Differentiating Hardware- and Software-based Data De-duplication

By Lauren Whitehouse, Analyst

February, 2008
# Table of Contents

Table of Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table of Contents</td>
<td>i</td>
</tr>
<tr>
<td>Introduction</td>
<td>1</td>
</tr>
<tr>
<td>Data Protection Challenges</td>
<td>2</td>
</tr>
<tr>
<td>Data De-duplication Adoption</td>
<td>4</td>
</tr>
<tr>
<td>De-duplication Delivery</td>
<td>4</td>
</tr>
<tr>
<td>Other Considerations</td>
<td>5</td>
</tr>
<tr>
<td>Symantec NetBackup PureDisk</td>
<td>7</td>
</tr>
<tr>
<td>Use Cases</td>
<td>9</td>
</tr>
<tr>
<td>Distributed Data Protection</td>
<td>9</td>
</tr>
<tr>
<td>Data Center Data Protection</td>
<td>10</td>
</tr>
<tr>
<td>Virtual Server Data Protection</td>
<td>12</td>
</tr>
<tr>
<td>Summary</td>
<td>13</td>
</tr>
</tbody>
</table>

All trademark names are property of their respective companies. Information contained in this publication has been obtained by sources The Enterprise Strategy Group (ESG) considers to be reliable but is not warranted by ESG. This publication may contain opinions of ESG, which are subject to change from time to time. This publication is copyrighted by The Enterprise Strategy Group, Inc. Any reproduction or redistribution of this publication, in whole or in part, whether in hard-copy format, electronically, or otherwise to persons not authorized to receive it, without the express consent of the Enterprise Strategy Group, Inc., is in violation of U.S. copyright law and will be subject to an action for civil damages and, if applicable, criminal prosecution. Should you have any questions, please contact ESG Client Relations at (508) 482-0188. This ESG White Paper was developed with the assistance and funding of Symantec.
Introduction

IT organizations are plagued with problems in ensuring that critical business systems and data are available. The safety net most often applied is maintenance of a secondary or backup copy of data so that business operations may continue even if the primary copy of the data is lost. The difficulty for IT is ensuring that duplicates are recent, readily available and can be recovered rapidly. Disk-to-disk (D2D) backup responds to these complexities and—when combined with data de-duplication—its implementation is more economically feasible.

Data de-duplication eliminates redundant data, saving valuable storage space. It may be deployed via hardware- or software-based approaches—each having advantages and trade-offs, such as ease of deployment, flexibility and scalability. Symantec NetBackup PureDisk is one such software solution enabling D2D backup with disk capacity and network bandwidth optimization.
Data Protection Challenges

ESG Research found that IT organizations are most challenged with keeping pace with the growing volumes of data that need to be protected1. As the capacity of data grows and regulatory mandates dictate longer retention periods, the amount of data under management may exceed the time allocated for backup. In an effort to reduce backup times, IT organizations are deploying disk in their backup process at an increased rate. However, ESG found that the cost of storage systems is another concern creating a conundrum for IT organizations (see Figure 1). How can IT provide adequate service level agreements for data protection, keep pace with data growth and keep spending in check? One way may be to implement a data de-duplication solution.

FIGURE 1. DATA PROTECTION CHALLENGES

Which of the following would you characterize as the primary challenge with your organization’s current data protection processes and technologies? (Percent of respondents, N = 398)

- Keeping pace with capacity of data to protect: 14%
- Need to reduce backup times: 13%
- Cost of storage systems: 11%
- Need to improve recovery reliability: 8%
- Need to improve backup reliability: 8%
- Meeting regulatory compliance requirements: 7%
- Difficult to validate backup/recovery success: 7%
- Need to reduce recovery times: 7%
- Limited or no ability to classify data for data protection purposes: 4%
- Searching for data across backups: 4%
- Unacceptable level of data loss: 3%
- Tape media management: 3%
- Cost of data protection software licenses: 2%
- Meeting e-discovery demands (i.e., litigation support): 2%
- Meeting corporate governance requirements: 2%
- Poor service and support from vendor(s): 1%
- Tape media costs: 1%
- None of the above: 6%

Source: Enterprise Strategy Group, 2008

Data de-duplication is the process of examining data to identify and eliminate redundancy. It can have a significant impact on the capacity of data stored, which, in turn, can deliver significant economic benefits.

Data de-duplication is being adopted by companies of all sizes, allowing IT to optimize the capacity of data stored in the backup process—significantly reducing the size of data stores and saving additional investments in disk storage. The popular route to de-duplication has been disk-based backup to “intelligent” disk targets, such as VTLs or NAS storage that employ some form of de-duplication for capacity reduction. These all-in-one hardware solutions are easy-to-implement and are non-disruptive to existing backup processes. However, as

capacity growth continues, a target-side solution’s ability to scale is limited to adding more appliances—often creating infrastructure sprawl and management headaches for IT administrators.

Data de-duplication is applicable to multiple backup scenarios including backup consolidation for distributed environments such as remote and branch offices (ROBO), capacity reduction of backup data stores within corporate data centers and capacity reduction in virtual server environments. Consolidating backup in a distributed environment involves transferring backup data from the application server to the backup storage device across a LAN or WAN. In this scenario, client-side de-duplication aids IT by lowering bandwidth requirements for backup data transferred across a LAN or WAN and requiring less storage capacity once the backup data is aggregated and de-duplicated.

Backup within the data center involves greater volumes of overall storage capacity, especially for larger companies. Disk-based backup offers faster, more reliable backup and recovery of data, allowing IT to deliver higher levels of service to its constituents. However, the cost of disk storage systems could spiral out of control if users do not leverage data de-duplication technology to eliminate redundant data in backup sets, thereby reducing storage capacity requirements. ESG Research found that of the organizations that have implemented data de-duplication, over 50% are utilizing it within corporate data centers (see Figure 2).

**FIGURE 2. DATA DE-DUPLICATION ADOPTION**

<table>
<thead>
<tr>
<th>In what areas of your organization's IT environment have you implemented data de-duplication technology?</th>
<th>(Percent of respondents, N = 58)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Within corporate data center(s), 52%</td>
<td></td>
</tr>
<tr>
<td>At remote / branch offices, 10%</td>
<td></td>
</tr>
<tr>
<td>Both remote / branch offices and corporate data center(s), 31%</td>
<td></td>
</tr>
<tr>
<td>Other/Don't Know, 7%</td>
<td></td>
</tr>
</tbody>
</table>

With the adoption of virtual server technology, IT organizations have seen a net increase in total storage volume, which has implications for data protection. ESG Research found that more than one-third of virtual server users had seen an increase in the total amount of data to back up after implementing server virtualization (See Figure 3).²

Server virtualization has also introduced a new backup challenge. While virtualization optimizes the use of processing, disk and network performance of physical hardware for application servers, it eliminates the idle cycles that were used for backup. New CPU and bandwidth optimized backup methods are required to backup virtual environments.

Data De-duplication Adoption

Data de-duplication is a feature available in backup applications and disk storage systems to reduce disk and bandwidth requirements in data protection processes. With less data to store, less disk storage is needed. That means that IT organizations will either save money by purchasing fewer disks or enable longer retention of data on disk. The latter provides faster access to data for recovery purposes, which can improve service levels.

Capacity optimization is not new; data compression has been around for years and is implemented by most companies in the data protection process. Newer capacity optimization technologies, such as single-instance storage and block-level de-duplication, are gaining in adoption. Small and medium-sized (SMB) companies that eliminate tape-based backup, replacing it with D2D backup, have embraced de-duplication to improve the economics of disk storage over tape.

Enterprise-class companies have shown greater adoption than SMBs. Data de-duplication can have a tremendous impact on the capacity glut afflicting these firms. ESG Research found that the majority of companies using de-duplication have seen a 10:1 or greater capacity reduction ratio. This means that for every single TB of primary data, which would typically generate 10 TB of backup data, organizations can use de-duplication to reduce that backup load and store it on only 1 TB of disk.

De-duplication Delivery

De-duplication software is gaining popularity as it is deployed in purpose-built appliances, such as virtual tape library (VTL) solutions and NAS target systems; or as an option to backup software solutions. There are several advantages and disadvantages to packaging data de-duplication in hardware and software solutions.
**Purpose-built Appliances**
Appliance-based solutions are pre-assembled, pre-configured and pre-tested—offering a more plug-and-play experience for the end-user and easier support and maintenance for the vendor. The drawbacks of an appliance model include reduced flexibility and scalability.

Intelligent disk with de-duplication is compatible with existing backup software—VTLs appear as tape libraries, drives and media, requiring little to no reconfiguration of backup processes, and de-duplication appliances appear as a disk target. Some appliances are bundled with storage hardware adding to the seamlessness of deployment and others can be coupled with external disk allowing the IT organization to decide on disk vendor selection. As capacity grows, additional units may be required to keep pace, as de-duplication is typically limited to a single unit. This will decrease the capacity reduction across the whole dataset and require careful configuration of the backup streams to achieve the best optimization. The result is added management complexity and overhead, especially if a consolidated view of appliances is not offered.

**Software Approach**
The software approach does offer more flexibility and scalability, which can be seen in a few ways. First, the IT organization can use any disk storage of choice, including existing disk in its environment. Second, disk capacity is more adaptable—the solution can be more easily sized to meet storage volume and budget requirements. Software-based solutions also scale capacity while maintaining de-duplication across the dataset, eliminating the management headaches that typically accompany management of multiple, individual appliances.

Software solutions may require more time and technical knowledge for installation and configuration. Additional hardware may have to be acquired, but, end-users have the flexibility to use either new or existing hardware with their configuration. Should new hardware purchases (servers and/or storage) be required, the aggregate footprint may be less than that of an appliance.

**Other Considerations**
To fully understand the trade-offs associated with either deployment approach, there is another factor that should be evaluated: where and when de-duplication occurs. If bandwidth reduction is of importance, de-duplication needs to happen on the application server before the backup data is transmitted across the network, which is referred to as client-side de-duplication. Alternatively, de-duplication of the backup stream can happen on the backup server which can be referred to as proxy-de-duplication or on the target device which is target-de-duplication.

If de-duplication occurs after data is written to the disk target, it’s considered a post-process de-duplication. Data that is de-duplicated before being written to disk is referred to as inline de-duplication.

Post-process de-duplication will write the backup image to a disk cache before starting de-duplication, which allows the backup to complete at full disk performance. Post-process de-duplication requires disk cache capacity sized for the un-de-duplicated backup data, plus the additional capacity to store de-duplicated data. The size of the cache depends on if the de-duplication solution waits for the entire backup job to complete before starting de-duplication and whether or not the de-duplication solution releases storage space as it de-duplicates. The latter could have some impact on performance.

Inline de-duplication won’t require a reserve of capacity, but could impact backup performance. Inline approaches inspect data on the way to the disk target. Performance degradation depends on several factors, including the:

- **De-duplication algorithm**
  The algorithm is the method of “fingerprinting” the data and checking that fingerprint in the master index to determine redundancy.
Differentiating Hardware- and Software-based Data De-duplication

- **Granularity of de-duplication**
  At what level is the data being examined? A file-level de-duplication checks file attributes and eliminates redundant files. Block (and sub-block) de-duplication carves the data into fixed- or variable-sized chunks, fingerprint the block and compare it to the index—the more chunks, the more comparisons.

- **Location where inline de-duplication occurs**
  Client de-duplication fingerprints data as it is sourced (at the host) through agent technology. The agent communicates with the server hosting the index/database. If the data is unique, it’s transferred to the storage device; otherwise, it’s ignored. This method optimizes storage capacity and network bandwidth, but could tax the host. Proxy de-duplication on a backup media server uses a similar approach, but eliminates the CPU overhead on the client. Both client and proxy de-duplication optimize performance by distributing the de-duplication processing of the aggregate backup load across a large number of clients or media servers. Target de-duplication requires powerful purpose-built storage appliances as all of the backup load needs to be processed simultaneously on the target (see Figure 4).

- **Scalability of the solution**
  How rapidly data can be inspected, compared with the central index and ingested depends on the architecture of the solution and how it distributes the workload.

**FIGURE 4. INLINE DE-DUPLICATION**
Symantec NetBackup PureDisk

Symantec NetBackup PureDisk is a software solution that offers disk-based data protection with global de-duplication to deliver bandwidth and storage optimization.

Initially introduced as a standalone platform for protecting the ROBOs of larger organizations, the solution has been more tightly integrated with Symantec Veritas NetBackup—the market-leading backup/recovery software—and now has applicability in the data center. PureDisk integrates with NetBackup to create tape media—with NetBackup’s media catalog tracking the operation. Further integration between products allow NetBackup-captured images to be de-duplicated. PureDisk replication allows for efficient transfer of NetBackup images to a remote site for disaster recovery purposes.

PureDisk’s architecture (see Figure 5) enables the solution to non-disruptively scale as backup growth dictates more capacity. PureDisk consists of two main components:

1. **De-duplication Engine**, implemented in either:
   - **PureDisk client/agent software** (installed on systems to back up), which handles de-duplication, compression, encryption and application integration.
   - **PureDisk Deduplication Option plug-in for NetBackup media servers**, which enables a media server to de-duplicate NetBackup backup images in-line.

2. **PureDisk Storage Pool**, consisting of one or more nodes, which are x86 servers with PureDisk operating system and application software. A Storage Pool has two types of nodes—the Content Router nodes store (and restore) the de-duplicated backup file data on third-party DAS, NAS or SAN storage and maintain the distributed fingerprint index. The MetaBase nodes form PureDisk’s distributed catalog/database containing metadata. The MetaBase and Content Router components can exist on one server or across multiple servers. All nodes form one large logical storage system for de-duplicated data, which is managed from a single storage pool manager through a Web interface.

Each MetaBase engine has a recommended limit of 400 million metadata entries. Database scalability can be achieved by adding another server, configured with a MetaBase engine, to the Storage Pool. Content Router scalability can be achieved in a similar fashion (and without disrupting existing backup procedures and policies) by adding additional servers configured as Content Routers to the Storage Pool. A distributed model aids in boosting performance and storage capacity while maintaining de-duplication across the whole dataset.

PureDisk Storage Pools support global de-duplication—aggregating only the globally unique data changes—which is useful when consolidating data protection for multiple sites or from multiple data sources.

PureDisk client-side de-duplication delivers bandwidth optimization benefits for backup data transferred over a LAN or WAN, or from a virtual guest. PureDisk clients installed on the host systems targeted for backup perform data de-duplication. Data can be optionally encrypted client-side for added security using embedded 128-bit encryption. When the data arrives at the PureDisk Storage Pool, no re-encryption need occur.

The PureDisk agent segments larger files into small chunks; fingerprints each segment (performing an algorithm to generate a unique identifier for that segment); and compares the unique ID with each fingerprint in the distributed index on the Content Routers, which contains all fingerprints of stored segments of previously examined files. Even identical files with different attributes—names, locations, creation dates, etc.—will have the same fingerprint. In files that have been modified slightly, the unique changed segments will be identified. Only new and unique segments are transferred over the network, reducing bandwidth by a factor of 50x or more.

---

3 Available in PureDisk 6.5
4 An increase from 100 million to 400 million occurs in release 6.5.
5 Content router capacity increases from 4 TB to 8 TB in release 6.5.
Similarly, the PureDisk Deduplication Option plug-in on the media server de-duplicates NetBackup backup streams from NetBackup clients by splitting them into small segments and identifying the unique segments based on their fingerprint. This can be done in-line during the backup or as a post-process duplication from staging to a PureDisk storage pool.

As data is transferred to the PureDisk storage pool, it is not limited through client or proxy de-duplication as PureDisk supports running hundreds of backup streams simultaneously to a single storage pool.
Use Cases

NetBackup PureDisk offers significant benefits over hardware-based de-duplication solutions in a number of environments.

Distributed Data Protection

One example of a distributed data protection configuration consists of PureDisk clients at the remote location (ROBO or department) transferring data directly over a network connection to a PureDisk Storage Pool at the central data center (see Figure 6).

In this scenario, the ROBO or department may have previously implemented a traditional tape backup solution—one that collected and transferred data over a network connection to the central backup server. With remote backup, it’s likely that the network transfer of data could not complete within a prescribed window of time. Too much data had to travel through too small a pipe, causing bottlenecks in the backup process.

FIGURE 6. PUREDISK CONFIGURATION FOR DISTRIBUTED DATA PROTECTION – NO ONSITE STORAGE

Through PureDisk’s client-side de-duplication, the capacity of data is reduced before being transferred to the storage pool where data is consolidated from multiple locations. PureDisk’s 50x or more capacity reduction solves the distributed data protection problem and makes remote backup of data economically feasible for companies with bandwidth challenges. Should a tape copy be required, PureDisk can export the data to NetBackup, which can create a copy of the data on tape media.

Another example of backup consolidation for ROBOs or departments can be a configuration with local storage for local operational recovery of data and centralized storage for disaster recovery. The configuration would consist of PureDisk clients at ROBOs/departments storing data locally in a PureDisk Storage Pool. The local Storage Pool would be replicated to PureDisk storage in the central data center (see Figure 7).

In this scenario, the ROBO or department may have previously implemented an on-premises traditional tape backup solution—one that performed backup to local tape hardware, with tape media being sent offsite for disaster recovery purposes. A local tape backup solution requires onsite tape hardware, tape handling and management.

The PureDisk implementation provides a D2D backup strategy—improving backup performance and reliability over the previous tape-based solution—and removes the administrative burden associated with managing tape devices and media. The ROBO or department still has local storage for rapid recovery. However, the storage is
Differentiating Hardware- and Software-based Data De-duplication

optimized by PureDisk’s de-duplication, allowing for increased retention or more frequent backups—greatly improving recovery time and recovery point objectives (RTOs/RPOs). Client-side de-duplication allows for multiple backups to be performed simultaneously, without compromising the LAN. Global de-duplication (for data replicated to the central storage pool in the data center) adds to the efficiency of the implementation. The remote storage pool can be managed through a Web interface by the central IT staff, eliminating the need for local IT staff and reducing the management costs for the remote backup solution, as it can be fully monitored and operated from the data center.

**FIGURE 7. PUREDISK CONFIGURATION FOR DISTRIBUTED DATA PROTECTION – WITH ONSITE STORAGE**

Data Center Data Protection

A data center data protection configuration consists of PureDisk Storage Pool as a disk target for the NetBackup backup application (see Figure 8).

Today, NetBackup supports tape and disk targets for backup. Disk targets can include disk arrays such as DAS, SAN and NAS for which NBU offers a broad range of flexible disk options, and intelligent disk appliances through the OpenStorage API such as VTL and NAS with capacity reduction solutions. Similar to OpenStorage, NetBackup can use the PureDisk Storage Pool as the intelligent disk target for backup sets through the PureDisk Deduplication Option. This option is different in that the de-duplication is performed in software on the media server, rather than on the target device.

Deploying disk arrays as a backup target for NetBackup requires a lot of capacity as every backup (full or incremental) is stored as a backup image, which represents a lot of redundant data. The lack of capacity reduction will affect the volume provisioned, the footprint of disk storage in the data center, as well as power and cooling requirements. Because the data being stored on the disk is not optimized, IT or storage administrators will need to use the flexible disk options in NetBackup or manually manage the disk (de-fragmenting and grooming disk as backup sets expire) to make best use of the available disk capacity. This complexity might cause backup administrators to become stingy with disk—decreasing retention times or the frequency of backups for D2D, which can affect RTOs and RPOs.

Conversely, deploying PureDisk Deduplication Option decreases disk storage capacity while supporting longer retention times and more versions through de-duplication. The capacity savings allows better recovery objectives due to increased retention and/or frequency of backups. A PureDisk Storage Pool easily scales beyond the boundaries of physical disk volumes by adding Content Router nodes—without impacting ongoing backup and recovery processes or requiring re-configuration. Intelligent Disk and VTL appliances do not offer this flexibility and are limited in scalability as they typically only de-duplicate within the device, not across devices. This requires backups from a particular client to be stored to the same device at all times to benefit from de-duplication, which then limits the flexibility of using NetBackup load balancing, capacity management, backup spanning and failover features.
PureDisk Deduplication Option de-duplicates the NetBackup data streams on the NetBackup media server. Multiple de-duplication processes can run on parallel backup streams on a media server, and multiple media servers can share a single PureDisk storage pool. This architecture is set up to scale the de-duplication performance beyond the boundaries of a single processing head, while maintaining de-duplication across the whole dataset. As de-duplication only happens while data enters the appliance, target de-duplication devices are limited by the aggregate de-duplication performance of the device.

To improve DR processes, the built-in replication capabilities of the PureDisk Storage Pool can be controlled from NetBackup to create an off-premise copy. In addition, the PureDisk Storage Pool nodes can be configured as a cluster for High Availability.

With PureDisk storage as a target for NetBackup, IT has more choice. In addition to allowing flexibility for disk type, vendor and capacity, IT has the ability to set inline or post-process de-duplication strategies on a per backup policy basis (client-side inline for distributed backup, inline for LAN rate backup and post-process for SAN rate backup, for example). Data de-duplication on the media server allows throughput to be scaled with the number of media servers. Importantly, PureDisk Storage Pool, as its name implies, allows for storage to be pooled and de-duplication to occur across ROBO, department and data center backup sets.

![FIGURE 8. PUREDISK CONFIGURATION FOR DATA CENTER PROTECTION](image)
Virtual Server Data Protection

A virtual server data protection configuration consists of PureDisk clients residing in virtual machines backing up to a local and/or remote PureDisk Storage Pool (see Figure 9).

In server virtualization configurations, IT organizations may have previously deployed traditional client/server backup products, backing up to either tape or disk. The comfort level with implementing traditional backup in physical environments is mirrored in how backup is handled in virtual environments—deploying client agents in virtual machines to collect and transfer data to the LAN-based backup server. Since a virtual machine is represented as a single file on the host system, traditional backup applications will target it for backup when an incremental strategy is employed. This can result in a dramatic increase in the size of the backup and the required storage capacity. A traditional backup agent in the virtual guest (host OS) will have a high impact on the CPU utilization of the virtual server, which typically does not have enough CPU cycles left for completing the backup.

PureDisk offers significant value in VMware environments because it removes multiple instances of redundant data within and across virtual machines, effectively eliminating the backup load of traditional full backups. De-duplicating data will significantly reduce the strain on shared server resources and applications, as well as the amount of data copied, transferred and stored. PureDisk also creates a full backup image that can be immediately restored—no multi-step recovery of full and subsequent incremental/differential backups is required.

PureDisk virtual machine backup complements the NetBackup VCB backup solution, which offers granular, single file recovery from image-based backups for SAN environments. While the VCB backup is targeted at large SAN-enabled environments, the PureDisk virtual machine backup offers a solution for application backup running in the virtual machine and for environments where virtual machines are not running on SAN- or ESX-enabled virtual servers.

**FIGURE 9. PUREDISK CONFIGURATION FOR VIRTUAL SERVER PROTECTION**
Summary

Data growth, the need to protect data and the desire to save money will not change any time soon. All of these issues are increasing, placing tremendous demands on IT organizations. D2D backup solutions are taking hold and those that leverage data de-duplication are delivering tangible benefits.

There are distinct differences between hardware- and software-based D2D de-duplication products. While NetBackup supports both approaches, a software solution—such as NetBackup PureDisk—can offer more flexibility, including the type of disk storage hardware and the size and scale of the solution over time. When PureDisk is deployed in typical user configurations such as distributed, data center and virtual server environments, the benefits versus those of hardware-based solutions become evident.

IT organizations considering data de-duplication need to look past the immediate benefits of hardware-based approaches—mainly, the ease of deployment—and consider the long-term viability of the solution. As requirements evolve, issues of scalability, flexibility and manageability of the solution, as well as applicability to multiple use case scenarios will take precedence. ESG believes that software-based approaches provide benefits that can future-proof data de-duplication decisions.