Celeros

Bringing SANity to Storage Costs

White Paper

Maximizing Disk's Benefits in Backup and Recovery

The New Corporate Requirement For a Scalable Disk Cache in the Backup Process

February 2008
By: Jerome M Wendt
Lead Analyst & President
DCIG Inc

Maximizing Disk's Benefits in Backup and Recovery

The New Corporate Requirement For a Scalable Disk Cache in the Backup Process

That Nagging Backup Issue

Solving backup remains one of the nagging issues in corporate data centers. Enterprise companies can create configurations that support mission critical databases and handle tens of thousands of transactions. However they cannot seem to get their arms around backing up and restoring database and network file servers in an efficient, economical manner.

More companies now recognize that using disk in some way as part of their backup infrastructure is an imperative if they ever hope to shrink growing backup windows, expedite recoveries and get this issue under control. However companies are rapidly discovering that introducing disk into the backup equation is not as easy as just pulling tape out and putting disk in.

Before companies can confidently introduce disk-based storage system into the backup process, companies need to ascertain the benefits and drawbacks that tape inherently provides. The principle reasons companies typically continue using tape in their backup process include tape's cost, density and portability as well as intangibles like their existing investment in tape cartridges and libraries.

Despite these benefits it is tape's downsides that are driving companies to explore using disk as alternative to tape as a primary target in backup. Two distinct problems that tape creates in the backup process include lengthy recovery times and tape media management. These issues impact businesses in ways that often do not easily show up in their bottom line but impact them when it matters and costs them the most – during data recoveries. Routine recoveries – such as recovering accidently deleted files – may take hours to complete depending on how long it takes locate the date on the tape and retrieve it from the tape.

If the needed data is not on tape cartridge in the tape drive or tape library, then the merry-go-round of finding the right tape cartridge begins. Finding the tape cartridges assumes that internal tape management procedures are followed, such as correctly labeling tape cartridges, tracking where tape cartridges are stored and the rotation of tapes occurs as scheduled. Even then, the chance always exists that tape cartridges are not where they are supposed to be because they were lost, misplaced or fell off of the back of a truck.

Problems with Using Tape as Primary Backup Target

- Tape drives can "shoe shine" or go back and forth without steady stream of data from backup software
- Tape drive can only be controlled by one backup

software product at a time

- A server's data can span multiple tapes requiring the unloading and loading of tape cartridges into tape drive
- Difficult to recover small amounts of data files or directories since need to reposition tape cartridge in tape drive to locate the specific file
- Tape cartridge media film can fray, break or wear down resulting in data loss

These types of issues are leading companies to introduce disk into the backup process. However when companies begin to look at using disk in the backup process, they rapidly encounter a myriad of choices that all sound equally viable. Of these, virtual tape libraries (VTLs) and deduplicating backup appliances often sound the most appealing because of how they deliver the combined benefits of disk and tape while mitigating some of their downsides. However prior to deploying these new disk-based technologies, companies should consider how or even if these technologies are the right fit for their environment.

Disk as a VTL

Virtual tape libraries (VTLs) are simply disk-based storage systems to which software is added so that they function and appear like real, or physical, tape libraries to backup software. This software gives VTLs most, if not all, of the same characteristics as physical tape libraries so that backup software manages and treats them in a similar manner. For instance, administrators can create and present virtual tape drives on the VTL's network interfaces to the backup software. On the back-end of the VTL, administrators can create virtual tape cartridges with specific capacities (20, 40, 100, 200 GB, etc.) that match that of physical tape cartridges as well as virtual tape drives that appear in a format, such as 3590, T10000, LTO or DLT, that the backup software recognizes or is accustomed to using.

Configuring a disk library as a VTL now gives companies the benefits of using disk while keeping their backup infrastructure in place. Companies can continue to use their backup software exactly as before but now direct the backup jobs to the VTL. Policies they have previously set in the backup software, such as retention of specific tape cartridges, sharing of specific tape drives and management of tape library partitions, remain in place. However since disk, not tape, is the target for backups, companies should immediately start to see the benefits of using disk as a primary backup target, including shorter backup windows and faster recovery times.

The problems that start to surface when obtaining and using VTLs are less intuitive. An initial concern with obtaining a VTL is that it is a dedicated disk appliance that is exclusively used by backup software. This precludes companies from using the spare storage capacity in the VTL for other general purposes, such as archiving or general purpose data

storage. Creating a VTL also adds to the cost of the appliance since the software required to convert it into a VTL typically adds significantly to the cost of a disk library running without it.

Some Hidden Problems with Using VTLs

- Only have finite capacity since virtual tape cartridges can not be ejected
- Creating infinite capacity requires the reintroduction of tape and the problems associated with tape
- What manages the migration of data from the VTL to tape – the backup software or the VTL?
- Need to ensure the backup software catalog tracks what virtual tape cartridges are exported to physical tape cartridges
- Are features like compression and encryption turned on in the tape drive but not the VTL? If so, will all of the data on a virtual tape cartridge fit on its matching physical tape cartridge?

Another major issue with VTLs is that the virtual tape cartridges do not possess the same portability characteristics of physical tape cartridges. When companies are ready to move data offsite, they now have one of only two options, deploy another VTL at another site and then replicate the data to that secondary VTL or export the data from the VTL to a physical tape library with tape cartridges. Neither option is particularly desirable. Deploying a secondary VTL requires the purchase of another VTL, network bandwidth between the two sites, floor space at the other location to house the VTL and replication software.

Moving data from a VTL to a physical tape library is equally problematic. Companies need to determine what will manage and perform the migration of data from disk to tape – the backup software or the VTL? If it is the backup software, how does it optimize the movement of unencrypted, uncompressed data from VTLs to physical tape libraries that frequently encrypt and compress data as it stores it? If the VTL, not the backup software, handles the data movement, the VTL also needs to account for these sorts of challenges plus it needs a means to update the backup software catalog otherwise the backup software, when it needs to recall a specific tape cartridge, will not know where the data is located.

Of course, this previous scenario assumes that the VTL only manages disk, not physical tape libraries. This is not true either. While the majority of VTLs only manage disk, a small number also manage physical tape libraries. This creates a situation where companies are now virtualizing physical tape libraries so the VTL presents a virtualized version of a physical tape library to the backup software so the backup software only thinks it is seeing one image. If that configuration sounds complicated, that's because it is.

However, the biggest downside to most VTLs is that they do not offer natively infinite capacity. Because companies can insert and remove tape cartridges

from physical tape libraries, it never fills up so tape libraries can manage hundreds of terabytes or even petabytes of data. VTLs have no such luxury. Once a VTL is full of data, companies either need to start deleting data or buy another one which creates a whole new set of management headaches. To try to address this need and give backup appliances in general and VTLs in particular the attributes of "infinite capacity", a new data reduction technology called deduplication has emerged.

Deduplication to the Rescue?

The recent introduction of deduplication into the disk storage systems ranks as one of the more significant storage innovations in some time. Much of the momentum behind deduplication is attributable to its ability to introduce to disk storage one of tape's more desirable features: infinite capacity. Deduplication works by identifying similar blocks of data in backup streams and then only storing that block of data once.

The reason deduplication has met with such enthusiastic customer response is that the data change rate (new and changed files) in many customer environments is typically about 5% per week. Storing only new or changed data from each backup, these deduplicating appliances may deliver data reduction ratios of 20:1, 30:1 and 50:1 or greater (essentially infinite capacity) using finite amounts of disk capacity.

The challenge is that the measuring the true benefits of deduplication is not cut and dry. Part of the difficulty in establishing its true value is that the actual data reductions each company will experience may vary widely – from as low as 2:1 to as high as 500:1. In order for companies to realize the full benefits of deduplication is predicated upon:

- Companies running full backups on a weekly basis
- Data changes rate that average about 5% or less
- Retaining data for at least 90 days
- Deduplication appliances with scalable architectures

Even assuming all of these factors hold true, companies still should only realistically expect to see data reduction ratios in the range of about 20:1.

The performance overhead associated with deduplicating the data during backups and then reconstituting the data during restores are two other factors that companies need to account for before deploying deduplicating backup appliance. This performance overhead can eventually mitigate whatever benefits that introducing disk into the backup process initially provided. As more and more data is backed up, the deduplication appliance requires more time to deduplicate and reconstitute the data unless the appliance is architected to scale to meet these specific requirements.

Backup appliances also implement deduplication at different points in the backup process which impacts the performance of backups and recoveries. Some backup appliances deduplicate data as it is backed up and reconstitute it during restores while others

deduplicate data after the data is backed up. In the former case, called inline processing, companies need to ensure that the appliance can scale the backup appliance's CPU and memory resources otherwise it will become a bottleneck. In the latter case, called post-processing, companies need to ensure that the appliance has adequate storage capacity to hold each daily backup and then a sufficiently large enough window to process and deduplicate this backed up data.

Inline versus Post-Processing Deduplication

Using inline deduplication, backup appliances deduplicate data as it enters the backup appliance, analyzing the incoming backup stream of data and storing like chunks of data together. The problem with this is approach is that more and more memory, processing power and storage is needed as the deduplicated backup store grows and can result in ever lengthening backup and recovery times. Adding another backup appliance doesn't help either since the new appliance needs to start from scratch in the deduplication process.

Post-processing deduplication would seem to address this problem. It permits backups to complete in their entirety and only starts to deduplicate data after the entire second backup is complete. However the problem with this approach is that it requires a sufficiently large enough deduplication window between the completion of one backup and the start of the next. Without this window, all of the data from the previous backup will never be completely deduplicated resulting in an ever growing cache of raw backup data.

Solutions for both of these problems are in the works but are still in the early stages of testing with enterprise beta customers.

However all of these gyrations by VTL and deduplicating backup appliance vendors to use disk in the backup process beg the question, "What ever happened to the idea of just using disk storage system without creating a VTL or deduplicating the data right away?" Neither VTLs nor deduplicating backup appliances are cheap as they can cost as much as 15 - 20X of a disk storage system without the software. Besides, VTL and deduplicating backup appliances can take the focus off of one of the fundamental reasons companies wanted to use disk in backup process in the first place: to expedite backups and recoveries.

Creating a Disk Cache

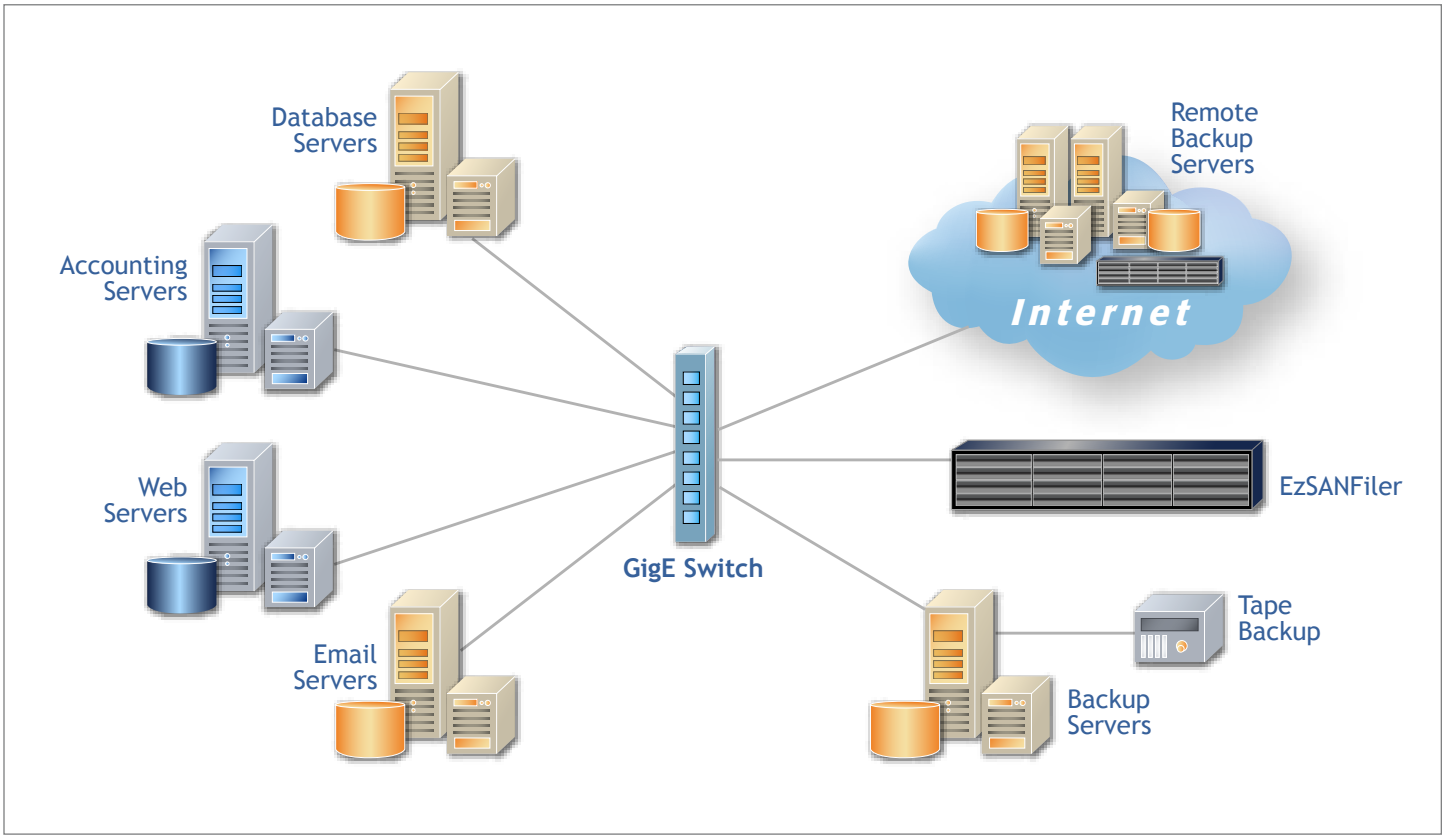
Avoiding VTLs and deduplicating backup appliances means re-introducing a basic concept into the backup architecture: the creation of a disk cache using an ordinary storage system that the backup software uses as a backup target. In this design, no expensive VTL or deduplication software is used on the storage device. Instead, storage systems with high capacity disk drives are used to stage the data before moving it off to tape.

Deploying this basic storage system array eliminates most if not all of the downsides that storage sys-

tems enabled with VTL and deduplication software typically introduce. Since disk is used as disk, it appears as a large storage pool to the backup application. In this configuration, there are no extraordinary requirements to appropriately size the controllers on the disk system to account for the specific performance overhead that deduplication introduces.

There is also no need to try to match the sizes of virtual tape cartridges to that of physical tape cartridges or ensure that you have the right feature set in the backup software to manage the virtual tape library. As long as the backup software can backup and store data to this cache of disk, which most backup software products can, companies only need to configure their backup software to recognize and use this disk cache as the primary target for backup.

The creation of this disk cache in the backup process begins to immediately deliver the benefits that companies want from disk, shorter backup windows and faster recoveries. It also enables them to keep their current backup infrastructure in place, specifically backup software and existing tape libraries, so they can continue to extend the life of these technologies while immediately gaining the backup and restore benefits of using disk. However there are right and wrong ways to introduce a disk cache into the backup process and before introducing any disk cache, it should possess certain characteristics to ensure it can cost-effectively scale to meet your corporate requirements.



Ideal Disk Cache Characteristics

Once companies recognize that using disk-as-disk will meet their pressing backup and recovery needs does not mean they should just deploy any disk system. Corporate backups create enormous amounts of data and using storage systems that are undersized, contain expensive components or the wrong sort of disk drives and the costs can rapidly mount up. The following are some key features to keep in mind when choosing the right storage system to use a disk cache in backup:

- Uses high capacity, low cost disk drives. Look for support of serial ATA (SATA) disk drives in these storage systems. SATA drives are generally available in sizes that range from 300 GB to 1 TB in size so even a basic storage system could provide as much as 15 TBs of raw capacity. Even accounting for data protection methods like spare disk drives and RAID configurations, such a storage system should still deliver at least10 TBs of useable capability.

- High performance expansion slots. No matter how much capacity disk drives offer, corporate data growth still seems to stay one step ahead of the growth. As a result, even storage systems with 10 TB of data can quickly fill up. To scale the storage system, look for storage systems that support SAS (Serial Attached SCSI) interfaces. Multi-Channel SAS is a high-speed interface so companies can add more disk capacity on to their existing storage system while experiencing minimal or no degradation in performance.

- The storage system is configurable as NAS or SAN. Not all servers backup their data in the same way. Some backup their data over storage area networks (SANs) while others backup their data to network attached storage (NAS) over LAN or WAN connections. The system should support either type of environment. To support NAS environments, it should offer some type of standard network operating system such as Linux or Windows so LAN and WAN attached servers can see its storage on the network and backup data to it.

- Constructed with off-the-shelf components. Companies want to keep the backup costs as low as possible. However storage systems are typically constructed with proprietary and expensive processors, memory and interfaces to meet the more rigorous requirements of high performance, production applications. When selecting storage systems to use as targets for backup, identify systems that are constructed with the off-the-shelf components. They provide ample performance and availability for backup at a substantially lower cost.

- Support for multiple network interface connections and protocols. In order to connect to either NAS or SAN environments, the storage system needs to provide the appropriate type of network interfaces – whether that is Ethernet or FC. The storage system should provide support for either multiple 1 Gb or 10 Gb Ethernet speeds as well as support for protocols like CIFS, iSCSI and NFS. FC interfaces should also be available in 2 or 4 Gb now and provide a roadmap for 8 Gb support.

- Offers technologies complimentary to backup. Technologies like RAID 6, MAID and NDMP may never be on the tip of anyone's tongue but they are still important technologies for a storage system intended as a backup disk cache to possess.

- RAID 6 protects against the failure of not one but two disk drives in a storage system. Due to large capacity 750 GB and 1 TB disk drives, it is possible a second drive can fail before the storage system has fully recovered from one failed disk drive.
- Since disk drives are always powered on, they consume significantly more power than tape and degrade more quickly. MAID (Massive Array of Idle Disk) technology takes on increased importance in storage systems used as a backup disk cache since it spins down disk drives when they are not in use. This saves power and extends the life of the disk drives.
- Using NDMP (Network Data Management Protocol), backups can go directly from the file server to the network attached storage system. This eliminates the need for data to go from a file server to the backup server and then back out to the storage system that is acting as a backup target.

While there is nothing necessarily remarkable about any one of these features, what is noteworthy is the lack of deployments of storage systems that support these features in enterprise data centers. However for companies willing to look beyond storage systems supporting VTLs and deduplication will find that these storage systems meet their immediate, tactical backup needs without requiring a major overhaul of their backup environment. By initially introducing a cost-effective, high capacity storage system into their existing storage infrastructures, companies can free up the time they need to more carefully examine which of the emerging disk-based data protection technologies is most appropriate for their backup environment long term.

The Celeros EzSANFiler

The Celeros EzSANFiler is a combination iSCSI (IP-SAN) and NAS products that meets the immediate tactical needs that companies are trying to address in their backup environment without requiring the expenditures of extraordinary amounts of money. By using industry standard hardware and software, Celeros provides companies an economical yet scalable product that meets the specific needs for corporate backup.

Available as a disk target in either FC or Ethernet iSCSI SANs or as a NAS Filer in LAN environments, the Celeros EzSANFiler line of storage systems give companies the flexibility to start small and scale as large as their backup environment grow. Available in configurations as small as 4 TB, the Celeros EzSANFiler product line can scale out to support over 100 TBs of capacity.

The Celeros EzSANFiler offers a SAS expansion module, RAID 6 and MAID features to meet immediate and longer term corporate backup requirements. The RAID 6 protects against unexpected data loss should not one but two disk drives unexpectedly fail in a short amount of time. The SAS expansion module and MAID features serve to future-proof the storage system. Companies can add more storage capacity using the SAS expansion module should data backup loads

or data retention requirements unexpectedly increase. Should this occur, the MAID functionality also kicks in spinning down disk drives not accessed for extended periods – a distinct possibility for storage systems designated for hosting backup data.

This set of features provides the economical entry point into disk-based backup that many companies are initially looking for when looking to dip their toe into the disk-based data protection water. But because of its ability to scale, Celeros gives companies the freedom to scale out to support tens of TBs of data without requiring entirely new storage systems.

Maybe most importantly, storing data in its native format eliminates many of the concerns companies now have about storing data on a VTL or in a deduplicated format. Rather, backing up data to a Celeros EzSANFiler product configured as disk cache keeps the data in a format that is accessible now and into the future by any system. In so doing, companies solve their immediate problems with backup and recovery while putting in a place a system that can scale with them regardless of how they decide to backup their data long term.

Jerome M. Wendt

President & Lead Analyst of DCIG, Inc.

Jerome Wendt is the President and Lead Analyst of DCIG Inc., an independent storage analyst and consulting firm. Mr. Wendt founded the company in September 2006. Since founding the company, Mr. Wendt has published extensively in data storage publications and journals covering all facets of storage. Mr. Wendt was nominated for ComputerWorld's Storage Innovator of the Year of 2003 for his initiatives in bringing storage virtualization into First Data. Mr. Wendt regularly speaks at storage and records management conferences across the country including PRISM International, Storage Decisions and Storage Networking World.

About Celeros

At Celeros our mission is to make ***reliable***, high performance storage solutions that are easy to operate and ***affordable***. While we do not have any religion with respect to the type of technology that can help our customers, we are ardent believers in choosing appropriate technologies that cost effectively solve today's problems and scale to address tomorrow's needs.

Celeros

1170 Hamilton Court
Menlo Park, CA 94025
888-306-0646

For more information go to www.celeros.com
info@celeros.com
sales@celeros.com