

Data Domain: De-Duping Near-Line Data and Enabling T2T Storage

Date: October, 2007

Author: Heidi Biggar, Analyst

Abstract: Data Domain's new release of its operating system—version 4.3—adds near-line support for persistent data. Users can now leverage the Data Domain DD400 and DD500 series systems for more than just backup data—and they can get further efficiency by applying data de-duplication to both data types. In doing so, Data Domain doesn't just improve its efficiency story; it also becomes a key player in tier-to-tier storage (T2T).

Overview

It's a fact: All data starts out dynamic and ends up persistent (or unchanging) at some point—usually within 90 days of its creation. The problem is that this data—and there's tons of it—is sitting on high-end primary disk, and it shouldn't be. Primary disk is expensive to buy, manage, power and cool—using it for persistent data is overkill.

ESG contends that the volumes of persistent data stored on Tier One storage should be moved to an appropriate tier of disk storage. While any number of lower-cost disk systems could theoretically serve as another tier of storage, we recommend that users implement some type of storage system offering enhanced capabilities such as data de-duplication, replication, snapshot, scalability or ease of use functionality. These systems are the foundation of what ESG refers to as an Enhanced Tier of Storage, and are a critical enabler of an efficient tier-to-tier (T2T) storage environment.

For years, Data Domain has provided enhanced capabilities (e.g., data de-duplication, replication, etc.) for managing backup data. Now, the company is extending these capabilities to persistent data. At the heart of the Data Domain solution is its data de-duplication technology, which scans for redundancy at the sub-file level. The ability to scan all near-line data (backup and persistent data) for redundancy makes Data Domain's de-duplication story even more compelling. By applying de-duplication to persistent data, Data Domain provides another level of efficiency for organizations. Consider the following: A user backs up DOCUMENT_A to the DDX and also moves it off its tier one disk to the same Data Domain system. The system will keep only one copy of the data (instead of two) on disk. This simple example illustrates the benefits of performing "cross de-duplication" (of two entirely different applications) from a single system.

While a number of disk backup vendors have hinted at plans of extending their support of near-line data types to include persistent data, Data Domain has turned words into an actual shipping product. In doing so, they have put another stake in the ground. Data Domain, which completed an IPO this summer, currently boasts more than 1,000 customers and is the current leader in D2D backup data de-duplication adoption.

Benefits of Data Domain's Strategy

Using Data Domain disk as a target or repository for both backup and persistent data has several significant benefits:

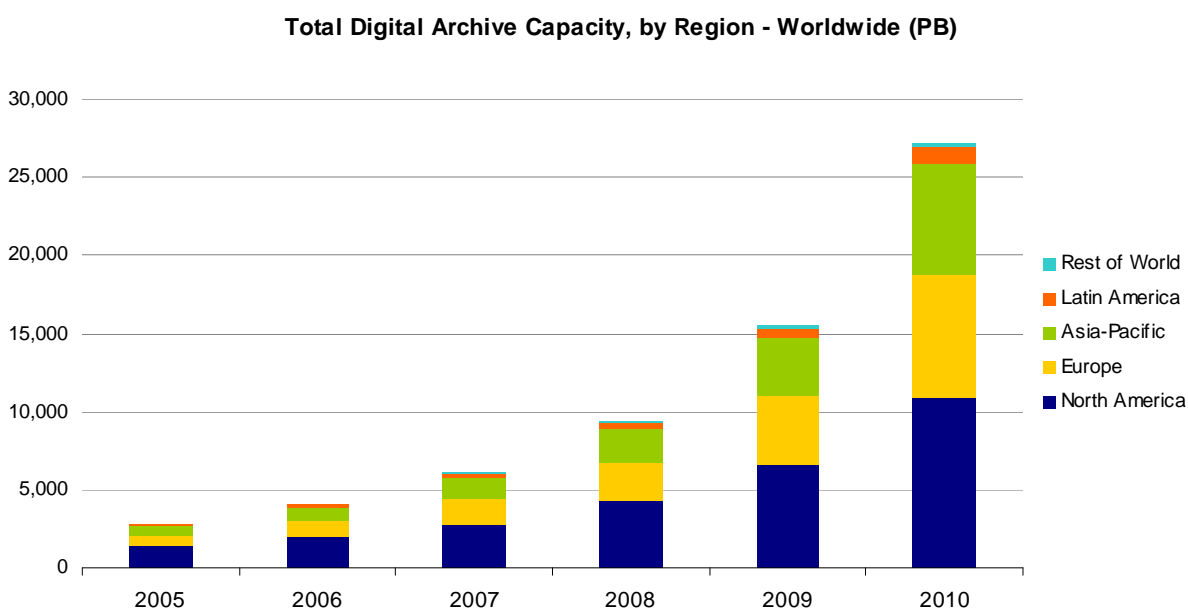
- It facilitates moving data from primary storage systems onto more appropriate storage tiers. This means:
 - Less hardware to buy and less "stuff" to manage. Data de-duplication is applied to backup and persistent data—eliminating redundancy between the two types of near-line data.
 - Significant savings in power and cooling consumption as well as floor space. Backup and persistent data is consolidated onto one efficient storage tier. Energy efficiency is quickly

becoming one of the biggest issues in data centers today, but for some reason, Data Domain doesn't highlight this benefit often. ESG views it as one of the company's most important value propositions.

- Also, because both types of data are housed on the same (tier two) disk, search and indexing should be streamlined in eDiscovery, regulatory, DR, etc., situations.
- It eliminates the confusing "Is it archive" or "Is it backup" line of questioning. It also puts customers in a better position to meet backup and recovery objectives, especially if they're not already backing up to disk.
- End-users get the advantage of combining functionality with field-proven Data Domain data de-duplication and remote replication technology in a single system.

According to ESG Research¹ digital archive capacity (persistent data) will exceed 26,000 PB over the next few years (see Figure 1). Clearly, this much data cannot continue to be stored on traditional storage systems. It requires enhanced tiered storage that provides users with capacity efficient, scalable, reliable, intelligent and easy to manage platforms for the long-term retention of this type of data.

FIGURE 1. DIGITAL ARCHIVE GROWTH



A Closer Look at Data Domain's OS 4.3

The Data Domain OS 4.3 includes a new version of its directory manager, which supports several orders of magnitude more files, de-duplicated snapshots and de-duplicated snapshot replication. Snapshot and snapshot replication provide users with more granular backup and DR options. Today, multi-directional cross-site replication, which is performed via in-line de-duplication, is a key differentiating feature of the Data Domain platform.

¹ ESG Research: *Digital Archiving: Market Trends and Forecast 2006-2010*, January 2006.

Data Domain has also added snapshots, which ESG believes is critical to a T2T environment. The company uses a logical snapshot methodology, which requires very little overhead and is capacity efficient. Combine logical snapshots with data de-duplication and you've got an "uber"-efficient solution—a next-generation snapshot, if you will.

As for the directory manager, it is the part of Data Domain's file system that manages the directories (namespace, etc.) on the platform. It is what keeps track of the files that are written to the disk back-end. The new version of the directory manager is able to keep track of both backup and persistent data in the same system at the same time and find duplicates across data sets. Previous iterations supported backup only. To bring persistent data onto its platform, Data Domain had to tune its file system so that it could handle normal-sized files (e.g., persistent or archive) in addition to large backup files. Data Domain says there is now no architectural limitation to the number of files its OS can support, and they have tested up to 100 million on a DD580 class system. One customer using the refreshed system stored ten million files on the platform during the first month.

The Data Domain OS upgrade is available to existing and future Data Domain DD400 and DD500 Series users, on support, at no charge. Data Domain currently supports third-party archiving applications from Arkivio, Atempo, CommVault, EMC and Symantec. ESG also believes that an integrated and robust search capability will be valuable to Data Domain, especially as the size of its systems increase. The ability to perform detailed searches quickly and completely in an eDiscovery, regulatory or DR situation will be essential. Data Domain has also recently certified its disk system with Index Engines.

Data Domain still needs to go further up the food chain in terms of the scalability of its platform—with the goal of supporting 100s of TBs to PBs of capacity. With its data de-duplication technology, it can support a ton of data in a single system. At 20x data reduction, a single DD580 can provide over 400 TB of capacity and with a 16-controller DDX, it can support more than 7 PB, but ESG knows of environments with requirements 10 times that. We're confident that Data Domain will get there, but they are not there yet. Data Domain also has to be more vocal about their impact on power, cooling and floor space efficiency. It should be one of their main talking points. Finally, they need to do more educating about how they integrate with virtual machines.

The Bottom Line

This announcement is a big deal for Data Domain and end-users on a number of fronts:

- From an end-user perspective, it shows that Data Domain is committed to going beyond backup. This is about taking the huge quantities of data that doesn't belong on tier-one (or primary) storage and moving it to an enhanced tier of secondary storage. This tier is optimized for backup, disaster recovery and persistent data (i.e., how it is managed, stored and accessed)—allowing for more efficient use of disk capacity on both tiers.
- From a market trends perspective, this release is further evidence of what ESG has always believed: Data de-duplication should exist at multiple tiers. By providing this solution, Data Domain proves an understanding of its value, remaining ahead of the curve by extending near-line support to persistent data.
- This allows Data Domain to broaden its market and give customers the opportunity to extract even further value from their Data Domain systems. ESG is a strong proponent of consolidating storage and believes that a single system capable of being used for D2D backup and persistent data (as an enhanced tier of storage) in a T2T fashion is a good move.

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.