THE CLIPPER GROUP
Navigator™
Navigating Information Technology Horizons℠

Published Since 1993 Report #TCG2013020 October 18, 2013

Deploying Data Protection the Dell Way — Dedupe Your Data on the Fly

Analyst: David Reine

Management Summary

As we progress through the next hurricane season, we can look back to the last big one. You know the one, the hurricane that knocked down trees, felling power lines all around you. What happened in your house? Well, first, you lost all power, eliminating all lights, air conditioning (or heat), the nightly news or anything else that might be on TV. Worse, even though the storm only lasted a few hours, the after effects lasted for days. A mad dash to the hardware store for a generator proved futile as most dealers ran out before you got there. Man cannot control Mother Nature and natural disasters. No electricity, no refrigeration; all of your food spoiled – although that gave you a good reason to eat all of the ice cream! How long did it take you to recover? Well, three days later when the electric company finally reached your street and turned the power back on, you resolved to head right to that same hardware store and order that generator in order to be prepared for the next one. Did you? Well that's another story.

Natural, and manmade, disasters are all too common in the data center of every enterprise, large or small. That same hurricane that knocked your house offline likely did the same thing to your office. Major enterprises are usually prepared with emergency power for the enterprise data center. Unfortunately, not every small business has the same resources as their larger brethren with generators or high availability solutions that can transition their operations to a remote site. Then there are the minor disasters that can afflict every enterprise, no matter the size of the budget. You know the kind: someone accidently deleted the customer file or, worse, payroll. How long did it take you to recover from your most recent disaster? Did the outage caused by the loss of data result in lost business, or worse, corporate failure? How quickly can your data center staff recover a lost file, and at what cost? With data doubling in size every 12-to-18 months, the responsibility to backup all of it can become overwhelming. Worse, it can destroy your IT budget. How quickly can you recover that lost file or database? At what cost? Sometimes, the need to do an immediate recovery can require the staff to re-think their backup and recovery paradigm. Tape may have been your backup medium of choice for a decade, but today the need for data immediacy may force you to look in a new direction. One solution is to transition your backup requirements to disk, while resolving your concerns over storage growth with data reduction techniques. One solution that can resolve these concerns is a backup/recovery appliance.

There may be a plethora of solutions for the enterprise with the deepest pockets, unfortunately that may not you. One company that has a solution that can fit every enterprise requirement and budget is Dell. To learn more about Dell and its *DR4100* appliance, please read on.

IN THIS ISSUE

Business-Critical Data Explosion

During the last century, it was still said that the sun never sets on the British Empire. Today, the sun never sets on your files or those people trying to access them, both legally and illegally. Enterprises of all sizes are required to operate on a 24x7 schedule, 365 days a year. The business activities of your customers and partners around the world depend heavily upon continued access to your enterprise files, with an integrity level matching the highest levels of availability. It is the responsibility of the data center to ensure that access, while at the same time, providing data protection, for an ever-expanding storage environment. However, that is difficult under the best of circumstances. It is only complicated further by the constraints of the IT budget.

Data growth is one of the tallest hurdles facing the enterprise data center, and managing that data has become an increasingly difficult proposition. It is not a question of "if" a disk drive will fail; it is a matter of "when" will a drive fail. It is an issue of how much data has been lost and how fast the staff can recover it. How many backups can be taken and preserved before the data center runs out of backup capacity? What is the recovery point objective (RPO) of your enterprise? How quickly does the data need to be recovered in order to satisfy your recovery time objective (RTO)? Due to the growth of backup data and the time that has been committed to recover it, in many instances tape may no longer be a viable alternative for backup in your data center.

While the amount of storage that needs to be protected is growing at a seemingly inexorable rate, more than 50% annually, the backup window provided to your data center staff is not. In fact, it is shrinking! The data center staff has to find a way to reduce the amount of backup data being stored without affecting data access. The staff must find a way to squeeze ten pounds of data into a five pound (or less) bag. One solution to this dilemma is to reduce the amount of data actually being stored in a timely fashion. The first question that needs to be answered is "*how*". The next question is always "*how much will it cost*". In order to reduce the amount of data that needs to be saved, replicated, and protected, the enterprise data center must take advantage of every affordable technological advancement available to it. The staff can deploy a system that uses the traditional compression algorithms, but that will not eliminate the storage of duplicate records or redundant data. In order to accomplish that, the data center needs to deploy the latest storage-saving technology, *data deduplication*. The implementation of both of these techniques can significantly reduce your storage footprint. This will enable the data center staff to keep critical data on disk longer, providing faster recovery times and reserving more of your tape resources for archiving and the long-term preservation of your data.

The cost to accomplish this, however, is a separate story. With an unlimited budget, much can be accomplished. Unfortunately, you do not have unlimited resources. Your budget is shrinking, or, at best, static. You do not have the resources to configure your own solution, combining the best hardware and software system from vendors around the storage universe. This would be complex at best and very difficult for a data center with limited staff and resources to maintain. Furthermore, it is unnecessary, as Dell has already done it for you with the *DR4100 Disk Backup and Disaster Recovery Appliance* to compress and deduplicate enterprise data.

Let's look at what typically happens when you make a copy of a file and send it to a remote repository for "safekeeping". First, you need to move it to that repository, whether nearby or remote. This consumes network bandwidth. Then you need an equal amount of storage at the target repository to hold it. Now, you have potentially doubled your storage cost, consuming scarce resources. *While that is bad enough, it can get worse!* What if that file had been backed up previously? Now, you may be paying repeatedly to transmit and store the same file in that repository. *And, it can get even worse!* What if that attachment, presentation or spreadsheet, had been distributed to multiple employees, which they each then stored in their personal directory on a server? You may be transmitting that same file dozens or hundreds of times. Now, you can see how the costs of network bandwidth and storage capacity can increase many, many times, protecting exactly the same file.

What if you made a small change to that spreadsheet or presentation that had already been protected? If you just changed a couple of cells in the spreadsheet (representing, say, 2% of it), do you want to retransmit the same 98% again, and again? This is where data deduplication, at the sub-file level, becomes important. If you modify only a small part of a file, then only the changed and/or new portions would need to be transmitted.

Exhibit 1 – Inline vs. Post-Process Data Deduplication

Post-process dedupe processes the data after it has been written to disk. At that time, the dedupe process begins and reclaims redundantly-used storage. Because the data is not processed in real time, computational demands are lower. However, post-process does have its drawbacks.

- There are greater, short-term space requirements as storage must be reserved for both the original and the deduplicated data. This could lead to more complicated sizing and configuration calculations.
- Greater bandwidth is required as you must first transmit the duplicate data and then read it back to be processed and then manipulated what is redundant.
- Post-process dedupe will take more time. In this case, the staff must complete two separate and sequential processes, limiting the capability to perform a backup without affecting operational efficiency.

Inline dedupe processes the data before it is first written to disk. This limits the amount of data being stored. It delivers a more predictable performance level, enabling the data center staff to properly size the required solution. Inline processing, however, can affect ingest speed. Because of the requirements to process data in real time, maintaining acceptable performance levels can be very compute intensive.

What is Data Deduplication

Before I review what you can get from Dell, let's take a closer look at what the data center can accomplish with *Data Deduplication*. So, what is Data Deduplication, or simply, *dedupe*? Dedupe is a technique used to reduce the volume of data being stored by eliminating duplicate copies of the same files or same data, structured or unstructured. It proves very effective in environments where the majority of data being stored is an exact copy of data that has already been stored, which in typical enterprises is more than 50%. Examples of this would be attachments to email that are sent to multiple people in the organization or backing up a file that has not been changed since the last backup. In a virtual environment with multiple VMs on the same system, each clone represents a copy of the same operating system image and is very space consuming. Even though only one copy is being retained, the system makes it appear to the user that it is his/her personal copy.

There are two questions that then must be answered: The first is: *when to do the dedupe*, and the second is: *where to do the dedupe*. Because the data deduplication process needs to be totally transparent and can be performed "*post-process*"[1], i.e., from a buffer in a dedupe appliance after the transmission, or *inline*[2], reducing networking costs. (See Exhibit 1, above, for a more complete description.) As to where to do it, the data center staff can deploy a *target-based dedupe* or a *source-based dedupe*.

With an inline data deduplication process, the data is deduplicated before it goes over the WAN, reducing transmission volumes and costs. With a post-process deduplication, the data is deduped after the transmission, reducing the processing burden on the server, or client, being backed up. This method is transparent to existing workflows, minimizing disruption to the process. However, it does consume more network resources because all of the data, including redundant data, must be transmitted. Choosing the correct method for your application set can be critical to the success of the process.

What is the impact of this? Simply put, data deduplication enables the data center to deploy a system specifically designed to save a tremendous amount of time and storage, significantly decreasing the amount of space required to do a backup, and thus speeding up the backup process, enabling the data center to create or maintain a disk-to-disk backup architecture without fear of totally exceeding the storage budget. This is especially true for backups from branch offices or remote locations where the use of wide area networks for transmission can significantly increase the cost of the short-term preservation of critical enterprise assets.

Dedupe enables the data center staff with the flexibility it needs to deploy an automated back-

[1] Also known as "out-of-band", i.e., not in real time during the process of first storing it on disk.

[2] Also known as "in-band", i.e., as part of the process of first storing it to disk.

up/recovery environment, increasing the frequency of replication in order to improve both RPO and RTO. Dedupe provides the vehicle that the data center needs to change the backup/recovery paradigm for both local and remote backups, replacing manual intervention with automation, reducing the amount of supervision required, and lowering the TCO of short-term data preservation. Dell's DR4100 Disk Backup and Disaster Recovery appliance is a preconfigured solution for this process.

Dell DR4100 Disk Backup Appliance

The DR4100 is a high-performance, disk-based backup and recovery system that enables the enterprise data center to implement a full data reduction solution mitigating risk with both compression and data deduplication built-in, enabling a data reduction level of up to 15:1, depending upon the data makeup. The DR4100 is easy to deploy and manage, while lowering the total cost of ownership of your storage environment. With technology obtained with the acquisition of Ocarina, Dell has been able to deliver a best in class in-line appliance to handle your data deduplication needs. The DR4100 enables all of the enterprise backup data to remain on disk for a longer time as a result of the data reduction, with shorter RTOs and more frequent RPOs, providing fast and reliable restores while reducing data management complexity.

Dell has designed the DR4100 to enable target-based deduplication at both the local and remote site, offering the greatest flexibility and highest predictable performance while reducing WAN traffic. By deploying DR4100s at both sites, backup data is deduplicated at the target in the remote office. Once that is complete, the deduped data can be replicated back to the primary site, such as the data center. This design facilitates the restore locally, both at the data center or remote location, significantly reducing the time to recover in order to meet, or exceed, existing RTOs.

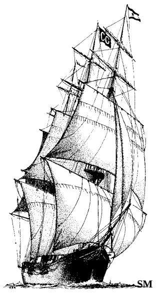
Dell has implemented innovative firmware, with optimal functionality and all-inclusive licensing, thus eliminating the cost uncertainty associated with future feature upgrades. The DR4100, powered by *Dell PE 12G* hardware, with a 2U format, is available in a complete range of options from 2.7TB of after RAID physical capacity to 81TB (1.2PB logical), with flexible and seamless capacity expansion as the enterprise grows, deferring capital costs until the demand requires expansion. It has been designed to handle streaming backup workloads expeditiously via a pair of 10GbE interfaces per node, reducing media usage, network bandwidth usage, and power and cooling requirements while improving overall data protection and lowering the TCO. The DR4100 also resolves backup issues for multi-site environments through the DR4100 replication function enabling improved tolerance for potential disasters. Replication can be scheduled to occur during non-peak periods and data ingestion can be prioritized over replication data to help ensure optimal backup windows.

A convenient graphical user interface (GUI) provides the staff with an overview of the system, including status, hardware and software alerts, and storage capacity and *savings* from the data reduction functionality. The GUI provides the staff with the health of the hardware and the integrity of the system software, automatically.

Conclusion

As the data center continues to gather and retain more and more information in order to improve enterprise responsiveness to an ever-changing customer environment, the costs associated with saving and protecting that data continue to rise. The data center simply has to gain control and reduce the amount of data being accumulated. Dell's DR4100 changes the economics of data protection by reducing storage costs and mitigating risk through integrated data deduplication. It is a simple and easy-to-use disaster recovery solution that helps the data center to reduce both capital and operational expenses, lowering the TCO of the IT infrastructure. The DR4100 provides exactly what the data center needs –a timely and accurate data recovery to enable business continuation in the face of a potential disaster, including the protection of up to 32 remote nodes, simultaneously.

If your enterprise data center is experiencing the kind of growth that can bust any IT budget, then check out the Dell DR4100. It may be the solution that you have been seeking.



About The Clipper Group, Inc.

The Clipper Group, Inc., now in its twenty-first year, is an independent publishing and consulting firm specializing in acquisition decisions and strategic advice regarding complex, enterprise-class information technologies. Our team of industry professionals averages more than 25 years of real-world experience. A team of staff consultants augments our capabilities, with significant experience across a broad spectrum of applications and environments.

- ***The Clipper Group can be reached at 781-235-0085 and found on the web at www.clipper.com.***

About the Author

David Reine **is a Senior Contributing Analyst for The Clipper Group.** Mr. Reine specializes in enterprise servers, storage, and software, strategic business solutions, and trends in open systems architectures. In 2002, he joined The Clipper Group after three decades in server and storage product marketing and program management for Groupe Bull, Zenith Data Systems, and Honeywell Information Systems. Mr. Reine earned a Bachelor of Arts degree from Tufts University, and an MBA from Northeastern University.

- ***Reach David Reine via e-mail at dave.reine@clipper.com or at 781-235-0085 Ext. 123. (Please dial "123" when you hear the automated attendant.)***

Regarding Trademarks and Service Marks

The Clipper Group Navigator, **The Clipper Group Explorer**, **The Clipper Group Observer**, **The Clipper Group** *Captain's Log*, **The Clipper Group Voyager**, **Clipper Notes**, and *"clipper.com"* are trademarks of The Clipper Group, Inc., and the clipper ship drawings, *"Navigating Information Technology Horizons"*, and *"teraproductivity"* are service marks of The Clipper Group, Inc. The Clipper Group, Inc., reserves all rights regarding its trademarks and service marks. All other trademarks, etc., belong to their respective owners.

Disclosures

Officers and/or employees of The Clipper Group may own as individuals, directly or indirectly, shares in one or more companies discussed in this bulletin. Company policy prohibits any officer or employee from holding more than one percent of the outstanding shares of any company covered by The Clipper Group. The Clipper Group, Inc., has no such equity holdings.

After publication of a bulletin on *clipper.com*, The Clipper Group offers all vendors and users the opportunity to license its publications for a fee, since linking to Clipper's web pages, posting of Clipper documents on other's websites, and printing of hard-copy reprints is not allowed without payment of related fee(s). Less than half of our publications are licensed in this way. In addition, analysts regularly receive briefings from many vendors. Occasionally, Clipper analysts' travel and/or lodging expenses and/or conference fees have been subsidized by a vendor, in order to participate in briefings. The Clipper Group does not charge any professional fees to participate in these information-gathering events. In addition, some vendors sometime provide binders, USB drives containing presentations, and other conference-related paraphernalia to Clipper's analysts.

Regarding the Information in this Issue

The Clipper Group believes the information included in this report to be accurate. Data has been received from a variety of sources, which we believe to be reliable, including manufacturers, distributors, or users of the products discussed herein. The Clipper Group, Inc., cannot be held responsible for any consequential damages resulting from the application of information or opinions contained in this report.