White
Paper

Leveraging EMC Deduplication Solutions for Backup, Recovery, and Long-term Information Retention

By Brian Babineau and David A. Chapa

January, 2011

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

Contents

Introduction

ESG started tracking the use of disk in backup processes in 2005. At the time, a small percentage of organizations were considering tape replacement strategies because data growth had elongated backup windows so greatly that information was often left unprotected and vulnerable to corruption, deletion, or disaster. Over the past five years, disk usage within backup and disaster recovery processes has dramatically accelerated as companies continue to fight the same battle: protecting more and more data within the same, if not shorter, periods of time. The performance, reliability, and total cost of ownership of disk as a backup target delivers meaningful benefits to many organizations—so much so that ESG's most recent data protection research indicates that the average amount of on-site secondary data stored on disk now exceeds that of tape. Just a few years ago, tape was the predominant on-site backup media—now, it is the minority.

One of the big reasons disk-based solutions have been able to have such a positive impact on operations and budgets is the introduction and use of data deduplication technology. In short, systems utilizing data deduplication only save unique data—all redundancy at the file, byte, or block level (depending on the solution) is removed. Since there is so much duplicate data in secondary storage environments, users are able save much more logical capacity in a single physical device. This has allowed disk-based backup systems to meet or beat the acquisition and operating expenses—indeed, the total cost—of tape.

Data deduplication is not just altering how backup targets are used; it dramatically affects operating efficiencies, simplifies remote office data protection, and makes disaster recovery significantly more affordable and realistic for a much greater percentage of the overall market. This is driving deduplication solution adoption at a pace very similar to that of disk-based backup solutions: ESG research shows that more and more companies are evaluating deduplication, with 38% of 2010 respondents using a data deduplication solution compared to only 11% in 2008 (see Figure 1).[1]

Figure 1. Data Deduplication Solution Adoption, 2008 vs. 2010

Is your organization currently using any data deduplication solutions? (Percent of respondents)
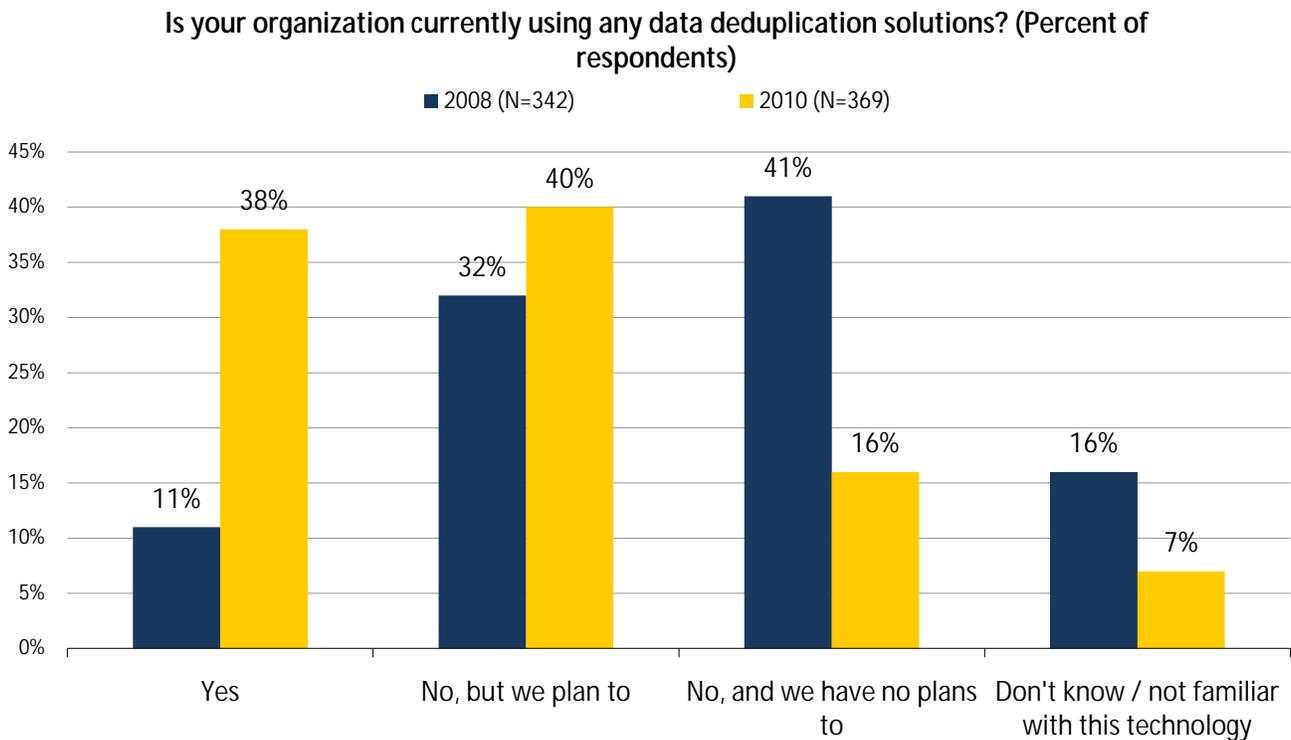
Source: Enterprise Strategy Group, 2010.

---

[1] Source: ESG Research Report, *2010 Data Protection Trends*, April 2010.

The advent of deduplication is not unlike other storage innovations where market leadership was not necessarily determined by a capability, but rather by the true achievable business benefits an entire solution brought about. Twenty years ago, EMC convinced customers that storage was more than a "mainframe peripheral" and has since had a profound impact on the entire IT industry. EMC has continued to stay ahead of the technology curve, recognizing the business value of data deduplication as organizations continue to generate volumes of information. EMC Data Domain systems are now at the forefront of its backup and recovery solutions, helping customers attain higher data protection service levels with the use of disk in backup, disaster recovery, and archive environments.

The Impact of Deduplication

A Better Choice

In the most recent iteration of ESG's annual IT spending intentions survey, senior IT executives cited "managing data growth" as a top technology investment priority for the upcoming 12 months. Data growth is an enduring issue, but the sources of that growth vary from year to year. In 2009, ESG estimated that Microsoft SharePoint capacity was increasing at 25% per year.[2] In mid 2010, growth was seen in primary e-mail systems, with 40% of organizations experiencing greater than 20% growth.[3]

Growth in data usually leads to tradeoffs. IT is forced to keep buying storage—which means other budgeted items go unfunded—and deal with the increased operating costs associated with managing more devices. The only other option is to reduce the amount of data retained, which could impact compliance, recovery service level agreements, and business intelligence initiatives. Data deduplication offers a better alternative by removing redundant content before it is ultimately stored within secondary environments—eliminating most of the negative downstream effects capacity growth would cause.

Gains in capacity savings lead to much more optimistic outcomes, such as the ability to retain more information online for longer periods of time. This can lead to huge benefits, such as moving corporate archives from tape to disk to facilitate compliance and discovery processes. Keeping data on disk longer also expedites recovery times, enabling IT to deliver better data protection service levels without increasing costs.

It should also be noted that deduplication allows many data protection and archive processes to be achieved with fewer devices. By reducing the need to buy additional storage, organizations cut operating expenses associated with energy (power and cooling), labor, and floor space. These savings may provide the biggest incentive to deploy a deduplication solution: 54% of the organizations surveyed by ESG indicated that a reduction in operating costs was the top consideration in justifying a technology investment (see Figure 2).[4]
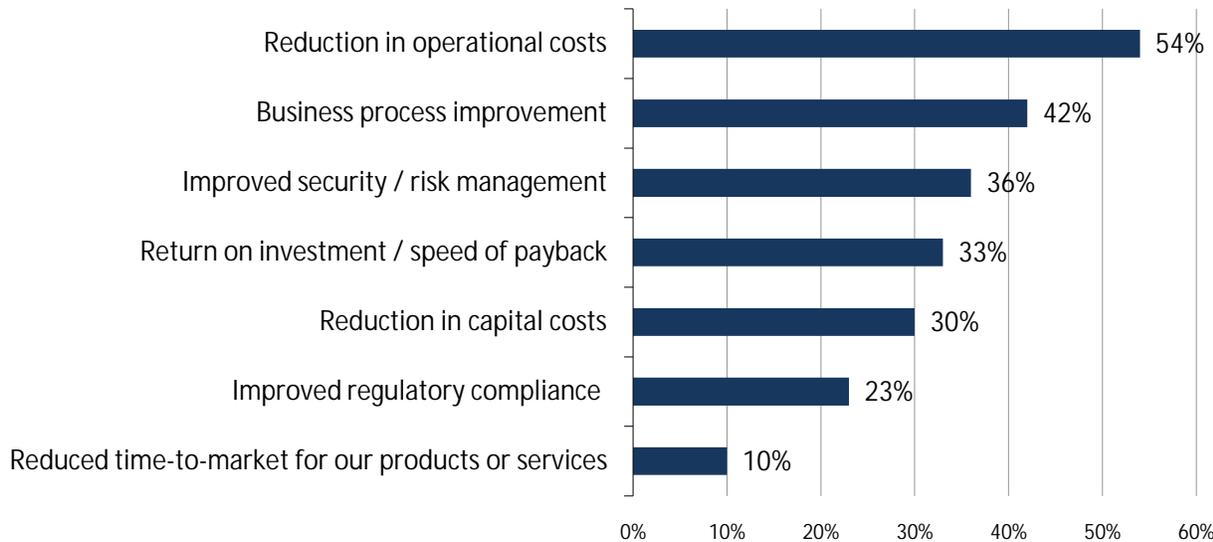
---

[2] Source: ESG Research Report, *Microsoft SharePoint Adoption, Market Drivers, & IT Impact*, March 2009.
[3] Source: ESG Research Report, *E-mail Archiving Market Trends*, May 2010.
[4] Source: ESG Research Report, *2010 IT Spending Intentions Survey*, January 2010.

Figure 2. Important Considerations in Justifying IT Investments Over the Next 12-18 Months

Which of the following considerations do you believe will be most important in justifying IT investments to your organization's business management team over the next 12-18 months? (Percent of respondents, N=515, three responses accepted)

Consideration	Percent
Reduction in operational costs	54%
Business process improvement	42%
Improved security / risk management	36%
Return on investment / speed of payback	33%
Reduction in capital costs	30%
Improved regulatory compliance	23%
Reduced time-to-market for our products or services	10%

Source: Enterprise Strategy Group, 2010.

A New Storage Tier for a New Stage in the Data Lifecycle

Over the last ten years, IT departments witnessed an increase in the reliability of mid-tier storage systems. RAID technology, snapshots, and other features once only found in monolithic enterprise devices became standard capabilities. As modular storage systems improved in availability, vendors introduced less expensive ATA drives into their architectures. IT started to build tiered storage deployments, keeping mission critical applications on traditional fully redundant Fibre Channel storage systems and placing less critical applications on denser ATA-based devices. ESG prefers to categorize "tiering" through the lens of the data itself via the Universal Data Lifecycle. In short, data exists in four stages:

- Stage 1: Dynamic Active Online Data
- Stage 2: Persistent Active Online Data
- Stage 3: Persistent Inactive Online/Nearline Data
- Stage 4: Persistent Inactive Offline/Deep Archive Data

Stage 3 represents the overwhelming majority of data under management in most organizations: data that is unchanging and infrequently accessed. The overall attribute requirements for data in this stage generally include lower-cost, higher scale levels and the highest possible operating efficiencies. Data deduplication is designed to meet these challenges for all data types that exist in a non-changing, infrequently accessed state.

Now, tiered storage deployments with ATA-based devices are commonplace and a new category of storage systems exist that fulfill some, or all, of the requirements for Stage 3 data: systems with deduplication are emerging. In some instances, deduplication storage emulates tape so customers do not have to change their backup processes, while other implementations serve as a target for bulk online or nearline storage. With seamless integration into existing operations, the deduplication system tier is quickly becoming the place to retain all types of persistent, inactive information such as backups, archives, or other infrequently accessed data assets. Although still in the early stages of adoption, ESG believes that storage systems with deduplication have taken the concept of tiered storage—along with the potential savings that such a strategy can generate—to a new level of operational efficiency.

EMC Data Domain Alters Data Protection

"Data protection" incorporates both on-site operational recovery and disaster recovery. Any given organization may use several hardware and software products for these processes, creating complexity and driving up costs associated with the operation and maintenance of many different solutions. EMC Data Domain systems have the potential to change the status quo when it comes to data protection by reducing the number of products needed for operational and disaster recovery.

Disk-Based Backup

There are two primary implementations of Data Domain deduplication systems. The first is the more common approach where a Data Domain system is deployed as a backup target. In this configuration, customers can continue to use their existing backup software as Data Domain presents a standard NAS file system (NFS/CIFS) or virtual tape library (VTL) interface, which allows the system to be connected to a Fibre Channel SAN. Data is deduplicated as it is being copied to the Data Domain system, a process that is referred to as "inline." Customers only have to buy the capacity they need to complete backup operations, avoiding additional storage requirements to hold data for deduplication that occurs after information has been copied (referred to as "post process").

The second configuration option leverages EMC Data Domain Boost software (DD Boost). DD Boost integrates directly with the backup software and distributes parts of the deduplication process to the backup media server. This increases the performance of Data Domain systems, reduces the amount of data transferred over a backup network, and lessens the load on the backup server because it is transferring less data. With Data Domain systems included as part of the backup and restore process, jobs complete faster and there is less risk of media failure as the system is protected using RAID 6 and other data integrity processes. Deduplication minimizes the amount of information actually stored, driving backup infrastructure consolidation. In some situations, customers have used Data Domain's combination of disk-based systems and deduplication to significantly reduce or completely eliminate tape infrastructure. In other scenarios, organizations use tape solely to keep older data longer: for example, when utilizing the "daily, weekly, monthly" backup schedule, monthly copies are moved to tape and stored offsite. If data loss or corruption does occur, IT restores information from the Data Domain system and only goes to tape in a worst-case scenario (if they still use tape at all).

Business Continuity for the Masses

Right now, many organizations have selective disaster recovery processes in place where a few specific applications are replicated to another site using a disk-based solution to facilitate near-instantaneous recovery. Other applications are copied to tapes that are sent offsite, creating a less predictable recovery time objective. The problem is that if some of these "other" applications are unavailable for even just a short period of time, the organization experiences a negative business impact such as revenue loss. In fact, 42% of ESG survey respondents said they could only tolerate three hours or less of downtime for their <u>business</u> critical, not mission critical, applications before such a negative impact occurred.[5]

The reason for the selective disaster recovery approach is cost. Until recently, it was too expensive to replicate data between sites as the capital expenses of two storage systems, replication software, and network bandwidth running between them proved unaffordable for most. Data Domain systems mitigate this challenge with replication software: as information is backed up to a Data Domain system, it can be replicated across a WAN to another Data Domain system in a geographically separate location. Additionally, customers can consolidate disaster recovery operations by having multiple Data Domain systems replicate to one larger Data Domain target device (many-to-one replication) or they can implement a "multi-site" business continuity plan where data is replicated to one site and then on to another (cascaded replication). With the introduction of the EMC Data Domain Global Deduplication Array and its ability to handle up to 14.2 PB of logical storage, customers can extend backup consolidation efforts, with some using this sizeable system as a replication target for up to 180 remote sites.

---

[5] Source: ESG Research Report, *2010 Data Protection Trends*, April 2010.

Regardless of the configuration an EMC Data Domain customer chooses to use, they do not have to worry about exorbitant network bandwidth costs as information is already deduplicated during the backup process and only unique new data is replicated thereafter—this is what allows Data Domain systems to execute disaster recovery over existing WAN resources in contrast to other replication solutions which may require a dedicated network.

Also, by copying only new bytes of backup information, more applications can be added to business continuity plans as the storage and bandwidth expenses associated with replication will not increase in proportion to the information to be protected or restored if a disaster does occur. Organizations can affordably add new applications to their replication schemas, improving recovery time objectives for a majority of business information.

Remote Offices Join the Data Center

Remote office employees' primary responsibilities are, as productivity workers, focused on sales, customer service, and other tasks. Yet, many of them have had to perform their respective jobs and assist with IT functions, including backup. Remote offices often contain messaging applications, file servers, and, of course, desktops that need to be protected. Corporate data center staffs, recognizing the limited choices available for protecting remote data, often put tape in the remote office and either hire IT specialists to manage the media rotation or call upon general employees to insert and remove tapes. The former adds significantly to operating costs and the latter increases the risk that backups do not get done because it's not anyone's primary job. Regardless of who handles the tapes at a remote office, there is always the possibility of human error resulting in incomplete backup jobs and unrecoverable data.

IT departments know the costs and risks of running data protection at remote sites. They need to move information back to a central data center. Data movement during a backup requires bandwidth between the remote sites and the data center. Network capacity is not always available and, if it can be obtained, it is usually expensive.

An ideal remote office data protection solution is one that is minimally invasive while still facilitating low cost data movement. Data Domain systems can address both criteria: a small Data Domain system at the remote office is used as a backup target instead of tape. Backup data is deduplicated and stored locally and can then be replicated to a larger Data Domain system at the corporate data center. Remote office employees do not have to handle tapes and the bandwidth required for data transfer is kept to a minimum because deduplication occurs beforehand. Most importantly, remote office information is protected according to corporate IT policies.

Software Simplification

The Data Domain software portfolio extends beyond replication, further reducing the number of products needed to protect information. Customers can create a point-in-time copy of information within a system via the company's snapshot solution. An initial snapshot does not consume any additional storage capacity and each incremental snapshot is deduplicated to contain only the new bytes added to the system. This extremely cost-effective data protection solution enhances Data Domain replication and remote office offerings.

Data Domain systems integrate with common backup software applications, such as Symantec NetBackup via OpenStorage (OST), enabling a direct connection between two solutions. This integration continues as DD Boost also works with Symantec's OpenStorage capability. The OpenStorage plug-in enables NetBackup to maintain a consistent catalog of information's location within the EMC Data Domain infrastructure so data can be easily recovered from any local or remote copies.

ESG expects EMC to continue its integration efforts to simplify data protection environments. One recent example of this is the announcement of DD Boost support for EMC NetWorker.

Investment Protection Considerations

Maintaining Scale

The primary problem EMC Data Domain is solving—protecting valuable corporate data within an allotted timeframe (the backup window)—is not going to disappear anytime soon. In fact, it is likely to get worse as data growth and retention requirements expand. Disk-based solutions will continue to help address this issue so long as they are affordable when compared to other media in terms of total cost of ownership (TCO). Deduplication changes the economics of disk, making it an affordable backup storage media.

The real value of Data Domain systems lies in their ability to maintain performance while executing data deduplication. The real question becomes one of whether or not these systems continue to do this as the amount of data continues to increase. This is relatively easy for EMC to answer thanks to Data Domain SISL (Stream-Informed Segment Layout) scaling architecture. In short, Data Domain systems identify 99% of the redundant data in RAM before any is written to disk. When data is written, the actual block segments and fingerprints are stored together, facilitating faster access when it needs to be read. As a result, when EMC adds faster processors to its Data Domain system portfolio, the systems can receive and deduplicate more data faster. Put another way, a Data Domain system is not capacity-restricted when it comes to performance. So long as faster processors are available in the marketplace, Data Domain systems will be able to back up and deduplicate more and more data.

Extend the Value of Deduplication

One the best benefits of deduplication is that that the more data that is analyzed, the higher the probability that the data reduction ratio will become more impressive. In short, this means that the more information a customer sends to a Data Domain solution, the more logical capacity that system is likely to store. ESG research indicates 58% of current deduplication solutions are experiencing at least a 10x data reduction ration. Translated, the worst case scenario is 100 TB of information can be stored with 10 TB of capacity. Further extrapolated, a 100 TB system could hold nearly a petabyte of data.

Data Domain systems make it easy for customers to take advantage of this powerful benefit. Systems can be configured to run Retention Lock software, which saves information in a non-erasable, non-rewritable format. This software, which can run on most Data Domain systems, is an ideal target to use in conjunction with several leading archive software solutions including Symantec's Enterprise Vault, CommVault's Data Archiver, and the EMC SourceOne family of products. The combination of archive software and a scalable and affordable system will help companies retain more information online for compliance, electronic discovery, and business reference purposes.

Ongoing performance enhancements to Data Domain systems and their distribution of deduplication processes lead ESG to believe that the solutions may be able to handle almost any Stage 3 workload. Although the systems are optimally designed to write data, it is feasible that a company could run a non-critical, non-performance intensive application or file share on a Data Domain system.
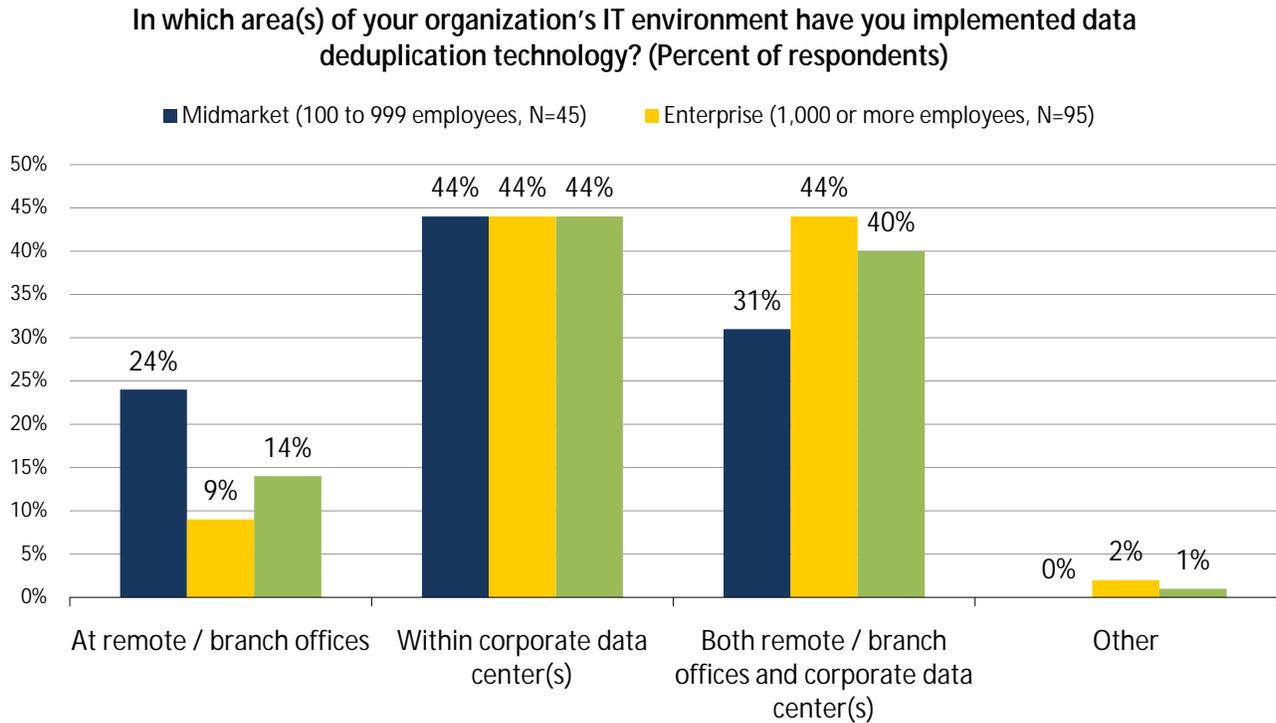
Take Advantage of the EMC Backup and Recovery Solutions Portfolio

ESG has referenced a few examples where Data Domain systems have been—or have the potential to be—integrated with other EMC data protection solutions. This presents the greatest opportunity for EMC on a near term basis as many organizations are interested in turn-key solutions to minimize operational costs. ESG estimates that 58% of data protection expenses are operational (not relating hardware or software) in nature. And there are many assets within EMC's backup and recovery solutions portfolio, including EMC NetWorker and EMC Avamar, that when deployed with an EMC Data Domain solution will solve more than a point data protection problem. As some of this integration matures, customers will move beyond "tape replacement" and "backup consolidation" projects—they will be thinking about designed data protection service levels and meeting them at the lowest possible cost.

The Bigger Truth

Figure 3 highlights the rapid adoption of data deduplication in large and small companies across a variety of locations (data centers, remote offices, etc.).[6] ESG expects usage to expand—companies are always going to create more data and it will always need to be protected.

Figure 3. Data Deduplication Across the Organization

In which area(s) of your organization's IT environment have you implemented data deduplication technology? (Percent of respondents)

■ Midmarket (100 to 999 employees, N=45)   ■ Enterprise (1,000 or more employees, N=95)

Source: Enterprise Strategy Group, 2010.

Aside from adding more hours to the day, deduplication is one of the only ways organizations can deal with the data growth within tight budgets. EMC Data Domain brings data deduplication to the mass market and continues to innovate and integrate this technology to make its adoption even more possible and affordable.

---

[6] Source: ESG Research Report, *2010 Data Protection Trends*, April 2010.