

White Paper

EMC Data Domain Global Deduplication Array (GDA)

Delivering Massive Backup Consolidation

By Brian Babineau and David A. Chapa

August, 2011

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

Contents

Introduction	3
Understanding the Global Deduplication Array	3
The Basics	3
The Value of DD Boost Software	4
A New Perspective on Scalability.....	4
Network Efficient Replication.....	4
Centralized Management	5
Consolidate for the Future	5
Mitigating Risk	5
Investment Protection.....	5
Flexibility for the Future	6
The Global Deduplication Array in Action - One Customer’s Experience.....	6
Overview.....	6
Selection a Solution	6
Justifying the Investment	6
Plans for the Data Domain Global Deduplication Array.....	7
The Bigger Truth	7

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.

Introduction

Every senior IT executive is interested in undertaking consolidation projects that provide immediate opportunities to reduce capital and operating expenses. The math is simple: costs decrease when there are fewer systems to buy and manage. Achieving such benefits has been relatively straightforward in certain areas of IT infrastructure such as the server domain, where the combination of virtualization and greater computing power has enabled rampant consolidation. However, in other areas—particularly secondary storage, the home of backup and disaster recovery (DR) operations—massive data growth and distributed operations make significant consolidation a challenge.

Backup consolidation efforts usually include many different projects such as: centralizing remote office backups to minimize risk of data assets being out of corporate control and reducing and standardizing the number of backup storage systems. Any of these projects can run into the same constraints: performance (completing backup and replication jobs in the allotted window), cost (capital outlays for software and systems), and management complexity (configuring and monitoring data protection policies). What's worse, these forces often offset one another. An organization may choose to leverage disk to speed up backups, but this typically requires a higher upfront investment than tape, which used to be the backup medium of choice. To achieve cost parity or a reduction compared to tape, companies may turn to disk systems with deduplication. Unfortunately, some deduplication-enabled systems have performance and capacity limits, forcing deployment of multiple devices to meet backup objectives. Organizations are then left with additional management complexities as they monitor, tune, and service several devices.

[EMC's Data Domain](#) Global Deduplication Array (GDA) can help customers avoid some of the aforementioned consolidation challenges thanks to its massive scalability in terms of both throughput (how *fast* data can be stored) and capacity (how *much* data can be stored); the latter is enabled by Data Domain's well-known data deduplication capabilities. This paper will discuss how the GDA enables several levels of consolidation from local backup to multi-site DR. For further substantiation, this paper also includes the experiences of a large US-based health care provider that selected the GDA over an exhaustive tape infrastructure upgrade. This user testimonial from a senior IT executive exemplifies the cost reduction and consolidation potential of GDA. Specifically, the organization was able to reduce the amount of backup systems required and eliminate ad-hoc tape media purchases previously needed to accommodate data growth. In short, ESG's discussion with the health care provider shows that EMC's latest Data Domain system is more than a backup solution—it changes the economics of backup consolidation.

Understanding the Global Deduplication Array

The Basics

The EMC Data Domain Global Deduplication Array presents a backup application with a single logical storage pool that can scale up to 768 terabytes (TB) of raw storage capacity. However, this number represents only a fraction of the data that this system can protect when one takes into account the effects of its deduplication capabilities. Specifically, the maximum logical capacity (which incorporates a deduplication factor) of the GDA can reach over 25 petabytes (PB). This massive scalability is enabled by the fact that data reduction ratios typically increase as more data is sent to a single system (the larger the data set, the more likely duplicates will be found). In fact, in ESG's most recent data protection research study, over 50% of respondents currently using deduplication solutions were achieving 10-20x data reduction ratios.¹ Probably the most significant benefit is that the GDA stores all of this data, no matter how you measure it, within two standard data center racks.

The GDA consists of two Data Domain DD890 controllers delivering an aggregate throughput of up to 26.3 TB per hour when implemented with the EMC Data Domain Boost software option and up to 10.7 TB when connected to a backup application via a virtual tape library (VTL) interface. This performance, combined with its storage density, enables customers to back up massive amounts of data in a short period of time—such as 200 TB in an eight-hour backup window—to a single system.

¹ Source: ESG Research Report, [2010 Data Protection Trends](#), April 2010.

Customers can purchase an entry-level configuration of 64 TB split across two controllers and seamlessly add capacity (up to the 768 TB limit of the combined systems).

The GDA supports all leading backup application and can connect via VTL or DD Boost interfaces. This enables customers to seamlessly introduce GDA into existing backup environments, allowing for consolidation without major software configuration alterations.

The Value of DD Boost Software

The key to GDA's industry leading performance is Data Domain Boost. DD Boost is licensable software available across the Data Domain product line that enables advanced integration with backup applications (current GDA support is for EMC NetWorker and Symantec NetBackup and Backup Exec). The software serves two functions: first, it enables the backup application to connect to the GDA unified storage pool over existing networks without passing through any specialized interfaces such as a VTL connection. Second, it distributes parts of the data deduplication process to the backup server or application client rather than having the Data Domain system handle the entire workload.

Distributing deduplication processing between the backup server or application client and the GDA improves performance by up to 50%. In addition, less bandwidth is required because less data is sent over the network from the backup servers to the GDA. Although DD Boost consumes some processing power and RAM on the backup server, the overall load on the server is actually reduced due to the reduction in resources required to send data to the GDA since it is moving less data overall.

A New Perspective on Scalability

The GDA offers extreme scalability without adding complexity—a textbook definition of what a consolidation solution should accomplish. This scalability is especially important for many large enterprises with complicated backup processes designed to protect many clients (“nodes”), often in the thousands. The only way these companies can complete backups within constrained windows is by running multiple backup jobs simultaneously, which usually involves numerous backup servers and backup storage systems. While some companies may not reduce the number of backup servers they maintain, GDA will enable them to consolidate to one backup storage system and a single backup application.

The GDA can also be a replication target for other Data Domain deduplication storage systems, which is especially useful for consolidating remote site backups. The GDA can serve as a replication target for up to 270 streams from other devices. This enables a fan-in ratio of 270 remote Data Domain systems to a single GDA or a combination of remote site and local backup streams adding up to 270.

Bi-directional replication is supported, as are any number of cross-site protection scenarios. For example, data centers in three cities can have a GDA serving as a local backup target and then replicate to one another in a circular pattern: site one replicating to site two, site two replicating to site three, and site three replicating to site one. This enables a streamlined multi-site disaster recovery strategy along with consolidated local backups.

Network Efficient Replication

Broadening DR to remote locations via the GDA is also likely to raise the question: Will I have enough bandwidth to completely replicate all of this data? This is an easy challenge to address with any Data Domain system, including the GDA, thanks to network-efficient replication. When any Data Domain system replicates to another, only unique data is sent, which substantially reduces the amount of network bandwidth needed to complete the operation by up to 99%. Even though a GDA is likely to store large data sets given its throughput and scale, it's likely that it won't have to continuously send much data to the remote site to satisfy DR requirements. Replication with Data Domain systems offers significantly improved automation and speed compared to what is done today: a backup application creates a full copy to a local disk or tape device and a subsequent full copy is made and sent offsite. In addition, Data Domain Replicator software offers flexibility to meet multiple use cases including system-to-system, many-to-one, one-to-many, bi-directional, or cascaded replication approaches.

DD Boost also enables NetWorker, NetBackup, and Backup Exec users deploying GDA to manage the replication process from the backup application. Users can establish an “offsite copy” command within the backup software applications and execute it via Data Domain Replicator software as opposed to routing data from the GDA through the backup server, eliminating 50% of the burden on the network. The backup application catalog tracks both the onsite and offsite copies to ensure that IT can execute a restore through the backup application from either copy if necessary.

Centralized Management

After connecting the backup servers to the GDA via DD Boost or VTL, users can immediately start backing up data to the system. Backup and retention policies are mapped to a single storage pool (as opposed to the individual controllers) and this single storage pool serves all backup and replication policies.

The GDA is managed through the EMC Data Domain Enterprise Manager, and those that deploy the GDA and other Data Domain systems can manage up to 20 devices from a single console. Since the GDA is made up of two DD890 controllers, they both appear in the DD Enterprise Manager. One controller is displayed as the master and shows the combined metrics for both systems including utilization rates, capacity and performance metrics, deduplication ratios, configuration details, and any system alerts. The other controller is displayed as a worker and primarily shows its own diagnostics and component health information.

DD Enterprise Manager can also be used to configure and monitor all replication activities. Users can see if a system is a source or a destination, how much data is being sent or received, and if operations complete successfully.

While customers are likely to be drawn to the GDA’s speed and scalability, the ability to quickly deploy this centrally managed solution into production may be the ultimate reason it becomes the centerpiece for backup consolidation. ESG research indicates that ease of implementation ranks as one of the most popular selection considerations for a disk-based deduplication solution.

Consolidate for the Future

Mitigating Risk

Consolidation does mean putting all your backup eggs in one basket, so customers need to be assured that if part of the basket breaks (which is a possibility with any device), the system has safeguards. Like all Data Domain systems, the GDA has the Data Domain Data Invulnerability Architecture built in, which executes continuous recovery verification, fault detection, and self-healing to keep data protected during the initial backup and throughout the data lifecycle.

Several additional features make the GDA ultra-safe, ensuring maximum uptime and reliable recovery. In addition to redundant, easy-to-replace fans and power supplies, GDA is configured with RAID 6, so even a double disk failure won’t bring the system down. Lastly, non-volatile RAM ensures a fast reboot if needed.

Investment Protection

Once the GDA is installed, expansion capacity can be added non-disruptively online. This is critical for customers experiencing rapid growth in data that must be added to the data protection schema or those who simply want to extend retention policies. Regardless of how backup and disaster recovery capacities increase—it is pretty much a given that they will—GDA enables those needs to be met without having to add new backup storage systems. This also removes the need to set up new backup servers or migrate backup policies to handle the incremental data. In addition, customers can be wary the impact upgrades will have on data; EMC addresses this concern by enabling data in place upgrades to the latest controller. This is a true in place upgrade, which is made possible because no data resides on the controllers and therefore controllers can be upgraded without compromising data stored on the system.

Flexibility for the Future

Unlike other inline deduplication systems, the GDA was designed to balance and maximize performance and scalability, facilitating multiple levels of backup consolidation. A GDA can be deployed in any location and serve as a replication source or destination as well as a backup device; fast, efficient, reliable replication occurs without extra bandwidth. Organizations can also easily implement any number of multi-site DR configurations.

The GDA in Action - One Customer's Experience

Overview

In an effort to validate the direct user value of the Data Domain Global Deduplication Array, ESG interviewed a senior IT executive at a large health care provider based in the southeastern United States. In serving over 350,000 patients per year, the capacity required to protect medical records continues to increase. This growth is exacerbated by retention regulations (retention periods vary by age of patient, type of study, etc.) and DR requirements outlined in HIPAA. When the IT department of this 24,000 employee provider calculated its data growth over 2009 and 2010, the final figure was 85%.

This annual growth rate was a critical calculation in planning and budgeting for a backup infrastructure upgrade in 2010. Tape was the incumbent backup media for essential medical records applications being used for both local and offsite disaster recovery copies. During prior years, the organization continually had to buy extra media to keep up with data growth. In addition, continued data growth was applying pressure to the backup windows and, subsequently, so is the production window for end-users including the 2,000-plus medical staff responsible for providing patient care.

Selecting a Solution

Having budgeted for the upgrade, the team set out to explore the options available to keep pace with growing backup demands. It came down to two options: adding more tape drives, tape library frame expansions, and additional tape media or implementing a disk-based backup solution. To help determine what would best address its requirements, the IT team outlined several goals and initiatives it needed to accomplish in its request for proposal (RFP), including:

- Regaining control over the increasing backup window
- Reducing management costs
- Cost savings through more efficiencies
- Encryption to meet industry requirements

During its research, the IT team also considered a few other factors to help narrow the focus for a go-forward backup solution—specifically, IT staff and budget would remain flat even in the face of massive data growth, the existing 1.4 PB of backup data currently under management on tape, and the ongoing costs of a tape-only solution as well as the current vendor's inability to support encryption on the tape drive. The customer turned to EMC to present a solution to meet these needs. Already a trusted primary storage solution provider, EMC recommended an architecture that would replace the tape libraries with a local GDA that would replicate to another Data Domain system at the DR site. As a long-time NetBackup user, the hospital would be able to leverage DD Boost software to help improve backup times and offer efficiencies between the backup servers and the GDA to reduce the amount of data sent across the network.

Justifying the Investment

After a review of the technology, the customer decided to test the proposed Data Domain solution to see if it would meet expectations. One of the key benefits for this customer was the ability provide a single, very large backup storage pool to simplify backup administration and management.

The hard dollar investment justification came when the organization compared the upfront investment of the GDA with the cost of the tape upgrade, ongoing media expenses, and management/administration of the tape environment. The tape infrastructure upgrade, plus management costs, would have far exceeded that of the proposed GDA.

Also supporting the decision was the fact that the tape solution being evaluated did not support encryption, a necessary requirement in the health care industry given privacy regulations. The Data Domain GDA met all security requirements outlined in the original request for proposal.

Plans for the Data Domain GDA

The organization has been using this system for less than a year, but has already experienced immediate benefits that include operational efficiencies through the use of DD Boost software and the reduction of ongoing tape media costs by going nearly “tapeless.” The hospital plans to move even more backup jobs to the GDA and subsequently replicate data to an offsite DR location where an additional Data Domain system resides, further reducing dependence on tape. The additional replication portion of the project is likely to include adding capacity to the remote Data Domain system as more data is copied to it. An additional key benefit of the new infrastructure is the ability to have all data online, ready, and accessible in the event that a major recovery effort is required, providing a level of resiliency the organization didn’t have previously with a tape-only solution.

The Bigger Truth

Whether deploying an enormous robotics-enabled tape library or taking advantage of the latest deduplication innovations, backup consolidation is not new to IT. The issue has always been that no one solution could enhance protection without compromising something else. For example, before deduplication, when disk was first adopted in the backup process, IT still utilized tape for DR because replication wasn’t feasible for most organizations. And, despite its popularity, deduplication is not ubiquitous—customers still face the management issues that arise from dealing with cumbersome interfaces, lack of scalability, and poor replication feature sets. In short, no matter how companies have tried to consolidate, there always seem to be a trade-off.

The GDA facilitates consolidation with an architecture that can perpetually address constant data growth. It optimizes the data deduplication process throughout the backup infrastructure and allows multiple physical backup controllers to appear as a single system to be addressed (by the backup application) and managed (by IT) as one. By only sending unique data over the network, it also replicates data extremely efficiently. With its massive performance and scalability, customers can choose where and how they want to consolidate backup processes and then expand these efforts into other workloads as warranted. The benefits experienced by the aforementioned health care organization clearly show the value the GDA can bring to an existing backup environment by presenting a single disk-based backup system and reducing management and administration overhead.

Like any backup system, GDA cannot do everything on its own—it needs backup software to initiate and manage the backup process. The interaction between the backup software and GDA is enhanced with advanced integration via DD Boost software. To get the maximum performance throughput, users must also upgrade the backup network, or the backup server connection from the GDA itself, to 10GbE. However, after some stalling due to budget constraints, most organizations are already commencing network upgrade projects. As such, companies are running out of reasons not to take the next step in backup consolidation—the GDA eliminates the majority of tradeoffs that have hampered these projects in the past.



Enterprise Strategy Group | **Getting to the bigger truth.**