Data Backup and Recovery Options
by Jeff Drake

One of the most fundamental elements of any disaster recovery plan is data backup and recovery. If an IT staff lacks the ability to recover lost data following an incident or disaster, redundant hardware or hot sites are of only limited value.   Relatively few years ago, planning a backup strategy was reasonably straightforward for the small-office-home-office (SOHO) and small-to-medium-sized-business (SMB). Home users were limited to floppies and then more recently writable compact disks. Small to medium sized business owners typically maintained several servers, each with its own tape drive system that administrators manually changed on a regular basis. In more recent years, the options for data backup and recovery have increased for both of these markets to integrate new technologies, greater bandwidth and the need to integrate the data backup process across the entire server room to increase the overall efficiency of data backup operations and reduce the number of labor hours necessary to manage it. Additionally, the types of devices and data needing to be backed up have also grown exponentially and have increased the complexity of data backup and recovery strategies. This paper will describe some of the backup and recovery strategies and options available to the SOHO and SMB IT staff and explore ways to improve upon a traditional one tape drive per server backup solution specifically, as well as considering some of the newer sources of data that must be considered when a backup strategy is being developed.

What is a Backup?

The increased complexity of today's information technology environment has led to the development of new terms and processes in the area of disaster recovery. A backup today is not necessarily what was conceived of as a backup ten years ago. Part of the reason for this is the existence of a sister term/practice of archiving data. Historically data backups were copies of files stored on tapes and stuck in a cabinet in some back room in case of system failure. Now, backups are more specifically designed for immediate recovery of data while archiving of data is more for long term storage of data that may for whatever reason be called for at some later date, typically after its end of useful life (Hope, 2006). So, archiving could be thought of as long term storage of data backups while backups are thought of as short term storage of data for the purpose of disaster recovery. There can be fine line of distinction between to the two depending upon how backups are implemented. It could be, for example, that a disaster wipes out an entire server room through fire.   If all backups are located in the server room then disaster recovery would have to call on the archived data, hopefully stored off-site, to perform disaster recovery operations. So, an archive could be used for disaster recovery in the scenario, but the intent of it is still towards long term data storage in case some audit for example requires administrators to retrieve data from four years ago. This paper will consider to some degree both of these topics, data backups and archiving.

Why Backup?

The question of why to backup data or systems seems a useless question to begin with at this point in IT history. Few would argue with the need to backup data and would have historically cited the potential for hardware failure or user error resulting in data loss as being the reason for data backups. These reasons are no less true today and perhaps are truer than they ever have been considering the value of information to today's businesses and the cost of recreating it. For example, one study reported that if a company experiences a server failure an average of $10,000 in lost revenue is experienced (http://www.symantec.com/smb/library/article.jsp?aid=backup_solution, 2005). Additionally, another study by the US Bureau of Labor Statistics found that of those companies experiencing what was referred to as "significant data loss", 93% went out of business within a five year period (http://www.symantec.com/smb/library/article.jsp?aid=importance_of_disc_based_backup, 2006). This being the case, surprisingly, the study cited above indicating a $10,000 loss in revenue per server crash also found that 35% of its respondents did not backup data on a regular basis. Apparently there is still a disconnect between an understanding of the need for data backup and actual implementation of backup practices.

In addition to the financial implications of data loss, the answer to "Why backup?", has now grown in length to include a long list of legal requirements that have placed strict requirements on the handling of information assets. The most recent legislation to impact many corporations is the "Public Company Accounting Reform and Investor Protection Act of 2002". It is more commonly referred to as "Sarbanes-Oxley" reflecting the names of the two congressmen who drafted the legislation (Cole & Spears, 2006). Sarbanes-Oxley 'impacts all firms publicly traded in the United States and specifically holds executive management accountable for the integrity and availability of all information in financial reports' (Cole & Spears, 2006, p. 2). Notice here the term 'availability'. This means different things to different departments in a publicly traded business, but for the IT department it means that data must be available and old data at that, regardless of fire, server failure or if a wild herd of rhinos prances through your server room; the data must be available. The Sarbanes-Oxley Act more specifically requires that 'all records, including emails, be kept for at least five years' (Hope, 2006, p48).

One article on disaster recovery partially defined a disaster backup and recovery plan in this way:

> ' a sufficient hardware and software configuration to restore the critical
> services/applications within a prescribed period as specified by top management'
> (Lui, Ma & Tu, 1989, p.1).

We must now add to this definition "what the law requires" in addition to 'top management'. It will be interesting to see where the legal obligation falls when and if an IT person must choose between following the data recovery guidelines of the law versus those laid out by management. In addition to the Sarbanes-Oxley laws we can add HIPPA, The Fair Labor Standards Act, The National Labor Act, The Americans with Disabilities Act and others that dictate how long certain types of information must be maintained (Hope, 2006).

Finally, in addition to the financial implications of losing data and the laws that mandate the ability to retrieve data in some cases indefinitely, there are other organizational and auditing policies that may dictate the need for data backup practices. Many organizations such as educational institutions fall under the auditing power of their home state or some other accreditation body which may now include IT audits in their evaluation or may at a minimum audit or review areas that imply an area of IT responsibility when it comes to data backup and retrieval. Failure to comply with these policies can be just as damaging because it can result in loss of employment, loss of accreditation or loss of funding.

What to Backup?

The financial and legal implications of data access have added to the reasons for backing up data. Similarly, the increase in the quantity of applications, data and devices that are used in organizations has increased the complexity of answering the "What to Backup?" question as well. Historically user created files and perhaps a small database of customer information was the extent of what needed to be protected from loss. Although these items still exist, we have added to the list of "data" and sources of that data as well as repositories where that data may reside.

In order to evaluate what needs to be backed up on your network it could useful to consider three different areas. First, what applications are being used in an organization. One article divides these into four different application areas reflecting their function and importance in the day to day operations of the organization as well as the impact of not having access to that data (Cegiela, 2006). Applications themselves, apart from the data they create and manage are often vital for an organization's day to day survival as in the case of what Cegiela called 'Front Office Systems'. These are apps directly used by an organization's customers such as online transaction processing. Loss of these applications results in perhaps immediate revenue loss and must be recovered after a system failure as quickly as possible. Often times, mission critical applications such as these should not only be backed up in such a way to allow for quick and easy system rebuilds but also a strong set of redundancy and fault tolerant technologies should be implemented to reduce or virtually eliminate the possibility of downtime to prevent loss of revenue. Determining what applications are most mission critical can aid in creating a priority list and thereby aid in determining where to allocate scarce backup and recovery resources.

After considering applications, organizations can consider what data they are using and producing. This should not be done in a vacuum by an IT department but rather be a process involving key players from across the organization, not only here but in many areas of disaster recovery planning. What data is being created and utilized that is of significance either financially or legally to the operations of the organization? This can vary widely from one organization to the next and even within organizations different departments or personnel may have varying opinions about this. Typically most can easily agree that information such as health records, employment data such as payroll information, customer databases and official letters are vital pieces of information that require protection. However, to a teacher, created exams and presentations they have created are mission critical because they represent many hours of work and thus many hours of rework if they are lost. These documents are the output of one worker and no other worker comes in contact with them but they represent mission critical information for a teacher. Similarly, a salesperson's contacts list may be vital to them. It may hold contact information for hundreds of current and/or potential contacts specific to their account management. The loss of this list could have a severe impact on their ability to easily conduct business. Again, this is information that is not necessarily seen or utilized by any other organizational member but it is vital for that user and their function on behalf of the organization. Emails are also considered business communication as well as instant messaging if this is a tool a company uses to interact with customers. The Enron court case as well as many others constantly make use of email correspondence in litigation and a courts can and do subpoena these emails and unfortunately they can subpoena emails from five years ago which makes the issue of backing up email and archiving it all the more important.

Finally, after considering mission critical applications and data, reviewing the types of devices where these applications and data reside is needed. Again, historically, IT personnel backed up servers. However, servers are no longer the single repositories for data. For example, PDAs are now common place amongst many enterprises. They can serve functions from email to inventory

control. Yet, one study found that only 8% of respondents backed up their PDAs on a daily basis (Latamore, 2006). This article revealed that in 2-3 years, 75%-85% of all "knowledge workers" would be using some sort of handheld smart phone device. So we will have wide spread usage with low frequency of backups which seems a recipe for data disaster. Before we collectively panic, it is important to realize that just because these handheld devices are used to access data does not mean that data exists on the device. However, this must be determined and if important data does exist, some strategy for data backup should be implemented. Along with the mobility of handheld devices is the mobility provided by laptops to remote or traveling users. If a user is mobile then it is more likely that data will not be saved on network located equipment.   A user flying on an airplane is more than likely not working via a VPN connection back to the main office. Where does the information they are working on reside? On that laptop's hard drive more than likely. These mobile employees with mobile data increase the complexity of implementing sound backup strategies. And, although you may setup untold numbers of network shares, mission critical information may still reside on user's workstations either be design or by accident. Some users may, for example, create very large multimedia files which they save to their high performance desktops which may even be equipped with redundant disks. These disks are still subject to failure, buildings are subject to fire, etc. So, is this data being backed up? One study found, surprisingly to this author, that 22% of IT departments perform daily workstation backups (Jacobi, 2006). But equal amounts performed workstation backups one time a month or less which indicates a wide variety of behaviors in regards to backing up desktop data.

Backup Hardware

Gone are the days (thankfully) when backing up data consisted of a stack of floppy disks onto which users may have copied folders from their desktop computer. These floppies ended up in drawers of briefcases or attics. Now floppies are no longer standard equipment on computers, reflecting a change in technology on a small scale which is reflective of larger scale considerations as well. Most users have now moved away from floppies and their initial replacement, compact disks, in favor of USB flash drives which offer increased portability and features, increases storage capacity and increased speed over floppies and CDs. USB flash drives do not require special software because they are viewed by systems as simply an additional disk and copying files to them is the same as copying data to internal hard drives. These devices may even come with built-in software that automates or at least simplifies the backup process for the end user.

External USB hard drives work similar to USB flash drives but allow end users to backup larger amounts of data. These external USB hard drives may also come with built-in backup software which can simplify the backup process for the non-technical end user. Many external USB hard drives are themselves reasonably portable and can be used for backing up multiple workstations fairly easily for end users and could be an especially cost effective solution for small businesses where information resides on the desktops of the end users and there is no centralized network storage area. There are also online backup and data storage options that permit users and businesses to backup there data to an offsite location. This process can be completely automated and typically would require only the installation of a small piece of agent software on the target systems and perhaps some firewall adjustments.

These solutions are all useful and practical for small businesses and home offices however they do not provide the scalability, speed, availability and storage capacity that many small and medium

sized businesses require. Typically most small and medium sized devices have relied on some direct attached storage (DAS) such as a tape drive system installed in a server to backup data. This has been the industry standard for many years and has been an effective solution for many and will continue to be. Tape drive systems are readily available and supported and make it simple to carry backups off-site for storage. Backup hardware can be single tape devices or "robot" devices that maintain multiple tapes so that administrators limit the number of times they must manually load and unload tapes.

Increasingly though, organizations are looking at other options such as Network Attached Storage (NAS) or a Storage Area Network (SAN). A NAS is a LAN connected device running its own form of operating system which is communicated with via the TCP/IP protocol and simply represents another node on the network, whereas a SAN is typically connected to one or more servers via fibre channel connection and resembles more an external hard drive (http://www.nas-san.com/differ.html, 2003). The advantage of the NAS is that most any network attached device can connect to it because it will support different file system types such as NFS for Unix/Linux system access and CIFS for Microsoft system access. The SAN option is typically a much more expensive solution but does have the advantage of speed using fibre channel technology. (Keep in mind that for the purposes of this paper these devices are potential repositories for data backups but in reality may be used for real-time data access as in the case of a cluster sharing a SAN.) As the above mentioned article also indicates, these technologies are coming to look more and more alike with the release of technologies such as iSCSI which allows a SAN to be accessed via TCP/IP. Although this level of hardware may seem unreachable to some SMBs, newer low cost options are available.

Companies that target small-office-home-office (SOHO) and SMBs are releasing network attached storage solutions up to a terabyte in size in the $700.00US range. These devices can permit SOHOs and SMBs to more easily backup data to a centralized location. This is not necessarily the final step in a well developed backup/archival strategy but it is vastly superior to no data backup at all.

For those organizations that need NAS type capabilities for data backups there is also the option of a home grown solution. An interesting argument is made for the use of the slightly older EIDE technology in building a disk array system (Gao, Pan, Pei & Xu, 2004). Although EIDE does not represent the fastest disk technology currently available, it is argued that it does provide increased speed over tape drive systems while being cheaper to build and support than SCSI or fibre channel systems. Arguably, a home grown device such as this could be loaded with an open source operating system and provide a viable alternative in some cases.

Increasingly, backup solutions are utilizing a disk and tape hardware combination for data backup and archiving referred to as disk-to-disk-to-tape or D2D2T. (http://www.symantec.com/smb/library/article.jsp?aid=backup_solution, 2005). In this hardware configuration, backups are routinely made to a disk device such as a SAN or NAS and then these backups are periodically archived to tape.

Backup Software

As mentioned already, some storage devices, especially those targeting SOHOs have their own built-in backup software solution. For most SMBs these SOHO type products do not offer the features, redundancy, speed or storage capacity to meet their needs. So, they must select hardware solutions that do not have built-in software solutions. This then compels the IT staff to choose from a variety of software packages which have an increasingly vast array of feature sets.

Most operating systems come with some sort of built-in backup software solution. Nix type systems come with various dump commands and archiving commands that can be used to perform data backups. Windows desktop and server products come with the NT Backup utility giving Windows admins an easy to use gui backup tool without additional costs. It could be though that these built-in tools do not provide the level of functionality and flexibility desired by and IT staff.

Backup software usually makes a backup copy of data or applications and performs a proprietary archival-compression function on the data. It allows the scheduling of this task and provides at least modest logging capabilities to permit administrators to review backup sessions for success or failure and allow for troubleshooting any problems. But other considerations must be made. For example, does the backup solution support backing up open files? Does the messaging server backup software permit the recovery of a single mailboxe or single email message? Does it effectively copy data or applications or does it image them in some way permitting faster system recovery because images effectively return machines or data sets back to their exact configuration and status. If the software is backing up a database how long does the database stay locked or offline during the backup process. Another consideration for those environments that like to maintain software homogeneity is whether or not the application runs on multiple operating systems. It should be understood that some backup software is extremely complex in its feature set so proper review of its setup and configuration is critical.

Backup Strategies

The four traditional backup strategies employed by administrators are full, incremental, differential and copy backups (Barrett & King, 2005). Full backups, as the name implies, backup up an entire dataset regardless of whether it has been altered or not. Upon completion of the backup procedure it then clears any archive bits that were turned on. Many organizations consistently perform full backups on a daily basis. Backup times are longer because an entire dataset is being backed up even though much of it may not have changed but since the entire dataset resides potentially on one tape in the case of tape systems, restorations are quicker. Incremental or differential backups may be used in conjunction with full backups to reduce the amount of data and thus time that is used to backup data.

An incremental backup only backs up data that has been changed, or more specifically, that still has an archive bit set on the file. This might be used so that one time a week a full backup is performed and then on subsequent nights that same week incremental backups are done. Thus, incremental backup would have a different dataset on it and different revisions of the same file

potentially, necessitating the use of all tapes during a restoration process. Incremental backups then clear the archive bit on files.

Differential backups are used in conjunction with full backups much like incremental backups are. The difference being that differential backups do not clear the archive bit so the effect is that after say a full backup on Monday night, a differential backup continues to backup any file that has been changed since the last full backup whether it was backed up on that day or not.

Finally, copy backups are full backups without a reset of the archive bit. This might be used prior to an operating system patch installation. If the patch were to cause a system failure requiring a reload or just simply data loss, the dataset could be immediately restored to its pre-patch status without changing the archive bit and potentially impacting the nightly backup routine.

Although these types of backup strategies implemented in various forms have been the staple for many years, variations on these themes are in place now. For example, strategies may not call for a dataset backup but an entire system image. This creates a recovery scenario where administrators do not have rebuild the entire software set on a machine prior to restoring data. The image is a low level picture of the disk itself which permits the image to be applied to a repaired or new system. Images or snapshots may be of entire systems, a partition or a particular dataset. Imaging is becoming increasing popular and simple as companies like Symantec target end users with their Norton Ghost products. This may seem a constant full backup process but some imaging software can effectively perform incremental snapshots of datasets.

In some cases data backup strategies can include what is called 'continuous data protection (CDP)' (Jacobi, 2006). With this technology/strategy, data can in some cases be backed up as it is changed.

Another consideration for data backup is where the responsibility will be placed. Obviously network servers remain under control of the IT department. However, as we have discussed, increasingly, other devices such as laptops and PDAs store organizational information and applications. Who will take responsibility for backing up these devices? It is an added burden to an IT staff but offloading the responsibility to the end user may not be a viable business option if these devices contain sensitive or critical information. If for whatever the reason backup responsibility is passed on to end users then automating that process as much as possible as well as proper training are essential if success is to be expected.

In terms of offloading backup responsibility, it is also possible to offload the responsibility entirely. Many companies offer completely online backups that remove the workload and thus much of the responsibility of data backups as mentioned already. If this is implemented, it does not relieve the administrator of the need to periodically test the backups to make sure they are recoverable.

Regardless of the backup strategy that is utilized, the handling of the archives of that data, typically in a tape form, is extremely important. Off-site data storage is critical because unless an offsite NAS or SAN is being used as the repository for the backups, data is not protected from structural disasters such as fire, flood or tornado. It is best to have archived data copies in two locations that are unlikely to be affected by the same disaster scenario. A tornado or flood could easily impact two building side by side in a campus type environment.

Finally, do not be lulled into a false since of security with backups. A backup strategy should always include periodic restoration checks to make sure that data can actually be retrieved in the event of an emergency. A backup that cannot be recovered is the same as no backup at all.

Bibliography

Barrett, D., & King, T. (2005). *Computer Networking Illuminated.* Sudbury, MA: Jones and Bartlett Publishers.

*Cegiela, R. (2006, May). Selecting Technology for Disaster Recovery. *Dependability of Computer Systems, 2006. DepCos-RELCOMEX '06. International Conference on* , pp. 160-167.

*Cole, R., & Spears, J. (2006, Jan 4). A Preliminary Investigation of the Impact of the Sarbanes-Oxley Act on Information Security. *System Sciences* , p. 218c.

*Gao, K., Pan, L., Pei, J., & Xu, H. (2004, August 2). Implemenation of EIDE Disk Array System for Mass Data Backup. *Aerospace and Electronic Systems Magazine* , pp. 24-29.

Hope, M. (2006, October 16). Messaging. *Computerworld* , pp. 46-48.

Jacobi, J. L. (2006, October 16). Desktops. *Computerworld* , pp. 40-42.

Latamore, G. B. (2006, October 16). PDAs. *Computerworld* , p. 44.

NAS-SAN.com Technology Overview. (2003). Retrieved June 21, 2007, from www.nas-san.com: www.nas-san.com/differ.html

Selecting a Backup Solution for Your Critical Information. (2005, January 24). Retrieved June 21, 2007, from www.symantec.com: http://www.symantec.com/smb/library/article.jsp?aid=backup_solution

The Importance of Disk-Based Backup. (2006, July 24). Retrieved June 21, 2007, from www.symantec.com: http://www.symantec.com/smb/library/article.jsp?aid=importance_of_disc_based_backup

*To, C., Ma, V., & Lui, N. (1989, Jan 3). Knowledge engineering approach to data centres disaster backup/recovery planning. *System Sciences* , pp. 241-248.