Mesabi Group

Commentary

June 18, 2008

Choosing the Right Hardware and Software for Data Protection Solutions

With the increasing importance of disk-based data protection, the question arises on what are the relative choices of disk and tape for basic operational recovery. The answer lies in what data protection software technologies are used and for what reasons. Businesses can use a combination of disk-based backup, copy, and replication strategies in addition to basic RAID and standard backup/restore software to tape. Disk now takes its place along tape as secondary storage and tape can serve a complementary role to secondary storage disk as tertiary storage.

Start with a Focus on Operational Recovery for Data Protection

Data protection seems to be getting more complicated each day. The complexity arises for two reasons. First, the *breadth* of data protection is greater due to compliance and governance (say data security and civil litigation) demands. Secondly, the *depth* of data protection is greater due to expanding high availability (24 x 7) demands coupled with a seemingly never-ending increase in storage requirements.

Still companies of all sizes have to remember that data protection starts with the business continuity function inherent in their risk management responsibilities.

Although the business continuity responsibility extends to disaster recovery sites and may well include remote data site protection, the primary responsibility for data protection starts with the local site for operational recovery. Operational recovery is local recovery from a service-level impacting event to one or more applications.

With the emergence of disk as a more common target for data protection, businesses need to review their options for when to employ disk and when to employ tape as a part of their overall data protection strategy.

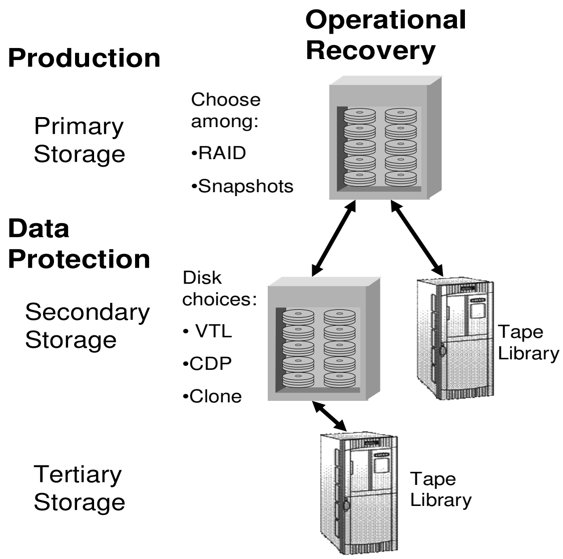
So businesses should review their data protection choices from the ground up and they start by looking at operational recovery with which they have the most familiarity because of the familiar backup/restore software data protection technology that permeates companies.

When reviewing their choices however, smaller companies have smaller IT resources (and consequently smaller IT budgets). Consequently they have to be especially sensitive to the ease of use of any data protection technology as well as its price. Therefore smaller organizations may not be able to afford the highest availability or functionally-rich data protection solutions.

Hardware Storage Tiers for Data Protection

Some data protection is performed at the primary storage level where production applications get their data, but most data protection occurs at the secondary (as well as tertiary) storage level (Figure 1).

Figure 1: Tiers of Data Protection Storage



Source: Mesabi Group June 2008

Tape has traditionally been the secondary storage that backup/restore software used to protect primary, online storage.

However, disk storage has taken a significant role as a secondary storage target for data protection, including disk-based backup and replication strategies.

The use of disk for secondary storage for data protection does not mean that tape goes away. As tertiary storage tape can well provide a complementary role to disk to provide additional copies for data protection processes, say for the movement of tape media to a remote site for disaster recovery.

Logical and Physical Data Protection

Data protection for operational recovery involves one or more additional copies of the data to be stored in one form or another.

Data protection copies defend against physical problems, logical problems, or both. Physical data protection guards against physical failure of a storage device, say a hard disk or a piece of tape media failure as well as fire, flood, hurricanes, and the like. Logical data protection guards the data from change through unauthorized or erroneous I/O requests, such as a virus or accidental file deletion.

Each data protection technology has its own characteristics as to whether or not it protects against physical problems, logical problems, or both as follows:

Primary Storage

- *RAID* — provides physical data protection only against disk failure
- *Local replication or snapshots* — provide logical data protection only as the snapshot is on the same array as the production data, i.e., no additional disk copy

Note: The combination of RAID and snapshots provides both physical and logical data protection.

Secondary Storage

- *Backup/restore software* — provides both physical and logical data protection, whether backup is disk-based (disk-to-disk or virtual tape library) or tape based; the copy is outside the production I/O path so is not subject to logical data corruption and provides logical data protection; the copy is on a piece of disk or tape media separate from the production storage media so it also provides physical data protection
- *Continuous data protection (CDP)* — provides both logical and physical data protection; the fact that the I/O to a CDP appliance is outside the production I/O path delivers logical data protection; the fact that the CDP copy resides on a separate disk array delivers the physical data protection
- *Remote Replication (such as cloning)* — the separate physical copy delivers physical data protection and the point-in-time nature of the copy delivers logical data protection.

High Availability and Low Data Loss

In the context of operational recovery, the focus is often on RTO and RPO. RTO (recovery time objective) is the time that is required to return an application to a working state after a failure occurs. RPO (recovery point objective) is the difference between the time when a failure occurs and the previous time when a set of data was available copied (such as a tape from a previous day) from which recovery is made. Such a recovery results in a potential basic loss to all changes to data for the intervening time.

Keep in mind in the following discussions that the goal of RTO is high availability (conversely little or no downtime) and the goal of RPO is little or no permanent data loss.

Degrees of Data Protection

Data protection comes in degrees (which also can be thought of as layers). The first degree where data protection can be provided is for primary storage. Built-in data protection of primary online storage can help prevent data loss or downtime from service-level threatening events (such as a single disk failure).

However, at least one secondary copy — a full copy of the data that is physically separate and distinct from the original — is necessary in case primary storage is unusable.

One degree of data protection means that one failure is tolerable; data is recoverable. If a failure should occur, data protection falls to zero degrees. Zero degrees of data protection means no more failures can be accommodated without total and permanent data loss. This is a level of exposure that IT organizations find unacceptable.

That is why additional degrees of data protection are necessary. The question is how many are appropriate. The minimum acceptable number is two. If one failure occurs, the degrees of protection are down to one. Given that technology is not perfect; having only one extra degree of freedom to fall back upon is not advisable, so most users should consider a minimum of three degrees of data protection. Each additional layer beyond three adds expense, but one or more additional layers may still be justified.

Tape Is Here to Stay

Let’s get one thing straight. Tape is here to stay. Other technologies may front-end or complement tape, but organizations that have an existing tape infrastructure are likely to maintain it. Let’s see why.

Unlike disk, individual pieces of tape media are easily removed from one tape drive and can run in any compatible drive. This capability is important because it allows tape media to be transported to and put into use at remote sites independent of the primary datacenter. Thus, movement of data is dependent upon the availability of transportation, but not upon the availability of a network. Tape drives can operate independently, but are often embedded in tape automation solutions (such as an autoloader or tape library).

This process actually provides a great deal of both physical and logical data protection for operational (as well as disaster) recovery. Since each tape copy is on a copy of physical media other than the primary disk copy, tape delivers physical data protection. Since any tape cartridge that is not in a tape drive is not in the I/O path, tape media also deliver logical data protection. Since a tape copy can be physically transported to a disaster recovery site, tape provides both physical and logical data protection for disaster continuity.

Since several generations of tape copy are typically available, no single tape represents the only copy, and no single point of failure exists — thus, tape can provide more than one degree of data protection.

RAID Is a Basic Necessity

For any production data whose loss a company simply cannot tolerate, a RAID (redundant array of independent disks) configuration is a necessity. RAID 5 tends to be the preferred choice although RAID 1 (a.k.a., mirroring) is also popular.

However, these current RAID technologies, while quite good, offer unpalatable choices: either hoping that a rebuild of a failed drive will complete before a second disk drive in an array fails or investing in a costly extra mirrored array. Given that the wrong disk drive can be pulled from an array and cause an unexpected second failure, and that the disks in an array may be from the same batch of disks (and thus may be more likely to suffer from the same problem that led to the first failure), the organization should be a bit anxious about the possibility of a second failure.

Typical RAID (excluding mirroring) is based upon a single-parity protection scheme. Multiple-parity RAID requires the equivalent of multiple drives assigned for the parity function. Several companies now offer dual parity RAID 6 solutions. Dual parity solutions can tolerate the loss of two drives before a RAID group is rebuilt. Since the extra drive can essentially take the place of the hot spare in an array but is also more of a working drive, then the move to RAID 6 may not even increase costs.

Disk-based Backup

When using disk-based backup in conjunction with traditional backup software, backup jobs copy data to a disk array for data protection rather than to a tape drive. A set of disk drives has to be reserved for this process, and the cost of that system as well as any software that is necessary to process the data is an incremental cost to an IT organization since the existing tape automation infrastructure is typically not replaced. Moreover, some change in operational procedures as well as retraining of staff might be necessary. The question then may very well be why so many IT organizations are so interested in disk-based backup.

The answer lies in two words — reliability and speed. The former refers to improving the reliability of the data restoration process, and the latter to shortening the length of time that a backup or data restoration job takes.

Speeding up the Backup/Restore Process

A key justification for incorporating disk as an additional layer in the backup/restore process is to reduce the time required to create a backup copy and to restore a given set of data. Although there are different ways to perform backups at any time (such as from a point-in-time copy of the data), many backup jobs are still run after a production application has been shut down at night. The problem is that the ever-increasing amount of data many businesses have to manage takes longer to back up, but the number of hours in a night has not changed. This is the “running out of night” (a.k.a. “shrinking backup window”) problem. On the restore side, improving the time to restore data in order to meet quality of service objectives may be equally important.

Improving Restore Reliability

Another justification for disk-based backup that many IT organizations offer is improving the reliability of restore. Many IT organizations have a concern that the potential failure rate on data restorations with tape automation infrastructures is higher than they find acceptable.

Keep in mind the caveat that the data must be available on disk (not staged off to tape) for the restoration process. Disk space should be able to accommodate, say, a weekly full backup as well as all the daily incremental backups for a week. This should suffice for most circumstances.

Virtual Tape Library

Today’s backup/restore software is designed to minimize the impact of these processes on the existing policies, practices, and procedures of an IT organization. Standard backup/restore software packages can target disk as well as tape. Another option is a virtual tape library; software that runs on a disk array to emulate a tape library.

However, simply retargeting standard backup/restore software from disk to tape requires that each backup job be manually retargeted to disk. That is not true of a virtual tape library. If the number of backup jobs that must be changed is manageable, straight disk-based backup may be a feasible alternative. A second concern is that there might be a two-terabyte (TB) file system limitation, which would apply to straight disk-based backup, but not to a VTL.

The two primary issues of concern regarding VTL are integration and scaling. More complex backup environment and/or large capacity backup requirements (say, six TB or greater, as a rough measure) would tend to favor a VTL. Otherwise, straight disk-based backup might be a reasonable choice.

Data Deduplication

VTLs are becoming very popular for improving the overall manageability of the backup/restore process.

But businesses typically want to have a number of generations of backups available to them in case old information has to be recovered. In fact, legal requirements are driving even more retention. The problem is that storing a full copy of each backup on disk quickly becomes economically unfeasible, but trying to recover old information from tape may take much longer than is acceptable.

Data deduplication solves the problem. The difference between two full backups is small so saving the small number of changes is easy. Storing even a large number of backups on disk really does not become a burden in terms of additional disk.

Businesses have a choice of inline or post-processing deduplication. Inline deduplication is performed during the backup process as data is written to the storage on the VTL. Post-processing is performed after the data is written to storage managed by the VTL. More systems resources are used to perform the full deduplication operation in real-time using inline, but post-processing requires scratch disk for the deduplication process.

Smaller business would tend to favor the inline method as the impact on system resources is not likely to be onerous, but buying additional disk might be too costly. Conversely, larger businesses may prefer post-processing as they cannot afford to slow down the backup process, but can afford the extra disk.

Copy Strategies

For purposes of classification, a copy is a reproduction of data either on an array that is logically (if not physically) local or on the same array as the original. (For low RTO/low RPO purposes, disk — not tape — must be used.) Point-in-time copy capability was the first disk copy capability commonly deployed. Continuous data protection is a newer copy capability that is attracting a lot of attention.

Point-in-Time Copy

The ability to create a point-in-time (PIT) copy of a pool of data has — and will continue to have, through ever more ingenious uses — a major impact on data protection. A PIT copy is a “copy” of a pool of data at a chosen instant in time. The advantage is that the PIT copy is frozen in time. Although there is no guarantee that the copy itself does not suffer from data corruption, there is a guarantee that changes after the time of the copy will not change the original pool of data; thus a PIT copy provides logical data protection.

The two basic “flavors” of PIT copies are clones and snapshots. PIT clones are an exact *physical* copy of a pool of storage and thus deliver both physical (as of the time of the cloning) and logical (from cloning time onwards) data protection. The price that is paid (other than the cost of the software) is a doubling of the amount of disk storage required by a storage pool. The cost of such a doubling (as well as manageability) typically limits the number of clones to only one or very few. The advantage is using a clone for backup to tape or for production testing does not impact the performance of the production disk array.

Table 1: Where and When Data Protection Technologies Are Best Used

	Where	**When**	**Comments**
Choice	**Production Copy**		
RAID	Primary storage, i.e., production disk array (although can be used on an array for disk-based backup)	Provides protection against physical disk failures	Does not protect against disaster recovery (DR) type problems (such as a fire) or against logical problems (such as data corruption)
Snap-shot	Primary storage, i.e., production disk array	Provides protection against logical data protection problems	Multiple point-in-time snapshots may be taken, but data between last snapshot and time of failure is lost; complements the physical protection of RAID
	Data Protection Copy		
Clone	Secondary storage — uses disks that do not store production data	Provides protection against both logical data protection problems and hard disk failures on the production array	Can provide fast restart if primary production storage fails and the clone has to be used as the target for production applications; costly as it requires a second full disk copy
Tape	— Secondary storage (if backup is directly from production disks) —Tertiary storage (if behind disk-based backup disks)	Tape media provides both logical and physical data protection for primary storage	Can provide local operational recovery if tape media are on site or disaster recovery support if tapes are stored remotely; time to recover data may be quite long
Virtual Tape Library	Secondary storage on disk	Provides both logical and physical data protection for primary storage	Improved speed of recovery over a tape library especially for single files; additional cost as tape is still likely to be used
Data De-dupli-cation	Secondary storage on disk that is used by a VTL	Significantly reduces the amount of storage that a VTL needs to provide a given level of data recovery protection	Much backup data is redundant; non-redundant data can be stored on disk and reduce the amount of tape that has to be stored
CDP	Secondary storage on disks separate from production disk	Provides both logical and physical data protection for primary storage	Little or no data loss; very fast restoration to primary storage as only corrupted data has to be replaced; may serve as temporary primary storage

Source: Mesabi Group June 2008

A snapshot is a software image of data as of a predefined instant, taken on the original disks where the data is stored. At the time the snapshot is taken the original production data and the snapshot are identical. That means that no additional physical space is required at that instant. The original production data and the snapshot data diverge as writes change the original production data.

PIT copies can serve many roles, including serving as a starting point for making a backup copy or for application production testing, but a key PIT role is in providing high availability logical data protection.

Continuous Data Protection

With continuous data protection (CDP), a company can create a data protection copy (typically on a disk-array-based data protection appliance) that can recover to *any* point-in-time. Typically, changes are recorded continually by the CDP appliance, which uses a non-invasive journaling technique that does not require even the momentary halting of an application's I/O processing that has to take place when creating a snapshot PIT copy). The journal can be rewound to any point-in-time as the basis for creating an any-point-in- time (APIT) copy of the data without having to know at what point-in-time a copy should have been taken.

When integrated in a data protection appliance, CDP offers today's only up-to-the–moment logical data protection and physical data protection with high availability. CDP is not a new backup approach, but an alternative (or, more likely, a complement) to the traditional backup software approach. CDP provides fine granularity over the data restoration process in that logical unit volumes (LUNs) or individual files can be restored.

Conclusion

Businesses now have *a number of good choices* for protecting their data; the challenge is now to decide what combination of software-hardware best meets their storage and business needs. Table 1 helps with this process by summarizing when and where data protection technologies are best used.

Each technology accomplishes a specific purpose, and, in the appropriate combination for each business, they provide an integrated data protection infrastructure that delivers the necessary level of both physical and logical data protection for operational recovery.

David Hill

Analyst Name: David Hill
Topic Area: Data Protection

Mesabi Group LLC
26 Country Lane
Westwood, MA 02090
www.mesabigroup.com

This document is the result of research performed by Mesabi Group LLC. Mesabi Group LLC believes its findings are objective and represent the best analysis available at the time of publication. This Commentary is based upon research sponsored by Hewlett-Packard Company.

Phone: (781) 326-0038
email the author: davidhill@mesabigroup.com