Storage Requirement Trends

Lucid Information Systems

http://www.lucidsystems.org
lucid@lucidsystems.org

March 18, 2006
Revision 1.2e

Author : Henri Shustak
Editors : Stuart Page, Margo Shustak, Yoko Shinohara

Contents

1 Storage Reliability

We live in the information age, where Information is power and wealth, just as it has been throughout history. Storing information digitally has associated advantages and disadvantages. As digital information storage continues to drop in price, the storage requirements seem to rapidly increase. Twenty years ago, a floppy disk could contain a typical operating system, and a word processor. Over the last two decades the expectations of computer users have changed. Today a typical operating system will just fit on a single DVD, and many applications are now distributed on Multiple DVD's.

Applications are still used to make documents, as they were 20 years ago. However, today data is being piped into computers from DNA sequencing devices, video cameras, satellites, and high energy physics experiments. The vast information stored in computers is beyond our perception, yet it affects the fabric of society. Global network technology has linked computers together, allowing information to be accessed from anywhere on the earth.

Much of the information stored in computers is trivial, however a small percentage is worth more than gold. Just like gold in a vault, information is only useful when it is able to be accessed and used when required.

1.1 Importance of Reliable Storage

Striking a balance between reliability, usability and cost is an ongoing process. As increasing numbers of people work together on ever larger projects, storing the information pertinent to the project also becomes more important.

When you etch ideas into stone with a chisel it will be stored there until it is weathered away, cracked into pieces by natural event, or deliberately altered or destroyed. Information manipulated via computers is far more susceptible to alteration or deletion. Human beings for the most part no longer use stone when writing. Today paper and ink is far more common, because it provides adequate reliability for most purposes, and it is due to industrialization paper and pens are inexpensive.

The majority of computer based information systems are different to paper. They are not currently as reliable as paper, however they are slowly approaching a similar level of reliability, for small amounts of data. An example is a word processor; they can generally be set to save the state of your work every minute. However it is still very simple to accidentally delete half a paragraph unknowingly, carry on only to discover that a section is missing once the document is printed out to paper. There are systems to avoid these situations, and for small amounts of data they are getting better. However as the amount of data increases, the ability to store it in a way which is as reliable as paper becomes more of a challenge.

For more than twenty years, magnetic floppy disks have been a common information storage medium. A floppy is far from the reliability of paper. As I write this document PC manufacturers, with the exception of Apple[1] are only starting to ship standard PC systems without floppy

[1] Floppy drives were discontinued from standard PC configurations by Apple Computer over five years ago, in January 1999

drives.

Today most computers store digital information magically on a disk. Many disk drive manufacturers now, include SMART[2].This provides realtime access to the drive condition. These technologies are a good idea, however such systems only reduce the risk of data loss due to hardware failure. Another common cause of data loss is due to errors within the file system[3], however file systems and operating systems are evolving[4] and becoming more reliable.

It is more common for the user of the computer to modify or delete something accidentally. This may be to do with poor interface design or a clumsy operator, but the fact remains that it is exceedingly easy to unintentionally modify or delete a digital document or even worse, multiple documents. Although in many cases the document is recoverable[5], the cost in time or money of such recovery is often prohibitive.

Although having redundant information storage systems and backups is the best way to avoid such data loss, it is unlikely that such loss can be completely avoided. Channeling unlimited resources in an attempt to avoiding data loss is futile, because in most circumstances you will eventually reach a point where the effort used to protect the information will be more than what was used to generate the information.

Reasonable steps can be taken to protect data. However the key question is *How important is the data to you or your organization?*

1.2 Availability

Future information availability requirements can be calculated by studying current requirements and using forward planning techniques. Seven primary factors affect the cost of information storage :

1. When data must be available
2. How long data must be available
3. The required data access rates
4. The number of people who require access to the data
5. The number of people who control access to the data
6. The importance of data Integrity
7. The importance of data security

Although this is the most common prioritization of factors contributing to cost, it is probably not the best. Data security and integrity should almost always have top priority. However, priority is rarely given to security or integrity issues.

Priorities will vary depending upon the uses of your data. Generally data availability receives a high ranking. People prefer to have highly available information access even if it is not required. This causes a dilemma, because the cost of information storage increases exponentially with regard to availability. Other unlisted factors will probably also need to be taken into account, and to add further complexity to data storage, all seven factors affect each other. *Ultimately you*

[2]SMART : Self Monitoring and Reporting Technology. Refer : *http://www.drivehealth.com/faq.html*

[3] File systems keep track of information on storage devices

[4] Journaling is one example of such evolution *http://en.wikipedia.org/wiki/Journaling_filesystem*

[5] Deleted documents can often be recovered by a data recovery expert

or your organization must find a reasonable balance.

Digital data can be cataloged and searched. Today cutting edge search systems can effectively catalog and search though text very effectively. This is one of the prime advantages to storing information in a digital format.

2 Storage Requirements

2.1 Decreasing Costs

As more people use digital information systems, the costs per unit of storage is reduced. This is a byproduct of information storage research and industrialization.

2.2 Increasing Requirements

As more people use digital information systems, more space is required, however the real requirements increase comes from the kinds of information we manipulate. Digital images, audio, and video, are just a few of the expanding list of data types stored in computer systems.

Compression reduces the storage requirements, however information if it is to be stored, must somehow be represented. Ultimately, if the system is digital, this information is represented as bits. Looking to the past for future trends, reveals that the requirements are increasing, even though compression techniques continue to advance.

2.3 Ability to Expand

As the requirements grow, so do the options for digital storage. For most organizations it is advisable to consider the many available options, and then make a decision. Rather than being locked into a closed system by a sales representative. The long term operational costs from a decision made during the planning stages are difficult to determine. An expandable system which has the ability to interconnect with other solutions is an advantage.

If possible avoid being locked into a single non-expandable system. Otherwise there is a greater probability of ending up with high expenditure over the long term. As an example; if a vendor offers a solution with a feature which precisely meets your requirements, and this vendor's solution is compatible with other vendor's solutions, which have yet to implement that particular feature, it is possible that the compatible vendor will offer the feature at a lower cost in the future. Should this happen you will be able to expand at a lower cost. In addition, as your requirements change a compatible vendor may appear to have a more attractive solution. Having these options open is advantageous.

3 Backup

Often until you personally lose data of importance, do you even think about backup.

3.1 Reason for Backup

People make mistakes, this is a major reason why we are able to learn. From previous experience of making mistakes, systems have been developed to make them less of a problem. That is why we have erasers and "undo" commands to name just a few. When dealing with computers, a simple mistake can lead to major data loss.

One such system to minimize the impact of such an event is to have an operational backup and restore strategy in place before such an event occurs.

Depending upon your information storage system, a back up may also be used to recover from a hardware or software failure.

Backup solutions consist of many considerations, some are listed below.

3.1.1 Human Error

The most common form of data loss in a reliable system is a mistake. If this is happening, it is time to look at the interface, so these mistakes can be minimized.

3.1.2 Hardware or Software Failure

Hardware systems eventually fail. Software systems often fail because of complexity. Both situations can be reduced by not having a *single point of failure*. However *single points of failure* are costly to remove. Often the risk of down time will be accepted because the complexities of removing a *single point of failure* are too great.

3.1.3 Intentional Tampering

Some people will do nasty things to information just because they can, or because they may stand to gain something.

3.1.4 Major Disaster

In events such as natural disasters it is important that peoples lives are the first thing considered. However some people spend their lives working on information which will save lives. If this life saving information is destroyed in such an event then so will the peoples lives which depend upon that information.

3.2 Speed of Data Recovery

The speed and ease of data recover from a backup, is important. If data recovery from a backup is overly complex or expensive, then in most situations there is no point in having a backup.

3.3 Incremental Backup

An incremental backup can provide a rudimentary versioning system. Allowing you to retrieve an earlier version of a document. The grain (time between each backup) of the backup is an important consideration. If the grain is not fine enough then there may be nothing to recover from the backup. Yet if the grain is too fine, then the backup may cause a performance hit to the users of the system, or require too many resources.

3.4 Resources associated with Data Recovery

If data is to be recovered from a backup, then it must not be resource intensive. If is takes more resources to recover from a backup than it took to create the data to be recovered then the backup is pointless.

3.5 How long is a back up Required

The length of time backed-up data is kept is an important consideration. Storing data longer will require more resources. At some point the data in the backup will become stale (useless), but will continue to consume resources. After data becomes stale, it should be discarded. Unless this is done manually or by an AI system then

a time limit will have to be set. If this time is too short, then important data could be lost and no longer be available for recovery. However, if the time is too long then resources will be wasted.

3.6 Archiving Valuable Data

An archive is a set of information which is to be stored for some period of time and is generally accompanied by an index. Archival of information is often refereed to as data vaulting, or grandparent backup. The word 'archive' is used loosely even within the information storage industry.

Archives may contain data that is offline and when possible should be cataloged to simplify the retrieval process. Important data should be archived and a copy of the archive should be stored offsite. An archive is an indexed backup of key data. In certain circumstances an archive is a backup of a backup.

4 Redundancy

4.1 Importance of Redundancy

When a device fails, you can have another device ready to take over. This is the basic principle of redundancy. A favorite example is a RAID[6]. A RAID can be configured in many different ways. Various configurations have different advantages and disadvantages.

Perhaps the most basic to understand is *mirror*. This is where two or more disks are involved in the RAID. All disks contain the same information all the time. If one disk fails due to a hardware failure, the other disk becomes available and the system continues to work. Depending upon the subtleties of the configuration, a two disk mirror has the following advantages :

- Some protection against hard disk failure
- Relatively simple to understand
- Relatively easy to implement

The Primary disadvantage of a mirror is that your storage space is reduced to the size of the smallest disk (depending upon the configuration).

Disk drives are not the only hardware in a computer: There are power and network, connections to name just a couple. What about the computer? Maybe the entire computer becomes unusable. Then you can have multiple computers configured to form a cluster.

4.2 Single Points of Failure

If you do not have redundancy you have a single point of failure. If a *single point of failure* fails then something is going to stop working. Removing all *single points of failure* can be prohibitively complex and expensive.

4.3 Online Media

Online means available and is a relative term. Hard Drives are the most common form of online media, they are accessed locally or though a network. However as different kinds of storage solutions emerge and the prices of these drop more viable alternatives become available. Online media is good, because it provides quick backup and quick recovery of data. However depending upon the kind of online media and your archival policies, it may not be a good choice for an archive.

[6] RAID : Redundant Array of Inexpensive or (Independent) Disks

4.4 Offline Media

Offline Media is also a relative term. Tapes are the most common offline media, but floppy disks, CDs, DVDs and hard drives all fit into this category, although some can also be used as online media. Offline media is perfect for data archival.

4.4.1 Optical

Optical technology is advancing rapidly. Currently it is not common to use optical media for online storage, however this is set to change. Even CD's can be utilized as online media. Knoppix[7] is an actively developed operating system which runs from a CD.

4.4.2 Magnetic

Magnetic technology is also advancing rapidly. Magnetic metals, plastics and papers are being actively developed. Magnetic technology is currently being used as online and offline storage.

4.5 Redundancy Conclusions

The risks associated with data loss is different for various organizations, and for the many kinds of data within an organization. Allocating data into Tiers[8] is a first step to towards appropriately protecting valuable data.

Different configurations of hardware and software will provide varying levels of digital asset security. There are extreme variations in the costs between the various solutions available for storing and accessing data. These various configurations also offer varying advantages. Therefore it is advisable to take time comparing the costs of the different redundancy solutions against their associated risk of data loss, or down time.

5 Conclusions

Whether a computer stores a movie, a cure for a disease, a recipe or perhaps a novel. It is clear that computers are no longer just for computation, they have become our societies information access medium.

People store and manipulate information in a digital format. Whether you own a computer and use it to store your digital assets, or depend upon the bank's computers to keep track of money in an account, the storage requirements and reliability expectations continue to grow. Effectively storing this digital information requires balancing many factors.

People managing and working with digital information systems may become excited by an upgrade which fixes a known issue. However sometimes such a fix causes unforeseen problems. When this happens, the upgrade must be removed. These information technology complexities have made me grateful that the information in each of our heads, is being perpetually upgraded without too many major complications.

If this document has been an update for you, I trust it went smoothly. I also wish that any upgrades to computer systems you deal with are smooth.

[7] Knoppix CD/DVD based operating system *http://www.knoppix.net/*

[8] Tiered Storage : *http://images.apple.com/xserve/raid/pdf/Tiered_Storage_Whitepaper.pdf*