How deduplication helps reduce the cost of backup and **disaster recovery**

By Kay Benaroch and Shane Jackson

Dell and EMC have teamed up to deliver efficient disk-based backup and recovery with integrated deduplication technology designed to make disk-based backups as cost-effective as tape-based backups.

Deduplication pays off

Dell/EMC DD Series appliances incorporate deduplication and other key technologies to help meet storage and backup challenges.

- **Up to 99 percent** bandwidth reduction for replication
- **90–97 percent** data reduction for backups
- **Leading defense** against data integrity issues with the Data Domain Data Invulnerability Architecture

Today's explosive data growth is prompting many organizations to look for ways to increase the efficiency of their storage, backup, and disaster recovery processes. The growth of data is straining capacity, but at the same time, cost and complexity frustrate many IT managers seeking to make improvements. For example, tape-based approaches that worked well when originally implemented can prove to be slow, expensive, and unreliable as the organization grows and as tapes must be physically transported from remote sites to a central storage location. Disk-based backup is inherently faster and more reliable than tape, but has traditionally been more costly, and large disk arrays can be difficult to manage.

Dell and EMC are addressing these challenges by leveraging the data reduction and cost-saving advantages of deduplication storage systems from Data Domain, which was purchased by EMC in 2009. Rather than adding deduplication to an existing storage platform, Dell and EMC have introduced the Dell/EMC DD Series—purpose-built appliances that incorporate disk-based storage and deduplication. These appliances are designed to tame data growth as effectively as possible, helping organizations to reduce complexity and costs.

Overcoming challenges to effective data protection

Cost pressures can keep many organizations from making needed storage and backup improvements to cover the variety of operating systems, applications, and geographic locations that produce critical

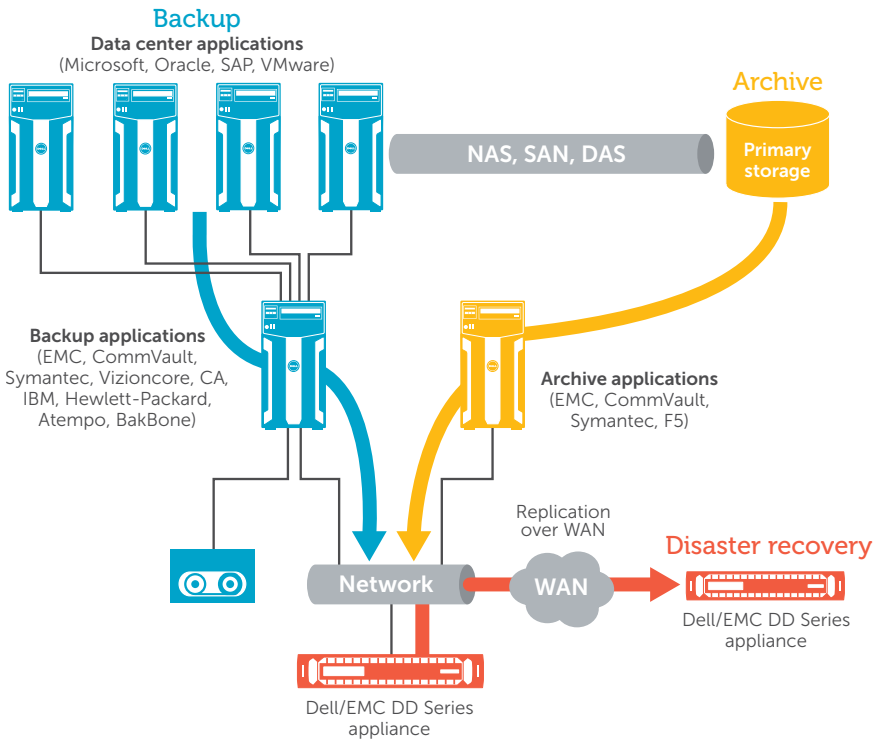

Figure 1. Dell/EMC DD Series infrastructure and ecosystem

information. IT managers may be forced to deploy systems piecemeal, requiring additional management by IT departments that are already stretched, or they may have resources for headquarters or central sites but lack the budget to create a comprehensive plan that encompasses all types of servers and applications at local and remote sites.

Organizations may also be exposed to risk by doing without essential protection such as off-site disaster recovery. Off-site processes for disaster recovery have historically induced latency, prompting many organizations to simply create local copies to speed local recovery—an approach that exposes the organization to data loss in the event of accidental deletion or site disasters. In addition, the cost of bandwidth to replicate the growing volume of critical information over a wide area network (WAN) is often prohibitive, especially for remote or branch offices.

Complexity in the backup environment is another factor that can delay storage and backup improvements. Many organizations have multiple backup software instances or heterogeneous environments. This complexity can increase operational and procedural errors, leading to unprotected, corrupted, or lost backup data. As a result, IT managers may not be able to meet business needs for data restoration and recovery.

To address these challenges without requiring a costly redesign of backup procedures, IT managers are exploring ways to remove duplicate data from the backup process, reducing data volume to the point that they can use fewer hard drives than they had been. Finding an approach that works with a flexible range of backup and archive software is also desirable. Dell and EMC are working collaboratively to create storage and backup systems that include integrated deduplication functionality.

Examining technical requirements

Building a storage system that delivers the full potential of deduplication requires a flexible approach that addresses variable-length duplicates, multiple formats and protocols, and other technical considerations.

Variable-length duplicates. Conventional storage systems, whether network attached storage (NAS) or storage area networks (SANs), store data in fixed-size blocks. Some deduplication systems simply look for identical fixed-size data blocks, an approach that is inadequate for maximizing deduplication results; to be effective, a system must be able to identify variable-length segments as well. This capability is necessary to maximize the amount of redundant data that can be found in incoming data blocks, regardless of small changes in those blocks compared with previous backups.

Format parsing. Data comes in many formats generated by many different applications, and the same duplicate data is often embedded in those different formats. The sheer number of these formats and the speed at which they change make it impractical for a storage vendor to support them all. Parsing the formats requires substantial overhead. A

Enhancing logical capacity

Comparing physical capacity to logical capacity illustrates the value of integrated deduplication in Dell/EMC DD Series appliances. The physical capacity is the amount of raw storage provided by the disks in a particular system. Some of this storage is consumed by appliance operations such as RAID and spares; subtracting this amount from the physical capacity yields the usable capacity, to which the expected deduplication ratio is then applied. Multiplying usable capacity by the deduplication ratio yields the expected logical capacity—the amount of backup data an administrator can keep on the system after deduplication, depending on change rate, retention policies, and other factors.

For example, 1.5 TB of physical capacity in the Dell/EMC DD140 appliance ultimately provides up to 17.0 TB of logical capacity with a typical 20:1 deduplication ratio (see Figure A). That increase can help deliver substantial additional terabytes of logical backup capacity to the organization, allowing administrators to increase the scope of information backed up for replication and disaster recovery purposes.

	Physical capacity	Logical capacity	Maximum throughput
Dell/EMC DD140	1.5 TB	Up to 17.0 TB	Up to 450 GB/hour
Dell/EMC DD610	Up to 6.0 TB	Up to 75.0 TB	Up to 675 GB/hour
Dell/EMC DD630	Up to 12.0 TB	Up to 165.0 TB	Up to 1.1 TB/hour

Figure A. Capacities and throughputs for Dell/EMC DD Series appliances

Calculating the benefits

Based on a few simple questions, the Data Domain Deduplication Calculator can help organizations evaluate how deduplication could help reduce storage and bandwidth requirements in their specific environments.

dedupecalculator.com

storage-based deduplication engine should be data agnostic, and find and remove duplicates in data no matter how it is packaged or stored in the system.

Multiple protocols. Many standard access protocols are used in storage systems today, from Common Internet File System (CIFS) and Network File System (NFS) to block-based and virtual tape library (VTL) access methods. For example, user directories may be in NFS, the Microsoft® Exchange server may need to run in data blocks, and backups may require VTL. To be efficient, a backup storage system should support all of these protocols, and the deduplication approach should be able to remove redundant data no matter how it is stored.

Processor-centric versus disk-intensive algorithms. Over the last two decades, processor performance has increased dramatically compared with disk performance. Today, processor performance takes another leap with every doubling of the number of cores in a chip. Algorithms developed today for deduplication should take advantage of the growth in processor performance instead of being tied to disk performance.

Deduplicated replication. When it comes to disaster recovery, true data protection requires storing a copy of the data safely at a remote location. Replication has long been used for the relatively small volume of mission-critical or high-value data in an organization, but many organizations find that the cost of replication can be too high for the remainder of their information.

Backup storage systems designed with integrated deduplication and replication can help reduce the bandwidth required to replicate large quantities of data from one site to another, thus enhancing the cost-effectiveness of the replication process. However, not all deduplication systems can replicate, and even those that can have wide gaps in capabilities. The system must be implemented in a way that runs fast enough for deduplication and replication with low overhead across a

comprehensive range of topologies and infrastructure typically found in the distributed enterprise.

Combining disk-based backup with integrated deduplication

In an effort to bring the cost of disk-based backup close to that of tape-based backup, Dell has been incorporating deduplication into an increasing number of storage technologies. The Dell/EMC DD Series of storage appliances increases the breadth of options for disk-based backup with deduplication.

The Dell/EMC DD Series consists of three models: the Dell/EMC DD140, the Dell/EMC DD610, and the Dell/EMC DD630. Each offers different levels of capacity, throughput, and scalability. (For capacity and throughput details, see the "Enhancing logical capacity" sidebar in this article.) Each self-contained appliance has its own connectivity and power cable, so it can simply plug into an existing data center environment.

Dell/EMC DD Series storage appliances incorporate leading Data Domain disk-based deduplication storage systems. Data Domain systems are "CPU-centric" and therefore are not dependent on disk drive performance advances or an increase in the number of disk drives to speed throughput over time. They are also designed to identify duplicate data regardless of how the data is packaged and regardless of changes in block size.

Because these appliances are built for deduplication from the ground up, they can run fast enough to apply the benefits of deduplication to replicated data streams in real time, so that the lag until data is safely off-site can be as small as possible. Systems that have add-on deduplication can impose unnecessary delays on the replication process.

Assessing Dell/EMC DD Series backup and recovery features

Deploying disk-based backup using Dell/EMC DD Series appliances with integrated deduplication can provide benefits that address many IT concerns regarding backup and recovery and help provide favorable return on investment.

Complexity reduction. Dell/EMC DD Series appliances provide a deduplication approach that can be implemented easily in heterogeneous environments (see Figure 1). They can connect to storage using CIFS and NFS over Ethernet, Symantec™ OpenStorage over Ethernet, and VTL over Fibre Channel, and can work with many common backup and archive software applications—enabling organizations to consolidate backups but maintain current data management schemes. By simplifying backup and recovery processes, organizations can save time and money while freeing staff resources to pursue other tasks that help increase productivity and profitability.

Data reduction. Dell/EMC DD Series appliances with integrated deduplication are designed to avoid the storage of redundant data, helping reduce the disk capacity required for backup. Reducing the number of hard drives needed to store data then helps reduce the costs of power, cooling, space, and management in the data center. Space that was allocated to the physical footprint for extra drives can be available for other uses. Using Dell/EMC DD Series deduplication technology, organizations can realize up to a 99 percent bandwidth reduction for replication in a typical backup environment with a 5–10 percent incremental change rate, and can achieve 90–97 percent backup data reduction with a typical 10:1–30:1 deduplication ratio.

Low overhead and high speed. Because Dell/EMC DD Series appliances use in-line deduplication, they do not require the overhead of extra disk capacity found in post-processing applications of deduplication. Performance is not dependent on adding disk drives, and equipment, management, and operating costs can be reduced compared with other approaches to storage backup. Deploying Dell/EMC DD Series appliances also helps speed recovery times beyond what tape-based backup can deliver, helping reduce the impact to an organization if something goes wrong.

Cost-effective disaster recovery. Replication to a remote disaster recovery site can be highly cost-effective with Dell/EMC DD Series appliances, because deduplication reduces the amount of data that needs to be transmitted and the required bandwidth. Bandwidth is one of the major expenses involved in replication, and by helping reduce bandwidth requirements, deduplication can make replication a viable

Taming the unstructured data beast

This blog post outlines three primary ways that organizations can address out-of-control data growth, including deduplication with Dell/EMC DD Series storage appliances.

en.community.dell.com/dell-blogs/b/direct2dell/archive/2010/03/24/taming-the-unstructured-data-beast.aspx

99% Dell/EMC DD Series deduplication technology can help organizations reduce bandwidth for replication by up to 99 percent.

disaster recovery method for a wide range of data sets and applications, and for many organizations that may have previously ruled out robust protection as a cost-prohibitive option.

Data integrity and protection. The Data Domain Data Invulnerability Architecture in Dell/EMC DD Series appliances is designed to ensure data integrity and protection through continuous fault detection and healing, end-to-end verification of data recoverability at time of backup, redundant system components, and other data safety techniques. Unlike some other deduplication systems or file-based backup systems, Dell/EMC DD Series appliances help ensure that recoverability is verified and then continuously re-verified.

Tape avoidance. By enabling organizations to avoid redundant data, Dell/EMC DD Series appliances with integrated deduplication help to eliminate space-consuming tape libraries. These appliances are designed to deliver all the benefits of backup to disk, including accelerated backup and restore times and enhanced reliability and data integrity compared with tape. More data can be stored with reduced media resources compared with tape, helping to speed the recovery process.

Optimizing the storage infrastructure

Deduplication has the potential to reduce data backup volume, enhance data protection, and, in general, simplify data management across the storage infrastructure. Dell/EMC DD Series appliances offer a cost-effective deduplication and storage backup option that enhances data protection and accelerates recovery time. The appliances implement deduplication in a way that combines performance and low overhead across a comprehensive range of backup environments. IT managers can easily deploy mature, flexible, high-throughput deduplication without needing to change current backup processes.

Dell/EMC DD Series appliances offer organizations a cost-effective storage, backup, and recovery approach designed to reduce complexity through a single point of contact at Dell. In addition, Dell has a suite of fixed-duration professional services that help organizations identify the optimal approach to archiving, backup, and deduplication to meet their specific needs. PS

Kay Benaroch is a senior marketing consultant at Dell focused on data protection, deduplication, and object-oriented storage, and has more than 30 years of experience in marketing technology products.

Shane Jackson is senior director of product marketing for the EMC Backup Recovery Systems division, and has more than 17 years of experience in the storage industry.

Learn more

Dell/EMC DD Series: dell.com/emc

Dell™ storage solutions: dell.com/storage