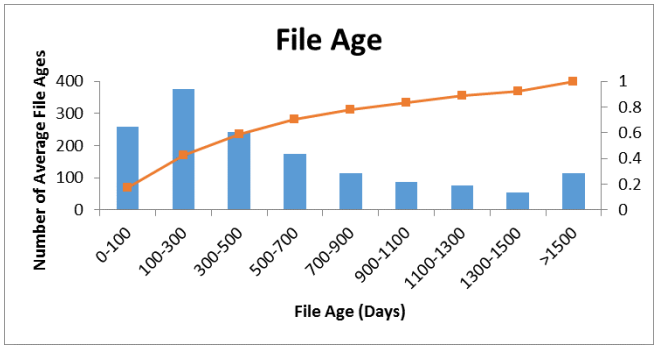


Figure 10: Previous Papers' Findings For File Ages

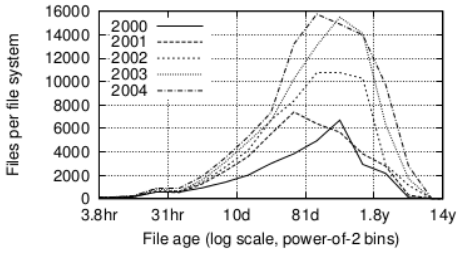


Figure 11: Five-Year Paper, p. 5

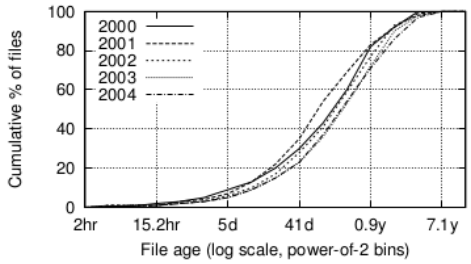


Figure 12: Five-Year Paper, p. 5

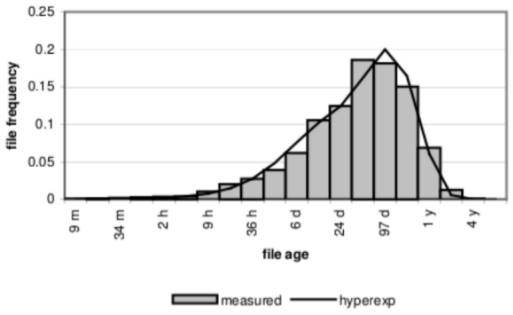


Figure 13: Large Scale Paper, p. 7

This section will focus on the distribution of file ages, the average file age, and the evolution of these values over the last 15 years. I define file age as the time since a file's last modification.

Two of the previously examined papers measure file age. The earlier paper, the Large Scale study, finds that a file's average age is 48 days. The distribution shown in the paper demonstrates that this median may be too low to accurately describe all file ages. The paper's distribution of ages shows that many files are older than 97 days including a significant number between 1 and 4 years old. [5, p.7] The Five-Year Study paper finds that file ages are distributed around 100 days old. The median is between 80 and 160 days and their distribution shows a significant number of files that have a lifetime of greater than a year. The CDF in the paper shows that about 20% of files have a lifetime of greater than 330 days. Finally, the distribution of file ages remains relatively constant between 2000 and 2004. However, they do find that the deviation of ages decreases over time. The number of old and young files decrease relative to those between 90 and 500 days old. [1, p. 5]

My data follows that of the Five-Year Study paper. The histogram shows that a large group of files continue to be between 100 and 300 days old. Additionally, there is a long right tail of files older than 500 days.

The major shortcoming with my data is that I did not individually record the age of every file. Instead, I measured the average age of files of each type on each computer. Then, I reported all file types that made up at least 0.1% of the data on the file system. Therefore, the above distribution is a histogram of averages. Additionally, these averages are not weighted by the number of files that made up each average. Therefore, if one average is based on many young files and another one is based on one old file, the two values are given the same weight. This potentially overrepresents very old files. I hypothesize that this is the reason that my data shows significantly more older files than the other papers.

I conclude that file ages are generally distributed around 100 to 300 days. I guess that the average age is in that range and that file ages remain static throughout time. However, I lack sufficient data points to state with certainty that those last two claims are correct.

Figure 14: This Paper's Findings for Spaced Used By Popular File Extensions

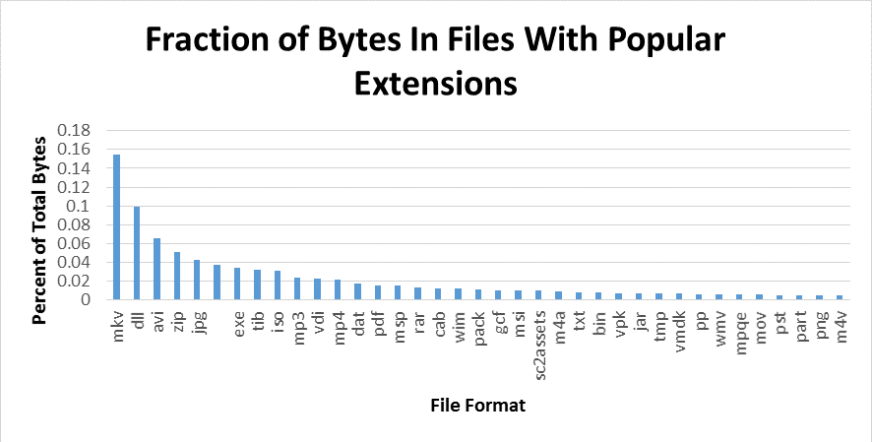


Figure 15: This Paper's Findings For Most Common File Extensions

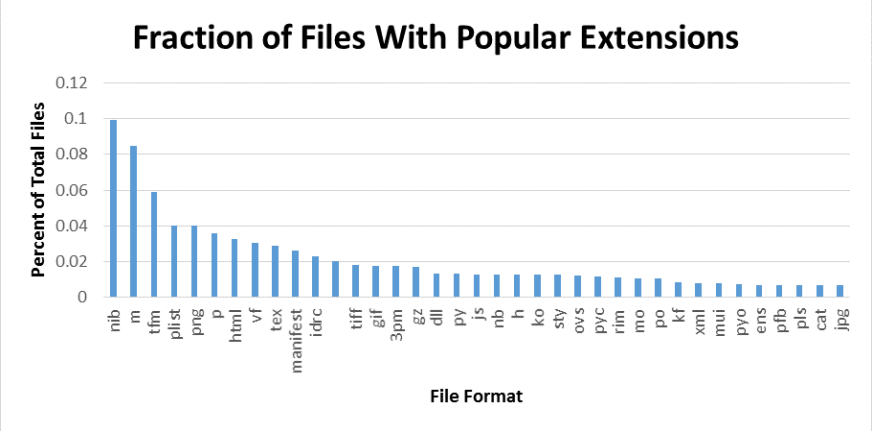


Figure 16: Previous Papers' Findings For File Ages

Rank	Ext.	Lg Size		Lg Age	
		mean	std dev	Mean	std dev
1	.gif	10.4	2.26	21.4	2.42
2	.h	11.8	2.26	22.5	2.12
3	.htm	11.4	1.74	21.6	2.46
4	.dll	16.0	2.00	22.4	2.53
5	–	8.5	3.62	22.5	2.38
6	.c	12.8	2.52	22.8	2.05
7	.exe	15.7	2.07	22.5	2.49
8	.ini	7.3	1.36	22.7	2.04
9	.cpp	12.7	2.21	22.4	2.11
10	.inf	12.9	2.63	22.6	2.59
11	.obj	13.7	2.65	20.4	3.15
12	.txt	10.2	3.17	21.5	2.65

Figure 17: Large Scale Paper, p. 8

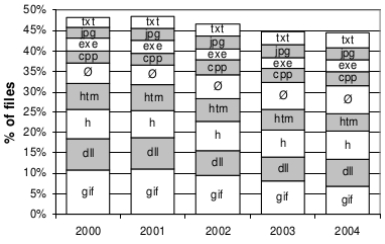


Figure 18: Five-Year Paper, p. 6

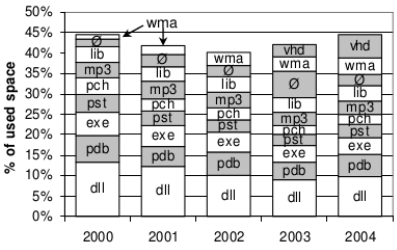


Figure 19: Five-Year Paper, p. 6

File Types

This section will focus on the most common file types, the types that consume the most space, and the evolution of these values over the last 15 years. Since the data collected by my study and the previous ones is in different formats, they are not directly comparable and my conclusions include

some speculation.

The earliest paper, the Large Scale study from 1998, finds that gifs were the most common file, followed by h, htm, dll, c, and exe files. Dll and exe files had the largest average size. However, they are of the same magnitude as other file type sizes. The Five-Year Study paper initially displays a similar distribution to the Large Scale one. Gifs are the most common type in 2000, followed by dll and h. Dll files consume the most space on user hard drives, followed by other program files and mp3 music files. [1, p. 6] By 2004, the only change in distribution of file frequency is a decreasing number of gif files. While wma music and vhd virtual image hard drive types are never among the most common, they consume significant amounts of space in 2003 and 2004. [3] [1, p. 6]

My data shows that the most common files are program files, followed by text and image files. The program files are nib, m, and plist, Objective-C files used for programs on Apple laptops, tablets, and cellphones. [3] Tfm is a tex font file and png contains images. The files that consume the most space are videos, mkv and avi; programs, dll and exe; and images, jpg.

I conclude that programs and media files are the most common and consume the most space. Over the past 15 years, the ratio between these two groups of file types has remained constant although the types of media files have changed from images to songs and videos. In each of the data sets from around 2000, the image type gif is most common. It is closely followed by dll and h types. With some speculation, it is apparent that these types also consumed the most space on disk. The data from the Five-Year paper directly shows this. The data from the Large Scale paper requires some speculation. In that paper, the average sizes of all the types are of the same magnitude. Therefore, the most common file, gif, probably appeared enough to consume the most space on disk even if its average file size is not the largest. The later data from 2004 and 2013 shows the same mix of program and media files. However, there is a decrease in the number of image files as there are fewer gifs. Wma's first appear in 2003, showing the rise of the music file. By 2013, mkv and avi video files consume large portions of users' disks. Thus, the file type distribution remains static

[3] I hypothesize that the vhd files appeared in 2003 because that year Microsoft acquired a company that used the file type. [12] The Five-Year Study paper is an examination of Microsoft employees and so they probably started using vhd files at the time of the acquisition.

except that image files are replaced by music and video ones.

File Count

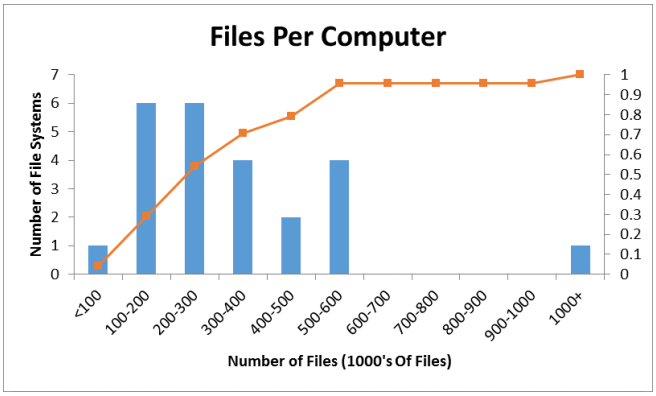


Figure 20: This Paper's Findings For File Count

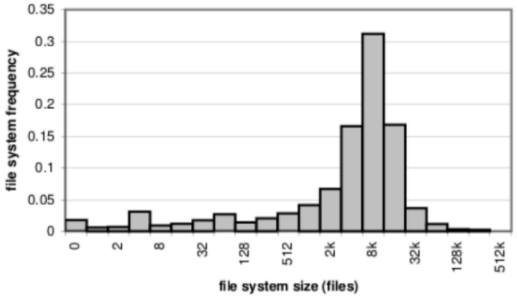


Figure 21: Large Scale Paper, p. 9

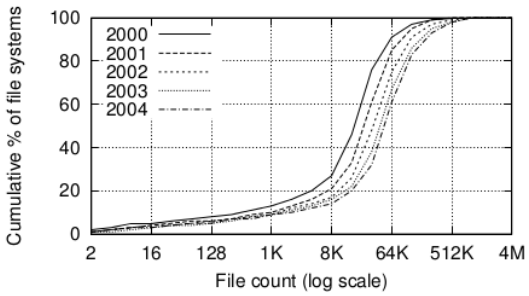


Figure 22: Five-Year Paper, p. 4

This section will focus on the distribution of file counts and the evolution of this distribution over the last 15 years. Since my sample size is significantly smaller than that of other papers and only two of the previous papers have data on file counts, my conclusions require some speculation.

The Large Scale paper from 1998 finds that 61% of the sampled file systems have less than 16000 files and that the mean number of files per user is 31835. Additionally, the distribution shows that most systems have around 8000 files and are between 2000 and 32000. [5, p. 9] The Five-Year Study finds that the average number of files per file system was 30000 in 2000 and increases to 90000 in 2004. Also, the study finds that the median number of files per system increases from 18000 in 2000 to 52000 in 2004. [1, p. 3] Finally, the CDF presented in the Five-Year paper shows that very few files systems are empty and that 60% of file systems have between 8000 and 64000 files. [1, p. 4]

My data shows that the average number of files per file system is 353752 and the median number is 292439. The distribution of files per file system is bimodal with one peak at 100000 to 300000 and another at 500000 to 600000 files. There is 1 file system that has less than 100000 files and the vast majority are between 100000 and 600000 files. The CDF displays that 60% of users are between 100000 and 500000 files.

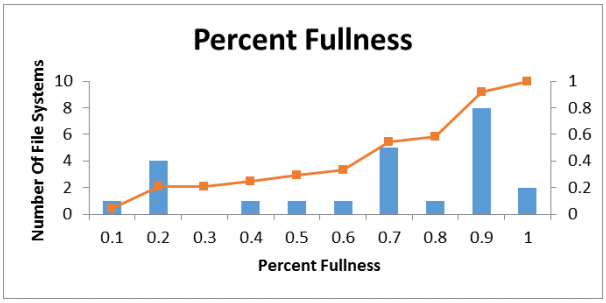
The major shortcomings with the data for this section are a small sample size and few previous examples. The small sample size may be responsible for the bimodal distribution. Each of the other two papers finds a unimodal distribution. [5, 1, p. 9,p. 4] The fact that I only have 24 subjects means that there are less data points to produce a smooth histogram. If I had more data points, I may have found a unimodal distribution. Below, I suggest that the number of files per file system is increasing at an increasing rate. I use three points, the three papers with data on file count, to draw this conclusion. This is the minimum number of data sets necessary to state the acceleration of a value. This hypothesis would be more conclusive if the other two papers also had data on file counts.

I conclude that the mean file count grew from 30000 files to around 300000 over the past 15 years and that this growth occurred at an increasing rate. Each paper finds a larger number of files per file system than the one before it. Additionally, this difference increases over time, suggesting that the second derivative of the number of files per file system is positive. The mean number of files per system increases by 60000 files between 1998 and 2004 and by roughly 260000 files from 2004 to 2013. This is a four times larger increase in less than double the time, 6 years versus 9 years. It is possible that some of this difference is due to different samples. However, since the previous sections' analyses show that my data is similar to that of the previous papers in other categories, I think that the rate is increasing and that this is not an aberration.

File System Sizes

This section will focus on the distributions of file system size and percent fullness and the evolution of these distributions over the last 15 years.

Figure 23: This Paper's Findings for File System Percent Fullness and Size



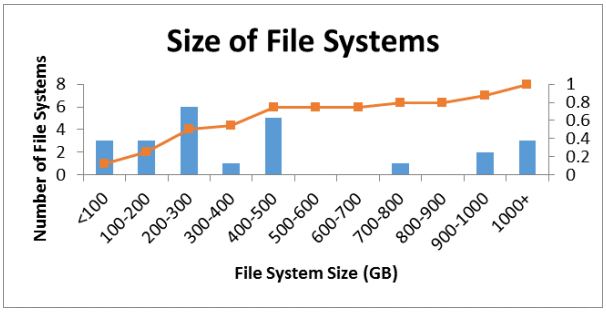


Figure 24: Previous Paper's Findings for File System Percent Fullness and Size

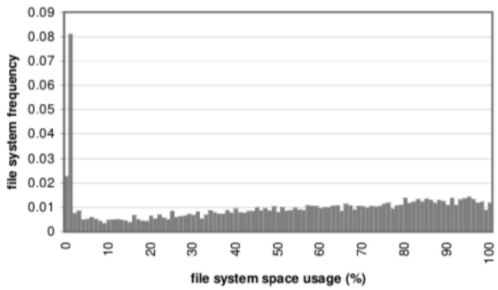


Figure 25: Large Scale Paper, p. 10

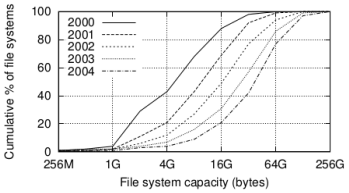


Figure 26: Five-Year Paper, p. 11

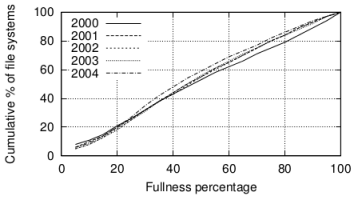


Figure 27: Five-Year Paper, p. 11

The Large Scale study finds that the most common percent usage for a file system in 1998 was between 0% and 5%. Starting at 5%, this number sharply decreases. Then, there is a slow linear increase in the number of file systems as the percent usage increases from 5% to 100%. [5, p. 10] The Five-Year Study paper also finds a linear increase in percent fullness from 5% to 100%. The paper finds few empty file systems. Additionally, most file systems in the study are between 1 GB and 64 GB in size. [1, p. 11]

My data shows a bimodal distribution of file system percent fullness and a unimodal distribution of file system size with a few outliers. The distribution of percent fullness has two peaks, one between 10% and 20% and another at between 80% and 90%. Most file systems are between 100 and 500 GB and a few are above 1000 GB.

The main shortcoming with my data, the small sample size, may be responsible for the unusual distribution of percent fullness. If I had more subjects, I may have found more values in the 80% full bin, leading to a more recognizable single peak over the bins from 70% to 90%.

Despite this shortcoming, I conclude that file systems are increasing in total size and are becoming more divided in terms of fullness percentage, being either almost completely full or empty. My

data shows larger file system sizes than the previous Five-Year Study. [4] For percent fullness, the two earlier papers show some nearly empty and full file system. However, there are also some files systems near 50%. I find very few file systems that are near 50% full. The gap seems large enough that it probably is not the result of the small sample size. Rather, file systems are becoming more divided over time with regard to percent fullness.

[4] I speculate that hard drive size standardization is the cause of the outliers in my file system size graph. Few manufacturers produce hard drives between 500 GB and 1000 GB. As of December 20th, 2013, Newegg.com has 720 hard drives for sale and 24 of them are between 600 and 900 GB. [14, 15] Thus, users can only purchase hard drives smaller than 500 GB or larger than 1000 GB.

Previous Research on Cloud Storage

This section address the following questions about potential consumer and enterprise users: desired use cases, concerns with the technology, and current adoption rates.

Consumer Cloud Usage

Consumers demonstrate their future desired use cases for cloud storage through their current activities: sharing content with others, accessing data from multiple devices, and backing up files. A 2012 report found that 90% of respondents use the cloud for photo sharing, 42% use it for sharing other types of data, 51% use it for accessing data from multiple devices, and 25% use it for backing up data. [7, p. 7] Since these are the most common use cases today, they provide a crude prediction of the capabilities that users will want from future cloud storage systems.

Data privacy is one of the most significant concerns. Consumers do not trust cloud storage providers and are willing to pay for a service that protects their data. Subjects in a 2011 experiment "almost unanimously believed that it would be 'easy' ... for a hacker to get their data from the cloud". [9, p. 6-7] Additionally, 79% of those surveyed would choose to pay for cloud storage rather than receive it for free if the paid plan guaranteed data privacy and the free one did not. [9, p. 9] Cloud storage services must ensure data privacy in order to achieve market adoption.

Consumers most often use cloud technologies on tablets and cellphones. In one study, 58% of subjects use the technology on their cell phone, 54% on their tablet, and 49% on their desktop. [7, p. 9] Therefore, new cloud technologies should run on mobile devices in order to reach consumers on the platforms where they are most comfortable.

Enterprise Cloud Usage

Companies can use cloud computing to replace three levels of infrastructure: end-user applications, software development tools, and hardware. [11] Each portion allows businesses to simplify their infrastructure. Companies can change computer capacity and costs without adding or removing hardware. They just pay more or less to the cloud provider. Software as a Service (SaaS) cloud

technologies replace end-user, externally developed programs such as Microsoft Word. They provide enterprise users with the ability to upgrade and manage access to applications without physically accessing their employees' hardware. [18] Platforms as a Service (PaaS) cloud technologies provide software developing, testing, and distributing abilities for a company's proprietary programs. This is more generalized than SaaS as enterprise users can run their own applications in the cloud; but, they must write and maintain their own code. [18] Finally, Infrastructure as a Service (IaaS) cloud technologies provide the most general cloud abstraction: computation and storage over the Internet without any extra software. [18] Users choose their own operating system and software packages without additional support from the cloud provider.

One of enterprise users' main concerns is the cloud's reliability. A 2013 survey found that 74% of companies that use IaaS have multiple multiple providers. [13] This redundancy allows for easier recovery in the event that one provider fails. However, it also suggests that enterprise users do not trust the cloud as otherwise fewer companies would feel the need to have multiple providers. Any technology, such as the hypothesized hybrid cloud-local file system, will need to address the issue of reliability before companies fully adopt it.

Businesses already have and will continue to cloud technologies. A 2013 study found that 75% of the companies surveyed are using cloud technologies and that this increased from 67% the year before. Additionally, the survey predicted that the market for cloud technologies would grow by 126.5% from 2011 to 2014. [17] Thus, despite their reliability concerns, enterprise users will adopt technologies that utilize the cloud, such as Internet-based storage.

The Hybrid Cloud-Local File System

I hypothesize that a file system consisting of both local and Internet storage components will be an improvement over current, local-only storage systems. First, I will explain the details of the proposed file system. Next, I will show how the file system satisfies the intersection between users' usage patterns found in section 3 and desires in section 4. The file system accomplishes this by providing the desired features while making acceptable trade-offs when compared to local storage file systems.

File System Specifics

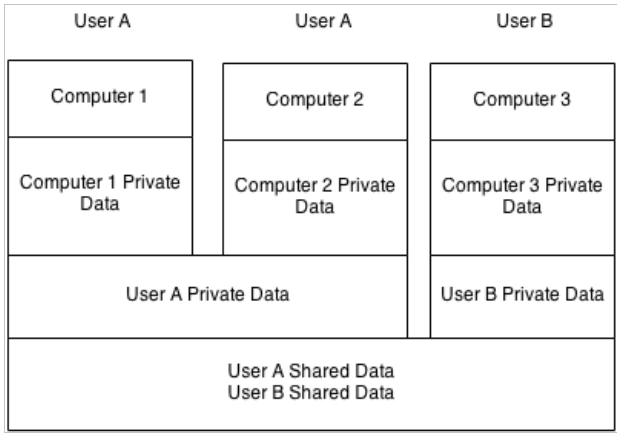


Figure 28: User's Perspective

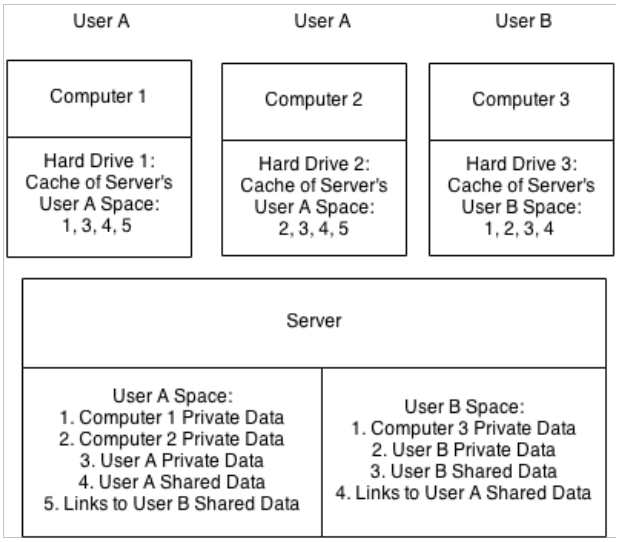


Figure 29: Implementation

The file system will create the illusion of infinite local storage that can be shared between different computers and users. Figure 28 shows the users' perspective; each computer appears to locally store all the data that a user has permission to access on the device. Figure 29 shows the implementation; all data is on the server and local computers only keep a portion. The system will store each file in blocks of a fixed size such as 4KB. Each file will have permissions for which computers and users can access it. The cloud will store the master copy of this data. Local hard drives will act as caches for cloud storage. The system files necessary to boot will remain on each computer's hard drive. The other files will be downloaded and deleted as necessary. If a file is needed, the file system will

download it and pause all access attempts until it is finished. Fixed size blocks mean that files can be partially downloaded so that the relevant parts are on disk in less time.

The file system will handle Internet connectivity issues through an update log. All changes to data blocks will be made on the local disk first and then written to a log. This log will be pushed to the cloud when the computer is connected to the Internet. The computer will be able to function even when it is offline as it will merge its changes with the online, master copy of the data once a connection is reestablished.

The hard drive is not a necessary portion of the file system but that it is a useful component. An alternative would be to just use RAM, ROM, and cloud storage. It is possible to store a bootloader on ROM, download the OS over the Internet on boot, store it in RAM, and keep all persistent data in the cloud. [10] The hard drive is a better choice for local storage than RAM due to the issues of cost, number of uncached data reads, and reliability of persistent storage. RAM is faster than a hard drive; but, RAM is also significantly more expensive. [20, p. 26] It may be prohibitively expensive to buy enough RAM to simultaneously store an entire operating system and all relevant data. A RAM solution would also require all files to be redownloaded after a restart. Hard drives can keep files in cache when the system is off. Redownloading a file is much slower than keeping it on disk. [4, 8, p. 4] Finally, the diskless system could not shutdown without an Internet connection. When the computer shuts down, all changes to persistent data must be uploaded to the cloud if there is no hard drive. With a hard drive, the log of changes to the file system can be uploaded after the computer restarts. This would allow the computer to function even if there is no Internet access.

Do Users Want It?

Section 4 made a crude prediction of users' desires. Consumers want the ability to share, back up, and remotely access data. Businesses want to simplify their infrastructure. Additionally, section 3.5 shows that users need more space. That section shows a bimodal distribution of file system percent fullness. A significant number of subjects currently use between 80% and 90% of the space on their drive. Additionally, the data from earlier papers suggests that this percentage is increasing over time

even as the size of hard drives increases. Therefore, users in the future will come even closer to using their entire hard drive and will benefit from having extra storage space in the cloud.

Will users adopt the hybrid cloud-local file system? First, the file system must provide users with their desired functionality. Data backups, storage space flexibility, and simplicity of infrastructure result from the cloud storage component. First, the provider must ensure data integrity so that backups are available in the event of a computer failure. Amazon S3 provides cloud data storage with 99.999999999% durability. [2] That should be sufficiently durable for most back up use cases. Additionally, cloud storage is flexible and simple due to the provider's web interface. Users can automatically increase or decrease the size of their drives using their provider's interface. As of December 2013, Amazon charges less than $0.01 per GB. [2] Changes in capacity require no hardware changes such as buying storage servers. Therefore, users can easily adjust the file system's storage capacity according to their needs while assuming that their data is data safe.

The file system will allow for users to share data and remotely access it from multiple devices. It will implement this feature by supporting multiple, simultaneous accessors to a file. These accessors could be several devices owned by one person or many people each using a different device. The system will accomplish this by having a master copy of each file that belongs to the original device and user that created the file. The master device may allow other devices to read and copy the file. Additionally, other devices that share the same owner may update the original copy if the master device does not have a lock on it. Other users' devices may make their own copy of the file. However, to make changes to the original file, they must request the owner's permission. The file system will attempt to merge changes automatically if it can recognize the file format, such as .txt. In other cases, the file system will warn the file's master of the potential issues and recommend that he reject the other user's changes. Thus, the file system can support multiple devices and data sharing with other individuals.

The file system must also provide this data in a fast enough manner so that users do not mind the download times. Data downloads from the cloud slower than it can be read from disk. [4, 8, p.4] Users that become frustrated with download times may chose to buy extra local storage and ignore

the cloud. I can optimize the solution so that it minimizes the users' download times. One approach is to guess which files will be accessed least frequently, such as those with the least activity. The findings from section 3.2 show that files are on average 100-300 days old. This means that the average file on disk is not touched for more than 3 months. Any file that has not been accessed in a long time probably can be deleted from the cache without causing a significant slow down. Since the file is old, it probably won't be touched for a long time and the delay in downloading it is unlikely to be realized. Another approach is streaming files of certain types before the entire download finishes. Programs can use the beginning portions of songs and videos while the rest downloads. The findings from section 3.3 show that movie file types are 2 of the 3 most common when weighting by bytes. Mp3 and mp4 files also appear fairly high on the list. This means that a significant number of the bytes can be streamed. The file system will always be slower than having all data on a local disk, but these optimizations may make it fast enough that users do not mind the speed difference.

Thus, I have designed a file system that may satisfy users' needs and desires for cloud storage while also being fast enough that they will not prefer local-only storage systems.

Conclusion

I collected my own data and compared it with that of previous papers to describe historical trends in and the current state of file system usage patterns. These data sets show that files are on average 6-100 KB in size and 100-300 days old. Additionally, I found that most file systems have between 100000 and 500000 files, are between 100 and 500 GB in size, and mostly hold media and program files. I presented a file system that stores data on a local drive and in the cloud to enable data backup and sharing. Finally, I used my research to analyze the hypothesized file system.

My research suggests that the hybrid cloud-local file system may satisfy users' needs, but the technology needs more research and a prototype before I can make a more definite comment. There are additional issues with the hybrid file system that are beyond the scope of this paper. Users may be concerned with the privacy of information stored in the cloud. It is conceivable that cloud providers would sell or give away this information against the wishes of the users. Additionally, all cloud providers may not build systems as reliable as Amazon's S3 service. I used S3's data durability to justify the cloud as a backup utility.

Future studies must have three components to address the above issues and take a definitive stance on the file system: better data, analysis of cloud storage providers, and a prototype file system. Superior data collection will solve the problems found in section 3. Specifically, the data will have more subjects and more precisely record file sizes and ages. This will probably be done by reporting data in a more raw form to the central server. For this study, I generated statistics summarizing each computer, reported these to the Splunk server, and then analyzed the statistics. The next study should individually report the size, type, and age of every file to the central server and do all processing once data collection is finished. Analysis of cloud storage providers will check the assumptions made in part 5. In that section, I assume that cloud storage is cheap and reliable enough to replace disk storage. The next study should survey cloud storage providers to check those assumptions. Finally, a prototype file system will provide real world data on the hypothesized system. Releasing a hybrid local-cloud file system to end users will enable researchers to study

how well the idea performs in reality. They will be able to ask users what they think of the product. Statistics from the file system will show if the users are using the extra space and multi-device features.

In conclusion, I hope that this paper provides data for future researchers and a new, multi-tiered approached to file system functionality.

References

[1] N. Agrawal *et al.*, "A five-year study of file-system metadata," *Trans. Storage*, vol. 3, no. 3, Oct. 2007. Available: http://doi.acm.org/10.1145/1288783.1288788

[2] Amazon.com, Inc., "Amazon s3, cloud computing storage for files, images, videos," 2013. Available: http://aws.amazon.com/s3/

[3] Apple Inc., "Property list programming guide: Quick start for property lists," March 2010. Available: https://developer.apple.com/library/mac/documentation/cocoa/conceptual/PropertyLists/QuickStartPlist/QuickStartPlist.html

[4] Bestofmedia Group, "Charts, benchmarks hdd charts 2013, [01] read throughput average: h2benchw 3.16," 2013. Available: http://www.tomshardware.com/charts/hdd-charts-2013/-01-Read-Throughput-Average-h2benchw-3.16,2901.html

[5] J. R. Douceur and W. J. Bolosky, "A large-scale study of file-system contents," *SIGMETRICS Perform. Eval. Rev.*, vol. 27, no. 1, pp. 59–70, May 1999. Available: http://doi.acm.org/10.1145/301464.301480

[6] K. M. Evans and G. H. Kuenning, "A study of irregularities in file-size distributions," in *In International Symposium on Performance Evaluation of Computer and Telecommunication Systems (SPECTS '02*, 2002.

[7] "Personal cloud services emerge to orchestrate our mobile computing lives," Forrester Consulting, July 2012. Available: http://www.sugarsync.com/media/sugarsync-forrester-report.pdf

[8] H. Hussain *et al.*, "The cost of connectivity 2013," New America Foundation, Tech. Rep., October 2013. Available: http://newamerica.net/sites/newamerica.net/files/policydocs/The_Cost_of_Connectivity_2013_Data_Release.pdf

[9] I. Ion *et al.*, "Home is safer than the cloud!: Privacy concerns for consumer cloud storage," in *Proceedings of the Seventh Symposium on Usable Privacy and Security*, ser. SOUPS '11. New York, NY, USA: ACM, 2011, pp. 13:1–13:20. Available: http://doi.acm.org/10.1145/2078827.2078845

[10] iPXE project, "ipxe - open source boot firmware [start]," April 2013. Available: http://ipxe.org/

[11] S. Koons, "How to use cloud computing to benefit from big data," *Penn State News*, December 2013.

[12] Microsoft Corp., "Microsoft acquires connectix virtual machine technology," Press Release, 2003. Available: http://www.microsoft.com/en-us/news/press/2003/feb03/02-19partitionpr.aspx

[13] Neovise, LLC. and Virtustream Inc., "Neovise and virtustream release results from research on public, private and hybrid cloud use by u.s. organizations," Press Release, April 2013. Available: http://www.virtustream.com/sites/www-dev.virtustream.com/files/Neovise-Cloud-Research-PressRelease4-16-13_0.pdf

[14] Newegg Inc., "Newegg.com - computer hardware, internal hard drives, all desktop hard drives," 2013. Available: http://www.newegg.com/Product/ProductList.aspx?Submit=Property&Subcategory=14&N=100007603&IsNodeId=1&IsPowerSearch=1

[15] Newegg Inc., "Newegg.com - computer hardware, internal hard drives, all desktop hard drives," 2013. Available: http://www.newegg.com/Product/ProductList.aspx?Submit=Property&Subcategory=14&N=100007603%20600003306%20600003312%20600237350&IsNodeId=1&IsPowerSearch=1

[16] J. Nielsen, "Nielsen's law of internet bandwidth," 2013. Available: http://www.nngroup.com/articles/law-of-bandwidth/

[17] North Bridge Venture Partners, "2013 future of cloud computing survey reveals business driving cloud adoption in everything as a service era; it investing heavily to catch up and support consumers graduating from byod to byoc | north bridge," Press Release, 2013. Available: http://www.northbridge.com/2013-future-cloud-computing-survey-reveals-business-driving-cloud-adoption-everything-service-era-it

[18] Rackspace Support, "Understanding the cloud computing stack: Saas, paas, iaas," Rackspace US, Inc., Tech. Rep., October 2013. Available: http://www.rackspace.com/knowledge_center/whitepaper/understanding-the-cloud-computing-stack-saas-paas-iaas

[19] SimpleHelp Ltd, "Jwrapper - overview," October 2013. Available: http://www.jwrapper.com/

[20] A. S. Tanenbaum, *Modern Operating Systems*, 3rd ed. Upper Saddle River, NJ, USA: Prentice Hall Press, 2007.

[21] VMware Inc., "Sigar api (system information gatherer and reporter) | hyperic," 2012. Available: http://www.hyperic.com/products/sigar

[22] W. Vogels, "File system usage in windows nt 4.0," *SIGOPS Oper. Syst. Rev.*, vol. 33, no. 5, pp. 93–109, Dec. 1999. Available: http://doi.acm.org/10.1145/319344.319158