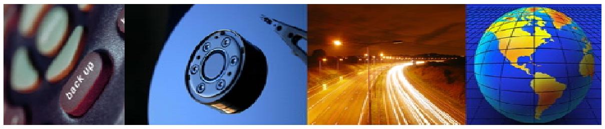


White Paper

Experiencing Data De-Duplication:
Improving Efficiency and Reducing Capacity Requirements

By Heidi Biggar
Storage Analyst, Data Protection
Enterprise Strategy Group

February, 2007

Table of Contents

Introduction

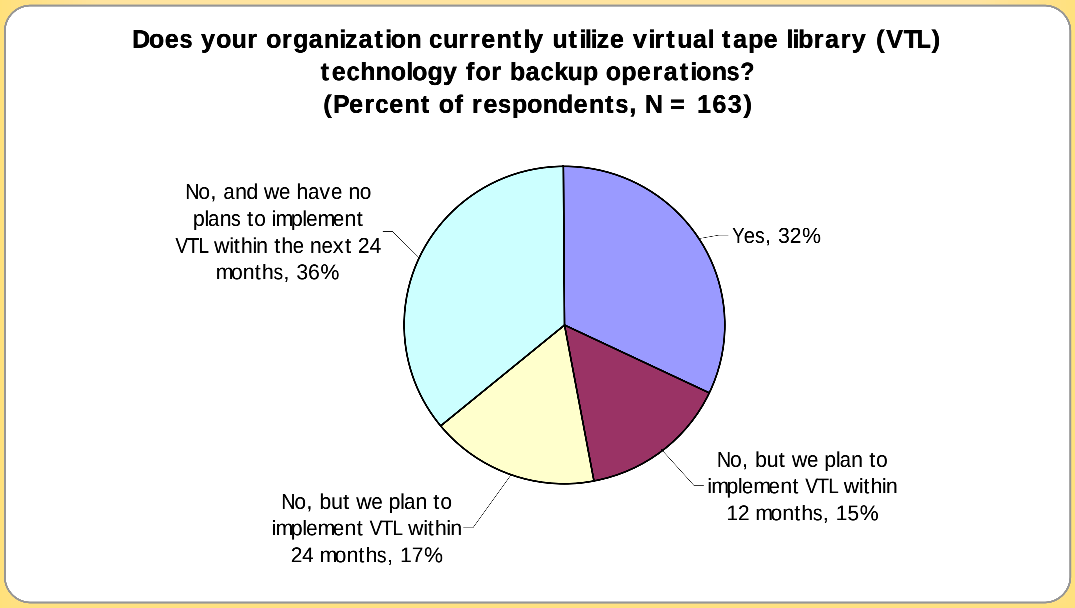
Old Habits Die Hard

For years, we were mired in a deep data protection rut. We backed up the only way we knew how: nightly to tape. As for recovering data, we crossed our fingers and prayed we'd never actually be asked to do a restore. After all, recovering from a system outage or some other type of data-loss event was uncertain and could take days, weeks or more to complete. Downtime wasn't just a potential risk of tape-based backup recovery. It was an everyday concern – one with potentially devastating business consequences. And for the many users who still back up solely to tape today, these risks still exist.

But that was before ATA/SATA disk and, importantly, well before data de-duplication. With the advent of ATA/SATA disk as a backup target, data protection got much better seemingly overnight. Backup speeds improved drastically and, importantly, so did recovery performance. For the first time, we started thinking about data protection in recovery, and not just backup, terms. And our recovery time objectives (RTOs) and recovery point objectives (RPOs) started getting a lot more aggressive.

Today, disk-based data protection solutions, such as virtual tape libraries (VTLs), are widely used by organizations of all types and sizes, and ESG expects adoption to continue to increase over the next 12 to 24 months. In fact, in a recent ESG Research survey[1], 32% of respondents said they already use VTL for backup operations and another 32% said they plan to implement the technology within two years (Figure One).

Figure One: Current VTL Use for Backup



However, while backing up to VTL or other disk-based backup targets has greatly improved our ability to meet, or even exceed, backup and recovery objectives, we're still dealing with an underlying problem ESG refers to as "capacity bloat." Blame this phenomenon on ever-increasing data volumes, regulatory or corporate governance mandates that require us to keep more data online for longer periods of time, more aggressive SLAs or shrinking backup windows, it doesn't really matter. The reality is we're backing up more and more data – and we're still not doing it that efficiently. That is, until data de-duplication.

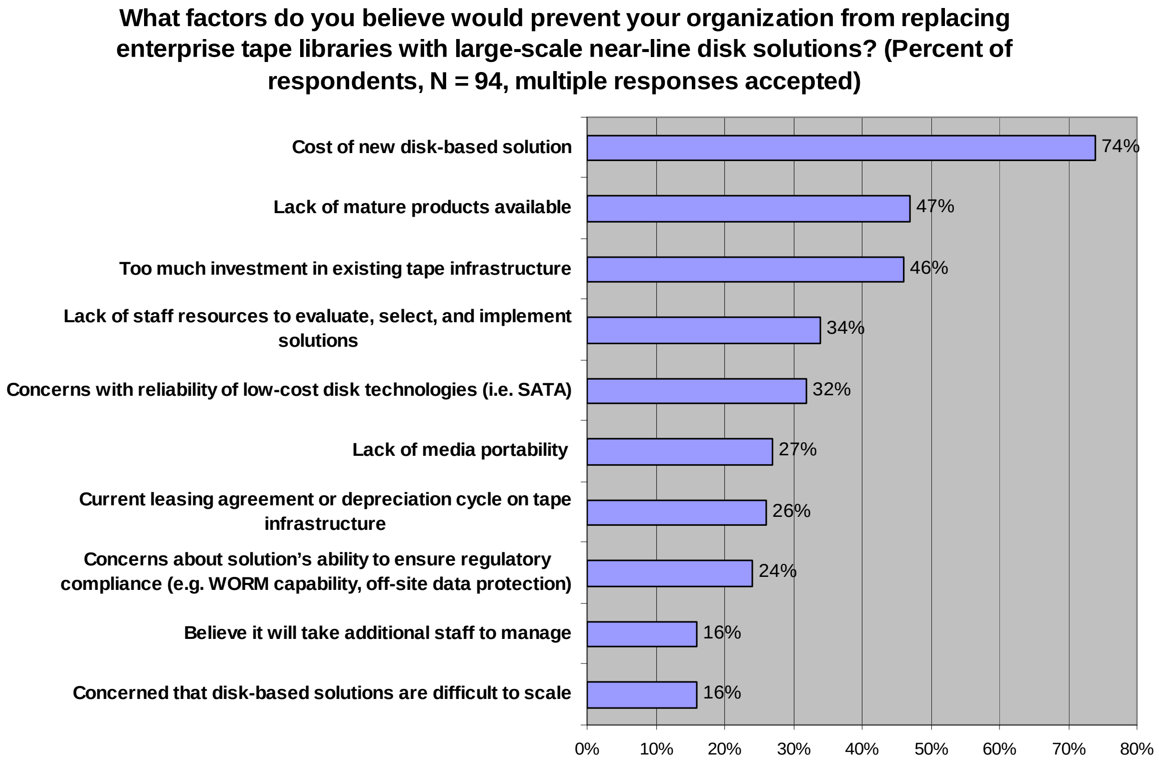
[1] ESG Research Report: *VTL Market Trends and Adoption*, October, 2006.

Enter Data De-duplication

ESG believes data de-duplication will be one of this decade's most important new data protection technologies. Why? Because data de-duplication has the ability to revolutionize data protection by making disk-based backup and remote backup and replication much more efficient than it is today. In fact, ESG expects data de-duplication to drive interest in and adoption of disk-based backup solutions, including VTLs, for the simple reason that it increases the value proposition of these technologies.

ESG Research consistently finds cost to be the number-one obstacle to disk-based backup adoption, and data de-duplication lowers associated disk costs by reducing back-end disk capacity requirements (see Figure Two).

Figure Two: Deterrents to Disk-based Backup Adoption



Overview

As with many new technologies, there is a lot of confusion in the market about data de-duplication. In fact, recent ESG Research[2] reveals strong interest in and awareness of data de-duplication among organizations of varying sizes and industries. ESG believes such strong interest in data de-duplication this early in the adoption curve could be indicative of either underlying confusion in the market about what constitutes data de-duplication and what doesn't *or* the compelling nature of data de-duplication, which sets it apart from other emerging technologies and enables it to break the rules of typical technology adoption curves. Data de-duplication is transparent. It doesn't rely on a number of dependent variables for it to be widely adopted. Either way, ESG expects data de-duplication to be implemented widely in the coming year, and beyond.

In the pages that follow, we will look at data de-duplication from all sides, answering such questions as:

- What is data de-duplication? Where does it fit in the data protection schema?
- How does data de-duplication differ from other backup approaches or technologies?
- What are the benefits of de-duplicating data?
- How can data de-duplication be implemented?

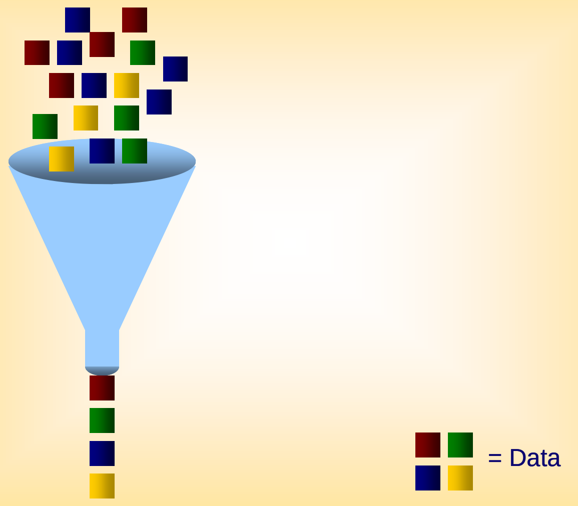
And, lastly, we'll take a look at Quantum's DXi-Series disk backup and replication appliances, the company's approach to data de-duplication and the benefits it can potentially bring to your data protection environment.

Data De-Duplication Defined

Let's start with a simple definition. ESG defines data de-duplication as the process of eliminating or removing redundant files, bytes or blocks of data to ensure that only 'unique' data is stored on disk. Data de-duplication is also an example of what ESG refers to as a capacity optimized protection, or COP, technology. COP technologies are designed to reduce data protection-related capacity requirements.

The potential benefits of data de-duplication are many, but the most notable advantage is that data de-duplication addresses the "capacity bloat" problem head-on by significantly reducing the amount of capacity required on the back-end. This capability is illustrated graphically in Figure Three.

Figure Three: The Power of Data De-Duplication



In this diagram, duplicate data is illustrated by the use of multiple boxes of the same color. While the granularity, or efficiency, (i.e., how much duplicate data it is able to detect) of the data de-duplication will vary

[2] ESG Research Reports: *Branch Office Optimization* and *VTL Adoption and Market Trends*, 2006.

by application or data type, the bottom line is that the number of like-colored squares (referring back to the illustration) should be significantly reduced.

The more granular the de-duplication process, the greater the capacity reduction. In general, data de-duplication that is done at the file level, while still effective, detects less duplicate data than de-duplication that is done at the byte or block level simply, and likewise, de-duplication that is done at the block level is generally more efficient at detecting duplicate data than data that is de-duplicated at the byte level.

This difference in granularity is illustrated in the following example: An end-user creates a 1MB PowerPoint presentation and then sends it out as an e-mail attachment to 20 internal people for review. In a traditional backup environment -- that is, one without data de-duplication -- each attachment would be backed up at the end of the day during the nightly full backup even though no changes were made to the files, consuming unnecessary disk capacity (20 X 1 MB). Even in a small organization, the cost of this type of redundancy can be significant in terms of physical disk capacity, power and cooling for that disk, etc.

With file-level data de-duplication, however, only one copy of the PowerPoint file is saved. All other attachments (i.e., the duplicate or repeated copies) are replaced with “pointers.” This frees up disk capacity for other applications and allows users to extend retention periods, should they desire to.

More granular de-duplication approaches, block- and byte-level, take the process one step further. They look at the pieces that make up those new 1MB files and compare them to elements that the de-duplication system has seen before, replacing repeated elements in the newer files with pointers rather than storing them again. (It should be noted that there are differences among vendors in how these processes are handled. Some performance differences between products may, in some cases, be related to the way that elements are compared and the way they are written to and managed on disk.)

Of course, there are other considerations besides the granularity of the de-duplication process, which can affect de-duplication ratios. For example, the type of data they generate (some data is inherently a lot more prone to duplicates than others), the frequency of the change rate, etc, all affect de-duplication ratios. That said, ESG Lab has tested several de-duplication technologies and believes a 10x to 20x-plus reduction is realistic, regardless of the granularity of the de-duplication process.

One other important note about data de-duplication before we move on: It is a feature or a technology; it is not a standalone product. Its first application is in the data protection and retention market. However, ESG also expects data de-duplication to be applied to other storage applications over time.

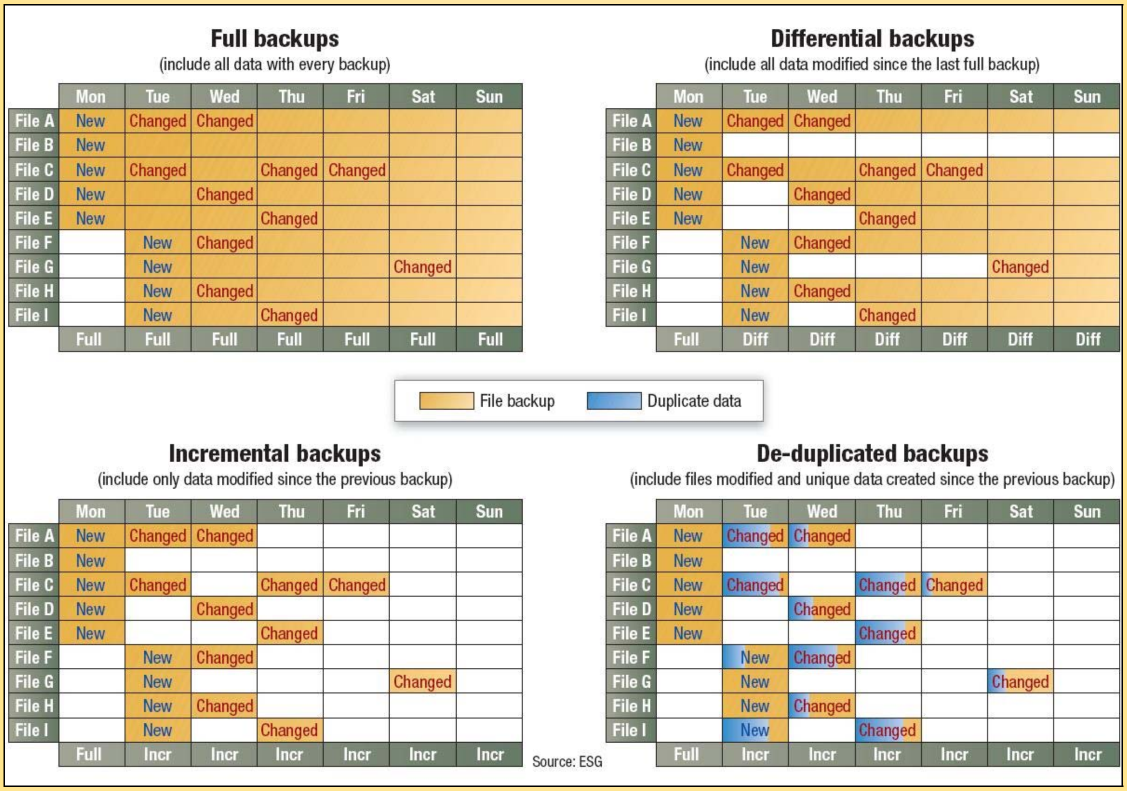
Applying Data De-Duplication to Traditional Backup

Data de-duplication, when added to traditional backup approaches (e.g., full, incremental, and differential backups), can have significant positive benefits, significantly reducing the amount of data that has to be backed up, as illustrated in Figure Four.

Let’s look at the following backup approaches more closely: full backups, incremental backups, differential backups, and what ESG refers to as data de-duplicated backups.

- **Full backups**: Full backups are generally performed on some type of regular basis (e.g., nightly, weekly, etc.) and involve taking a complete copy, or image, of an organization’s data. Full backups do not distinguish between “changed” data or “unique” data. They make a copy of all data with every backup. However, restoring from a full backup is generally more streamlined and less-time consuming than some other backup approaches.

Figure Four: Data De-Duplication Reduces the Size of the Backup



Full backups
(include all data with every backup)

	Mon	Tue	Wed	Thu	Fri	Sat	Sun
File A	New	Changed	Changed	🟧	🟧	🟧	🟧
File B	New	🟧	🟧	🟧	🟧	🟧	🟧
File C	New	Changed	🟧	Changed	Changed	🟧	🟧
File D	New	🟧	Changed	🟧	🟧	🟧	🟧
File E	New	🟧	🟧	Changed	🟧	🟧	🟧
File F		New	Changed	🟧	🟧	🟧	🟧
File G		New	🟧	🟧	🟧	Changed	🟧
File H		New	Changed	🟧	🟧	🟧	🟧
File I		New	🟧	Changed	🟧	🟧	🟧
	Full	Full	Full	Full	Full	Full	Full

Differential backups
(include all data modified since the last full backup)

	Mon	Tue	Wed	Thu	Fri	Sat	Sun
File A	New	Changed	Changed	🟧	🟧	🟧	🟧
File B	New						
File C	New	Changed	🟧	Changed	Changed	🟧	🟧
File D	New		Changed	🟧	🟧	🟧	🟧
File E	New			Changed	🟧	🟧	🟧
File F		New	Changed	🟧	🟧	🟧	🟧
File G		New				Changed	🟧
File H		New	Changed	🟧	🟧	🟧	🟧
File I		New		Changed	🟧	🟧	🟧
	Full	Diff	Diff	Diff	Diff	Diff	Diff

🟧 File backup 🟦 Duplicate data

Incremental backups
(include only data modified since the previous backup)

	Mon	Tue	Wed	Thu	Fri	Sat	Sun
File A	New	Changed	Changed				
File B	New						
File C	New	Changed		Changed	Changed		
File D	New		Changed				
File E	New			Changed			
File F		New	Changed				
File G		New				Changed	
File H		New	Changed				
File I		New		Changed			
	Full	Incr	Incr	Incr	Incr	Incr	Incr

De-duplicated backups
(include files modified and unique data created since the previous backup)

	Mon	Tue	Wed	Thu	Fri	Sat	Sun
File A	New	Changed	Changed				
File B	New						
File C	New	Changed		Changed	Changed		
File D	New		Changed				
File E	New			Changed			
File F		New	Changed				
File G		New				Changed	
File H		New	Changed				
File I		New		Changed			
	Full	Incr	Incr	Incr	Incr	Incr	Incr

Source: ESG

- Incremental backups: Unlike full backups, incremental backups copy only files that have changed since the last full or incremental backup. The main advantage of incremental backups is that they reduce the amount of files that you are backing up daily (versus a full backup), which allows for shorter backup windows. However, the restore process can be significantly longer since the last full and all subsequent incremental images, or copies, must be restored.

- Differential backups: Differential backups back up “all” data modified since the last “full” backup. This differs from incremental backups which include only data modified since the previous full or incremental backup. Once a file changes, it is backed up daily until the next scheduled full backup. So, clearly, the disadvantage with differential backups is that the size of the backup increases throughout the week as files are changed, becoming progressively larger until the next weekly full backup. However, on the recovery side, only the full backup image and most recent differential image need to be restored, potentially providing quicker restore than an incremental backup, depending on when the restore occurs.

- De-duplicated backups: By applying de-duplication to these three traditional backup approaches, users can significantly reduce the amount of non-unique data they backup. Full backups, incremental backups and differential backups do not scan for “uniqueness.” Again, the actual de-duplication rate depends on a number of variables (described above) but 10x to 20x-plus is typical.

Data De-Duplication's Benefits

Data de-duplication has several significant and immediate benefits for users. First and foremost, it can significantly reduce backup capacity requirements, which, among other things, can translate into cost-savings in a number of ways. It frees up capacity for backup data and enables longer retention periods, improves RTOs and reliability and can make WAN-based remote backup and replication more efficient. Let's take a look at each of these more closely:

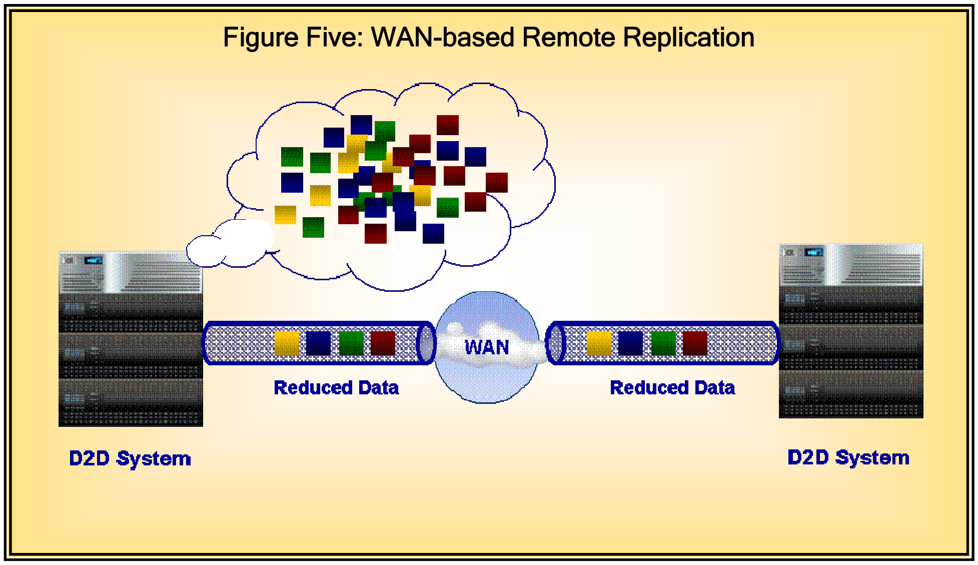
- **Reduced backup capacity requirements translate into cost-savings**. While the actual amount of capacity reduction varies from organization to organization depending on a number of variables, including the type of data that is being backed up, the change rate of the data and the frequency of the backup, ESG Lab finds 10x to 20x-plus reduction to be fairly typical. The ability to reduce disk capacity requirements by this ratio has some powerful cost-savings benefits for users, including lower disk and, perhaps equally important, lower power and cooling costs. In the case of disk costs, just consider the ability to store 20TB of backup data on 1TB of disk. The cost-savings are significant. In the case of power and cooling, which is becoming an increasingly important consideration in today's data protection environments, the ability to store more backup data on less disk (e.g., 20TB of backup data on 1TB of disk capacity) can reduce power and cooling requirements considerably.

- **"Freed up" capacity means more room for other backup data and longer retention periods with less media management**. Data de-duplication can reduce the amount of physical disk needed for backup. Users can use this "reclaimed" space for several purposes: 1) to bring other backup data onto disk and 2) to lengthen the retention periods of data that is backed up to disk. Bottom-line: De-duplication allows users to leverage disk as a backup target for more data and, importantly, allows data to be kept on disk for longer periods of time. Doing so has potentially huge benefits for users. Think about it. What if you could recover data that is three to six months old or even older – without ever having to go to tape? Without data de-duplication, doing so would not be economically prudent, but with data de-duplication, it is not only possible, it is cost-effective. Tape is reserved for long-term archival of data that is infrequently, if at all, accessed and for "doomsday" data recovery.

- **Data de-duplication enables better RTOs and improves reliability**. The more data users back up to disk, the better able they are to meet RTOs. Hence, data protection SLAs. Data de-duplication allows users to keep more data on disk and for longer periods, which enhances RTOs. The fact is recovery from disk is a lot fast than recovery from tape. As for reliability, again, data is retained longer on disk, which means users rely less on tape for recoveries.

- **Enables and expands WAN-based remote replication options for backup data**. Once again, the power of data de-duplication is in its ability to reduce the amount of data that is backed up. Because there is less physical data being trafficked over the WAN (see Figure Five), data de-duplication lessens the "cost" and/or "bandwidth" barrier of entry of WAN-based remote replication for many organizations, making it possible for some to do WAN-based remote replication for the first time and for others to cast a "wider net" of data protection around their remote data (i.e., include remote data previously not protected).

Figure Five: WAN-based Remote Replication



Implementing Data De-Duplication

There are many ways to implement data de-duplication - it can be done in software or through an appliance. As for the point of origin of the de-duplication process - that is, where the actual data de-duplication is done – it can be done *in-line* or *off-line:*

- In-line: The de-duplicating is done at the host by the backup application or by an appliance sitting in the data path.

- Off-line, or post-process: The de-duplicating is done by the system or an appliance sitting outside the backup path after the backup job is complete.

Both approaches are very effective in eliminating duplicate data and, again, have been proven in ESG Lab testing to provide enormous benefits. But as with any technology, there is a trade-off. In this case, it is performance and capacity. There is a performance impact to doing the de-duplication in the data path and there is a capacity hit to doing the process off-line since capacity has to be initially allocated to the backup process (this capacity is later released after the de-duplication process is complete).

Determining which approach is better for your environment requires a thorough capacity/performance trade-off analysis. If performance is critical, then the off-line approach may be the better route to take. But if it isn't – and you're looking for optimal disk capacity savings (throughout the entire process) – in-line may be the better approach. Of course, in-line versus off-line is just one consideration when evaluating de-duplication. As mentioned previously, technologies also differ in terms of the degree, or level of granularity, of de-duplication they do. All are important considerations when evaluating the different technologies that are available.

That said, it is important to note that while each approach has pros and cons in terms of performance, capacity and cost, ESG believes the benefits of data de-duplication – in particular, the potential disk cost-savings – are significant enough to warrant the technology's adoption.

Data De-Duplication the Quantum Way

In Quantum's case, data de-duplication can most closely be defined as "in-line" (or as data is ingested by the appliance), but in a two-stage "parallel" process to help speed performance. As data is written to disk (in this case, a Quantum DXi-Series appliance) in the native format of the backup application, it is de-duplicated. Unlike off-line de-duplication technologies, the DXi-Series appliance does not wait for the backup job to be completed before it begins the de-duplication process.

Also, for further cost- and capacity-savings benefits, and as part of the de-duplication process, the data stream is compressed. Quantum's compression is done in hardware, not software, in order to avoid the performance degradation common with software-based approaches.

A couple of things about Quantum's approach should be noted:

1) Quantum is the first vendor to ship an appliance that does *both* data de-duplication and hardware-based data compression. By de-duplicating data and then compressing it, Quantum makes optimal use of disk capacity, further reducing associated disk costs and maximizing WAN bandwidth for remote replication.

2) Doing the data de-duplication as it is being ingested by the appliance ensures better performance and data fidelity.

In terms of backup performance, Quantum claims a data throughput rate of up to 290GB/hour for its entry-level DXi3500 appliance and up to 800GB/hour for its high-end DXi5500 appliance. ESG Lab is planning to test Quantum's appliances to validate these performance claims. If true, the DXi-Series is significantly faster than competitive products in the midrange. On the recovery side, how quickly the data is restored depends on several factors, including where the data resides (i.e., whether in cache or on system disk), and whether the data has been de-duplicated or not. While recovering de-duplicated data may have slight performance impact versus recovering from native disk, the value compared to previous recovery approaches cannot be ignored. Data that has been de-duplicated must be "un-de-duplicated" before it can be restored. Typically, the "un-de-duplication" process is done using the same engine used during the de-duplication process.

In addition to data de-duplication, Quantum's new DXi-Series appliances have several other notable features:

- Organizations can replicate data asynchronously over the WAN from remote locations to their central location, or from a central location to a disaster recovery site, and because data at the remote sites is de-duplicated before it is sent, WAN traffic is minimized.

- The appliances can be presented to the backup application as NAS, VTL or both for optimal flexibility.

- The devices can be configured with Fibre Channel, GbE, or iSCSI connectivity. As iSCSI deployments increase, ESG believes it will be increasingly important for disk-based data protection solutions such as the DXi-Series to offer iSCSI connectivity, especially for their small to medium-sized products.

The DXi-Series consists of eight distinct models (divided between the DXi3500 and DXi5500 platforms). The DXi3500 scales from 1.2 to 4.2 TB of usable capacity and the DXi5500 scales to 11 TB. With Quantum's consultative sales approach leveraging advanced sizing tools, users can determine how much capacity they'll need to protect their applications today and over time. These tools provide guidance about expected de-duplication ratios, data growth rates, etc.

Currently, data de-duplication is only available with the DXi-Series appliances. However, Quantum does plan on integrating its own integrated software layer technology into future enterprise systems and data management software. Quantum also believes this technology is extensible into downstream applications as well.

Conclusion

As data volumes grow and business SLAs become more aggressive, users will increasingly find themselves in a Catch-22: they will need to keep more backup data online on disk longer to meet recovery objectives but they will also need to keep data protection-related budgets in check. Without technologies like data de-duplication – which make disk-based data protection more efficient – organizations will either find themselves faced with increasing storage-capacity and/or WAN-bandwidth-related costs in an effort to minimize potential negative business impacts of downtime (e.g., application downtime, dissatisfied users, lost data, direct revenue loss, etc.) or they will risk the exposure and limit what they actually back up to disk-based systems, such as VTL.

Data de-duplication significantly changes the economics of disk-based data protection and, as such, enables levels of efficiencies above and beyond what's possible today without it and eliminates problems that plague data centers today. Now companies can recover reliably and quickly, they can back up remote offices and they can minimize tape backups. For these reasons, data de-duplication is a very compelling technology.

As for Quantum, in introducing data de-duplication with its DXi-Series appliances, it joins the early ranks of technology vendors with this capability, and is in a good position – given its backup, recovery and tape archive heritage – to assume a leadership position in the data protection market, not just in terms of its product offerings but in educating organizations about the power of data de-duplication. Quantum understands data protection and the challenges it presents to users today.