White Paper

Improving Backup Effectiveness and Cost-Efficiency with Deduplication

By Lauren Whitehouse

October, 2010

This ESG White Paper was commissioned by Fujitsu and is distributed under license from ESG.

© 2010, Enterprise Strategy Group, Inc. All Rights Reserved
Introduction

IT professionals responsible for backup and recovery operations are seeking ways to make processes automated, faster, more reliable, and less costly via the use of disk-to-disk-to-disk (D2D2D) or disk-to-disk-to-tape (D2D2T) strategies. Inserting disk in the backup data path is a sure way to improve performance and reliability and reduce operator intervention. However, to affect cost, deduplication will be a necessary component of a disk-to-disk (D2D) solution. Fujitsu has a portfolio of D2D offerings with data deduplication that can introduce efficiency and improve the economics of data protection. This paper focuses on the ETERNUS CS800 solution, Fujitsu’s new file-, OST-, and VTL-interface disk target solution with deduplication, that is ideally suited for midmarket organizations (companies with 100-999 employees) and remote and branch offices (ROBOs) of larger organizations seeking to reduce or eliminate their dependence on physical tape media for tape backup.

Why is Deduplication Needed?

Data Growth

Companies large and small rely on digital information to conduct business. Financial applications, human resource systems, customer relationship management software, and supply chain management solutions are the backbone of most organizations. These systems, as well as increased reliance on collaboration, corporate Web sites, and messaging systems, also contribute to information growth. Data growth, which most ESG research respondents reported in the 11% to 30% range per year, is most often complicit in introducing challenges for data protection. ESG research found that the top ten data protection challenges¹ (see Figure 1) are related to the volume of data that IT organizations have to manage. The more data under management, the greater the impact on IT’s data protection resources including operational staff, the time required to back up data, storage hardware, and network bandwidth as well as IT’s ability to meet recovery time and recovery point objectives (RTOs and RPOs).

Figure 1. Top Ten Data Protection Challenges for Midmarket Organizations

Which would you characterize as the primary challenge for your organization? (Percent of respondents, 100-999 employees, N=206, top 10 responses)

- Keeping pace with capacity of data to protect: 13%
- Need to reduce backup and recovery times: 9%
- Unacceptable level of data loss or downtime: 7%
- Backup hardware costs: 6%
- Remote site backup: 6%
- Desktop/laptop/end-point backup: 5%
- Backup software costs: 5%
- Lack of disaster recovery plan or process: 5%
- Poor service and support from vendor(s) management: 4%


Increased Use of Disk in Backup and Recovery

When it comes to data protection, one remedy to this situation is the introduction of disk storage in the backup data path. As shown in Figure 2, ESG found that nearly three-quarters (71%) of midmarket organizations employ disk in a disk-to-disk (D2D) or disk-to-disk-to-tape (D2D2T) fashion. By doing so, IT organizations can accelerate backup to complete processes within the prescribed window of time, increase reliability of backup and recovery processes, reduce operator intervention, and improve operational recovery time.

Figure 2. On-site Data Backup Process – Midmarket Organizations

Which of the following best describes your organization’s on-site data backup process? (Companies with 100-999 employees, N=183)

- Back up directly to tape (no disk-based storage systems), 24%
- Back up to disk-based storage systems only (no tape), 17%
- Back up to disk-based storage systems and tape, 58%


Increased Retention on Disk

While the use of disk solves some of the challenges introduced by data growth, it also introduces new problems. First, the increased use of disk may increase spending on storage hardware. Over the last few years, ESG has seen a greater tendency to increase retention time of backup data on disk. In a 2008 study, ESG research found that 66% of respondents retained data on disk for one week.\(^2\) More recently, however, ESG research found that 61% of 2010 midmarket survey respondents retain data on disk for one month or more (see Figure 3).\(^3\) This increase directly impacts the volume of data managed on-premises as well as disk capacity purchase requirements.


Consolidating and Distributing Data over Low Bandwidth Networks

Another problem is the transfer of backup data between corporate sites (in the case of centralized backup of remote and branch office data) or between the primary backup site and the remote disaster recovery site. Most companies maintain an on-site backup copy and a second copy of backup data at a remote location for operational and, importantly, disaster recovery. Of those surveyed, most (67%) maintain a local backup copy and a remote backup copy where the off-site copy is transported on physical media or transferred over a WAN to the second site (see Figure 4). Of those surveyed, nearly 10% are consolidating backup of distributed data at branch offices at a central location. Both of these scenarios will require adequate network bandwidth to enable data transfer within the time allocated.

Figure 4. On- and Off-site Data Backup Process – Midmarket Organizations

Which of the following best describes how the data backup process is primarily managed by your organization? (Percent of respondents, companies with 100-999 employees, N=206)

- Data is initially backed up to on-site storage & a copy is sent off-site via removable media (i.e., tape) (D2T2T or D2D2T) - 40%
- Data is backed up to on-site storage with no off-site copy (D2D or D2T) - 22%
- Data is initially backed up on-site & a copy is sent off-site via removable media &/or WAN (D2D2T, D2T2T, D2D2D) - 22%
- Data is backed up over the WAN directly to a secondary corporate site (no on-site storage) (D2WAN2D) - 9%
- Data is initially backed up to on-site storage & a copy is sent to a third-party online backup service provider (D2D2C) - 5%
- Data is backed up over the WAN to a third-party online backup service provider (no on-site storage) (D2C) - 2%
Deduplication Benefits

Deduplication introduces significant benefits in data protection processes since it identifies and eliminates redundant data segments, reducing the amount of data transferred and stored. In backup, data is initially seeded on the backup disk storage device. Deduplication technology examines subsequently written data to identify and eliminate duplicates so that only unique data is written to storage. When duplicates are found, only a pointer linked to the unique piece of data is stored. This pointer consumes significantly less space than storing the whole item multiple times. When applied to the backup environment, deduplication can not only change the economics of employing disk, it can enhance on- and off-premises processes to:

- **Optimize disk capacity.** Deduplication reduces disk capacity requirements, stemming or slowing purchases and potentially allowing for existing disk capacity to be “reclaimed.” By reducing the footprint of disk in the data center, efficiency is also gained in floor space, power, and cooling. The impact of optimized disk capacity may be seen in both capital and operational expense savings.

- **Make more efficient use of bandwidth.** Optimization of data also lowers LAN/WAN/SAN bandwidth requirements; however, the most impact may be seen in WAN bandwidth since capacity optimization can enable site-to-site transfer of backup data over existing bandwidth where it might not have been feasible before. In the case where this allows for the elimination of tape hardware and media, savings can be introduced by eliminating or reducing tape devices, tape media, off-site storage fees, and the intervention of operational staff. More efficient use of bandwidth creates new strategies for disaster recovery copies and consolidating data protection for distributed data.

- **Improve RTO and RPO.** Deduplication can increase the likelihood that more data can be rapidly backed up to and recovered from disk, meeting prescribed backup windows and providing an improvement in recovery time objectives (RTOs). Disk “savings” resulting from deduplication optimization may enable more frequent backup copies on disk. Performing backup at two or more points in a 24-hour period increases the number of recovery points, improving recovery point objectives (RPOs).

Deduplication can have a direct or an indirect impact on the aforementioned challenges noted in Figure 1. Most notably, it allows IT organizations to keep pace with the double-digit data growth they are likely experiencing. Since deduplication makes backup to disk more feasible, expanding disk-based backup to additional workloads in the environment can help operational staff meet backup windows. Similarly, recovering from disk helps IT staff meet RTOs while greater reliability and frequency in backup jobs aids in meeting RPOs. Disk capacity optimization and the potential to reduce or eliminate tape hardware can contribute to lowering hardware costs. Finally, bandwidth optimization contributes to improvements in remote site backup and disaster recovery strategies.

Making a Case for Deduplication in Midmarket Organizations and Remote Offices

Midmarket companies have several characteristics in common with remote and branch offices of larger organizations:

- Both have on-premises workloads in need of data protection.
- There is often limited IT staff to focus on backup and recovery operations.
- The IT environment may lack a SAN infrastructure.
- Local backup copies are transported daily between local and secondary sites to consolidate backup at a central location, consolidate physical tape creation, or as part of disaster recovery best practices.
- Backup to physical tape creates operational overhead, requiring local IT staff as well as physical transfer of backup copies for disaster recovery.
These characteristics highlight the need for optimization in backup and recovery. Deduplication introduces process improvements and cost efficiency, enabling disk-to-disk backup without increasing costs. By employing a local disk-based backup strategy, backup operations can be streamlined. Disk-based backup reduces operational intervention in backup and recovery processes and facilitates remote management from a central location. This allows IT organizations to reduce or eliminate local staff at branch office locations. Deduplication also lowers bandwidth requirements; physical tape hardware, media, and tape handling at edge sites can be eliminated in lieu of electronic site-to-site transfer of backup data. Aggregating backup copies at a central location enables physical tape media creation at a central location, maximizing the investment in tape hardware, eliminating security risk with tape media chain of custody, and introducing efficiency with operational staff.

**Fujitsu D2D Deduplication Solutions**

Fujitsu’s strength in enterprise-scale environments is well known. Fujitsu has a longstanding market presence with its virtual tape appliance supporting disk and physical tape holistically—for both open systems and mainframe environments. With a recent expansion of its portfolio, Fujitsu delivers the high-value features and benefits of disk-to-disk backup with deduplication, catering to the requirements of midmarket companies and the remote and branch offices of larger organizations.

**Fujitsu ETERNUS CS High End**

Fujitsu ETERNUS CS High End delivers high-end virtual tape technology with deduplication and operates in disk-to-disk or disk-to-disk-to-tape modes. ETERNUS CS High End offers additional advantages for organizations that want to enhance backup processes where physical tape is required for long-term retention.

With its latest revision, ETERNUS CS High End offers a file interface in addition to the VTL interface. This file interface can be used as a target for backup and archiving applications. It can automatically migrate backup or archival data from disk to tape or any other medium connected to its fully-virtualized back end.

With its grid architecture, ETERNUS CS High End provides two key benefits: redundancy and scale. Its high-level of redundancy provides resilience for unexpected interruptions and planned upgrades. Scale is offered in processing nodes: modular x86 server nodes for front-end host and back-end storage connectivity. At the back end, ETERNUS CS High End can support tape, disk, and/or optimized disk. For example, it could support up to ten ETERNUS CS High End deduplication stores with a usable capacity of 1.6 petabytes before deduplication.

Fujitsu ETERNUS CS High End stands out due to its support for mainframe host connectivity (FICON/ESCON) and open systems host connectivity (Fibre Channel) in a single, integrated system and interface. Another distinction is its direct, autonomous control of physical tape libraries: ETERNUS CS High End has full media management capabilities, enabling physical tape creation from virtual tape copies without the intervention of the backup initiator. Furthermore, the virtual tape appliance offers storage tiering, allowing the use of different types of disk storage—high RPM disk for performance versus low RPM disk for cost savings.

The high-end solution is positioned for autonomously managing an optimized media stack of disk, tape, and/or optimized disk. The aforementioned direct-to-tape integration and multi-tier data placement facilitate data lifecycle management. Policy settings dictate automatic movement of backup data through its lifecycle to the appropriate tiers of storage media. Device-to-device replication of data is optionally available to facilitate off-premises data movement. Data can be moved asynchronously or synchronously, which provides high availability to the restore infrastructure. In tape-centric environments, replication is ideally implemented at remote and branch office locations. Remote site backup data can be aggregated at a central location where physical tape devices, tape media, and operational staff are present.

**Fujitsu ETERNUS CS800**

Fujitsu ETERNUS CS800 is a disk-based data protection appliance for midmarket companies that want to replace tape backup and modernize tape backup processes with disk and deduplication. The ETERNUS CS800 supports open systems environments with 1 GbE, 10 GbE, or 8 Gb Fibre Channel host connections via file, VTL, or Symantec
OpenStorage (OST) interfaces. The file interface supports network file system (NFS) or common internet file system (CIFS) mount points. The VTL interface emulates tape to backup applications, while the Symantec NetBackup and Backup Exec backup software sees OST-enabled appliances as disk and enables intelligent capacity management, media server load balancing, reporting, and lifecycle policies.

Fujitsu offers several configuration options of the ETERNUS CS800, with its entry-level system delivering 4 TB of usable capacity and an ingest rate of up to 0.7 TB per hour. With NAS and VTL options, the capacity starts at 16 TB and scales up to 160 TB of usable capacity via 16 TB expansion modules. With reduction ratios of factor 10 to 20, the inline variable block size deduplication engine optimizes capacity to well over one petabyte. Dual parity RAID-6 support enhances data integrity. Performance scales up to 3.6 TB per hour with the VTL option.

When it comes to creating off-site copies, Fujitsu ETERNUS CS800 offers two methods: network-efficient replication and disk-to-physical tape duplication. The ETERNUS CS800 can be configured to asynchronously replicate deduplicated data uni- or bi-directionally between primary and remote sites. Data is encrypted with AES-256 strength cryptography to meet privacy/security requirements. The replication configuration can be one-to-one (disaster recovery scenario) or many-to-one (remote site consolidation scenario).

Direct tape creation can be facilitated in a few ways. When operating as a VTL or in conjunction with Symantec backup products via the OST interface, physical tapes can be created directly by the ETERNUS CS800 while ensuring that the independent backup catalog is aware of the physical tape copies created. For VTL mode, this includes mimicking bar codes of virtual tape with physical tape. This approach eliminates overhead on the backup media server and the SAN. It also ensures that recovery is streamlined when recovering from physical tape media.

The Bigger Truth

End-users transitioning from tape- to disk-based backup strategies have many choices. Fujitsu’s depth of experience in delivering disk target solutions for enterprise-scale environments is notable; Fujitsu is now addressing midmarket customer requirements with solutions that fit many of the segments’ needs:

- Deduplication to optimize disk storage and network bandwidth.
- Several interface choices for seamless integration with existing backup infrastructure and processes.
- Flexibility for creating off-site copies to facilitate disaster recovery or backup consolidation.
- Ease of scale to allow configuration customization to meet capacity and performance requirements today and in the future.

Deduplication in the backup environment can deliver tremendous financial, operational, and business impact. The most obvious is the cost savings associated with reducing the footprint of backup data on disk and over network bandwidth. However, the bolder benefits might be seen in operational areas: considering that staff costs represent nearly 30% of the data protection budget, solutions that automate and streamline operations and reduce or eliminate operator intervention can contribute more to a return on investment calculation. Even more impressive, however, could be the improvement in IT’s ability to meet backup and recovery objectives. The win-win-win scenario is that deduplication could introduce better backup and recovery service levels with a reduction in capital and operational expenses.

---
