

WHITE PAPER

Realizing the Benefits of Deduplication in a Backup and Restore System

Sponsored by: HP

Noemi Greyzdorf
November 2008

Robert Amatruda

INTRODUCTION

Executive Summary

Small and medium-sized businesses (SMBs) are seeking better, more cost-effective ways to store, back up, and safeguard their critical data. In the past, many companies relied solely on tape data storage for data protection. Increasingly, SMBs face the daunting task of managing the ever-increasing amount of data. Today, in addition to traditional tape storage backup, customers have more options to employ disk products to meet their backup and recovery requirements. Rather than relying solely on tape as a data protection medium — which can fall short in terms of recovery objectives — organizations are increasingly turning to the benefits of a disk-to-disk-to-tape (D2D2T) strategy. In addition, D2D products now feature data deduplication technology, which reduces overall solution costs.

Data Protection Challenges

The exponential growth of digital data is occurring in organizations of every size. The creation and retention of more data translate into more data to be protected. Increasingly, larger data sets, combined with a push to conserve facility and energy resources, have stretched traditional data protection methods to the limit. Some areas in which IT managers are facing new challenges include the following:

- ☒ Globalization of markets has put pressure on IT operations, driving IT managers to ensure greater uptime and availability of applications and data and quicker recovery in case of a disruption. As a result, backup windows are shrinking and recovery time objectives are becoming more stringent.
- ☒ On average, organizations have been increasing the amount of storage consumed 30% year over year. The shrinking backup windows and more stringent recovery objectives, together with larger data sets to be protected, are pushing the limits of traditional data protection methodologies.
- ☒ Recent spikes in energy costs and increased competition from emerging markets have put pressure on the margins of many enterprises. IT, often viewed as a cost center, has been forced to manage larger environments without increasing personnel. To achieve these goals, managers have had to seek ways to simplify operations wherever possible.

Larger data sets, shorter backup windows, more stringent recovery objectives, and the need to manage more with the same resources are just some of the challenges serving as catalysts for change to the traditional approaches to data protection.

The Evolution of the Traditional Data Protection Systems

To highlight the value of new technologies that have come on the market and address many of the challenges listed, we need to step back and look at the traditional data protection methodology. Medium-sized enterprises with over 100 servers and 20TB of data could deploy a backup application that would manage a number of media servers all writing to tape. Wherever possible, compression at the tape drive would be turned on to optimize cartridge capacity and performance. Features such as multiplexing and multistreaming were used to create greater throughput and efficiencies. The emphasis was on the backup job being completed. Once a backup was completed, tape cartridges were often removed from a tape drive or library and sent offsite to serve as a backup in case of a disaster or for longer retention to comply with some of the internal and government regulations. This is a typical scenario that still takes place today.

But now the same organizations are grappling with rapidly growing storage requirements. IDC has found that those organizations are experiencing 50% growth in data, adding as much as 4TB of data each year. In addition, datacenter facilities are under pressure to reduce physical and environmental footprints, especially in the era of high-cost energy and electricity. Despite the physical and environmental limiting factors, many organizations need to retain data for longer periods of time and be able to recover quickly in the event of a disaster. It's well-documented that every hour of downtime can cost significant amounts of money and put a drag on productivity. Organizations relying solely on tape may have difficulty restoring their business-critical data in a timely manner or to the defined recovery point objective, which would be measured in hours. As a result, many companies are increasingly integrating disk into their data protection strategy plans.

A typical enterprise customer must augment its existing data protection scheme in a way that will not require a complete overhaul of its infrastructure. However, relying on tape alone will not fully mitigate risk during a catastrophic failure and ensure a timely recovery. Therefore, the integration of disk as a target media for backups may significantly improve the performance of the backup as well as the reliability of and the ability to recover the data. While tape storage continues to be a very cost-effective way to safeguard business-critical data, it must coexist with capacity-optimized disk systems that have been declining in price aggressively in the past several years.

Data storage capacity optimization has been available for many years in the form of compression on tape drives and virtual tape libraries (VTLs). Data compression can be software or hardware enabled. Typical data compression rates can vary due to type of data and application, but 2:1 or 3:1 rates are commonplace. However, in recent years, a more sophisticated approach to data compression has been made available as an option for disk systems: data deduplication. Data deduplication is the process of identifying redundant data and removing it from the datastream. Furthermore, data deduplication stores only unique instances of a data set or object. Backup data tends to be extremely repetitive, with daily incremental or weekly

backups at the file level creating many redundant files. Now, utilizing data deduplication, an organization can significantly reduce the required physical capacity to store backup images. Another benefit is that more data can be stored on a disk system for a longer period of time.

A number of other characteristics differentiate deduplication schemes. Some approaches are inline, others are postprocess. Some algorithms are hash based, others are object based. Inline deduplication means that data is deduplicated before it is committed to a disk drive. Postprocess first writes data to disk, and once a job or a data set has been completed, the deduplication process ensues. Both approaches achieve the desired result of reducing the data footprint. What matters when deciding which approach is better suited for your environment is whether the job can be deduplicated and committed to disk in the allotted time.

HP D2D BACKUP SOLUTIONS

Products

The HP D2D2500 and D2D4000 are fully integrated appliances with a VTL interface with built-in dynamic deduplication. The HP D2D systems use dynamic deduplication, which processes the backup datastream at time of backup and does not need additional storage capacity to hold complete backups for postprocess data reduction. HP has found that dynamic deduplication is well-suited to work with a broad array of backup software and operate with many different data types.

The HP D2D2500 and D2D4000 are disk-based data protection solutions that are designed to integrate seamlessly into existing environments and work with existing backup software applications to automate daily backup of multiple servers or host systems. Both systems are rackmountable and are designed to automate and simplify backups and reduce the time spent managing data protection processes. The HP D2D2500 and D2D4000 systems can reduce potential failures associated with physical tape drives or tape media, thus alleviating the unnecessary staff time associated with daily unattended backups or backup failures.

HP has delivered two D2D systems that will address the majority of data protection needs of most SMBs and remote offices. The features are exclusive to the D2D2500 and D2D4000 systems and include the following:

HP StorageWorks D2D2500 Backup System

- 1U rackmountable D2D backup system with 3TB raw capacity
- Dynamic deduplication, which enables customers to retain up to 50 times more data on disk for longer
- Supports backup of up to six servers
- Up to 180MBps aggregate performance
- Two iSCSI interfaces

- Hardware RAID 5 security
- Physical tape support for HP rackmount LTO-2 or LTO-3 tape drives and export with direct copy facility
- Supports export to physical tape without impacting network bandwidth

HP StorageWorks D2D4000 Backup System

- 2U rackmountable D2D backup system with either 4.5TB or 9TB raw disk capacity
- Dynamic deduplication, which enables customers to retain up to 50 times more data on disk for longer
- Supports backup of up to 16 servers
- Speeds of more than 80MBps or 288GB per hour aggregate performance
- Two 4Gb Fibre Channel or two 1Gb iSCSI interfaces
- Hardware RAID 6 security
- Physical tape support to copy or duplicate or mirror to physical tape for archive and offsite vaulting

Another advantage of HP's D2D2500 and D2D4000 systems is seamless integration into existing environments without any additional initial investment. Also, unified management will resonate well with larger customers that value a Web-based browser interface, allowing for the systems to be monitored locally or remotely. In addition, customers can view results or change settings remotely.

Remote Office Backup

Distributed operations are much more commonplace in today's global business environment with offices in remote or multiple sites. Oftentimes, these remote or branch offices are equipped with basic office applications such as file and print; however, email and databases must be supported. Unlike the datacenter, branch or remote offices do not have the trained personnel or practices to manage the IT infrastructure. Thus, automating or eliminating many of the manual IT functions alleviates the costs and mitigates risks associated with failures or human errors. Virtual tape or other disk appliances with deduplication can really improve operations and increase automation. D2D backup systems with storage optimization such as data deduplication enable more backup data to be stored on disk for faster recovery without restoring from physical tape. However, D2D backup system integration with physical tape is still cost-effective and necessary for long-term archive and compliance. Additionally, backups that have been deduplicated and written to a D2D backup system can be replicated to a central location much more efficiently since less backup data needs to be transferred over the remote link. Once the backup data is moved to the central location, it can then be moved to physical tape based on corporate policies. For example, a remote office with a file and print server managing

a terabyte of data is backed up daily on an incremental level, and over the weekend a full backup is performed. Using traditional methods, with a rate of change at 10%, there would be 1.5TB of raw media consumed for backups. If this data then has to be moved offsite, tapes are moved by a third-party provider or must be replicated if disk is involved. If the data is compressed, there might be 750GB to replicate (assuming disk is used as the target media). If the backup is optimized or deduplicated, only 100GB may be replicated. The bandwidth required to move 750GB versus 100GB is very significant indeed.

Small and Medium-Sized Business Backup

SMBs continue to struggle with data protection strategies that align with their business requirements and their budgets. At the high end, there are systems that offer VTL or NAS interfaces with data deduplication and support for export of physical tape for archive and compliance. For businesses with less data and smaller IT budgets, these solutions are not as accessible. For smaller businesses, appliances with data deduplication enabled provide a less costly, more easily deployed and managed alternative. Using deduplication appliances with VTL interfaces packaged in capacities that reflect most SMBs addresses the need for faster and more reliable backup and restore and an efficient way to move backups offsite for longer retention and disaster recovery.

Disaster Recovery

Natural disasters such as Hurricanes Katrina and Ivan, forest fires in Southern California, and floods in the Midwest have focused IT managers' attention on establishing a disaster recovery plan in case they find themselves affected. In evaluating options, managers are discovering that recovering from tape has two potential weaknesses. First, it can be time consuming, causing the business to be down for a significant period of time. Second, unless the tapes are stored at a reasonable distance from the datacenter, getting the tapes to a recovery site can also be problematic as became apparent during the September 11th attack on the World Trade Center in New York City. Having the backups on a deduplication-enabled appliance enables administrators to replicate this data to a site at a comfortable distance while retaining the benefit of restoring from disk, which can be faster and more reliable.

Challenges

HP has introduced deduplication appliances focused on SMBs in a market already full of solutions offering capacity optimization embedded into disk systems or offered as software options. The market is feeling a bit confused and is looking for differentiation. HP is facing both a challenge and an opportunity to educate the consumer and communicate the unique value proposition of the D2D solutions.

CONCLUSION

The HP D2D2500 and D2D4000 are fully integrated appliances with a VTL interface with built-in dynamic deduplication. These disk-based data protection solutions are designed to integrate into existing environments and work with existing backup software applications to automate daily backup of multiple servers or host systems. HP has found that dynamic deduplication is well-suited to work with a broad array of backup software and operate with many different data types. As SMBs seek better, more cost-effective ways to store, back up, and safeguard their critical data, D2D solutions are well-placed to meet that increasing requirement.

Copyright Notice

External Publication of IDC Information and Data — Any IDC information that is to be used in advertising, press releases, or promotional materials requires prior written approval from the appropriate IDC Vice President or Country Manager. A draft of the proposed document should accompany any such request. IDC reserves the right to deny approval of external usage for any reason.

Copyright 2008 IDC. Reproduction without written permission is completely forbidden.