Software

Solving Data Growth Issues using Deduplication

Reducing Storage Costs and Speeding Backups with Dell Deduplication Solutions

Abstract

Data growth is increasing at record rates, and ensuring the integrity of that data has become a struggle in many environments. More data means backups take longer to complete and consume more storage. Since organizations generally have a finite amount of storage allocated for backup, many of them are forced to either perform fewer backup jobs or keep their backup save sets for shorter periods of time.

Together, the Dell™ DR Series appliances and NetVault™ Backup provide the perfect combination to address these issues, enabling you to keep systems backed up within maintenance windows while keeping storage costs low through deduplication and replication.

This paper discusses the basic principles of deduplication and the benefits of deduplication in the corporate environment. Then it highlights the advantages of using the Dell DR Series appliance and its Rapid Data Access (RDA) plug-in with NetVault Backup (NVBU) versus using Dell NetVault SmartDisk with NetVault Backup.

Introduction

The challenges of data growth

Data growth is constantly being reported as one of the top challenges for IT infrastructure by various analysts, such as Gartner. Because of explosive data growth, organizations are struggling to maintain even a couple of months' worth of backups on disk. In many cases, these backups contain large amounts of redundant data, such as database information, operating system files, medical records, presentations and images, that can easily consume hundreds of terabytes or petabytes of space. Purchasing this storage and the solutions to manage it quickly eats into IT operating expense (OPEX) and capital expenditure (CAPEX) budgets.

Deduplication reduces storage requirements and speeds backups by eliminating redundant data and storing only unique data.

There is also the issue of the ever-growing backup window. As servers continue to grow in size, and more and more businesses operate for longer hours, administrators find themselves faced with small—sometimes barely achievable—backup windows.

The challenges for key stakeholders

Another way to look at the challenges of data growth is to consider the various concerns of individuals in the organization. If you were to ask key stakeholders about the problems they experience with regards to managing their data, their answers would likely sound like this:

- **CEO**: "I don't want to pay any more than I do now for a 'what if' scenario."
- **CFO**: "We need to reduce OPEX, not increase it. I can't afford to keep spending on IT at this point."
- **CIO**: "I have a disaster recovery plan right now, but I want one that actually works—at the cost of my current plan."
- **IT manager**: "I need a better way to manage backups without the tape fiasco, and I need a better way to perform restores."
- **Data center manager**: "It is becoming increasingly difficult to meet my nightly backup window. Is there a way to back up my data faster without upgrading my entire network to 10 GB?"

Deduplication is the answer to most of these stakeholders' problems, because it allows the organization to maintain or improve its current disaster recovery solution without a large increase in OPEX or CAPEX.

Understanding deduplication

Eliminating redundant data reduces storage costs and speeds backups

Deduplication reduces storage requirements and speeds backups by eliminating redundant data and storing only unique data. Specifically, deduplication works by using pointers to eliminate duplicate blocks of data. Pointers are simply references to existing blocks of data, but they consume significantly less space than those blocks. Deduplication ensures that only unique data is stored on disk, reducing storage requirements. Deduplication requires the ability to read data from any part of the disk (random access), and not just sequentially, making it ideal for data stored on disk but not an option for data on tape.

To demonstrate the power of deduplication, let's consider a virtual environment with 10 virtual machines, each 50 GB in size and running Windows 2008 R2. If we assume that the installation size of Windows is 32 GB, this yields roughly 10 x 32 GB of duplicate data. In other words, of the 500 GB of space, approximately 320 GB holds duplicate data. Eliminating that duplicate data yields a 64 percent storage savings from deduplication alone, and even more savings is possible when compression is factored in. This reduction in duplicate file storage can save an organization upwards of 90 percent in their storage capacity requirements.

How deduplication happens: block-level or byte-level

Typically deduplication takes place at either the block level or the byte level. Block-level deduplication analyzes entire blocks of data, which allows for granularity without being overly time-consuming and resource-intensive. Byte-level deduplication is slightly more granular, but it introduces significant resource overhead and performance penalties that outweigh the space savings. The block-level deduplication technology utilized by Dell provides up to a 93 percent reduction in duplicate data in a fast and efficient manner.

Block-level deduplication can be further broken down into two methods, fixed-block mode and variable-block mode:

- **Fixed-block mode** – The fixed-block algorithm utilizes smaller blocks than variable-block mode, so it yields inherently better deduplication rates, but it isn't able to eliminate all duplicate blocks of data. The primary limitation to fixed-block deduplication is that when, for example, a single character is changed in a document, it is likely that document will appear as an entirely unique document to the algorithm.
- **Variable-block mode** – With the variable-block algorithm, the block length changes dynamically to better align to data boundaries, allowing for higher deduplication rates than fixed-block deduplication. For example, a single change to a character in a document won't cause the document to appear as a unique document. In most situations, variable-block deduplication is the optimal method, primarily because of its ability to provide high-performance deduplication while still isolating the smallest changes in incoming data.

Where deduplication happens: client-side, on the backup server or target-side

In a given environment, deduplication can take place at one of three locations: the client (or source), the backup server, or the storage (target) server. Which option is best for you depends on the characteristics of your environment. Table 1 details the pros and cons of each method.

Deduplication location	Pros	Cons
Client-side deduplication	• Significantly reduces backup windows • Reduces backup network bandwidth requirements • Great for bandwidth constrained environments • Great for agentless backup in virtual environments due to reduced backup traffic • Transmits only unique data between backups (so the first full backup might take three hours, while the second full backup might take only a few minutes) • Provides a common deduplication pool across the environment, allowing for high deduplication ratios and greater storage savings	• Uses additional resources on the client machine
Backup server deduplication	• Reduces network bandwidth requirements between backup server and storage • Great for bandwidth-constrained environments • Does not increase the load on the client CPU	• Uses additional CPU resources on the backup server • Does not reduce bandwidth requirements between the client and the backup server
Target-side deduplication	• Offloads all CPU-intensive processes to the deduplication appliance, instead of sharing CPU, memory and other client-side resources • Provides a common deduplication pool across the environment, allowing for high deduplication ratios and greater storage savings	• Does not reduce backup window • Does not reduce bandwidth requirements between the client and the appliance • Full backups take longer to complete in comparison to client-side deduplication

Table 1. The ideal deduplication location depends on the characteristics of your environment.

Deduplication can take place at one of three locations: on the source (client), on the backup server, or on the storage (target) server. Which option is best for you depends on the characteristics of your environment.

Client-side deduplication can easily take what would normally be a three-hour backup window and reduce it to as little as a few minutes.

When deduplication happens: post-process or inline

There are also different times when the deduplication process can take place: post-process or inline.

- **Post-process deduplication**, as the name suggests, happens after the data has been stored on disk. Although post-process deduplication provides storage savings, the process can be time-consuming, requires sufficient space to store the data before it is deduplicated, and must be scheduled.
- **In-line deduplication** happens while the data is being written to disk. This option is fast and efficient and does not require a temporary landing zone for data. For these reasons, in-line deduplication the preferred deduplication process.

The benefits of deduplication

Deduplication provides a cost-effective way for a business to maintain backups on disk, for longer periods of time, before staging them off to tape for extended retention. Deduplication also reduces expensive network and WAN traffic, increasing the speed at which backups can complete—a tremendous benefit for the administrator when only a few hours each night are allowed for a full backup. In fact, client-side deduplication can easily take what would normally be a three-hour backup window and reduce it to as little as a few minutes. This drastic reduction in backup time is possible because client-side deduplication eliminates the amount of data that needs to be sent from the client to the backup target, reducing what would normally be, for example, 1 TB of network traffic down to just a few gigabytes.

As Table 2 illustrates, implementing a deduplication solution can reduce 151 TB of data down to only 7.6 TB, an astounding 20x reduction. With deduplication, organizations can store six months of backups using less storage than previously required for just a week's worth of backups.

NetVault deduplication solutions from Dell Software

NetVault Backup

NetVault Backup (NVBU) is an enterprise file-based backup solution that supports a variety of platforms and applications. It supports a wide range of targets for storing backups, including local disk, network attached storage (NAS) devices, storage area network (SAN) devices, virtual tape libraries (VTLs) and shared virtual tape libraries (SVTLs), as well as Dell and third-party deduplication appliances, physical tape libraries, autoloaders and tape drives.

	Backup data	Cumulative logical data	Estimated reduction	Cumulative physical storage
First full	5 TB	5 TB	1x	4.0 TB
Week 1	5.2 TB	10.2 TB	3x	4.2 TB
Week 2	5.4 TB	15.6 TB	4x	4.4 TB
Week 3	5.4 TB	21 TB	5x	4.6 TB
Month 1	5.4 TB	26.4 TB	6x	4.8 TB
Month 2	22.6 TB	49 TB	9x	5.6 TB
Month 3	24 TB	73 TB	12x	6.2 TB
Month 4	25 TB	98 TB	15x	6.7 TB
Month 5	26 TB	124 TB	17x	7.2 TB
Month 6	27 TB	151 TB	20x	7.6 TB
	Total	**151 TB**	**20x**	**7.6 TB**

Table 2. Deduplication savings (assuming a 15:1 deduplication and compression ratio)

NetVault Backup leverages the NetVault Virtual Tape Library (nVTL), a backup target that emulates the behavior and appearance of an actual tape library, allowing functions such as loading and unloading of tapes. The nVTL can reside on either direct attached storage located on the NVBU server, a NVBU SmartClient or a network share.

NetVault SmartDisk

NetVault Backup in conjunction with nVTL uses compression to reduce storage requirements, but compression can only go so far in the quest to reduce storage footprints. Compression does not eliminate the number of duplicate files or blocks stored by the backup process; rather, it simply reduces the space consumed by the duplicate blocks.

To solve the problem of duplicate data, Dell offers NetVault SmartDisk (NVSD), a software deduplication solution that is integrated with both NetVault Backup and Dell™ vRanger™. This standalone solution allows for up to a 90 percent reduction in backup footprint when compared to not using deduplication. Therefore, IT can easily maintain backups of mission-critical systems on disk for a longer period of time, ensuring faster, easier and more reliable restores.

SmartDisk is a post-process, target-side deduplication technology that can be configured to perform deduplication either while data is being written to disk or on a predetermined schedule. NVSD is installed on a server with access to local, direct-attached, iSCSI or networked storage (see Figure 1). This allows the backups to flow at their normal speed and get committed to disk as fast as the disk subsystem and NVSD will allow. Once the backups have been written to disk, the deduplication task runs and reclaims the blocks consumed by duplicate blocks of data.

NetVault SmartDisk provides a hardware-agnostic solution for any environment, allowing you to use your existing storage from Dell or other manufacturers for the deduplication process, without causing any additional CPU load to the backup clients. This is a key benefit to any organization that is backing up servers that are heavily used and need to remain unaffected by backups.

Other benefits of NVSD over traditional backup methods or NetVault VTLs include:

- **Replication** – NVSD can replicate backups in one SmartDisk instance to another SmartDisk instance.

NetVault SmartDisk provides a hardware-agnostic solution for any environment, allowing you to use your existing storage from Dell or other manufacturers for the deduplication process, without causing any additional CPU load to the backup clients.

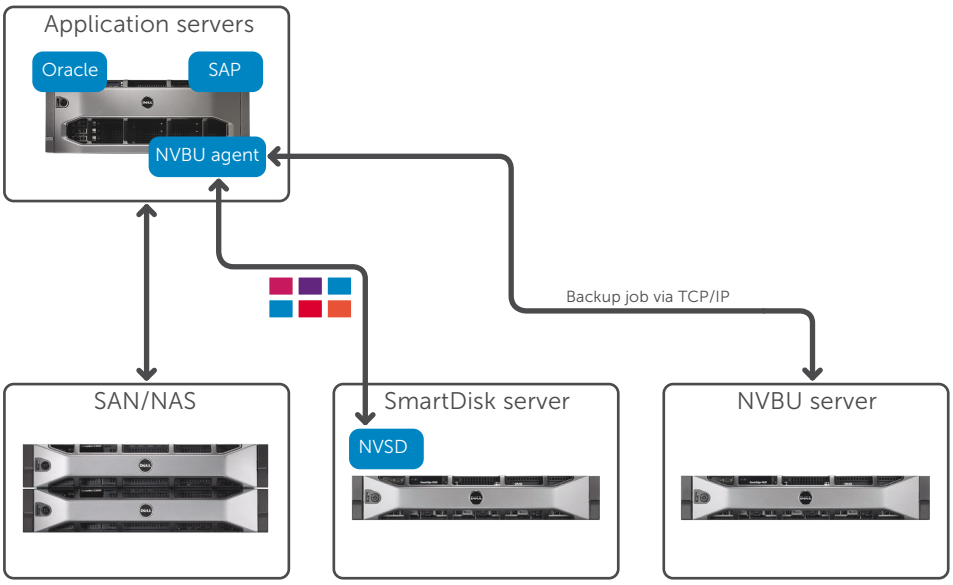

Figure 1. NetVault SmartDisk architecture

DR Series appliances are the go-to solution for anyone looking for serious deduplication performance.

- **Dynamic block size** – NVSD can dynamically change the block size to optimize the amount of data being deduplicated.
- **Global deduplication** – NVSD supports deduplication across multiple backup jobs and multiple backup servers.
- **Scalable architecture** - You can grow storage non-disruptively to maximum of 15 TB of deduplicated data per NVSD instance.

Dell's DR Series appliances

The DR4100 and DR6000 backup appliances

NetVault SmartDisk is a great start for any organization looking to reduce storage costs, but for customers seeking the ultimate in deduplication and backup speeds, Dell offers the DR Series appliances. DR Series appliances are the go-to solution for anyone looking for serious deduplication performance. Both the DR4100 and the DR6000 offer built-in, in-line and source-based deduplication as well as compression. Both appliances are built on the 2U Dell 12G 720xd platform. However, the similarities between the DR4100 and DR6000 stop there, as the DR6000 has a few extra goodies under the hood to allow it handle massive amounts of data:

- **DR4100** –The DR4100 disk backup appliance comes in 2.7 TB, 5.4 TB, 9 TB, 19 TB and 27 TB (after RAID6) configurations with support for up to two additional expansion shelves. At maximum capacity the DR4100 can provide up to 81 TB useable after RAID capacity—enough capacity for most of today's small to medium business (SMB) environments.
- **DR6000** – For today's enterprise environments, the DR6000 appliance comes in 9 TB, 18 TB, 27 TB and 36 TB (after RAID6) configurations with up to four additional expansion shelves, providing up to 180 TB of usable after RAID capacity.

Built-in deduplication engine using variable-block, inline deduplication

Under the hood of the DR Series appliances lies a powerful deduplication engine built by Dell, running on Intel processors and high-performance disks in a RAID6 configuration. The DR Series supports a wide variety of backup software solutions. As the data is ingested, the DR appliance inspects the backup stream, dynamically changes the block size according to the data used for deduplication (that is, uses variable-block deduplication), and then compresses these blocks, resulting in data reduction rates as high as 93 percent.

The inline deduplication process is further aided by a NVRAM (non-volatile RAM) card containing SLC NAND flash with low latency and high IOPS (input-output per second). This enables the DR appliances to quickly deduplicate data as it is ingested, avoiding the pitfalls and performance constraints of traditional post-process deduplication.

Integration with NetVault Backup and Rapid Data Access (RDA)

With the release of NetVault Backup 10 and the DR Series backup appliance 3.0 firmware, both NetVault Backup and the DR appliances support Rapid Data Access (RDA) plug-in. RDA is a protocol for the DR appliance that allows applications to have deep integration with the DR appliances. This exciting protocol allows NetVault Backup to seamlessly use the DR Series backup appliances as a target location for either source-side or target-side deduplication, allowing backups to be finely tuned to individual environments and requirements (see Figure 2).

With source-side deduplication, the DR Series appliances enable backup jobs to complete 200 percent faster than with target-side deduplication. Once the data is stored on the primary DR Series appliance, it can be replicated to a secondary site using deduplicated, compressed and optionally encrypted traffic.

Dell internal lab testing of NetVault Backup 10 and the DR6000 resulted an astounding 22 TB/hr peak ingest performance using RDA, and testing with the DR4100 resulted in 7.5TB/hr peak performance.

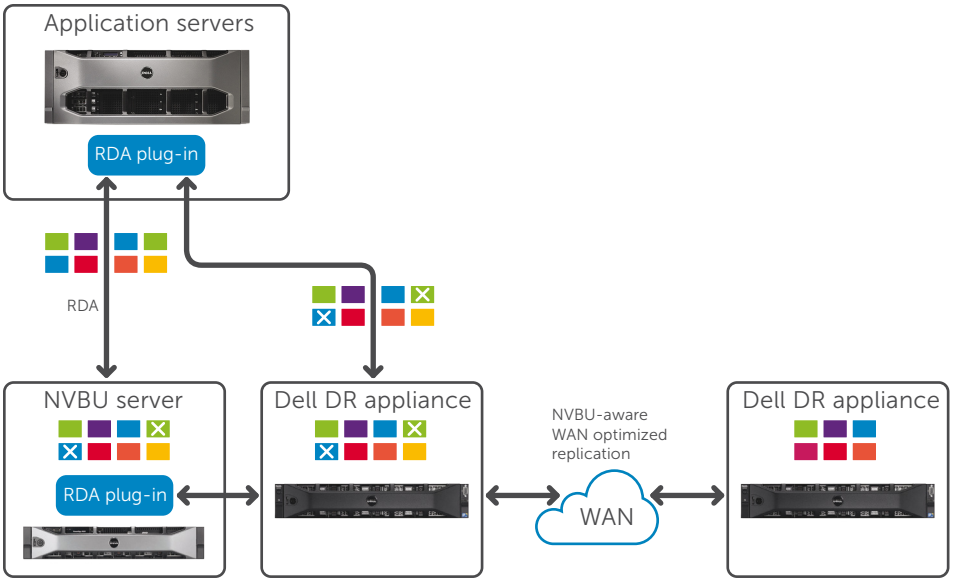

Figure 2. NetVault RDA plug-in and DR Series appliance topology

You can now easily meet your stringent backup windows—with RDA source-side deduplication, subsequent full backups can complete in seconds. This is a drastic reduction in backup time, since backups using NetVault SmartDisk generally take hours.

Benefits of the Dell DR Series appliances

Offering far more than NetVault SmartDisk

Organizations looking to take backup and disaster recovery to the next level should take a close look at leveraging the DR Series appliances in their environment. Despite the amazing benefits of NetVault SmartDisk, the DR Series appliances still have significant advantages over NetVault SmartDisk, including the following:

- The DR Series appliances are purpose-built for deduplication and have been highly optimized. This eliminates the need to custom-build a hardware solution, as is required to get maximum value from NetVault SmartDisk.
- The DR Series appliances deliver faster backup and recovery speeds than NetVault SmartDisk, along with over five times the backup capacity.
- The choice between client-side and target-side deduplication is at your fingertips. You can easily adjust backup deduplication to meet your performance requirements.
- The DR6000's 64:1 replication means that up to 64 remote sites can replicate to a single DR6000 from another DR Series appliance; the DR4100 provides 32:1 replication. NVSD is limited to 1:1 replication.
- Replication processing is offloaded to the DR Series appliances and not the NetVault server or client.
- The DR Series appliances deliver easy scalability and redundancy.
- The DR Series appliances have an all-inclusive license that includes the use of RDA and replication.

Intuitive global dashboard

A very important yet often overlooked aspect of the DR Series appliances is its ease of use. The DR offers a web-based console with a global dashboard that lays out all of its functionality in an easy-to-use and intuitive format, highlighting key information in a single view (see Figure 3). The dashboard presents detailed information such as the amount of remaining free space, total savings gained by deduplication and compression, and the current state of the system and hardware for one or multiple DR Series appliances. You can easily keep your eye on your environment either by keeping this dashboard displayed on a monitor in your data center, or by saving a bookmark to it for a quick glance as desired.

You can now easily meet your stringent backup windows—with RDA source-side deduplication, subsequent full backups can complete in seconds.

You have flexibility in choosing where and how data is maintained on the DR appliance, and how it is replicated to secondary sites.

Figure 3. DR4100 dashboard

Flexibility in data storage and replication, and in the choice of backup solution

Data is sent to locations the DR appliances known as containers. A container provides features such as deduplication and compression that work globally across all of the containers, while replication can be turned on or off on a per-container basis when you are not using RDA. When RDA and replication is being utilized, it is controlled via NetVault Backup using the optimized replication option, offloading all the workload of replication to the DR appliance while NetVault triggers the task.

A DR Series appliance can encompass multiple containers, each utilizing a different protocol (such as CIFS, NFS, OST or RDA) and features (such as replication). This means that you have flexibility in choosing where and how data is maintained on the DR appliance, and how it is replicated to secondary sites. This also means that multiple backup products can be used on the same DR Series appliance. For example, you can store NetVault data in one container while storing vRanger backups in a different container, and the appliance will continue to provide global deduplication across all containers (see Figure 4).

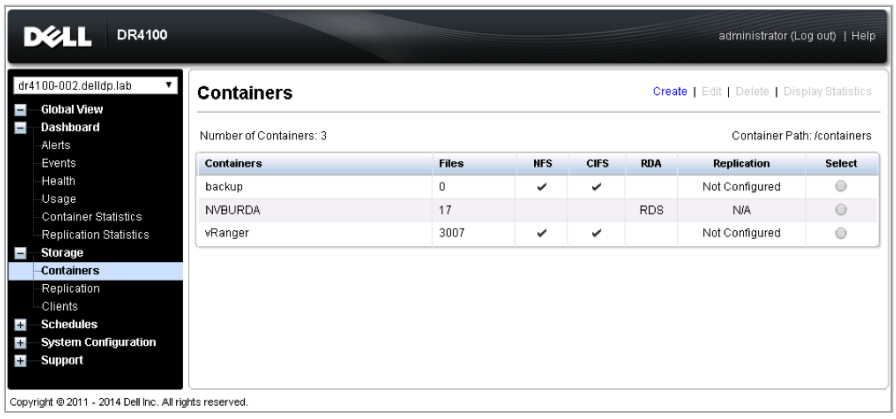

Figure 4. DR4100 RDA container

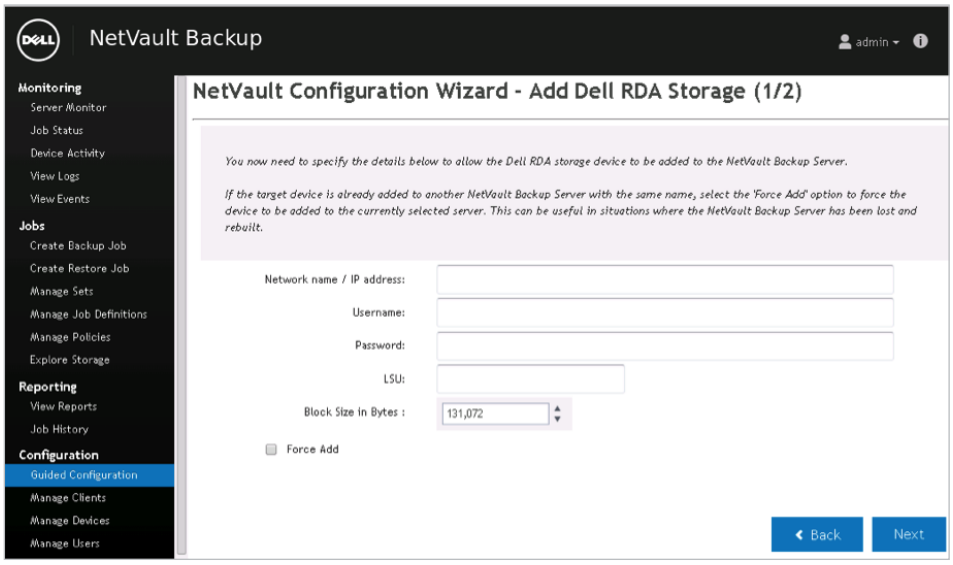

Figure 5. Adding an RDA device to NetVault Backup

Connecting NetVault Backup to the DR Series appliance

Creating an RDA container is simple—it requires only a few clicks. And connecting NetVault Backup to the DR Series appliance is even easier. With NetVault Backup 10, RDA is integrated into the console, providing a streamlined configuration experience. With just four simple pieces of information (see Figure 5), you can have an existing instance of NVBU backing up to a DR backup appliance in just a few minutes. To further simplify the process, the NetVault and DR solution defaults to using RDA source-side deduplication, speeding backups as well as configuration.

Conclusion

With ever growing data and increasingly shrinking backup windows in today's demanding organizations, deduplication is no longer just a nice-to-have; it's a must-have. Deduplication is a cost-effective way to adapt to data growth while drastically cutting backup times and keeping capital expenditures to a minimum.

The Dell DR Series appliances are easy to use, offer all-inclusive licensing, and can be deployed into an existing infrastructure with ease. With a Dell DR Series appliance, the hassles of finding the right server, disk, RAID level, processor and software are a thing of the past. Simply deploy NetVault Backup and the DR Series appliance, and watch your backup storage capacity needs, backup windows and network traffic shrink drastically.

For more information, please visit the Dell Software website:

- **The DR Series appliances** – software.dell.com/products/dr-series-disk-backup-appliances/
- **NetVault Backup** – software.dell.com/products/netvault-backup

You can have an existing instance of NVBU backing up to a DR backup appliance in just a few minutes.

For More Information

About Dell Software

Dell Software helps customers unlock greater potential through the power of technology—delivering scalable, affordable and simple-to-use solutions that simplify IT and mitigate risk. The Dell Software portfolio addresses five key areas of customer needs: data center and cloud management, information management, mobile workforce management, security and data protection. This software, when combined with Dell hardware and services, drives unmatched efficiency and productivity to accelerate business results. www.dellsoftware.com.

If you have any questions regarding your potential use of this material, contact:

Dell Software
5 Polaris Way
Aliso Viejo, CA 92656
www.dellsoftware.com
Refer to our Web site for regional and international office information.