



WHITE PAPER

What Data Domain is Doing to Storage

By Steve Duplessie
With Brian Babineau and Lauren Whitehouse

April, 2009

Table of Contents

Table of Contents	i
Introduction	1
The Impact of Deduplication	2
A Better Choice	2
A New Storage Tier for a New Stage in the Data Lifecycle	2
Data Domain Alters Data Protection	3
Disk-Based Backup	3
Business Continuity for the Masses	3
Remote Offices Join the Data Center	4
Software Simplification	4
How Did Data Domain Do It?	5
Solving the Right Problem	5
How Do Data Domain Keep Doing It?	6
The Next Challenges for a Leader	6
Corporate Archives and Other Stage 3 Data	6
Business Critical Storage?	7
Conclusion	8

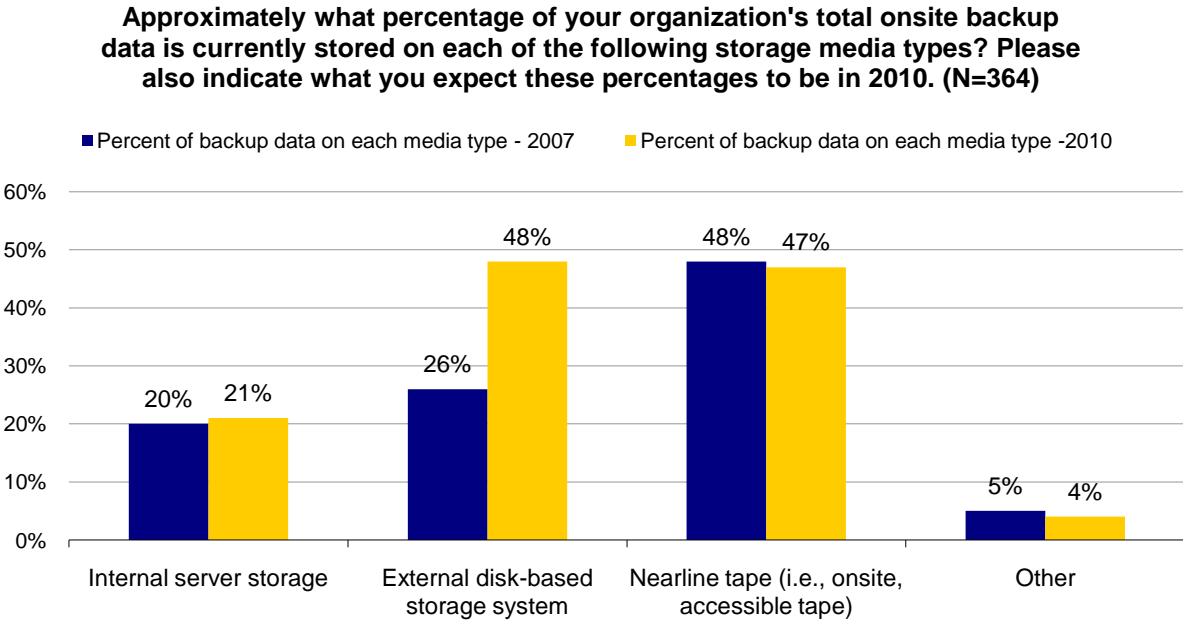
All trademark names are property of their respective companies. Information contained in this publication has been obtained by sources The Enterprise Strategy Group (ESG) considers to be reliable but is not warranted by ESG. This publication may contain opinions of ESG, which are subject to change from 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 Data Domain.

Introduction

In March 2005, ESG research indicated that 18% of the organizations we surveyed had already started to replace tape-based backup systems with disk alternatives. For those who had not done so but were intending to, three-quarters said they would begin within 24 months.¹ If you do the math, over eight in ten customers would have started tape replacement programs before June 2007. The primary driver for these projects is the performance boost that disk systems provide in the backup and restoration process. However, like many other changes in IT, infrastructure modifications occur much faster when there are clear cost savings involved. The same ESG research also stated that two-thirds of companies would accelerate tape replacement programs if disk-based solutions came within 10% of the cost of tape.² One technology, data deduplication, is responsible for driving the cost of disk systems to the same capital and operating cost range as tape. Data Domain pioneered the packaging of data deduplication for disk-based backup with storage system interfaces and software to make the solution consumable.

In theory, with data deduplication solutions, companies backing up a 10 TB file system may only need 1 TB of media. When a customer can reduce its backup capacity by 90%, the savings add up fast—and this is why ESG believes companies will increase the amount of backup data on disk in the next few years (see Figure 1).

FIGURE 1. EXPECTED INCREASE IN ONSITE EXTERNAL DISK SECONDARY CAPACITY BY 2010



Source: ESG Research Report, *Data Protection Market Trends*, January 2008

Data deduplication is not just altering what media companies use as backup targets; 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. Its advent is not unlike other storage innovations where market leadership was not necessarily determined by a technology capability, but rather the true achievable business benefits brought about by the entire solution. Twenty years ago, EMC convinced customers that storage was more than a “mainframe peripheral” and as such has had a profound impact on the entire IT industry. Cisco succeeded by standardizing the routing appliance and later, the IP protocol. Both of those have had hundreds of billions of dollars of global market impact. Data Domain is poised to make a similar

¹ Source: ESG Research Report, *Tape Replacement Realities*, March 2005.
² Ibid.

impact by illuminating a series of expensive problems within storage environments caused by an endless array of duplicate data sprawl. Customers now realize they do not have to keep buying more and more storage capacity as there are more efficient ways to store and manage information—especially in secondary storage environments.

The Impact of Deduplication

A Better Choice

ESG estimates that corporate data will grow at an estimated 25% in 2009 after several years of increases at two to three times that rate.³ When you combine this with flat to decreasing IT budgets, something eventually has to give. Companies will be forced to make a choice. They will have to either keep buying more storage—which means other budgeted items go unfunded—and deal with the increased operating costs associated with managing more devices, such as power, cooling, and data center space or reduce the amount of data retained, which could impact compliance, recovery service level agreements, and business intelligence initiatives. Data deduplication approaches offer IT a hybrid alternative, which is to remove redundant content before it is ultimately stored—eliminating most of the downstream negative effects that capacity would cause.

The gains in capacity savings provide customers with much more optimistic outcomes, such as the ability to retain more “virtual” and true information online for longer periods of time, dramatically lowering the operating impact of supporting that data and enhancing data protection operations through the use of disk. These outcomes can lead to huge downstream financial benefits, such as moving corporate archives from tape to disk to assist corporate counsels in responding to electronic discovery requests. In an August 2008 survey, approximately 60% of U.S.-based trial attorneys reported having cases that raise electronic discovery issues. Of that group, over 86% have issued or received a discovery request for electronically stored information since the new Federal Rules of Civil Procedure went into effect in December 2006.⁴ Corporate counsels need to be able to quickly run searches against centralized online archives in order to facilitate early case preparation and potentially avoid legal expenses as a result of reaching a settlement prior to trial.

There are several other examples where saving more data online can save companies money, but customers should also note that deduplication allows this to be achieved with fewer devices. By reducing the need to buy more and more storage systems, customers save money on power consumption, which 70% of business executives say is how they measure the success of corporate green initiatives.⁵

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

In the middle of this decade, IT departments witnessed an increase in reliability of mid-tier storage systems. RAID technology, snapshots, and other features that were once only found in monolithic 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 our Universal Data Lifecycle.⁶ In short, data exists in four simple lifecycle 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

³Source: ESG Research Report, *Enterprise Storage Survey*, November 2008.

⁴Source: American College of Trial Lawyers, *Interim Report on the Joint Project of The American College of Trial Lawyers Task Force on Discovery and the Institute for the Advancement of the American Legal System*, August 1, 2008.

⁵Source: ESG Research Report, *Global Green Business and IT Initiatives*, March 2008.

⁶Described in detail in the ESG Brief, *A Methodology for Driving Total IT Efficiency: The Universal Data Lifecycle*, June 2008

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 designed to fulfill some, or all, of the requirements for stage 3 data: systems with deduplication is 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 where companies retain all types of persistent, inactive information, such as backups, archives, or other infrequently accessed data assets. Although we are 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.

Data Domain Alters Data Protection

At ESG, the term “data protection” incorporates both onsite operational recovery and disaster recovery. Any given company may use several hardware and software products for these processes. This creates complexity and drives up cost as IT must know how to operate many different solutions. Data Domain provides ways to change the status quo when it comes to data protection by reducing the number of products needed for operational and disaster recovery. In some instances, companies can eliminate tape, as well as their backup software, by using a combination of Data Domain’s data protection hardware and software products.

Disk-Based Backup

The first obvious deployment is to use disk in the backup process. Because Data Domain deduplicates data as it is being stored (a process that is referred to as ‘inline’), customers immediately experience a capacity reduction. This means that they only have to buy the capacity they need to complete backup operations, avoiding additional storage requirements to hold data for deduplication processes that happen after information has been copied (an operation that is referred to as ‘post process’).

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. Symantec’s VERITAS NetBackup customers can leverage Data Domain’s OpenStorage solution to back up data directly to any Data Domain system. With Data Domain as part of the backup and restore process, jobs complete faster and there is less risk of a media failure as the system is protected using RAID6. Further, because backup jobs often involve a mix of incremental and full copies, data reduction rates are substantial. In some situations, IT has significantly reduced or completely eliminated tape infrastructures and in other scenarios, organizations use tape solely to keep older data longer. For example, when customers utilize the ‘daily, weekly, monthly’ backup schedule, the ‘monthly’ copies are moved to tape and stored offsite. If data loss or corruption does occur, IT restores information from Data Domain and only goes to tape in a worst-case scenario (if the customer still uses tape at all).

Business Continuity for the Masses

According to the United States National Weather Service, there were nearly 31,000 severe weather (hurricane, tornado, hail, etc.) storms in 2008 (through the middle of December 2008).⁷ There were far more storms across the globe and many other incidents occurred that would be considered disasters, but many organizations still do not have a reliable business continuity plan for mission-critical data solutions. The reason for the lack of preparedness is cost—until recently, it was expensive to replicate data between sites as the capital expenses of

⁷Source: http://www.spc.noaa.gov/climo/online/monthly/2008_annual_summary.html

two storage systems, replication software, and the network bandwidth running between them proved unaffordable for most companies. Offsite tape became the default disaster recovery plan; IT and the business dealt with its limitations, including long recovery times and periodic failed restores.

Data Domain mitigates the cost and complexities of disaster recovery with its 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 or they can implement a 'multi-site' business continuity plan where data is replicated to one site and then on to a third.

Regardless of how customers use Data Domain for disaster recovery, 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 to execute disaster recovery over existing WAN resources in contrast to other replication solutions which may require a dedicated network between two systems. Also, by copying only the 'new bytes' of backup information, Data Domain allows companies to add more applications 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.

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 handled the tapes at a remote office, there is always the possibility of human error, resulting in backup jobs that are incomplete and leave data unrecoverable.

Centralized IT departments soon figured out the costs and risks of running data protection at remote sites. They needed a way to move the 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.

The ideal remote office data protection solution is one that is minimally invasive for remote office employees and facilitates low cost data movement. Data Domain meets 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

Data Domain's software portfolio extends beyond replication, further reducing the number of products companies need to protect their information. Customers can create a point-in-time copy of information within a Data Domain 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 that were added to the system. This extremely cost-effective data protection solution further enhances Data Domain's replication and remote office offerings.

Data Domain has also made it easier to leverage its solution in Symantec Veritas NetBackup environments through Symantec's OpenStorage Technology (OST) software. The OST software plug-in runs on the NetBackup media server, a primary Data Domain backup target, and a replication target. The software allows Symantec to maintain a consistent catalog of where information resides within the Data Domain infrastructure so it can easily recover data. By maintaining a catalog similar to what is used when NetBackup creates local and

offsite tape copies, Data Domain reduces operational costs. Customers do not have to alter existing processes when introducing disk-based backup and disaster recovery.

How Did Data Domain Do It?

Solving the Right Problem

Data Domain's success follows that of several other technology patterns where a function is integrated into a purpose-built appliance to enable broader market consumption. In Data Domain's case, the function is deduplication. Cisco did the same thing for routing, which was once trapped inside proprietary operating systems. NetApp disaggregated the network file system function from workstations and centralized it on a storage device to facilitate file-based information sharing over the network.

In all of these scenarios, a specific problem became valuable enough to require "purpose built" solutions versus the typical "feature addition" method most often employed initially. If the problem is not valuable enough—where people will spend time and money to solve it—the feature addition approach can be more than adequate. When the problem is ubiquitous—and growing—it will normally require something designed from the ground up to solve. These "second wave" markets tend to be dominated by the first player with the proper solution who can grab the perceived and real market leading position. Further, those leaders tend to remain leaders as long as the current manifestation of the problem is solved by them—that, in essence, is the value of incumbency in a market. Cisco, EMC, NetApp—and even VMware—are all examples of this phenomenon. The problem Data Domain is solving is not of its own creation—it is the longstanding backup window problem caused by ever increasing data growth.

Several factors are common in hyper-growth second-wave markets, and Data Domain had them all.

First, these markets require a legitimate problem, which can be defined as a problem the market is willing to pay to solve. In this case, it was the backup window problem caused by never-ending data growth. Data Domain did not create this problem, but it was in a position to benefit from it. In the same way that VMware originally designed its server virtualization product for software QA and test/development environments, when the world changed and suddenly production servers required the same virtualization technologies for the same issues (utilization, consolidation, capital and operating efficiency, etc.), VMware in the perfect position to take advantage of the opportunity. VMware didn't create the production server market opportunity, but it was the first in line to take advantage of it.

Second, hyper-value markets need to be getting worse, not better. The amount of data being created is never going to decline, only increase, and the dependency on that data never abates. With time being fixed to 24 hour days, the problem (the shrinking backup window) will only get worse, not better. Cisco didn't need to create the market for routing—the fact that users were going to continue to buy more computers and not less made the market real. NetApp didn't need to create the networking market—the workstation market's success guaranteed that. The second-wave opportunities created out of those primary markets have been enormous from network backup, to storage, to switching, to management. In virtually every case, the first company to provide a legitimate sustainable solution to the problem tends to grab the preponderance of value created and also tends to dominate for as long as that market requirement is satisfied.

There are other similar characteristics between these analogies. For example, the function on the appliance had to execute flawlessly so it could scale. Solving any problem in IT invariably creates new, often unforeseen problems. Users are forced to choose between issues. When Cisco first succeeded with its routers, the unforeseen/unplanned problem of having to manage those routers was offset by the value created by utilizing the routers themselves. Those new problems are also second wave market opportunities. There would be no HP Openview if there were no Cisco, no IP networking, etc. Purpose built appliances tend to win when the problem needs to be solved specifically, not as an added feature onto an existing solution.

In all cases, unintended use cases create new opportunities and new problems. Cisco needed to develop a family of routers, each with the ability to optimize for different traffic patterns. NetApp has had to store more and more data while maintaining performance and compatibility and improving manageability over time.

Data Domain is no different. It must keep up with the increasing amount of data being copied and replicated for data protection. The core IP of the Data Domain solution (deduplication) has little to do with the original market opportunity, as simply backing up to disk versus tape can solve that problem. The reason for Data Domain being able to dominate the early stage of the market has been because of the fact that deduplication changes the *economic* realities of using disk based systems to solve the problem. Deduplicating data means users can keep more “virtual” instances on disk, avoid tape, and do so at significantly lower costs.

Now that this function has been proven, Data Domain, like Cisco and NetApp, gets to ride the commodity curve created by others. As disk capacities increase and processing power improves, the company benefits. Its IP—deduplication—simply becomes more effective and more valuable as IT advancements occur. If Data Domain wants to introduce a larger system, it simply needs to add more (or bigger) processors. The ‘in-line’ deduplication function will be able to handle more data because it leverages improved compute resources. Data Domain can scale capacity by supporting newer drives, which are typically bigger (from a capacity standpoint).

How Do Data Domain Keep Doing It?

In order to continue market leadership, every vendor mentioned needed to expand their portfolios, maintain quality, and make their solutions easy to consume. Data Domain systems support multiple interfaces (NAS, VTL, OpenStorage for NetBackup), which enables the solution to fit into most backup environments. It is also easy for customers to solve even more of their data protection problems as Data Domain can be deployed at a central data center, in a branch office, or in a combination of both—further centralizing the entire data protection process across the organization. Customers can become familiar with Data Domain’s solutions in a portion of their environment and then deploy it more widely for incremental benefits. Lastly, Data Domain can accept both backup and archive (information retained for compliance, electronic discovery, or business reference purposes) within the same system. This is a rarity as the two processes utilize different file and metadata formats, but customers do not have to worry as Data Domain has tested its solution with a variety of backup and purpose-built archive applications.

Similar to other market creators and leaders, Data Domain has to build on its advantages and the company appears to be doing so at a rapid pace. For example, through a partnership, Data Domain has added a mainframe interface that can run on any of its systems. There are other areas, including archiving and business critical storage, where ESG believes Data Domain can continue to exploit a function that continues to solve a recurring problem in every company: dealing with the effects of natural and inorganic (data sprawl/replicas) data growth.

The Next Challenges for a Leader

Corporate Archives and Other Stage 3 Data

Today, many organizations use tape media for archives as archive data is unchanging and does not have to be accessed frequently. This norm has changed dramatically—regulators and litigators are targeting archives as sources of evidence to support a wide range of investigations and legal matters. In addition, many companies believe that historical information can be an asset that enhances productivity, so they are choosing to make older information more accessible to employees. The need to keep corporate archives more accessible means that customers have to replace archive tape media with disk, but most companies cannot afford to save large stores of data for multiple years on expensive, primary storage. The advent of purpose-built information archiving solutions allows customers to move data between different storage devices while setting a retention period on the content so it can be managed in accordance with regulatory and legal mandates. But this requires storage

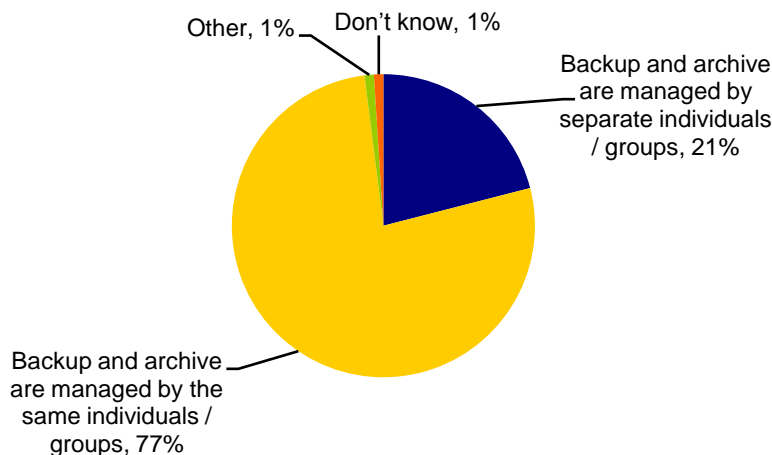
systems to be able to enforce retention requirements, including immutability, which prevents data from being altered or deleted from the device during the retention period.

Data Domain recently added Retention Lock software to its portfolio, which allows any of its systems to store files in a non-erasable, non-rewritable format. This software, which can run on any of Data Domain's systems, is also integrated with several leading archive software solutions including Symantec's Enterprise Vault, CommVault's Data Archiver, and others. Storage systems with data reduction capabilities will pay dividends in archive environments given that ESG expects companies to retain over 100,000 petabytes of data for compliance, electronic discovery, and business reference purposes over the next three years.⁸

Data Domain can also support all stage 3 data types within the same system. This may help further expedite operations as many IT departments rely on the same personnel to manage backup and archive processes (see Figure 2). Additionally, customers may experience greater data reduction ratios by consolidating backup and archive on Data Domain systems because there is a high likelihood that the same bytes sent by the backup application will be sent by the archive software. Simply put: the more data that is sent to a Data Domain system, the more opportunities Data Domain has to remove duplicates.

FIGURE 2. CONVERGENCE OF ARCHIVE AND BACKUP PROCESSES

In your organization, are the data backup process and the information archiving process managed by separate individual(s) or groups, or are both processes managed by the same individual/group? (Percent of respondents, N = 168)



Source: ESG Research Report, *Data Protection Market Trends*, January 2008

Business Critical Storage?

As customers become more familiar with and store more data on Data Domain solutions, the storage savings will take many forms. They will buy less primary storage capacity as older data is archived from primary systems to Data Domain solutions, freeing up valuable primary storage space. They will not have to worry about managing an abundance of tape devices and spending countless dollars on tape media. Network bandwidth costs associated with remote office consolidation and/or disaster recovery will not prevent companies from expanding business continuity projects. Some of these savings, however, may pale in comparison to the sheer reduction in the amount of storage needed to efficiently run data protection and archive operations. ESG research suggests that 58% of polled companies have achieved at least a 10x data reduction since deploying deduplication solutions, including those available from Data Domain.⁹ Over time, this ratio should increase—especially in Data Domain environments as more data is moved into them.

⁸Source: ESG Research, *Database Archiving Surveys*, December 2007.

⁹Source: ESG Research Report, *Data Protection Market Trends*, January 2008.

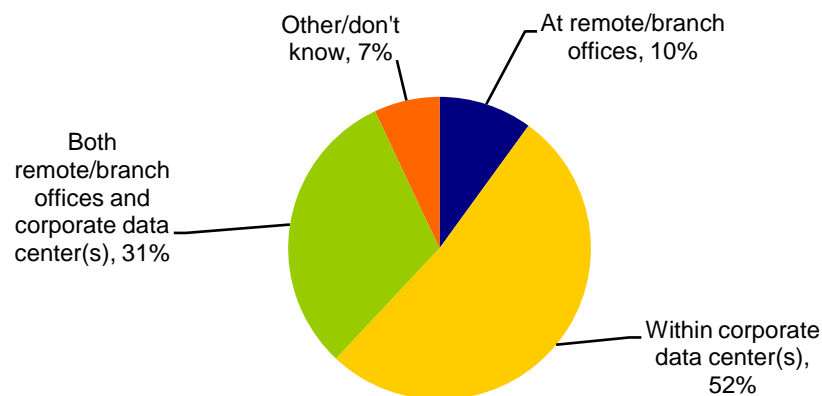
Such dramatic cost savings lead ESG to believe that customers will soon start to demand data reduction technologies in primary storage systems, making this the next possible opportunity for Data Domain. It is not typically recommended that mission-critical, highly-transactional applications be run on dense, slower, ATA-based storage systems. However, general purpose file shares and many other business applications do not need exceptional read/write performance from their underlying storage. If Data Domain's ability to move methodically into the archive market by introducing one software product is any indication, it is feasible to believe that the company could optimize its solution to become a primary, stage 3 storage solution.

Conclusion

Although data deduplication is still a relatively new capability, customers are deploying it in both data centers and remote offices, indicating that the benefits are real (see Figure 3). Some organizations have started with small implementations to measure the capacity savings. Others have witnessed the financial impact and are deploying it as fast as capital budgets allow. For those organizations that have not yet made an investment, it should be a top priority during the next project cycle as it is one where the return on investment (ROI) can be easily rationalized.

FIGURE 3. AREAS WHERE DATA DEDUPLICATION TECHNOLOGIES HAVE BEEN DEPLOYED

In what areas of your organization's IT environment have you implemented data deduplication technology? (Percent of respondents, N=58)



Source: ESG Research Report, Data Protection Market Trends, January 2008

Customers also have plenty of choice; several products include some form of deduplication and many more choices will exist soon. The overall market acceptance of deduplication is now mainstream, which will bring dozens of competitors. As always, ESG recommends that organizations evaluate the options in the marketplace and get vendor references to check on product quality, estimated data reduction rates in certain environments, performance, and overall manageability.

Plenty will argue for the vendor they believe to have the best deduplication algorithms or assert who was the first vendor with data reduction capabilities. ESG does not intend to take sides in this endless competitive banter, but we do believe there are two ways to identify a market leader. The first is revenue—customers ultimately vote with their dollars. Secondly, leaders also help the marketplace change the way business is done—for the better. Data Domain fits this criteria as all of the company's reported \$274M in 2008 revenue came from deduplication solutions and the company has spearheaded entirely new data protection architectures while delivering significant economic benefits. EMC arguably convinced customers to segregate disk storage from mainframe and open system servers, allowing companies to centralize information and improve disk utilization. VERITAS

(which is now part of Symantec) made several disk drives look like one logical entity, also assisting with capacity utilization. NetApp simplified network file systems so that customers could install storage devices on their own, making collaboration easier. There is no question that Data Domain's impact on the storage industry will be referenced with similar accolades.



20 Asylum Street
Milford, MA 01757
Tel: 508-482-0188
Fax: 508-482-0218

www.enterprisestrategygroup.com