Who Moved My Data? A Backup Tracking System for Dynamic Workstation Environments

Gregory Pluta, Larry Brumbaugh, William Yurcik, and Joseph Tucek – NCSA/University of Illinois

ABSTRACT

Periodic data backup is a system administration requirement that has changed as wireless machines have altered the fundamental structure of networks. These changes necessitate a complete rethinking of modern network backup strategies. The approaches of the 1980's and 1990's are no longer sufficient and must be updated. In addition to standard backup programs from vendors, specialized system administration tools are often needed. This paper examines one backup system and the major software components used to implement it. NCSA has developed a Backup Tracking System (BTS)[1] to perform backup operations based on knowledge of the network and when each machine was last successfully backed up. BTS can chronologically list all computers: from those currently attached to the network through those that have ever been attached over the life of the BTS program. BTS also provides information about all backup operations including the time of last attempt, success state, amount backed up, etc. The BTS database also contains the date of the last successful backup for each machine and whether it has at least one VIP user (to be given preferred status during backups) or all non-VIP users.

Introduction

Modern networks of end-user machines are becoming increasingly dynamic and heterogeneous. Operating systems come in various versions of Unix, Windows, or MacOS. Mobile hosts, which may only be available on the network rarely or on an intermittent basis, have become almost as common as desktop workstations. The data on individual hosts can be critical to the success of an organization (for cautionary stories of those who have been victims of data loss without backup see [19]).

A wide variety of backup and data integrity techniques exist, and they vary in cost, features, and effectiveness. Mirroring SAN systems are at the very high end. Such systems can provide real-time on and off-site mirroring and versioning of data as it is modified, and can allow quick recovery from both common and catastrophic failures. At the low end, users can individually manage backups to removable media, such as external hard drives, CD-R, or Zip disks.

While the techniques available to protect data have increased greatly, the management of such protection has not. Systems such as Amanda [14] expect collections of always-on, always-connected Unix workstations. Later commercial products like IBM's Tivoli Enterprise Management Suite [9] and Legato Systems NetWorker [11] have focused on extending support for the backend archive technology and new host operating systems.

Unlike previous systems, BTS is aware of the disconnected nature of modern networks and manages backup based on system and user priority. For example, a missed backup in most backup systems at best causes a host to be bumped in priority for the next scheduled backup. In contrast, BTS backups are on-demand, and if a host goes beyond the acceptable window without being backed up, a system administrator will be alerted to investigate the reason for failure.

Motivation

The differences in characteristics between older and more modern networks of end-user hosts necessitate revisiting the motivations and goals of data management. Modern networks are dynamic, and it is beyond the capability of current systems to cope with increasingly disconnected machines and backup latency.

The Reasons for Backups

There are many reasons why data backup is a crucial requirement for virtually every organization. The well-known, traditional reasons still hold. Disasters such as a flood and fire strike networks. Users inadvertently delete files and overwrite existing files. Hackers or disgruntled employees do the same intentionally. Disk drives, inherently fragile mechanical devices, fail, and lose all of the data they hold. Additionally, files become corrupted by bad disk sectors, magnetic fields, and improper system shutdown.

Beyond the traditional threats, there are new threats to today's systems. Thieves steal laptops, and the data contained on them, a threat which is much less applicable to traditional workstations and servers.

[1] Funded in part by a grant from the Office of Naval Research (ONR) under the auspices of the Technology Research, Education, and Commercialization Center (TRECC) established at NCSA/University of Illinois.

While the most skilled social engineer will have trouble convincing even naïve users to purposely delete data, even simplistic email viruses trick these same users into running hostile code with depressing regularity. Finally, the threat posed by modern worms dwarfs those of older worms [16], and they are able to compromise every vulnerable machine on the Internet faster than any manual response can prevent [17].

Organizations depend on their computer systems more than ever. Loss of data is therefore more expensive than ever in terms of lost work and downtime. Additionally, the public's increasing awareness of the importance of data security means that data loss has a large negative publicity component. With increasing threats and increasing costs, backups are more crucial than ever.

Lastly, we realize developing a backup strategy is an individual process specialized to specific network, data, and organizational objectives – different strategies work for different purposes. A survey of factors to consider such as contained in [8] provides an excellent planning tool for developing backup strategies.

Properties of Good Backups

In a well-managed network, backup operations are performed on a regular basis. Additionally a good recovery system is essential. During both normal use and recovery, backup operations should be transparent to users. Backup operations should be automatic and not be the responsibility of users. Instead, a system administrator should centrally manage backup and recovery operations. Since backups are a high priority, they should be managed by a person who understands their importance, rather than a new hire or intern. Finally, the scale of modern networks is beyond what can be manually managed. Good management requires human intelligence supported by automated information gathering and management.

Backup Nuances

Networks are categorized in various ways. A static network consists of physically attached workstations with a network structure that only administrators modify, and then only rarely. A dynamic network adds wireless machines and constantly changing physical structure. A homogeneous network consists of similar attached devices all running the same operating system. A heterogeneous network adds a mixture of various devices with different operating systems. Dynamic heterogeneous networks are a superset of static homogeneous networks, and performing backup operations in these networks is more complicated and requires additional tools. This paper discusses backup operations in the more general case, which applies to most modern networks.

We identify four distinct factors that account for backups being more complicated in a dynamic heterogeneous network. The proliferation of laptops exacerbates these factors, as laptops multiply the dynamism of the network.

- Networks consist of both physically connected and wireless computers. Both types need successful backups on a regular basis, yet each requires a different strategy. A physically attached workstation can be scheduled for backup when the network workload is light. A wireless computer requiring a backup must be scheduled while it is currently connected.
- Some computers may go days, weeks, or longer without logging onto the network. These machines cannot be backed up at a fixed scheduled time each night. To effectively backup these computers, a system must maintain information identifying the last time a computer logged on and the time of the last successful backup.
- Some users have multiple computers, such as several laptops and a workstation, which they periodically switch among. A person may use several machines in the same day and then use one machine exclusively for several months. All of the machines must be backed up.
- A machine can have multiple users who perform different types of processing. One user's job function may require preferred treatment of the machine during backups. Although some users are aware of the importance of backups, most are not and want no role in the backup process. Finally, some users' work habits are not conducive to good backup practices.

In a dynamic environment, the networked computers must initiate the backup operation, since the backup server does not know who is attached at a given time. Hence, software installed on each networked computer must coordinate data exchanges with the backup server. Whenever a new computer is added to the network, the backup client program should be part of the initial software load. Existing computers also need the client software installed.

Client software installation requires knowledge of which computers are actually present on the network. There may be no central point of control to identify when a new computer is added to the network. Likewise, existing computers can be permanently removed from the network without informing any authority. When a machine without backup software that has not been seen for months suddenly reappears, the machine's user needs to be contacted to install the software. Another computer not seen on the network for a comparable time may never reappear again. It would be a waste of time to contact its user.

Related Work

This section highlights backup systems or applied research with relevance to BTS. A comprehensive summary of all backup systems could not be included here, so we have selected a cross-section of the previous work. For a more comprehensive description of backup system issues and examples, see [8, 5].

Amanda (Advanced Maryland Automated Network Disk Archiver) is an early example of freely available backup management software [1, 14]. It uses a combination of full and incremental backups to concurrently backup networked clients to a single designated backup server and uses configuration files to determine the type of backup to perform. It has research significance in that it attempts to minimize cumulative overall backup per day in terms of number of backup runs, percentage change per backup run, and total amount of data [8]. Multiple commercial systems [6, 9, 11] now provide Amanda-like functionality; however, none deal gracefully with wireless hosts.

RAID [3] can protect systems against the failure of individual components. It provides no protection against unintentional/unauthorized modification of data, nor from catastrophic failure. Traditional RAID systems are impractical to field for mobile systems. However, a more recent RAID paper [15] applies RAID to a group of disconnected and distributed computers sharing storage remote from them. The aim is to provide a reliable RAID storage system that delivers acceptable performance while also providing a single coherent namespace for disconnected personal devices.

[4] proposes a taxonomy for backups, including categories such as full versus incremental, file versus device, online or not-in-use, snapshot, and copy-on-write etc. It then places well-known backup programs including xdump, tar, IBM ADSM, Legato Networker, Amanda, Plan9, and Andrew into the above categories.

Versioning file systems, such as Elephant [13], can protect against unintentional/unauthorized modification of data. However, a determined attacker can cause the history to be modified in undesirable ways. Even total versioning file systems like S4 [18] are no protection against physical failures.

[10] evaluates four backup algorithm strategies: (1) incremental, (2) daily-full, (3) mixture of full-incremental, and (4) concurrent backups using backup streams. The paper compares the efficiencies of these algorithms for both backup and restore operations.

In [7] a group of computers form a peer-to-peer network for backup operations. Data from one computer is distributed over other computers that have available capacity. The paper raises many non-standard backup issues related to confidentiality, integrity, authentication, and various other security issues. This is not currently a viable commercial solution but a very interesting paper nonetheless that may have future intranet applications.

The unique feature of BTS is its ability to prioritize backup based on system and user priority. The closest related work in the spirit of BTS is [2] which examines dependability in infrastructure systems by placing priority on components based on their utility in terms of economics and operations research. BTS carries this utility concept forward specifically as an ongoing backup process controllable by the user.

BTS System Description

To perform backup operations, a backup system must know which computers comprise the network and when each was last successfully backed up. Hence, in addition to the backup software and server, the BTS program monitors all networked computers and tracks backup status information, including the last successful backup date. BTS utilizes a database to manage this information on every computer that has connected to the NCSA network during the past several years. The relationship among these components is illustrated in Figure 1.

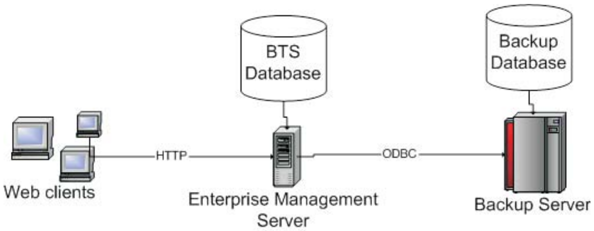


Figure 1: Relationship between Clients, BTS, and Backup Server.

BTS tracks whether users are logged-in via the NCSA authentication system and creates records of this activity at sixty-minute intervals. BTS also polls network machines to see if they are online. Then, for each online host, BTS uses algorithms based on system/user priority and time- since-backup to determine if an immediate backup is needed. If a backup is performed, the files are downloaded to a Mass Storage backup system containing 6 TB of disk cache and an ADIC tape library with six tape drives. An extensible database containing the specific information used to determine backup priority is used to coordinate information sharing for the different BTS components as well as for archiving and presentation to the user via a web interface. The data contained in the database includes the following: VIP users within the organizational hierarchy, problem machines, host-to-subnet mappings, and subnet-to-geographic location mappings.

BTS performs functions beyond helping with backup operations [12]. It provides information on network computers via the computer user name, IP address, and NetBIOS name. The Tivoli name is the identifier used for the backup processing. If the NetBIOS name and Tivoli node name are different, this will be indicated in the host list. BTS can chronologically list all computers, from those currently attached to the network through those that have ever been attached, over the lifetime of the BTS program. BTS also provides information about backups, including the time last performed, successful or unsuccessful, amount of data backed up, etc. Netview updates the BTS database every ten minutes. Hence only rarely will a network user go undetected. The BTS database also identifies each computer as having at least one

VIP user or only non-VIP users. Machines with VIP users have preferred status during backups. In practice, most computers have a single user categorized as a VIP or non-VIP.

Standard Types of Backup Operations

Historically three basic strategies have been used to perform backup operations, varying in the amount of data backed up and ease of restoration. A full backup backs up all data on a scheduled basis, and requires the most time and storage. However, it is the simplest to understand and the easiest from which to restore data. An incremental backup begins with a full backup. Subsequent backup operations copy only those files that have been modified since the last backup. Hence, every backup after the first includes a relatively small amount of data. Periodically, a new full backup is performed. Each of the partial backups is stored on a separate tape, so restoration involves processing all tapes from the current full copy up through the incremental backups to the most recent. Differential backup is similar to incremental, except that after the initial full backup, a single device is used for all of the incremental backups.

Progressive Backup Operations

None of the three basic strategies (full, incremental, differential) are well suited for a dynamic network environment since dynamics violate the timing considerations the standard techniques require. For example, a wireless user may only remain logged on for occasional short periods. To alleviate these issues a fourth backup strategy is used: Progressive Backup. Progressive backup initially copies all files on a computer and generates a summary report identifying when each file was last backed up and last modified. This report can also contain other file attributes such as size and creation date. The more information stored, the more efficient subsequent backup and recovery operations can be.

Each time a user logs on; a decision is made as to whether a backup operation is needed. Following the initial backup, subsequent progressive backup operations compare current file information with the summary report information. Based on this comparison, the backup only copies new and modified files. Unchanged files are not recopied. Determining the files that need to be backed up often requires more time than actually copying the data. Each backup operation also updates the summary report. Progressive backup can be extended to support versioning, where the most recent several versions and their summary information are saved.

During the backup, data is stored in the summary report relational database. SQL queries can be used to retrieve information about the backed up data associated with a given computer. Using the information about the backed up files in its database, it is possible for restore operations to be easily and correctly performed, something that does not always occur with incremental and differential backups. In some circumstances, these two strategies can restore redundant and even incorrect data.

Depending on the storage media, files copied during the current backup may not be contiguously stored with existing backed up files from the same computer. However, the backup systems' relational database identifies where each file from each computer can be located on the storage media. In this way, the database allows quick and easy restore operations to be performed.

In summary, progressive backups require less server time, minimize the required network bandwidth, utilize less storage media to hold backed up data, and are more accurate and efficient for restore operations than the other types of backups. When the host computer initially contacts the backup server, the server initiates a backup operation immediately for a laptop and schedules a later time for a workstation, typically after the workday ends.

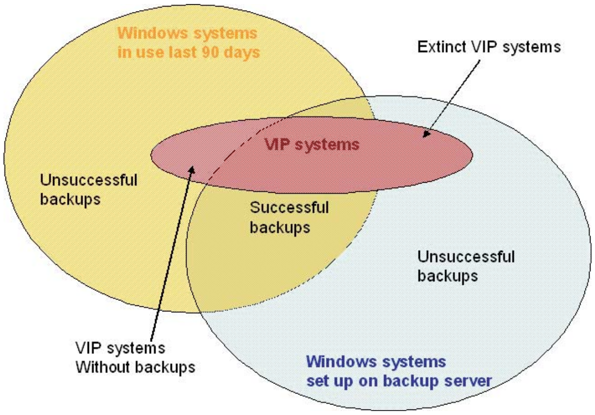


Figure 2: Venn Diagram categorizing backup status of clients.

Hierarchical Backup Strategy

Not all computers attached to a network need to have the same backup strategy – some computers and information are more important than others [2]. The Venn diagram in Figure 2 illustrates the various categories into which a computer can be placed. While performing progressive backups as described above, individual computers are prioritized, and time slots are assigned to each priority. Computers used by a VIP are considered more important and given more attention during backup operations than non-VIP computers. Part of this extra attention is currently provided manually, although the BTS program provides some help. The more important the VIP, the more important it is that timely backups are successfully performed on their machine(s). In the process of getting all existing users to install the client software on their computer, VIPs have been contacted first, in order of their importance. Computers used by the highest-level VIP are considered the most important computers. Computers used by VIPs reporting directly to this person are at the next highest level, etc. At the other extreme, the users classified as least important will be contacted last. The system administrator responsible for backup operations generates this user importance ranking and implements a strategy for contacting the VIPs.

The manner in which a "VIP" is defined will depend on the way that makes most sense for a particular organization. For example, the most important VIPs may be one or more key software developers rather than the organization's president.

A backup failure on a VIP's computer is given priority. Such a failure is investigated and the reason for the failure identified with an entry in the BTS report of all VIPs whose system has not been backed up in the last 10 days. The BTS Reports shown in Figures 4, 5, and 6 illustrate that it is possible to determine which VIP machines have been successfully backed up.

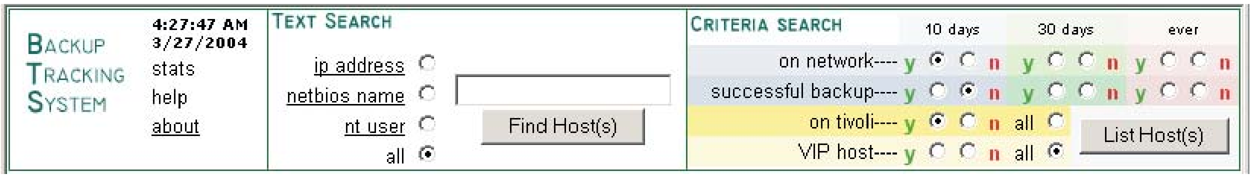
Reports Produced By the Backup Tracking System

When the BTS program starts, it displays a Menu screen that allows several types of reports to be generated. Figure 3 shows the Menu screen prior to entering any data. Default values are provided for every data entry field including the three text fields.

The central portion of the Menu (labeled "Text Search") is used to generate a listing of all computers on the network whose name contains the value entered in the Find Host(s) by Netbios name; Find Host(s) by its NT user name or Find Host(s) by IP address. A partial wildcard may be entered in all three fields. With the IP address, a partial wildcard value is considered the beginning of the address. If nothing is entered, a listing of all computers in the network is generated. Figure 4 shows the results of entering PC in the user name field.

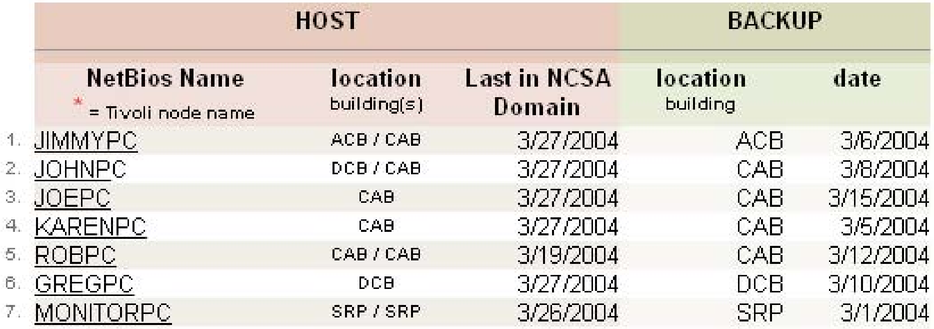
BTS generates reports that identify which machines have had a successful backup performed within the last N days, where N is 10, 30 or if it has ever been backed up. These reports can specify only machines belonging to VIPs, just non-VIPs or both groups. Three reports can be displayed that identify all computers successfully or unsuccessfully backed up within the last 10 days, the last 30 days and since the BTS started running.

Radio buttons on the right hand side of the menu screen allows a user to make selections in four categories to identify which computers will be included in the



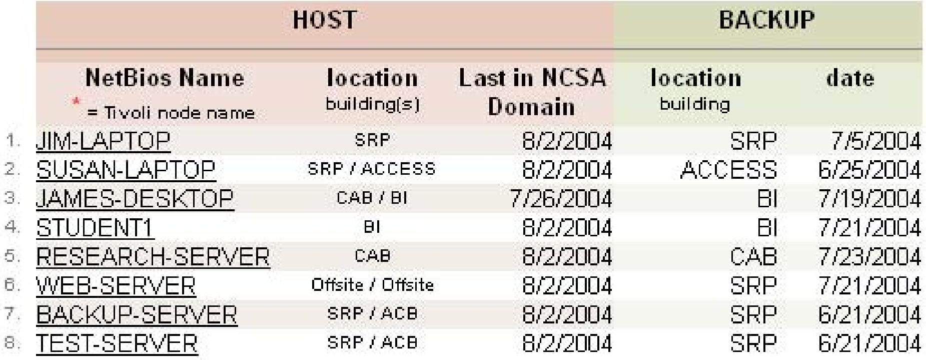
CRITERIA SEARCH	10 days	30 days	ever
on network----	y ● ○ n	y ○ ○ n	y ○ ○ n
successful backup----	y ○ ● n	y ○ ○ n	y ○ ○ n
on tivoli----	y ● ○ n	all ○	
VIP host----	y ○ ○ n	all ●	

Figure 3: Menu screen showing default settings.



	HOST			BACKUP	
	NetBios Name * = Tivoli node name	location building(s)	Last in NCSA Domain	location building	date
1.	JIMMYPC	ACB / CAB	3/27/2004	ACB	3/6/2004
2.	JOHNPC	DCB / CAB	3/27/2004	CAB	3/8/2004
3.	JOEPC	CAB	3/27/2004	CAB	3/15/2004
4.	KARENPC	CAB	3/27/2004	CAB	3/5/2004
5.	ROBPC	CAB / CAB	3/19/2004	CAB	3/12/2004
6.	GREGPC	DCB	3/27/2004	DCB	3/10/2004
7.	MONITORPC	SRP / SRP	3/26/2004	SRP	3/1/2004

Figure 4: Listing of all computers whose user name contains "wildcardPC."



	HOST			BACKUP	
	NetBios Name * = Tivoli node name	location building(s)	Last in NCSA Domain	location building	date
1.	JIM-LAPTOP	SRP	8/2/2004	SRP	7/5/2004
2.	SUSAN-LAPTOP	SRP / ACCESS	8/2/2004	ACCESS	6/25/2004
3.	JAMES-DESKTOP	CAB / BI	7/26/2004	BI	7/19/2004
4.	STUDENT1	BI	8/2/2004	BI	7/21/2004
5.	RESEARCH-SERVER	CAB	8/2/2004	CAB	7/23/2004
6.	WEB-SERVER	Offsite / Offsite	8/2/2004	SRP	7/21/2004
7.	BACKUP-SERVER	SRP / ACB	8/2/2004	SRP	6/21/2004
8.	TEST-SERVER	SRP / ACB	8/2/2004	SRP	6/21/2004

Figure 5: Listing of all computers meeting a specified criteria.

report. One of three choices is selected from VIP Host, Non-VIP Host or Both Categories. One of two choices is selected from Contains Client Software (Tivoli) or does not contain Client Software. One of two choices is selected from Successfully or Unsuccessfully backed up. One of two choices is selected from On or Off the network. For the previous two choices, an additional selection is made from one of three time intervals: last 10 days, last 30 days, never. Figure 5 shows the results of entering VIP host, Client Software Installed, and Unsuccessful backup during last 10 days

Any of the computers in the Figure 5 listing can be selected to have BTS generate a report providing additional general information and the status of recent backup operations for that machine. Figure 6 shows a report of the most recent backup operation results for computer NCSA-SERVER.

Unsuccessful Backup Operations

There are several reasons why some machines are not successfully backed up. Most commonly, the computer does not have a backup client installed on it. BTS is used to identify these machines. To resolve this problem with existing networked computers, it is necessary to contact the user of the machine and install the software, a time consuming process.

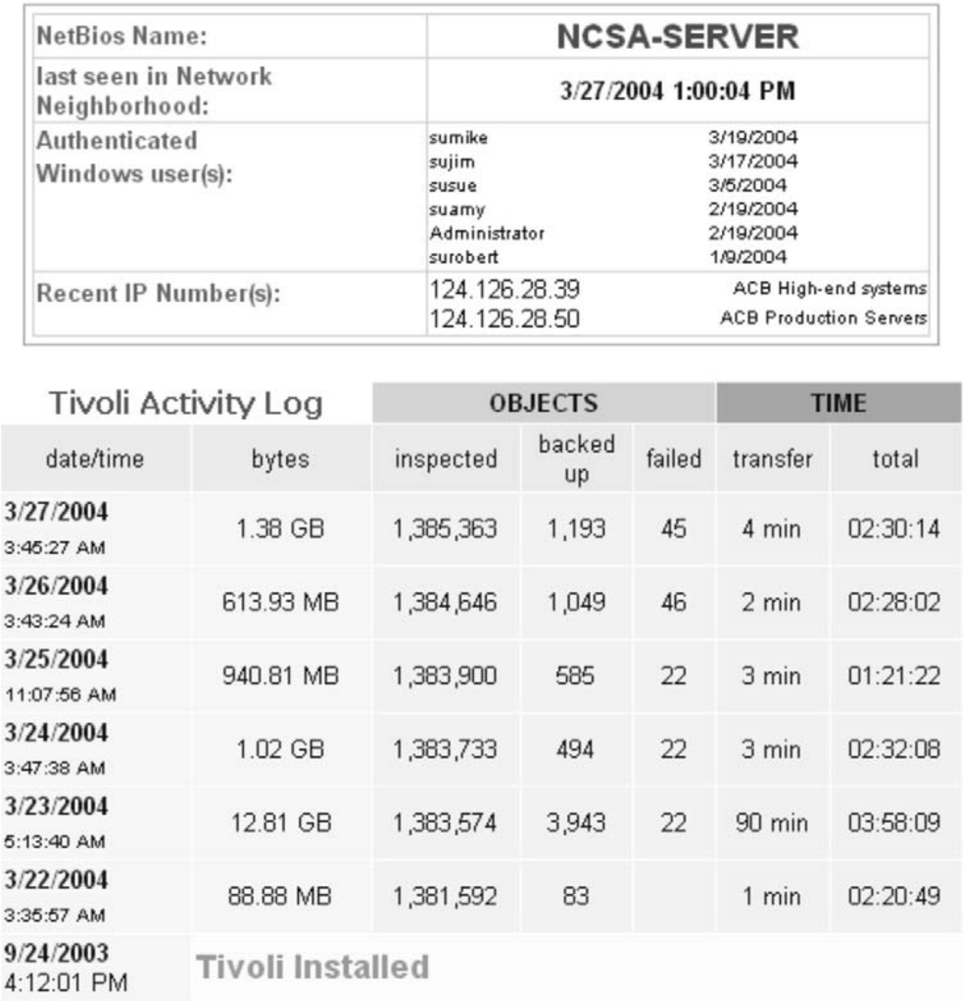
Another possible reason for failure is that the system has a backup client and is part of the network, but is unavailable to backup. BTS also identifies these machines. Many users have multiple machines and are currently using only one of them. If all the machines are not being used simultaneously, the machines not being used are not being backed up because they cannot be accessed. If a workstation is powered off at the end of the day, it cannot be backed up that night.

Failure can also occur when the backup client software is installed on the client, but is incorrectly configured on the server. The server must be scheduled correctly in order to backup the client computer.

Finally, a few backups may inexplicably fail and require a restart of the backup server.

Vendor Product Issues

There are several commercial products that can do the type of processing described in the Progressive



NetBios Name:	NCSA-SERVER	
last seen in Network Neighborhood:	3/27/2004 1:00:04 PM	
Authenticated Windows user(s):	sumike sujim susue suamy Administrator surobert	3/19/2004 3/17/2004 3/5/2004 2/19/2004 2/19/2004 1/9/2004
Recent IP Number(s):	124.126.28.39 124.126.28.50	ACB High-end systems ACB Production Servers

Tivoli Activity Log		OBJECTS			TIME	
date/time	bytes	inspected	backed up	failed	transfer	total
3/27/2004 3:45:27 AM	1.38 GB	1,385,363	1,193	45	4 min	02:30:14
3/26/2004 3:43:24 AM	613.93 MB	1,384,646	1,049	46	2 min	02:28:02
3/25/2004 11:07:56 AM	940.81 MB	1,383,900	585	22	3 min	01:21:22
3/24/2004 3:47:38 AM	1.02 GB	1,383,733	494	22	3 min	02:32:08
3/23/2004 5:13:40 AM	12.81 GB	1,383,574	3,943	22	90 min	03:58:09
3/22/2004 3:35:57 AM	88.88 MB	1,381,592	83		1 min	02:20:49
9/24/2003 4:12:01 PM	Tivoli Installed					

Figure 6: Report on backup information For computer NCSA-SERVER.

Backup Operations section including the Tivoli Enterprise Management Suite [9] from IBM, Retrospect [6] from Dantz Development Corp. and Networker [11] from Legato Systems.

However, some backup products are designed specifically for small networks and do not scale to a network the size of NCSA. At NCSA, Tivoli is used to perform the actual progressive backup operations. Tivoli performs all of the relevant backup processing and does not significantly interfere with regular network traffic. Several other products not mentioned were initially tried but they negatively impacted network traffic. In addition, with some earlier products the server initiated the backup rather than the client, a bad idea in a dynamic networking environment.

The Dantz Retrospect Professional product is designed for home and small offices using Windows and Apple Macintosh computers [6]. It does progressive backups with 100% correct restores and no redundancy (a significant feature). It uses compression and can backup data to any media. However this solution does not scale to larger networks the size of NCSA.

Other products can simultaneously backup dozens of clients. When the client contacts the server to determine whether a backup is needed, the server makes a decision based on the identity of the client computer. Criteria can include the following: the client is a wireless with a VIP user – backup immediately; the client is a server or a workstation belonging to an important VIP – backup every 24 hours; a user workstation – backup every weekday, but not over the weekend; and a VIPs computer not seen on the network for weeks – backup immediately.

Another significant backup issue is how to process the files that comprise well-known applications that are running on most computers. Examples include Microsoft Word, Excel, Access and even the operating system itself. It should not be necessary to copy these files from almost every machine. The backup software can be provided with a list of files to exclude during backups. Alternatively, application software and the operating system can be reinstalled rather than restoring from a backup.

BTS consists of an ASP application written in VBScript running with Microsoft IIS 5.0 on a Windows 2000 server. BTS uses an Access database containing information about the networked computers. The database is distinct from the relational database used by commercial backup software such as Tivoli.

Availability

Various statistics have been collected from the NCSA network, but the most interesting and valuable have been measurements of availability in terms of systems and users.

Let s represent the number of systems on the network at a given point in time, measured every ten minutes. Therefore, the normalized system and variation for a given time period can be respectively defined as:

$$N_s = \frac{\sigma_s}{\bar{s}} \qquad N_u = \frac{\sigma_u}{\bar{u}}$$

In an environment which is extremely static, and therefore the systems are on the network and each user logs in every work day, $N_s = 0$ and $N_u = 0$. As the number of systems and users per day varies, the values of N_s and N_u increase. For example, if the number of users vary an average of ±10%, then $N_u = 0.10$.

$$\bar{s} = \frac{1}{n}(s_1 + s_2 + s_3 + \cdots + s_n)$$

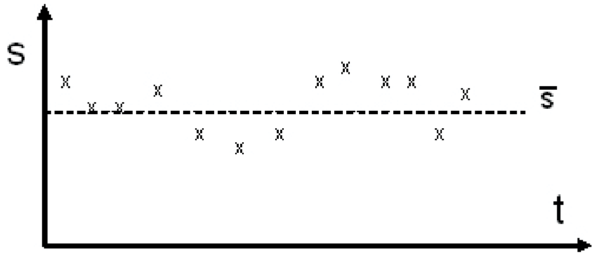


Figure 7: Hypothetical system availability.

Similarly, let u represent the number of users who authenticate on a given 24-hour period, recorded once per day at midnight.

$$\bar{u} = \frac{1}{n}(u_1 + u_2 + u_3 + \cdots + u_n)$$

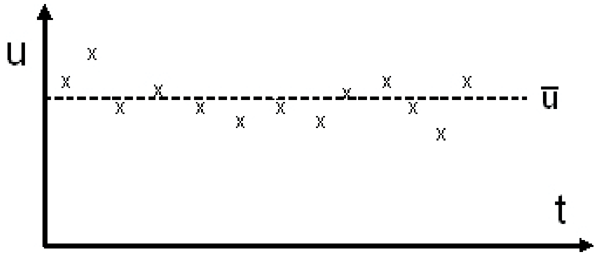


Figure 8: Hypothetical user availability.

It should be expected in environments which have high values of N_s and N_u that it would be considerably more difficult to backup, apply security patches, and track down systems than in an environment with lower values of N_s and N_u. High values of N_s and N_u would therefore imply a higher security risk for a given infrastructure effort, or to put it another way, a higher support cost for a given level of security and survivability. This ability to track usage trends has proved useful for capacity provisioning, security events, and equipment reliability failures.

Figure 9 is a sample of real measurements of system and user availability (respectively) on the NCSA network. Over a three-year period of measurement, the average number of systems available is 300 with 600 distinct systems, an upper limit of 400, and a lower limit of 100. The normalized system variation is 0.10. Over the same three-year period of measurement, the average number of users is about 190 users with 220 as the upper limit and 30 as the lower limit. The normalized user variation is 0.53.

Conclusions

Sharing the general class of backup problems we face at NCSA and our specific implementation solutions have proved to be valuable to peer organizations. We feel that the solutions we describe in this work are transferable to other environments even though this work was specifically targeted to the Windows environment. In fact, we already have a parallel project in progress transferring these same techniques to the Linux environment.

Future directions include examining the possibility of moving some of the functionality of Tivoli onto the client, and have the client perform tasks (such as determining the files to back up) via intelligent algorithms.

Lastly, more information about BTS, implementation instructions, and the software itself are available at this web page http://wegpublic.ncsa.uiuc.edu/bts .

Author Information

Gregory C. Pluta is currently Manager of the Windows Environment Group at NCSA. He is a graduate of the University of Illinois at Urbana Champaign (Aerospace Engineering, BS 1992, MS 1995). As a graduate student, Greg was responsible for the department's student computer scientific workstations, designed and built a nonlinear systems laboratory, and re-designed undergraduate engineering laboratories with modern data acquisition and analysis systems which he taught for two years. Greg then worked at Andersen Consulting as an analyst writing business and multimedia software for two years on more than a dozen different projects in almost as many languages, and as a software configuration management specialist for large software development projects. Greg then joined NCSA as a system engineer, then as manager of Windows desktop and server systems. Greg can be reached at gpluta@ncsa. uiuc.edu .

Larry J. Brumbaugh was born in Pittsburgh, Pennsylvania and graduated from the University of Pittsburgh with a BS in Mathematics in 1965. He obtained an MA in Mathematics from West Virginia University in 1968 and an MS in Computer Science from the University of Kentucky in 1976. He is ABD in Mathematics (Kentucky) and Computer Science (University of Illinois). In a long and varied career, he taught Computer Science for three years at Morehead State University in Kentucky and for 23 years at Illinois State University in Normal, Illinois. He is the author of two computer science books and numerous papers and presentations. His teaching and research

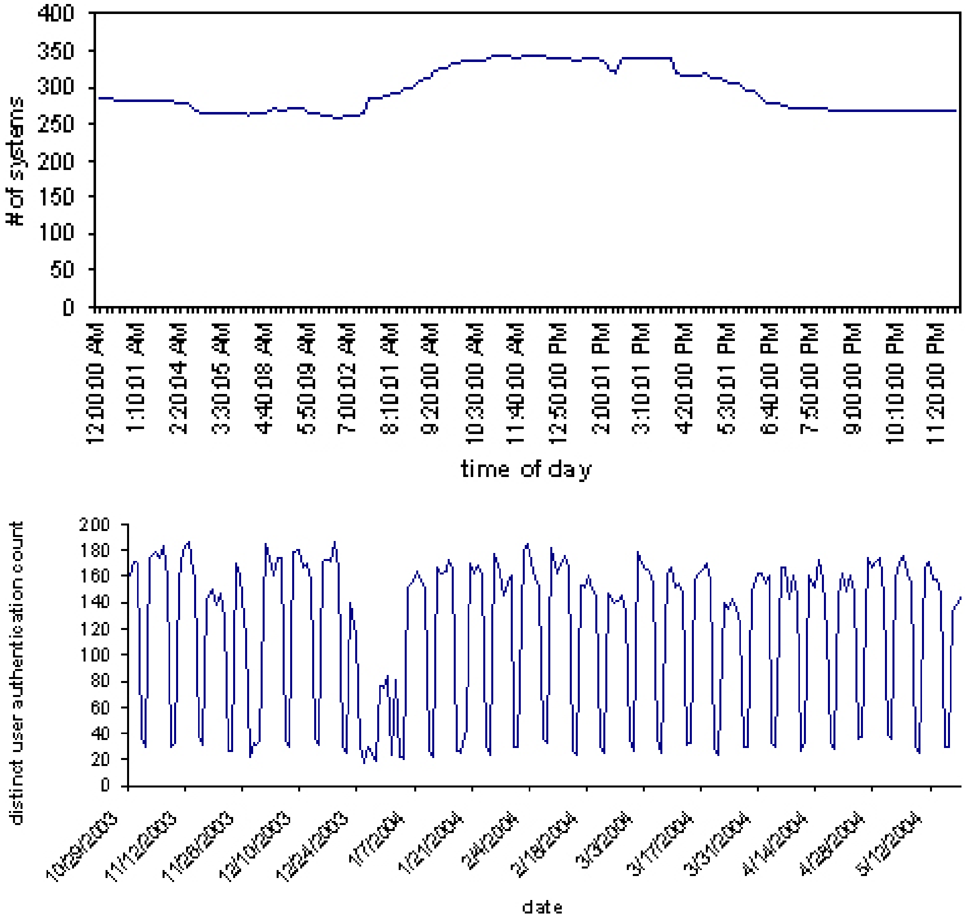


Figure 9: Representative system & user availability measurements.

interests have included mainframe programming, data communications and networking and computer architecture. He has worked as a consultant for many companies and government agencies. He joined NCSA in February 2004 as a consultant on Storage Security. Larry can be reached at ljbrumb@ncsa.uiuc.edu .

William (Bill) Yurcik is currently Manager of Security Research at NCSA. Priorto this he was Manager of Security Operations at NCSA, so he has both a theoretical and practical background in computer network security. He is a graduate of Johns Hopkins University (MS Electrical Engineering 1990, MS Computer Science 1987), the University of Maryland (BS Electrical Engineering 1984), and is ABD from the University of Pittsburgh (1994-99). He had 12 years of professional experience as a Network Engineer prior to joining NCSA (NRL, NASA, Verizon, MITRE). He has also been a Visiting Professor at Illinois State and Illinois Wesleyan Universities for three years, and since 2001 is an adjunct Professor of Computer Science at the University of Maryland. Bill can be reached at byurcik@ ncsa.uiuc.edu .

Joseph Tucek graduated from Washington University in St. Louis in 2003 with a BS in Computer Science and a BS in Computer Engineering. Although he played around with AI and robotics, he has found his true interest in systems. Joseph is currently a Ph.D. candidate in Computer Science at the University of Illinois at Urbana/Champaign, working in the Storage Security group at NCSA for support. Joe can be reached at tucek@ncsa.uiuc.edu .

References

[1] *The AMANDA Homepage*, http://www.amanda.org .

[2] Candea, George and Armando Fox, "A Utility-Centered Approach to Building Dependable Infrastructure Services," *Tenth ACM SIGOPS European Workshop (EW)*, September 2002.

[3] Chen, Peter M., Edward L. Lee, Garth A. Gibson, Randy H. Katz, and David A. Patterson, "RAID: High-Performance, Reliable Secondary Storage," *ACM Computing Surveys*, June 1994.

[4] Chervenak, Ann L., Vivekanand Vellanki, and Zachary Kurmas, "Protecting File Systems: A Survey of Backup Techniques," *Joint NASA and IEEE Mass Storage Conference*, 1998.

[5] Preston, W. Curtis, *Unix Backup and Recovery*, O'Reilly and Associates, 1999.

[6] Dantz, *Dantz Retrospect – Intelligent Backup and Restore*, http://www.nwfusion.com/whitepapers/dantz/whitepaper.html , June 2004.

[7] Elnikety, Sameh, Mark Lillibridge, Mike Burrows, and Willy Zwaenepoel, "Cooperative Internet Backup Schemes," *Usenix Annual Technical Conference*, June 2003.

[8] Frisch, Æleen, *Essential System Administration Third Edition*, O'Reilly & Associates, 2002.

[9] IBM Software, *IBM Storage Management Solutions*, http://www.nasi.com/tivoli_backup recovery.htm , 2004.

[10] Kurmas, Zachary and Ann L. Chervenak, "Evaluating Backup Algorithms," *IEEE Symposium on Mass Storage Systems*, 2000.

[11] Legato Software, *Legato Networker*, http://www.legato.com/products/networker/ .

[12] Pluta, Gregory, Larry Brumbaugh, and William Yurcik, "BEASTS: An Enterprise Management Tool for Providing Information Survivability in Dynamic Heterogeneous Networked Environments," *IEEE Local Computer Networks Conferences (LCN)*, November 2004.

[13] Santry, Douglas S., Michael J. Feeley, Norman C. Hutchinson, Alistair C. Veitch, Ross W. Carton, and Jacob Ofir, "Deciding When to Forget in the Elephant File *System," Symposium on Operating Systems Principles (SOSP)*, December 1999.

[14] da Silva, J., and O. Guomundsson, "The Amanda Network Backup Manager," *Proceedings of the Seventh Large Installation Systems Administration Conference (LISA)*, November 1993.

[15] Sobti, Sumeet, et al., "PersonalRAID: Mobile Storage for Distributed and Disconnected Computers," *Usenix Conference on File and Storage Technologies (FAST)*, January 2002.

[16] Spafford, Eugene H., "An Analysis of the Internet Worm," *Proc. European Software Engineering Conference*, September 1989.

[17] Staniford, Stuart, Vern Paxson, and Nicholas Weaver, "How to Own the Internet in Your Spare Time," *USENIX Security Symposium*, August 2002.

[18] Strunk, John D., Garth Goodson, Michael L. Scheinholtz, Craig A. N. Soules, and Gregory Ganger, "Self-Securing Storage: Protecting Data in Compromised Systems," *Fourth Symposium on Operating Systems Design and Implementation (OSDI)*, October 2000.

[19] *Tao of Backup Wailing Wall Homepage*, http://www.taobackup.com/wailing.cgi , June 2004.