2010

Feasibility of backing up server information in a distributed storage using client workstations hard drives

Raúl Rafael Acevedo Cárdenas

FEASIBILITY OF BACKING UP SERVER INFORMATION IN A DISTRIBUTED STORAGE USING CLIENT WORKSTATIONS HARD DRIVES

By

Raúl Rafael Acevedo Cárdenas

Thesis submitted in partial fulfillment of the requirements for the
degree of Master of Science in
Networking and System Administration

Rochester Institute of Technology

B. Thomas Golisano College
of
Computing and Information Sciences

Rochester Institute of Technology

B. Thomas Golisano College of Computing and Information Sciences

Master of Science in Networking and System Administration

Thesis Approval Form

Student Name: _Raúl Rafael Acevedo Cárdenas_____

Thesis Title: Feasibility of backing up server information in a distributed storage using client workstations hard drives

Thesis Committee

Name	Signature	Date

Charles Border, Ph.D.
Chair

Giovanny Heredia, MS
Committee Member

Arlene Estévez, MS
Committee Member

Thesis Reproduction Permission Form

Rochester Institute of Technology

B. Thomas Golisano College
of
Computing and Information Sciences

Master of Science in
Networking and System Administration

Feasibility of backing up server information in a distributed storage using client workstations hard drives

Date: December 10, 2010 Signature of Author: _____

Table of Contents

FEASIBILITY OF BACKING UP SERVER INFORMATION IN A DISTRIBUTED STORAGE USING WORKSTATIONS HARD DRIVES

1. Abstract

As a consequence of nowadays large hard disk capacities, we can frequently find many networks in corporate environment with a considerable amount of unused hard disk storage space dispersed among all its computers. In an immediate future, the purpose of this unused space is unclearly defined and represents a waste of resource. Several studies suggest and evaluate numerous ways to take advantages of workstation unused hard disk space in a network (1) (2) (3) (4) (5) (6) (7) (8). However, there are no evidences of studies that consider disk-based backup, distributed storage, and the unused workstation storage aiming at backing up server information in small business network. Determining whether it is possible to utilize these resources for backing up server information certainly can help small businesses to obtain a greater return of investment in their networks. In this paper, I present a case study in where I found out that under specific conditions there are resources that a backup system can utilize to back up server information by using workstation's unused hard disk spaces without significantly affecting normal operation of that network.

2. Introduction

Thanks to the development of new techniques, hard disk drives have been increasing their capacity in an exponential way over the years. As a result, nowadays computers data storage capacity is so significant that it usually surpasses the need any typical user might have during the computer life span. Hence, it is frequent to find many networks in corporate environment with a considerable amount of unused hard disk storage space dispersed among all its computers. From the perspective of the final user, this extra data storage space can be seen as something good; but, from the investment point of view, this situation is a waste of resources.

A study conducted by Microsoft Research (2) provides evidences about the magnitude of the storage space underutilization on computers. By examining 4,081 computers in a commercial environment, this study reveals that file systems are, on average, 47% empty. In this context, I found an opportunity to develop a new approach to take advantages of this resource that is, as mentioned before, increasing with technology development.

In this paper, I studied the feasibility of using a system to back up server information in a distributed storage conformed by unused workstations hard drives space in a small-size organization. The study provides statistical information that helps in the initial phase of the project development for a backing up system that would have the purpose of taking advantage of available computers' hard disk space in small-size organization network.

Having this information, one can decide whether the backup system will be a useful tool, under which conditions it actually works and, therefore, whether this tool should be developed or not.

In the following sections, I cover in more details the study I conducted and my analysis of the results. Section 3 states the motivation for this research. Related works are presented in section 4, followed by section 5 with the purpose statement. Section 6 presents the suggested backup system. Section 7 covers the methodology used for this research. Section 8 describes the network in where the measurements were taken. Section 9 then discusses the results obtained. Finally, section 10 gives the conclusions of this study and further recommendations.

3. Motivation

I have worked for more than ten years in a company dedicated to provide a diverse range of IT services to different kind of customers. This company provides system administration services to small business that do not have system administrator on-site due to the high cost of having a permanent staff for this purpose. During these years providing outsourcing system administration, my partners and I have had the mission to optimize the use of the customers' IT resources. One of the optimization tasks we have done is limiting the resource of hard disk space that users have available on the server and on their assigned computer. We have taken this arrangement because we have found that, eventually, some users

squander the resource storing personal files on them. In order to limit the space, we have used quota assignment for each user. Quota assignment has worked fine, and the arrangement has avoided, to certain extent, the misuse of storage resource by users. But this arrangement has come with another problem: unused disk space all over the network and unclearly defined purpose for this resource in an immediate future. I have been concerned about this situation for a long time. One solution I came out with is the one suggested in this proposal, which, among other things, might be particularly convenient for small business with limited financial resources for buying a tape backup system.

4. Literature Review

In this literature review, I have covered three aspects of the topic I am interested in. This literature review addresses alternatives for backing-up server information, explores studies concerned with the use of storage distributed among computers in a network, and show studies of the feasibility of using systems with similar purpose than the one intended in this proposal.

The traditional way to backup information from server has been using tapes. In recent years, improved technologies and its accessibility are creating other alternatives to backup. This is the case of the disk-based backup, which is becoming an alternative to protect server information.

There are several studies that address advantages of using disk-based backup. One of these studies was conducted by scholars from Shangai JiaoTong University (9) who concluded that using a better compression technique than the one used for tape-based backup, disk-based backup archived more efficient use of network bandwidth and storage. They reached this conclusion by conducting several experiments in which they compared different compression methods and network bandwidth usage. For these experiments, they utilized three kinds of files commonly found in computer environments: doc, pps, and pst. In order to resemble real world scenario, they made modifications to these files to simulate changes that occur on real environments where users create, edit, and modify these files. Both, disk-based backup and tape-based backup experiments were compared. The comparison shows that disk-based backup, using a compression method that cannot be used with tapes, known as delta compression, has a more efficient usage of storage and network bandwidth.

Another study also conducted by the same scholars from Shangai JiaoTong University (10) demonstrated that data restoration is more convenient and efficient using disk-based backup together with the delta compression technique that it is unique for disk-based backup. In order to get this result, they took an experiment to compare the time needed to backup and restore data. In this case, they compared their disk-based prototype system with a tape-based backup system; and they proofed that restoration from the prototype system is faster than the restoration using the tape-based backup system. Expressing this in numbers, the restoration of the disk-based backup took 49 seconds for 1.5GB of data; in contrast, the tape-based backup system took 12 minutes 45 seconds to restore the same data. In this last

case, the data came from three separate tapes and only 8% of the restoring process time was spent transferring data; the remaining time, 92%, was spent by the tape system "performing mechanical movement, file accessing, and loading and unloading data" (p. 458). They concluded that they can restore a file 16 times faster by using their disk-based backup system than by using tape-based backup system.

Despite the advantages presented by these studies, the authors think that disk-based back should be considered as a complement instead of substitute of the traditional tape-based backup system. Their argument for this last statement lies on the challenges of changing a backup system that is based on tapes and the economic implication of this change from tape to disk.

These backup systems studies have in common with my study that they consider the use of a disk-based alternative to backup information. Nevertheless, they do not consider a distributed storage system to backup data, and particularly data from server. By using a distributed storage system, we might take advantage of all the unused disk-based storage space that is normally available in all computers of a network.

Several researchers have pinpointed on the idea of creating a distributed storage system to backup purposes. One distributed storages system is Peer-to-Peer systems, but most of these systems are only intended to share files. However, some researchers have used these technologies to backup. Cox, Murray, and Noble (3) presented *Pastiche,* which is a peer-to-

peer backup system that uses unused disk space to perform backup among peers. In this study, they present the technologies in which *Pastiche* depends on and describe how *Pastiche* addresses the difficulties that are inherent in peer-to-peer networks to achieve data backup. *Pastiche* is very similar to my proposed system in that it takes advantage of the free storage space of computers, but it is focused on mutual backup among peers in a network instead of focusing on making backup of server information.

pStore (1) is another peer-to-peer backup system proposed by scholars from MIT. Like *Pastiche*, this system also exploits unused drive space on peers. Researchers in this study present how pStore addresses issues related to sharing the backup space among peers on collaborating computers network. By using quota policy, pStore establishes the amount of space available for a peer to backup in other peers. This quota policy consists on assigning to a peer an amount of storage space to backup in others peers equal to the amount of space the peer donates to the system. The mutual peer-to-peer backup and its quota policy differ from my study proposed system precisely in the policy to assign space. In my proposed system, I am suggesting to use a proportion of the space available on each client machines to backup.

Closely related to *pStore* and *Pastiche* backup systems is *PeerStore*, which is a peer-to-peer backup system proposed by Landers, Zhang, and Tan (6). *PeerStore* differentiates form other peer-to-peer backup systems technology by decoupling the data management from the backup data storage. In order to present the advantages of this approach, the researchers

conducted an experiment in which they found out that *PeerStore* generates less significant maintenance traffic in the network compared to *pStore*. Since *PeerStore* is a system similar to pStore, it has the same differences with my proposed system.

There are other distributed storage systems in the literature such as iDIBS (7), Hydra (11), Zebra (4), and Freeloader (8). These systems present differences in their purpose. iDIBS is intended to make backup, while Freeloader was created to cache information, but none of them have the specific purpose of backing up server information on available desktop storage space.

Feasibility studies about taking advantage of unused workstations' hard disk drive space can be also found in the literature. One of these studies was conducted by Microsoft Research scholars (2). In this study, they suggest a severless distributed file system where the workstations do not trust each other. In order to determine the feasibility of implementing their system, they measured and analyzed several metrics from workstations at Microsoft Corporation, including workstation availability, hard disk drive space usage, file activity, and loads. They found that 50% of the hard disk space is available and that 95% of time 50% of workstations are accessible. These scholars conclude that in an actual deployment of their proposed system, it would work satisfactorily. The system suggested in this study is considerably different from the one I propose. Their system is not oriented to give services to servers; instead, it is intended to provide storage for files of workstations. Another difference is that the study was conducted in a large organization as it is Microsoft

Corporation, in where, at the time of the study (1999), they have 64,610 workstations. My study was conducted in a small-size organization, which might have a different pattern of usage and different computer resources due to improvement of technology.

Another study about feasibility is from Virginia University (5), which is more similar to the study I conducted. Scholars from this university suggest a "distributed virtual storage system", which they called Storage@desk. This system is intended to make available, in an aggregate way, the workstations unused hard disk space to users and applications. To determine the feasibility of their proposed system, they collected statistical information about a real network in order to be used for the design of the Storage@desk. Their study was conducted using workstations of the Virginia University, which reach the number of 729 workstations. The metrics they measured were processor usage, available hard disk space, hard disk I/O load, available memory, computer availability, among others. These scholars reach the conclusion that Virginia University desktop infrastructure, at the time of the study, was capable of supporting the system they propose. As the title of the study suggests, the system was conducted in a large organization, and this is one of the differences with my study. Another difference is that the study does not clarify which kind of application can utilized the storage that the Storage@desk will make available. Furthermore, this study does not consider the usage of a system for applications, like backing up, which requires enough available space to store the information of the servers and a specific time to execute it without affecting workstations regular performance.

It seems that there are limited studies about systems that utilize distributed storage and scavenging workstations storage aiming at backing up server information.

In conclusion, there are evidences of technologies created to take advantage of the facilities provided by a disk-based backup. There are also technologies based on distributed storage for different purposes, included for backing up information on workstations and there are also studies of the feasibility about harnessing workstation hard disk unused space. However, there are no evidence of technologies that consider disk-based backup, distributed storage, and the unused workstation storage aiming at backing up server information. Hence, it seems that the study here presented contributes to enrich the available literature on the field.

5. Purpose Statement

This research has the purpose of studying the feasibility of utilizing unused workstations hard drives space as a distributed storage in a small-size organization to backup server information.

6. Proposed Backup System

In order to take advantages of unused disk space in a network, the system I proposed could utilizes two kinds of applications to coordinate the data transferring between server and the workstations.

One of these applications would reside in the server that would be backed up and its role would be to centrally manage the whole system. Using this application, the backup operator would be able to carry out the functions found on any backup application. The innovation in this application would be that destination storage for the backup would be the workstation unused hard disk space. The application would create a distributed file system with this space in order to follow the backed up files dispersed in the disks over the network. In addition, this application would be used to select the computers that would collaborate with the backup system. Knowing the most adequate slot time to accomplish the backup, the backup operator would select the information and would schedule the backup operation. In case of losing unused space on workstations, the application would be able to reconstruct lost information from the information found on the remaining computers that were collaborating with the system. A possible solution to accomplish this function could be utilizing Redundant Array of Independent Disks (RAID) technology, in which every workstation unused disk space would be a drive for the system. One possible RAID level that might be used is the hybrid RAID 50 or 51.

The second application has the main mission of working as intermediary between the server application and the file system in the workstation where it would reside. This application would be designed to run as a service in the workstation so that user intervention would not be required. Each computer participating in the backup system would have one of these client applications, therefore, the server application would be able to communicate with the file systems of these workstations through these client applications to request information about the hard disk space available and the data backed up in each computer. For security purpose, the client application might encrypt the data to protect it from unauthorized access.

The feasibility of implementing the proposed backup system depends on the availability of resources in the network to incorporate these two kinds of applications without affecting the network regular operation.

7. Methodology

In order to determine the feasibility of utilizing unused computers' storage in small business network for the proposed backup system, I measured whether there were enough resources in a small business network to successfully complete the process of backing up the server information.

The resources I measured were workstations availability, unused storage space for all hard drives on each workstation, processor load percent, physical RAM available, and network utilization. Additionally, I measured network utilization for the server network link and how much information the main server had and how it fluctuated during the study.

I consider those metrics above the most critical for assessing a workstation infrastructure that may be utilized by a backup system that take advantage of workstation's unused storage. Moreover, most of these metrics have been incorporated in previous feasibilities studies that have evaluated whether is possible to take advantages of hard disk space for diverse purposes (5) (2).

Scripts were configured to capture and store these metrics every 30 minutes, except for network utilization metrics, which were captured every 10 minutes as it is recommended by Priscilla Oppenheimer (12) for long-term network load analysis. All these measurements were taken from September 14th to October 16th, 2009. The frequency and duration of capturing provided enough information and granularity needed for this study without considerably affecting network and workstations performance.

The scripts I created were developed using Visual Basic scripts (vbScripts). In these scripts, I utilized Windows Management Instrumentation (WMI) queries to measure workstations, main server, and network resources. WMI is a mechanism that provides system administrators the ability to manage hundreds or even thousands of Microsoft Windows

computers in a structured, secure, and systematic approach (13). The WMI scripts were periodically executed on a server by using Windows Task scheduler. Since these scripts were executed remotely, I obtained a more precise measurement of computers' resources because the workstations resources were not significantly affected by the measuring process.

The network studied had 32 computers and 2 servers with different performance. The characteristics of these computers are shown in the following section, except the second server, which was only used for the capturing process. All the information collected from these computers is statistically analyzed in the subsequent sections.

8. Studied Environment

In order to have a better perspective about this study, it is necessary to know more about the environment in where it was conducted. The network from where I collected the data for this study belongs to a typical free zone company dedicated to cloths manufacturing in the city of Santiago, Dominican Republic.

This company executes operations from Monday through Fridays, from 8 am to 5 pm. As most users turn off their computer after work, the majority of the network resources are only available during the company's working hours. This company utilized 32 workstations

and 2 servers for their work. All these computers utilize Microsoft Windows operating systems. Workstations had installed Microsoft Windows XP Professional and the server Windows Server 2003. You can see in table 1 more details about the resources that these computers had.

Table 1 – Workstations Resources

Workstation	Processor	Memory (MB)	Hard Drive (GB)	Average Unused hard disk space (GB)	Network Connection Bandwidth (MB)
WS004	Intel Pentium III processor 1GHz	256	37.27	27.41	54
WS006	Intel Pentium 4 CPU 1.60GHz	768	38.28	24.56	100
WS007	Intel Pentium III processor 1GHz	512	18.65	7.19	100
WS015	Intel Pentium 4 CPU 2.00GHz	768	18.62	5.78	100
WS017	Intel Pentium 4 CPU 1.60GHz	641	37.27	25.25	100
WS018	AMD Duron 1.30GHz	768	28.63	15.66	100
WS019	Intel Pentium 4 CPU 2.80GHz	768	74.53	60.32	100
WS031	Intel Pentium III processor (0.70GHz	512	55.90	28.66	54
WS032	Intel Pentium 4 CPU 2.20GHz	640	38.29	17.04	100
WS035	Intel Pentium 4 CPU 1.80GHz	512	27.95	17.03	100
WS036	Intel Pentium 4 CPU 1.80GHz	1,024	18.65	4.19	100
WS037	Intel Pentium 4 CPU 1.80GHz	385	18.64	12.48	100
WS038	Intel Celeron CPU 2.66GHz	512	37.27	24.45	100
WS039	Intel Pentium 4 CPU 2.80GHz	512	74.53	60.93	100
WS040	Intel Pentium 4 CPU 3.00GHz	768	74.50	65.89	100
WS041	Intel Pentium 4 CPU 2.00GHz	513	18.64	8.98	100
WS042	Intel Pentium 4 CPU 2.00GHz	257	18.64	1.71	100
WS043	Intel Pentium 4 CPU 2.00GHz	513	18.64	8.47	100
WS044	Intel Pentium 4 CPU 1.80GHz	768	28.63	6.96	100
WS050	AMD Athlon 64 X2 Dual Core Processor 3800+	1,024	232.88	173.67	100
WS052	Intel Pentium 4 CPU 1.80GH	512	37.28	28.14	100
WS055	Intel Pentium 4 CPU 1.80GHz	512	37.27	23.91	100
WS057	Intel Pentium D CPU 3.40GHz	1,024	74.50	59.34	100
WS058	Intel Pentium D CPU 2.80GHz	512	74.50	61.78	100
WS059	Intel Pentium 4 CPU 3.20GHz	512	74.53	61.50	100
WS060	Intel Pentium 4 CPU 3.20GHz	512	74.53	61.49	100
WS061	Intel Pentium Dual CPU E2160 @ 1.80GHz	512	74.50	55.85	54
WS062	Intel Pentium Dual CPU E2160 @ 1.80GHz	1,024	149.01	116.81	100
WS063	Intel Core2 Duo CPU E4500 @ 2.20GHz	1,024	149.01	132.72	54
WS064	Intel Core2 Duo CPU E4500 @ 2.20GHz	1,024	149.01	131.47	100

WS065	Intel Core2 Duo CPU E8300 at 2826 MHz	2,048	232.83	216.59	100
WS100	Intel Pentium 4 CPU 3.00GHz	1,024	74.50	63.08	100

9. Data Analysis

9.1 Workstations Availability

In order to implement a backup system that utilizes workstation resources, we need first of all to determine whether workstations are available, when are they available, and in what quantity.

As users turn off their computers after work, workstations availability dramatically changed every day as it is shown in the Figure 1. Figure 1 also shows that Friday 9/25 was a day off and most of workstations were no available. Users in this network have to turn off their computer during not-working hours since this company has a policy that looks for energy savings that includes turning off most of its equipment. Visibly, despite of the policy, some users left some computers on as we can see in the Figure 1 and Figure 2. On average, 23.14% of computers, 8 workstations, were left on after working hours as we can see in Figure 2.

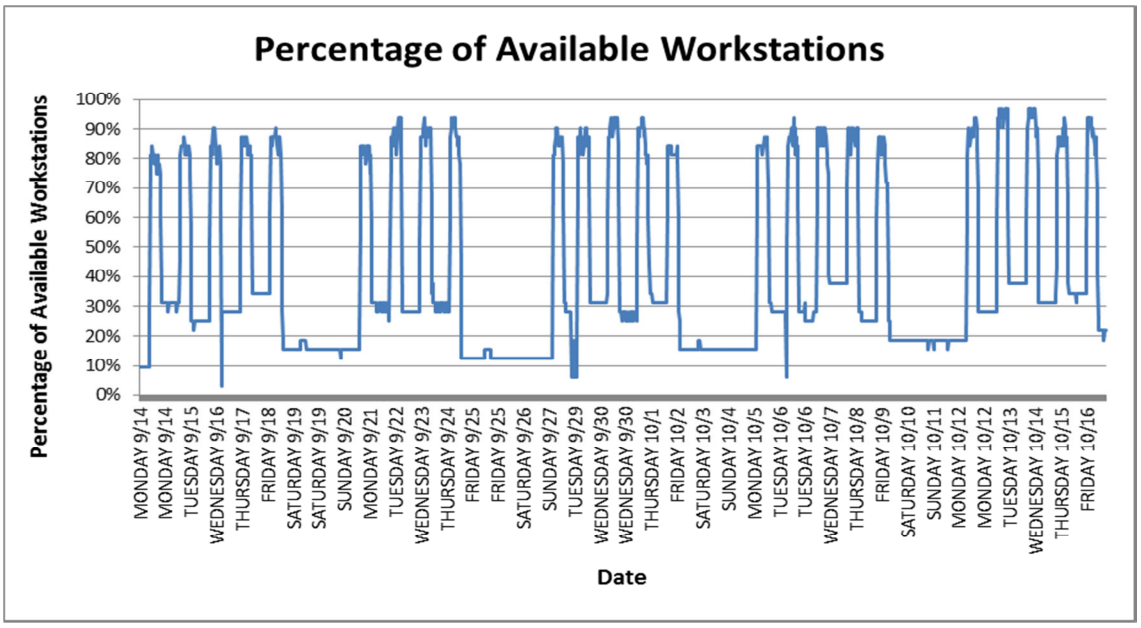

Figure 1 – Workstation availability

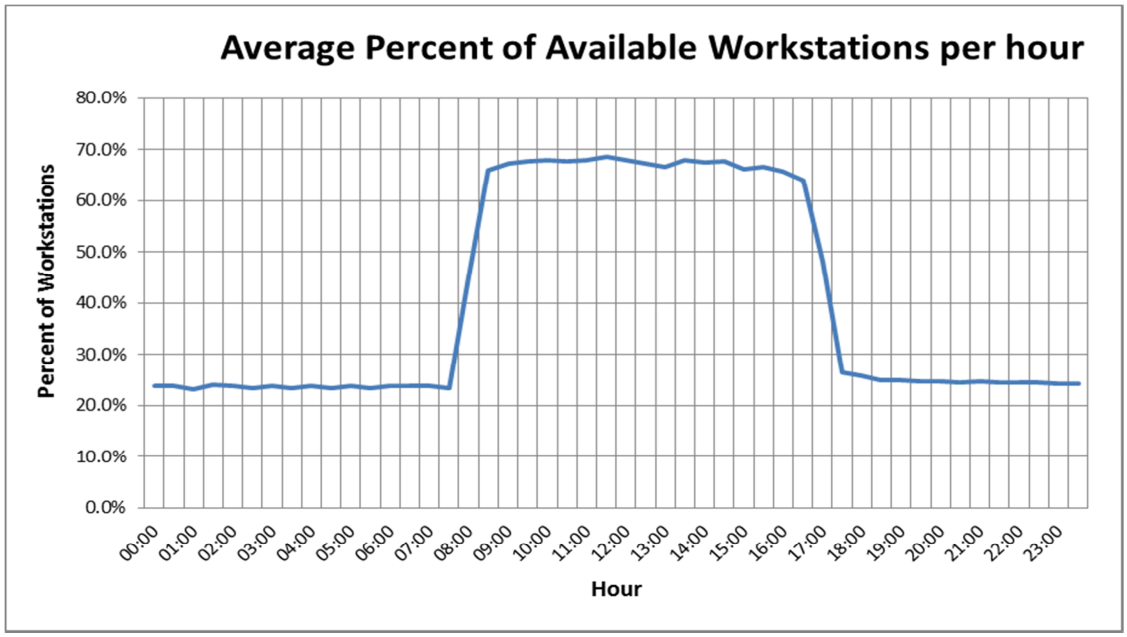

Figure 2 – Workstation Availabiltiy by hour

However, there is not guarantee that the same computers will be left on after work; therefore, we cannot count on remaining computers after working hours since these computers availability are very unstable as we can see in Table 2. This scenario is very

chaotic and certainly can negatively affect any system that wants to take advantages of computer's resources. Because of this, I decided to focus my study on the working days, working hours, and I excluded after-work hours' information. During working hours, we have a more stable environment and high likelihood of being successful in harnessing resources for a backup system as we will see in the following sections. Despite this company work from 8:00 am to 5:00 pm, I utilized the study data collected from 8:30am to 4:30pm. I removed the 30-minutes information after 8am and before 5pm because during these time slots the collected information change unsystematically each day due to the random process of tuning on and off workstations when users started to work and when they stop working. In order to make the naming of this time slot easier, I called it as Active Hours-Working Days (AHWD).

Table 2 – Workstations left on after work during the first six days of study

Day	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6
Workstation Name	WS017	WS017	WS017	WS017	WS007	WS017
	WS037	WS037	WS037	WS037	WS017	WS037
	WS039	WS039	WS039	WS039	WS018	WS039
		WS041	WS041	WS041	WS037	WS042
		WS043	WS043	WS042	WS039	WS044
		WS050	WS052	WS043	WS041	
		WS052	WS057	WS052	WS042	
		WS057	WS063	WS057	WS043	
		WS062		WS062	WS052	
		WS063			WS057	
					WS062	
Total	**3**	**10**	**8**	**9**	**11**	**5**

9.2 Workstation availability during AHWD

By using the information collected during AHWD, I was able to create Figure 3. Figure 3 illustrates how the computers availability acted during the study. On average, 86% of workstations were available during AHWD and never were available less than 69% of workstations.

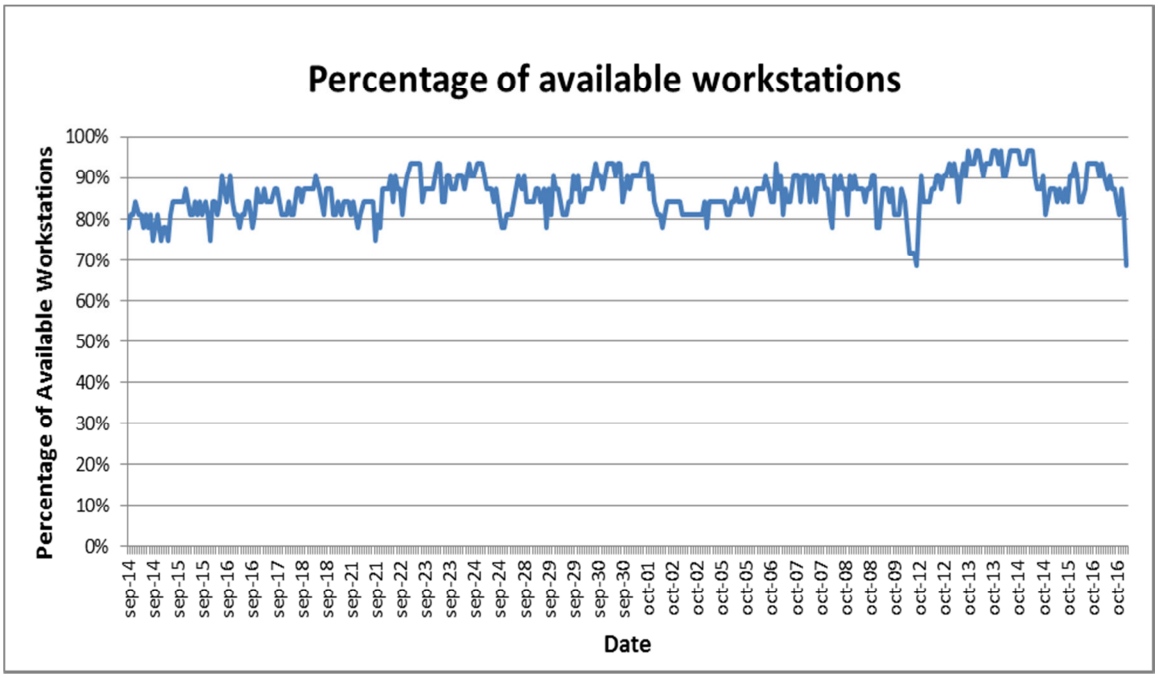

Figure 3 – Workstation availability from 8:30am to 4:30pm in working days

This information indicates that during AHWD, it is possible that a system take advantage of combining workstations resources because of its high availability; the workstations availability has a range of change of 17%, nearly 5 computers.

9.3 Workstations unused hard disk space

One important question is how much unused space had each workstation and whether this space was significant and persistent. Using Table 1, I was able to calculate the combined workstations hard disks capacities, which is 2.1 terabytes. At the beginning of the study, the total combined workstations unused space was 1,651,022 megabytes (1.574 terabytes). This unused space represents 76.4% of entire storage capacity; averaging, 62.7% of each workstation file system was empty, which reveals that a significant quantity of unused storage resources was available in this network. At the end of the study, the unused space was 1,645,606 MB (1.569 TB). The workstations unused storage only decreased 5,416 MB (0.32%) when the capturing of information was completed. We can anticipate and conclude that unused space is constant and stable; however, this calculation does not reveal whether the workstations gained or lost unused storage in the course of the study. It is possible that the combined unused storage could decrease significantly during several events and later increase hiding these changes that could be critical for the proposed backup system. Utilizing the maximum and minimum unused storage for each computer, I was capable of calculating how volatile were the unused storage for each workstations. I utilized the term volatility as was utilized by Huang, Karpovich, and Grimshaw (5), who define it as follows:

$$Volatility_{unused} = \frac{Max_{unused} - Min_{unused}}{Max_{unused}}$$

Figure 4 shows each workstation unused hard disk space. The barely visible red portion at the top of each bar represents how each workstation unused disk space varied in the course of the study. Visibly, the unused disk space on each workstation did not considerably change during the data collection.

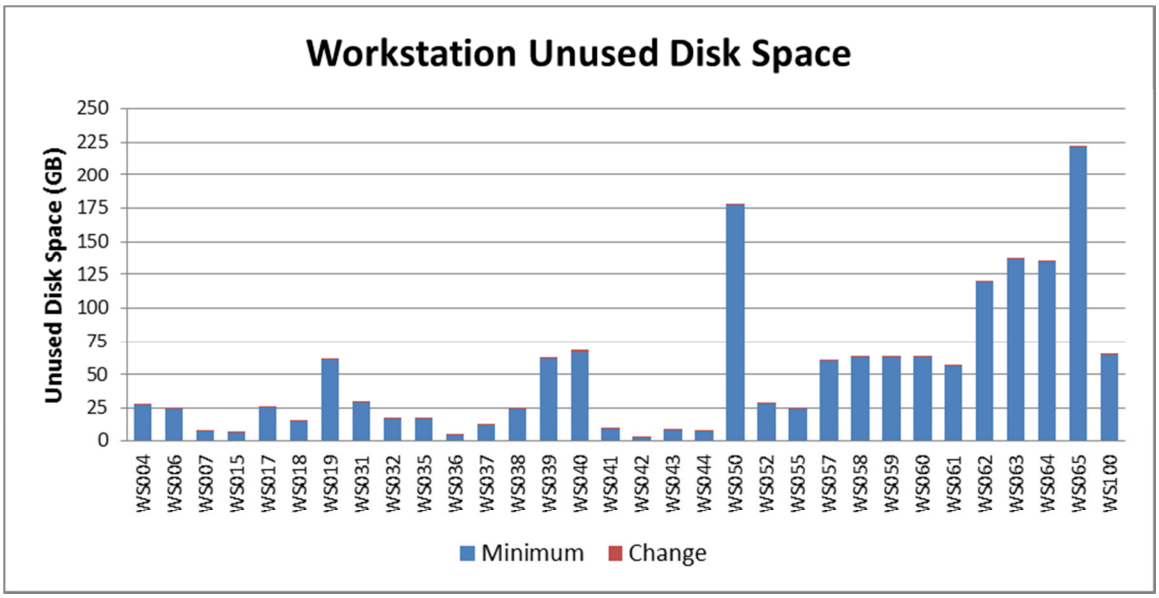

Figure 4 – Workstation unused disk space and change

Looking at Figure 5, we can appreciate the magnitude of the changes on each workstation unused disk space by giving the volatility calculated. The highest volatility was 10.57% and the lowest was 0.07%. Averaging, the volatility was 2.78% for every computer or 578.9 MB of change.

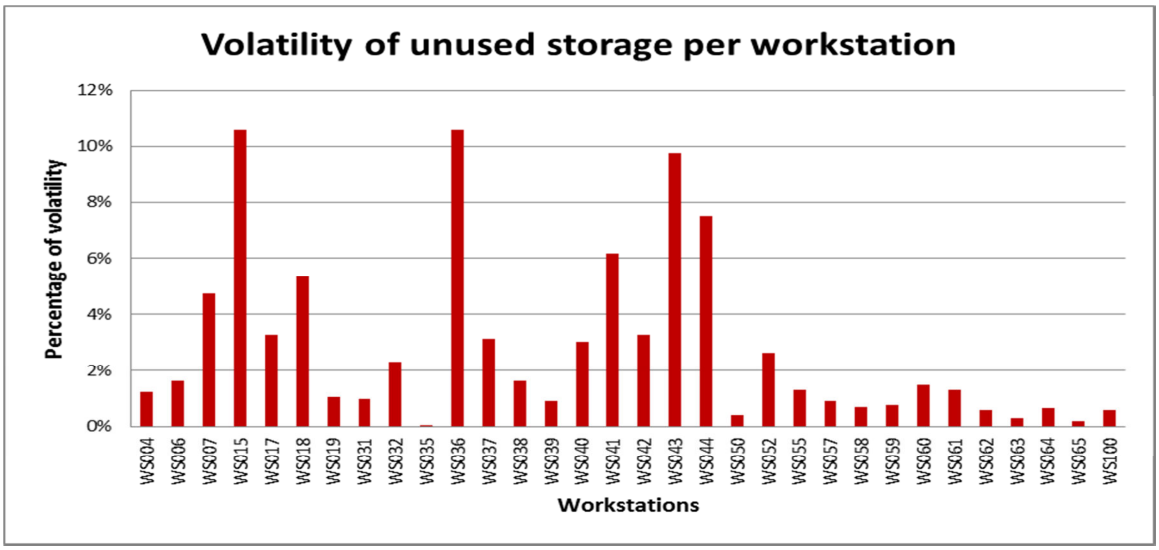

Figure 5 – Unused disk space volatility

The accumulated total change was 18GB, which indicates a volatility of 1% of all initial unused storage, which was 1.57TB. This point out that volatility of the unused storage was really low and we have an opportunity to seize this resource.

Even though the unused storage volatility in the network was very low, also we have to consider workstation availability to actually determine the unused storage available during AHWD. Figure 6 shows all the unused hard disk space available during AHWD. On average, the available combined workstations unused disk space was 1.46 terabytes, with a minimum of 1.12 terabytes, and a maximum of 1.59 terabytes. Using the volatility equation, I calculated that the volatility for the aggregated available unused space during AHWD was 30%. As volatility for each workstation unused disk space is less than the volatility for the combined workstation, I can conclude that in the studied network the workstation

availability had a higher impact in the aggregated available unused space than the operations executed on the workstations that utilized disk space.

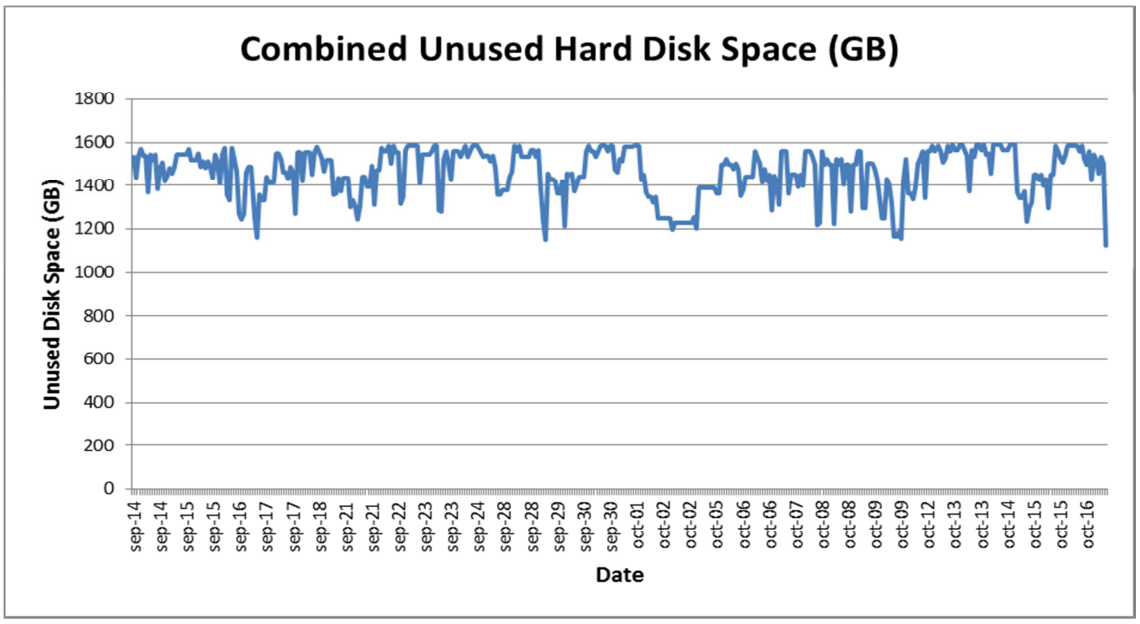

Figure 6 – Combined unused hard disk space

Despite the 30% of volatility caused by the workstations availability, there was always enough unused space because the minimum unused space available during the study was 1.12 terabytes, which is almost five times the space needed for the maximum information in the server to back up.

This evidence brought another question. Is this resource enough to back up server information? In the following section, I analyze how the server information behaved throughout the study.

9.4 Server Information

Since the proposed system is intended to back up server data, it is important to know how much data we have in the server. Thus, we can identify whether the unused storage in the workstations are enough for backing up server data.

As shown in Figure 7, server used space increased in the course of the study. On average, the total data was 236.25 GB, where the maximum was 251.37 GB reached at the end of the study and the minimum was 222.18 GB, which was captured on September 16th, two days after the study begins. The data in the server increased 26.91 GB in the course of the study and had a grow rate of 0.84 GB/day.

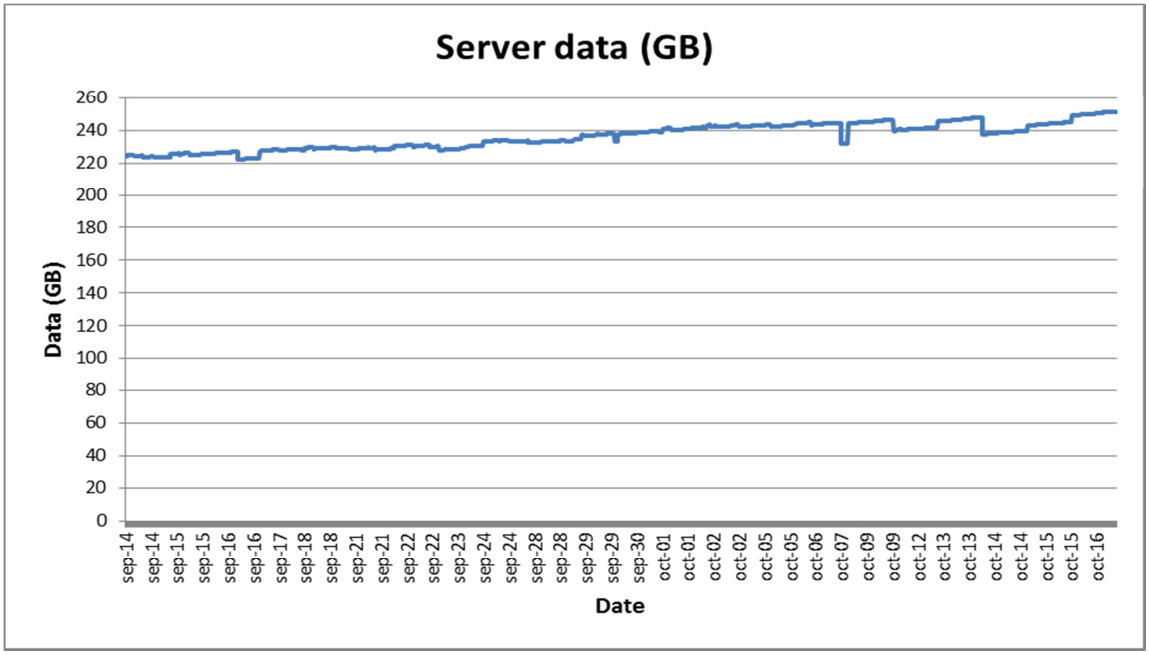

Figure 7 – Server information

The maximum information measured in the server only represented 22.36% of the minimum combined workstations unused disk space. If the information in the server continues

growing as it was growing during the study and no computers is added to the network, the quantity of information in the server will reach the minimum workstations unused disk space in 2.84 years. However, this event can be delayed by utilizing information compression techniques and, eventually, also new workstations will certainly be aggregated to the network, which will add more unused storage to the system.

As we have seen, there are evidences that we have enough and available storage to back up all server information and that this information is also growing at a rate that allows a system to seize the unused storage resources in the network. However, it is necessary to evaluate the remaining resources necessary for the system.

9.5 Workstations processors load

The next step I followed was to determine whether there are enough computational resources in the workstations that allow supporting the client application to be executed on them. Average processor loads on each workstation are illustrated in Figure 8. Averaging, all workstations underwent 6.5% of processor load. As we can appreciate, the workstation processors were not significantly loaded, except for workstations 36 and 35. I found out that both workstations are used for intense cloth label printing process. Due to their critical mission, these workstations are not appropriated to be used in the proposed backup system. Despite of this situation, the backup system would not be significantly impacted by excluding

these workstations; both workstation unused disk space represent only 1.33% of all combine unused disk space in the network.

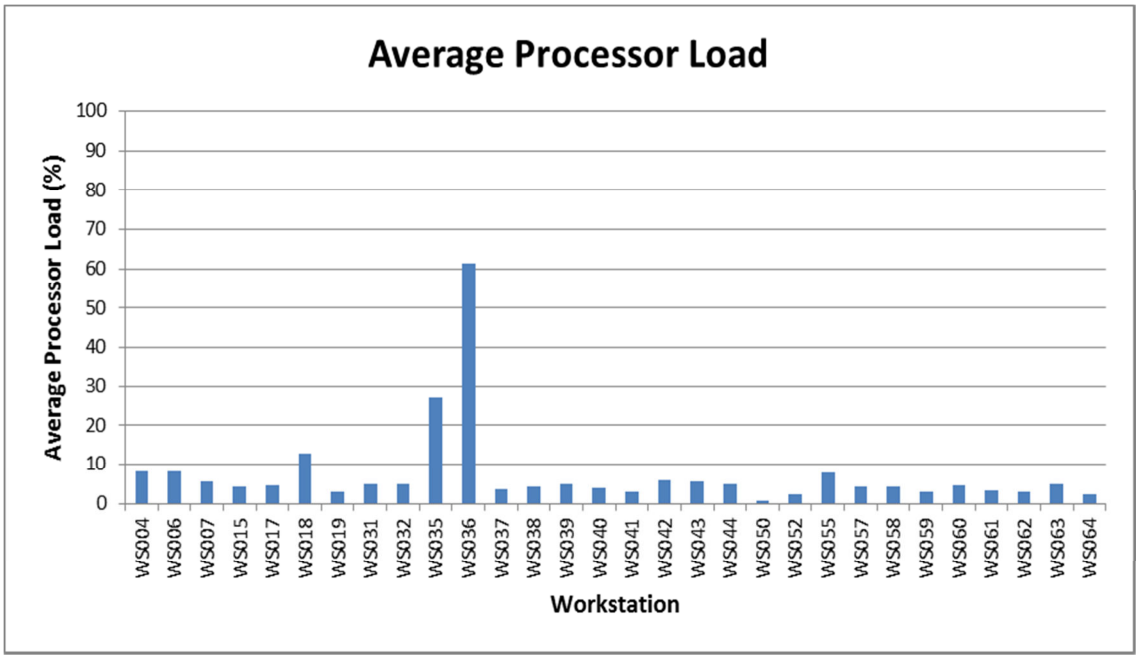

Figure 8 - Average processors load

9.6 Available RAM memory

Another resource needed to implement the agents in the workstations is RAM memory. As this agent would always reside in physical memory, the physical RAM memory available in the workstations determines whether there is the capacity to install an agent on the workstations. Figure 9 shows the average free physical memory available on each workstation during the course of the study. On average, 242.8 MB is available on each workstation as it is shown in Figure 10. I identified that lowest available memory measured during the study was 10 MB in the workstation WS060; however, this is an unlikely situation since this workstation averaged 110 MB of available memory with standard deviation of 31.8

MB. Certainly, there were enough RAM memory resources to implement agents in the workstations that manage the interaction between workstations file systems and the central backup application on the server.

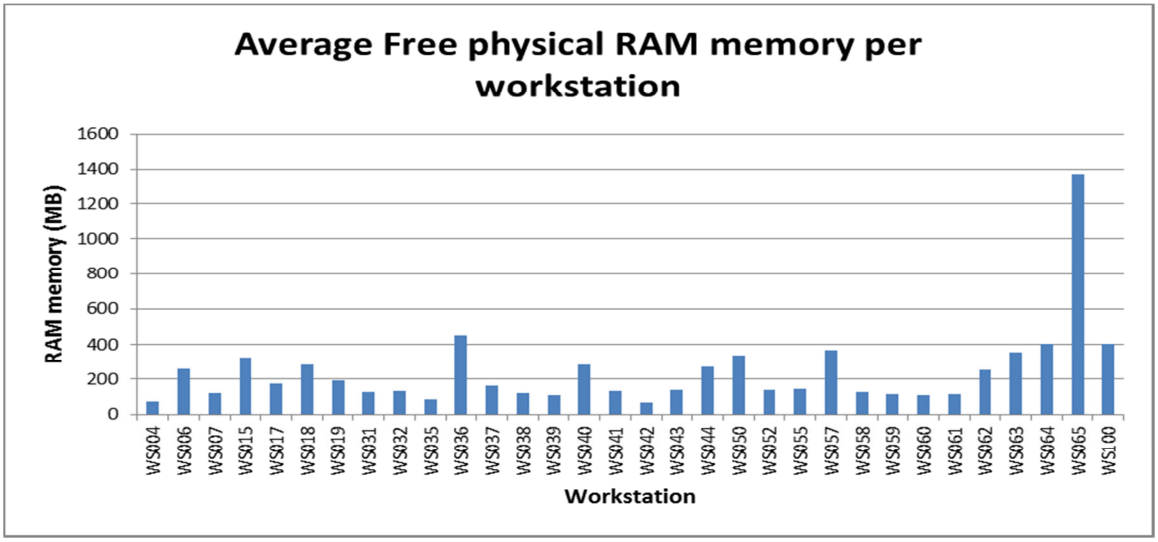

Figure 9 - Average available RAM

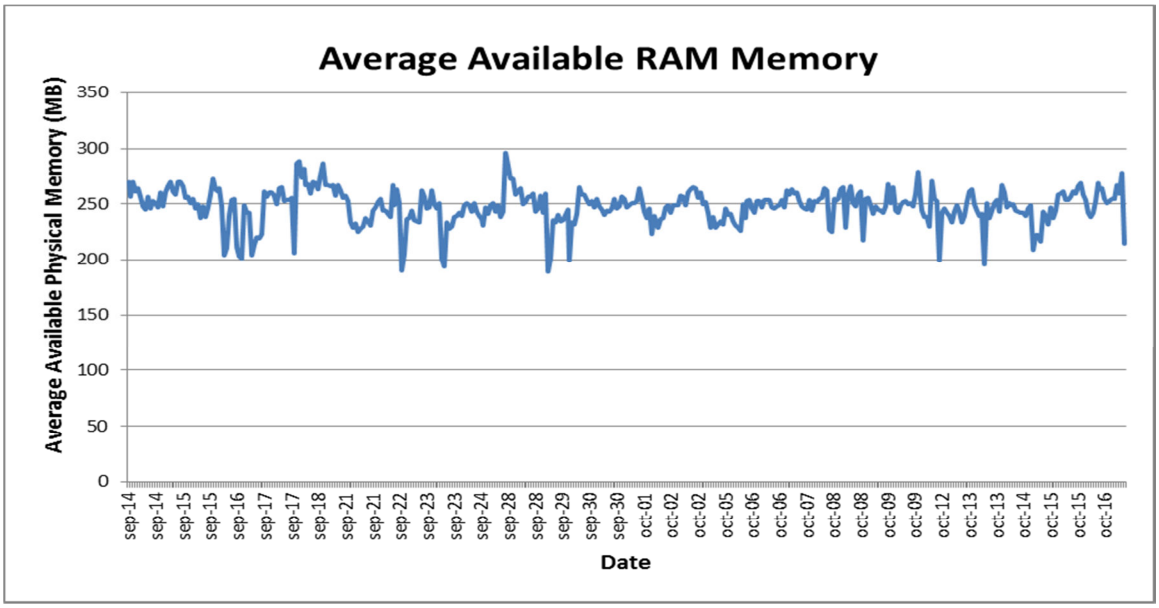

Figure 10 - Average available RAM

9.7 Network Utilization

The last resource I analyze in this study is network utilization. Network utilization is a critical factor for any system that utilizes network distributed resources to achieve its mission. Regularly, all communication in this network is between server and workstations. Communications between workstations are very unusual; therefore, the utilization of server network connection link was an excellent indicator of workstations network utilization. The server connection link had a capacity of 1Gbps. Figure 11 illustrates that the network utilization for server network connection link is very low; averaging, it was 0.153% of the link capacity, exactly 1.53Mbps.

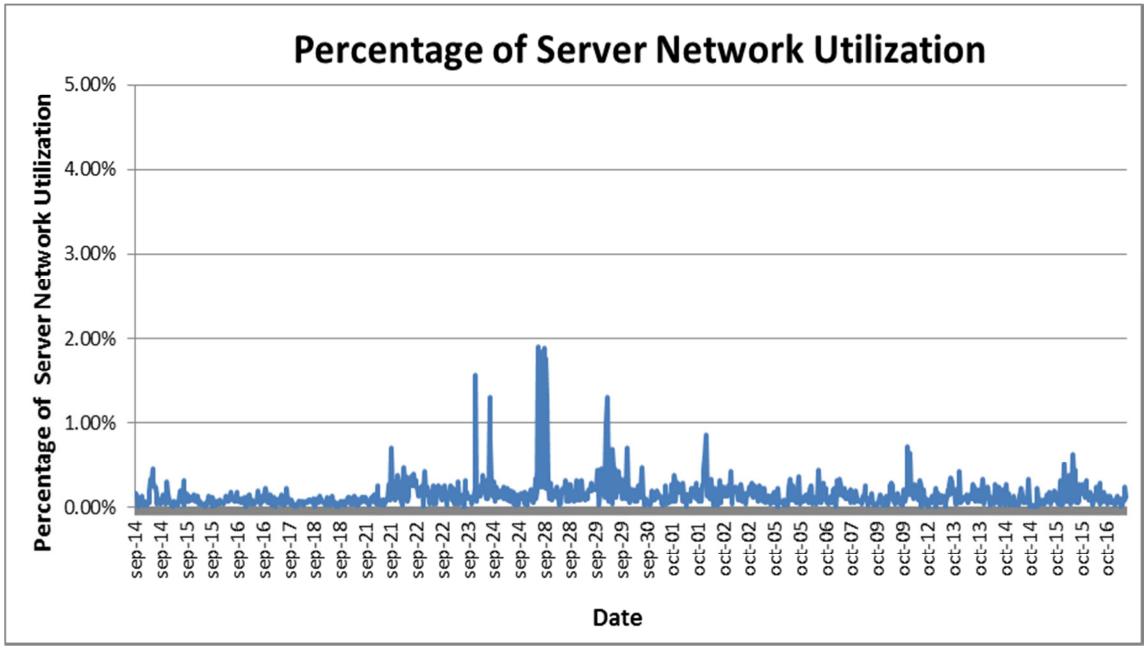

Figure 11 - Server network utilization

I also measured the network utilization for every workstation network link in order to have a direct source of information to determine how network connections had been utilized on each workstation. Results are shown in the Figure 12 and Figure 13.

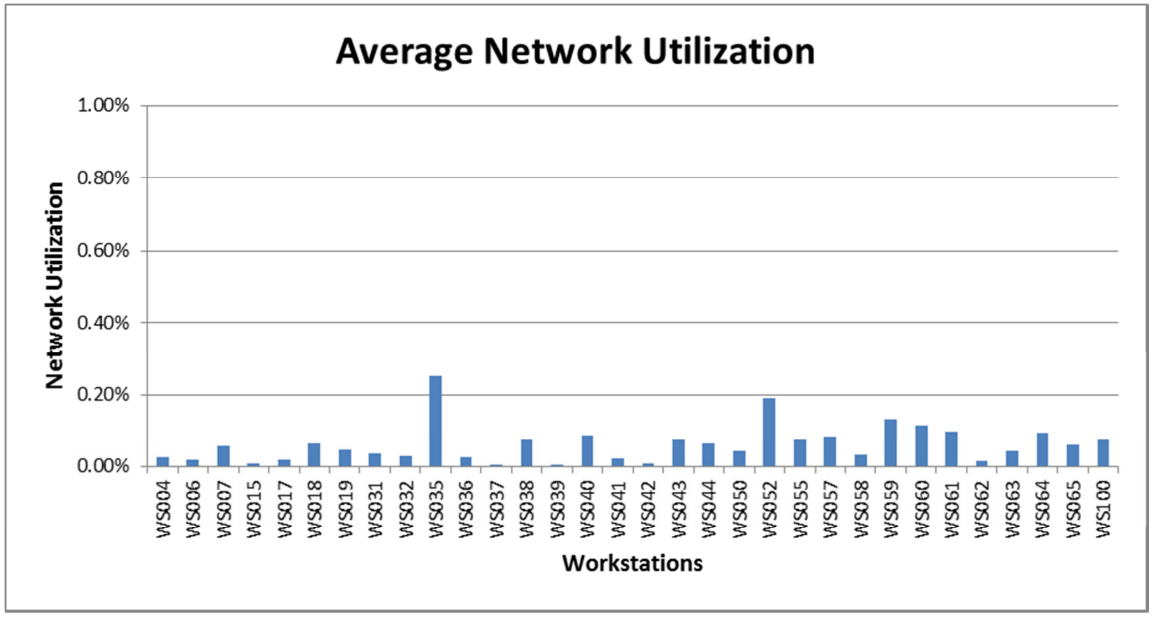

Figure 12 – Average Network Utilization

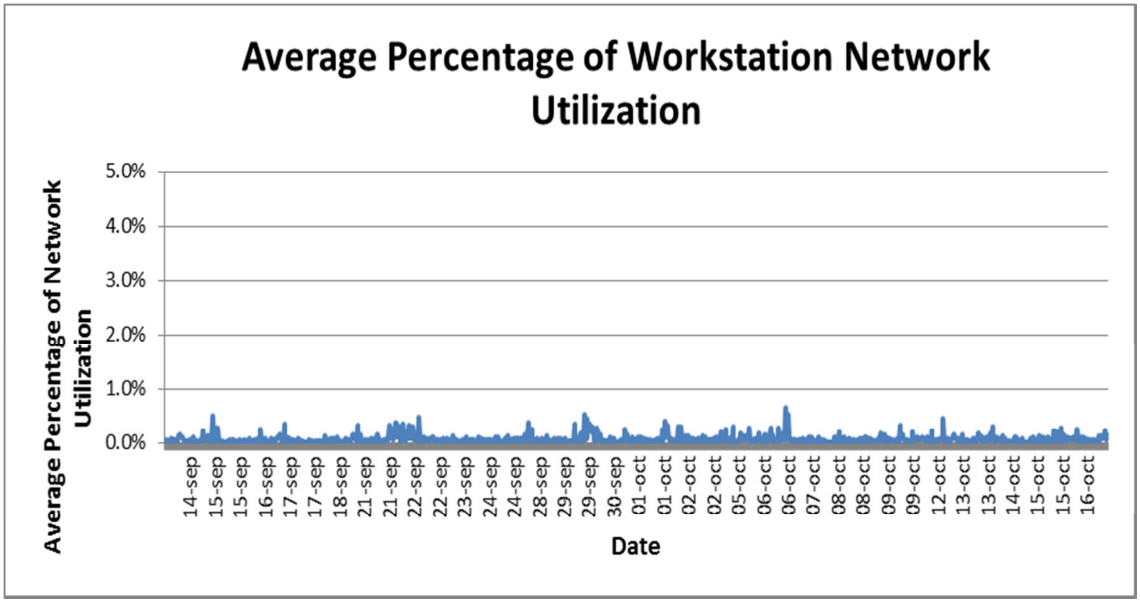

Figure 13 - Average workstations network utilization

In this regard, we can observe that network utilization in this network is considerably low and the proposed backup system could utilize this resource without affecting normal operation of network connections.

9.8 Using RAID Technology

In this session, I briefly analyze the implication of using Redundant Array of Independent Disks (RAID) technology in the proposed system. RAID technology might be an adequate approach to achieve the mission of the proposed backup system since this technology provides the redundancy and reliability required. Since RAID technology needs that each disk or partition that participates in the array must have the same capacity, we need to evaluate each unused space on every workstation in order to determine the total storage that would be available to RAID array. This storage would determine how big would be the total array, which would also depend on the RAID level utilized.

Figure 14 shows each unused space on each computer ordered from smallest to the biggest. As we can appreciate, if we want to use all the unused space we need to use a partition as big as the smallest unused disk space in the network. However, using all workstation with the smallest partition does not signify that we would have the maximum space available to the RAID array we can have. Figure 15 shows how the selection of partition sizes change the total combines partitions size available to the RAID array and how many computers would participate in the array. This brings a tradeoff between redundancy provided by the number

of workstations and array size, which are determine by the partition selection. As we can see in Figure 15, the best selection is a size partition of 55MB, which provide 770MB to the RAID array and in which participate 14 computers.

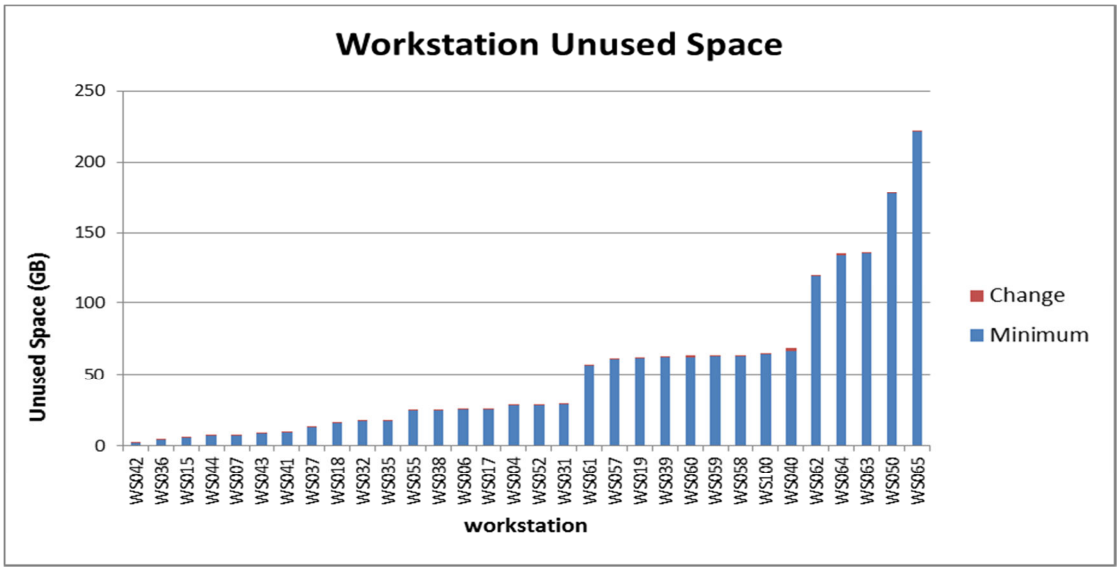

Figure 14 - Workstations unused space ordered from smallest to the biggest

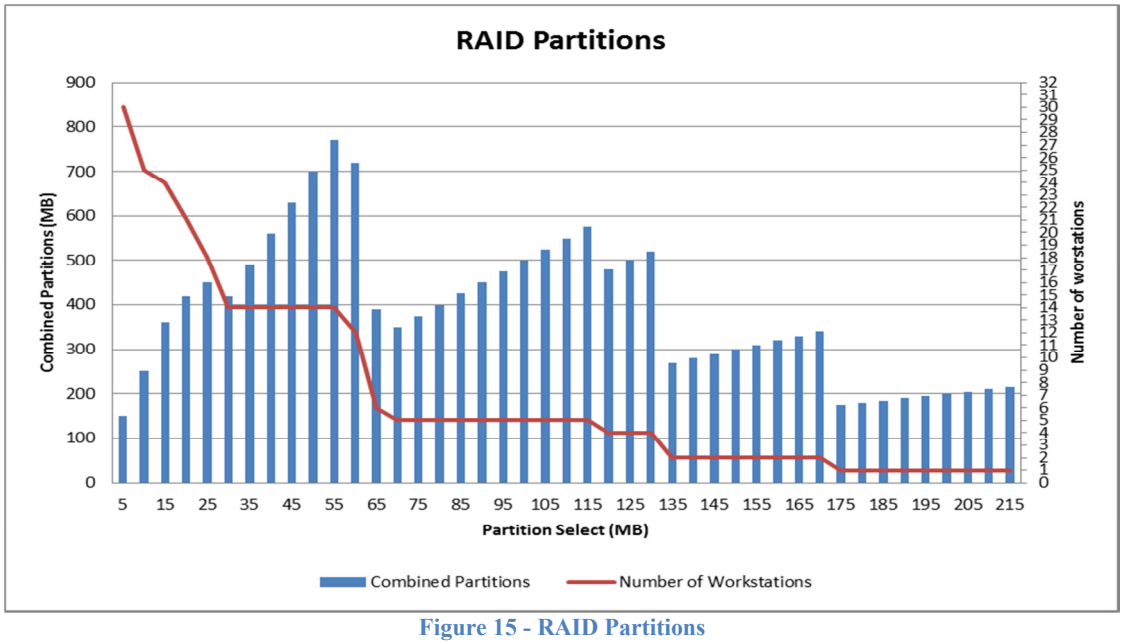

Figure 15 - RAID Partitions

10. Conclusion

In this research project, I analyzed the feasibility of utilizing unused workstations hard drives space as a distributed storage in a small-size organization to backup server information. I used a hypothetical proposed backup system in order to have an initial point of reference to conduct my study. This backup system would achieve its purpose by utilizing two kinds of applications: a server application and client application, also known as agent. The server application would have the function of central management tool. On the other hand, the agent application would be an intermediary between the server application and the file systems in the workstations.

By collecting several metrics, I was able to found out evidence that there were conditions and resources in the studied network to implement a Client-Server model based backup system that would allow to backup up server information in the workstation unused space.

The first evidence is that minimum combined workstation unused hard disk space was approximately five times bigger than the data stored in the server and this unused space is very stable because of the low volatility of the unused space on each workstation. Workstations Processors loads were another indicator in favor of the proposed backup system. Only two computers were heavily loaded during the course of the study; however, the remaining workstations barely reached the 15% of processor load as was shown in Figure 8. In order to overcome a possible processor overload some measure may be suggested. The client application might wait for slot time when the processor load is low and

stable or it might report to the server application that the computer is no appropriate for the system.

RAM Memory availability for agent application was also enough on every workstation, as well as network utilization, which never reached more than 0.3%, on average, on any workstation.

Additionally, I observed that the proposed backup system would find better conditions to operate as long as it is utilized in working days and from 8:30 am to 4:30 pm, when we have the highest workstations availability.

Nevertheless, another alternative to operate in different time slots than the suggested is that the backup system remotely turns on the needed computers to back up server information; however, this approach has an evident drawback. It will significantly increase company electrical expense.

RAID technology might be adapted to be used to develop the proposed backup system. But, it is necessary to take into account that the requirements of this technology introduce a tradeoff between redundancy and size of the array that is established by partition size selected.

11. References

1. **Batten, Christopher, et al., et al.** *pStore: A Secure Peer-to-Peer Backup System.* s.l. : MIT, 2001.

2. **Bolosky, William J., et al., et al.** *Feasibility of a Serverless Distributed File System Deployed on an Existing Set of Desktop PCs.* s.l. : SIGMETRICS, ACM, 2000.

3. **Cox, Landon, Murray, Christopher and Noble, Brian.** *Pastiche: Making Backup Cheap and Easy.* s.l. : USENIX Association 5th Symposium on Operating Systems Design and Implementation.

4. **Hartman, John H. and Ousterhout, John K.** The Zebra Striped Network File System. s.l. : ACM Transactions on Computer Systems, 1995, Vol. 13, pp. 274-310.

5. **Huang, H. Howie, Karpovich, John F. and Grimshaw, Andrew S.** *A Feasibility Study of a Virtual Storage System for Large Organizations.* s.l. : IEEE Computer society, Second International Workshop on Virtualization Technology in Distributed Computing, 2006.

6. **Landers, Martin, Zhang, Han and Tan, Kian-Lee.** *PeerStore: Better Performance by Relaxing in Peer-to-Peer Backup.* s.l. : Proceedings of the Fourth International Conference on Peer-to-Peer Computing, 2004.

7. **Morcos, Faruck, et al., et al.** *iDIBS: An Improved Distributed Backup System.* s.l. : Proceedings of the 12th International Conference on Parallel and Distributed Systems, 2006.

8. **Vazhkudai, Sudharshan S., et al., et al.** FreeLoader: Scavenging Desktop Storage Resources for Scientific Data. s.l. : Association for Computing Machinery, 2005.

9. **Chen, Yan, et al., et al.** *Data Redundancy and Compression Methods for a Disk-based Network Backup System.* s.l. : Proceedings of the International Conference on Information Technology: Coding and Computing, 2004.

10. **Qu, Zhiwei, et al., et al.** *Efficient Data Restoration for A Disk-based Network Backup System.* s.l. : IEEE, 2004.

11. **Xu, Lihao.** *Hydra: A Platform for Survivable and Secure Data Storage Systems.* s.l. : Association for Computing Machinery, 2005.

12. **Oppenheimer, Priscilla.** *Top-Down Network Design, Second Edition.* s.l. : Cisco Press, 2004.

13. **Wilson, Ed.** *Microsoft Windows Scripting with WMI: Self-Paced Learning Guide.* s.l. : Microsoft Press, 2006.

14. **Douceur, John R. and Bolosky, William J.** *A Large-Scale Study of File-System Contents.* Atlanta : Proceedings of the 1999 ACM SIGMETRICS international conference on Measurement and modeling of computer systems, 1999. pp. 59-70.