THE CLIPPER GROUP

Navigator™

Navigating Information Technology Horizons℠

Published Since 1993 Report #TCG2011022 June 8, 2011

Protecting Enterprise Data with ProtecTIER — Improving Recovery Time, and More

Analyst: David Reine

Management Summary

Madonna has said that “*We Are Living in a Material World*”, and she is right; however, we are also living in a *Virtual World*, one fraught with dangers over which we have virtually no control – the threat of terrorism hangs over us as we go about our daily lives. In addition, there are the natural disasters that threaten different parts of the country all of the time. Recently, we have seen devastating tornados wreak havoc, and claim lives, throughout the mid-west, practically wiping out one city in Missouri (Joplin), while flooding along the Mississippi has destroyed thousands of acres of farmland and several small communities, in order to spare larger cities such as New Orleans. Unfortunately, there was little that the government could do to spare New Orleans, or a large portion of the gulf coast, from Hurricane Katrina and oil spills, which flooded homes and ravaged a major portion of their tourism industry. It does not take long to destroy the property and years of work of many, but the recovery from these disasters can take months, if not years. The cost of these disasters is hard to calculate, not only as people try to rebuild their lives, but also on the impact of others who live nowhere near the scene, from seeing a rise in the cost of food to losing vacation homes on that coast.

Unfortunately, disasters are nothing new to the CIO and IT managers in enterprise data centers around the globe. For over sixty years, the IT staff has been carefully backing up mission- and business-critical data to all forms of removable media in order to transport it to some remote site to be able to recover when a disaster strikes, or when some misguided soul just deletes everything from the wrong folder. Performance for the backup process has dominated the conversation as data centers try to complete the procedure within an ever-shrinking window of opportunity, as the amount of data to be saved grows from terabytes to petabytes, and beyond. In fact, backups have transitioned from removable media to disk, and then to disk with data deduplication in an attempt to reduce the data volume and accelerate the backup process. Unfortunately, not as much attention has been given to the even more critical *recovery process*. *Recovery Time Objectives (RTOs)* are even more important than the time to do a backup. When systems go down, every hour, in fact every minute, of lost time can equate to thousands, or even hundreds of thousands, of dollars. The time that it takes to restore data is of the utmost importance. Now, IBM is focusing on the importance of the restoration time for this mission- and business-critical data.

With enhancements to the *ProtecTIER* deduplication platform, IBM has extended their data protection and retention portfolio to enable the data center with the capability to **accelerate the restore time** for an exploding data store, in order to meet the most stringent RTOs. To learn more about how improvements to ProtecTIER can get your enterprise servers and storage platforms back on-line faster, please read on.

IN THIS ISSUE

- Data Center Issues in 2011 2
- IBM ProtecTIER Enhancements 3
- IBM's ProtecTIER vs. EMC's Data Domain 4
- Conclusion 5

The Clipper Group, Inc. - Technology Acquisition Consultants ◆ Internet Publisher

One Forest Green Road ◆ Rye, New Hampshire 03870 ◆ U.S.A. ◆ 781-235-0085 ◆ 781-235-5454 FAX

Visit Clipper at *www.clipper.com* ◆ Send comments to *editor@clipper.com*

Data Center Issues in 2011

The data center of 2011 is faced with a series of mounting storage issues, the most serious of which is how to address the requirements for data protection and retention at a time when storage requirements are doubling every 12 to 18 months. In fact, the data center has progressed from a generation where storing a terabyte was a big deal to a time when storing many petabytes (PB) is a reality. What does the future hold for the storage of exabytes and zettabytes? The recovery time for this data becomes all the more important.

Data protection and retention issues fall mainly into two major categories: *archive* and *backup*.

Archive Data

Archive data, in most cases, is stored forever[1], with a disk-to-disk (D2D) architecture, in a disk-to-tape (D2T) environment or, perhaps, in a disk-to-disk-to-tape (D2D2T) configuration. Enterprises are afraid to delete anything for fear of losing data that *may* become useful five or ten years down the line, for fear of violating some governmental regulation, simply in fear of litigation, or fear from losing something that may some day provide a competitive advantage. The recovery time for archived data can be measured in minutes or hours, certainly not in sub-seconds, so high availability is not a critical issue. Cost and data integrity are more important in archiving than recovery time. In fact, we have seen that the total cost of ownership (TCO) for *long-term* storage requirements, where *immediate* retrieval is not a requirement, favors tape[2]. However, the immediacy characteristic of the recovery process for backup data often tilts the decision-making process back to disk.

Backup Data

Because of its use in the present, as opposed to some much more distant time, backup data tends to have a shorter lifespan in nature. The requirement to recover immediately often dictates that the backup process needs to be D2D for the performance levels intrinsic to the immediate restoration of data. High availability is a major issue here, as rapid access to the backup data is a critical consideration. Data deduplication also is becoming more critical to the backup/recovery process as the amount of duplicated data continues to grow.[3] By eliminating duplicate data *before it is replicated* (via backup or other process*)*, deduplication can reduce the bandwidth needed to transmit the data by up to 95% or more (but, most certainly, your "mileage" will vary depending on the nature of the data and levels of duplication). In addition to accelerating the transmission, deduplication also reduces the amount of storage needed to store and protect mission-critical data.

As we have seen with natural phenomenon, disasters will vary in nature from human error, to fire, to malicious destruction (by virus or worm) or accidental destruction of data. In fact, the same types of natural disasters that threaten the community at large can have a devastating effect on a physical data center. Each data center will have its own requirements for the retention period of backup data on disk before it is deleted or archived, but generally speaking, 180 days to a year (or even as many as five quarters) is common, and even longer depending on data center and enterprise policies. The entire backup and recovery process can be exacerbated by virtualization, with the deployment of many virtual machines (VMs) increasing the amount of data being processed by a single platform, through a single pipe, due to the number of partitions on the server and the potential for redundant system files being saved for each VM image.

That said, the data center is constantly in search of a data deduplication appliance or gateway with the scalability, reliability, and performance to handle current and future requirements, as well. *What are the real deduplication issues that cause the data center manager to lose sleep?*

When a vendor, any vendor, tells you that you can reduce your data backup capacity by 95%, or more, your budgeting antennae must go up. *How can I achieve such savings?* The details of the benchmark in terms of how much

[1] Even if it is stored for a known fixed time, say ten years, that is long enough to be treated similarly to data that truly will be kept forever.

[2] See the issue of Clipper Notes dated December 20, 2010, entitled *In Search of the Long-Term Archiving Solution – Tape Delivers Significant TCO Advantages over Disk,* and available at http://www.clipper.com/research/TCG2010054.pdf.

[3] There are two general reasons that data is duplicated, potentially many times. (1) It may have been distributed widely and stored by many of the recipients individually. Good examples are office documents and audio and video files. (2) Copies may have been made for special purposes (often these are *snapshots*) or the same data may have been backed up many times (i.e., over and over again, with each backup).

duplicate data is available and how frequently the backups are run will go a long way toward revealing how appropriate the benchmark is for your enterprise. If you run daily full backups, higher deduplication rates are achievable. Unless you have a clear idea about what data is stored and being added or changed, a realistic deduplication ratio is in the 20:1 to 25:1 range.

Are the published numbers applicable to a production environment or are they specific to some idealized state where everything is a duplicate and nothing new is being stored, except into a hash table? Let's all understand, however, that benchmarking is useful only when the test is being run against *your* data! Everything else is not benchmarking, it is *benchmarketing*!

Backup times are often quoted so that the IT staff can determine if the product will meet the enterprise backup window requirement. How often, however, can you ascertain the projected full recovery time? When you really need to recover data, you REALLY need to recover it, and quickly.

What is the scalable capacity of the appliance? What are the throughput expectations, for both backup *and* recovery? How robust is the solution? Is the system highly reliable? Can you trust the integrity of the operation? Is the platform flexible enough to operate in a heterogeneous environment? Apparently, IBM's customers are asking these same questions, as IBM recently has released an updated version *of ProtecTIER* to improve the throughput and scalability of their deduplication gateway and appliance solutions[4], to address these very issues.

IBM ProtecTIER Enhancements

IBM has made dramatic improvements in backup and restore performance across the IBM ProtecTIER deduplication family since the previous release. Taking advantage of software advancements made by IBM research and development scientists, ProtecTIER deduplication algorithms have been streamlined. These enhancements deliver improved backup performance and, more importantly, greatly increased restore performance. ProtecTIER's unique and patented deduplication technology is a non-hash-based approach to eliminating redundant data that avoids the possibility of data loss due to a hash collision.

What does this mean for the data center staff trying to deal with a never-ending growth pattern for mission- and business-critical data? IBM claims that ProtecTIER offers industry-leading inline deduplication performance and scalability up to one petabyte of useable storage capacity. This means that a single ProtecTIER system can store up to 25 PB (or more) backup data, assuming a 25:1 deduplication ratio. With the code enhancements, a single-node ProtecTIER gateway[5] can deliver sustained backup performance of up to 1400MB/second (5TB per hour), up from 900 MB/second, with a single-stream performance of up to 100MB per second[6]. A dual-node gateway cluster, providing high availability and global deduplication, with a single repository, can achieve a sustained rate for backup performance of up to 2000MB per second, or 7.2TB/hour, up from 1200MB per second. Even more significant are the improvements in restore performance, with up to 1800 MB/second (6.4TB/hour) for the single node, and up to 2800MB/second (10TB/hour) for the dual node cluster. Furthermore, ProtecTIER performance does not degrade as capacity increases. Performance remains constant, and can even improve as capacity is scaled upwards. **With ProtecTIER, backup is fast; but restore is even faster!**

IBM's ProtecTIER works with all major variety of backup and recovery software applications, including *Tivoli Storage Manager,* Symantec *NetBackup,* Symantec *Backup Exec*, EMC *NetWorker*, CommVault *Simpana*, and many others. IBM also offers a higher level of integration than before between its ProtecTIER deduplication solutions and the Symantec *OpenStorage (OST)* API.

So, how does this stack up against the other elephant in the room? Not too badly! In fact, IBM ProtecTIER solutions seem to compare very well when put head-to-head with EMC's *Data Domain* platform.

[4] See **The Clipper Group Navigator** entitled *Reversing the Requirement for Storage Growth – IBM Consolidates and Simplifies Tier-2 Storage* dated February 25, 2009, and available at http://www.clipper.com/research/TCG2009008.pdf.

[5] Single-node means one quad-processor System x3850 X5 or x3950 X5.

[6] Each backup job is a single stream assigned to one virtual tape drive. Each backup server can have dozens of backup jobs running at a time, with ProtecTIER capable of 256 virtual tape drives, accessible by multiple backup servers, simultaneously.

IBM's ProtecTIER vs. EMC's Data Domain

As mentioned above, you must assume that performance values cited by vendors are developed using a carefully tailored data set that will demonstrate their own superiority, unless specifically cited as results achieved from production environments. Before making any acquisition decision, you should review the performance numbers and deduplication rates from all vendors for their specific enterprise environment. That caveat being said, let's look at the claims.

With a claimed backup performance of **2250 MB per second**, or **8.1TB/hour**, EMC's *DD890* would appear to be a clear winner for single node backup performance as compared to ProtecTIER at **1400MB/second**. However, as we have seen above, figures may not lie, but many benchmarks often skirt the truth (or, at least, your situation's truth). EMC's benchmark is run using a VTL interface and 8Gb Fibre Channel (FC) controllers. *Does this match your environment?* Is the DD890 performance cited for *peak* performance or *sustained* performance? EMC quotes a throughput rate of **14.7TB/hr** for the DD890 using *DDBoost* to offload deduplication processing to the backup servers. *Does DDBoost work tightly with your backup application? What impact will performing deduplication have on your backup servers?* The IT staff also must carefully examine the benchmark dataset being backed up. Is it typical production data or has it been optimized so that there is 100% duplicate data? This would mean that the only data being transmitted (after the initial backup) is hash information going into the hash table. *Nothing is being changed*. I seriously doubt that the average enterprise would be running a backup with no changes.[7] With the ability to scale up to 1PB of useable capacity, ProtecTIER has significantly more scalability than Data Domain's DD890 whose raw capacity maximum of 384TB translate to approximately only 250TB of usable capacity. You may ask: "What about Data Domain's *Global Deduplication Array* (*GDA*)?" In fact, the GDA appears to be two DD890s tied together to share a single storage pool, the storage is not globally mapped, and each controller only has access to its own portion of the data. In terms of maximum capacity, IBM quotes the ProtecTIER deduplication ratio at 25:1, in line with industry estimates for data deduplication. Even though IBM and others quote deduplication ratios in the 20:1 to 25:1 range, EMC quotes their maximum capacity based upon a 37:1 ratio, putting their maximum logical capacity in doubt for many enterprises. (Again, your mileage may vary). In terms of high availability, the GDA does not support failover across controllers, creating a single-point-of-failure, putting access to all of your data in doubt in the event of a controller outage. Therefore, we will confine the comparison to the DD890.

Can you communicate with real users achieving these rates in a real production environment? What is a realistic maximum backup performance for these platforms? IBM has estimated throughput for the DD890 to be in the 800-1300MB/second range. Even allowing for competitive gamesmanship, this would put the DD890 and ProtecTIER in a similar position for backup. However, what about Restore?

IBM quotes a single node restore rate of **1800MB/second** (**4.7TB/hour**) for a ProtecTIER system in a production environment. The dual node cluster rate is quoted at **2800MB/sec** (**6.4TB/hour**). I would like to compare this to the DD890; however, I cannot, as EMC does not publicly quote a restore rate for the DD890. Therefore, you should ask EMC. Even better, you should run that platform against your data in order to know for sure (and that goes for ProtecTIER, as well).

Interestingly enough, an environment used to generate superb backup performance does not always provide the best restore results. For example, every read has to be un-deduped, resulting in no advantage for a system that has a very high rate of duplicated data. The DD890 may gain no restore advantage from DDBoost. In fact, DDBoost may actually hinder restore performance, as the restore process will probably be bottlenecked by disk I/O performance. A realistic maximum restore performance may be less than 1000MB/second, according to IBM.

In addition, while EMC's DD890 requires DDBoost to achieve an advanced integration with EMC NetWorker, Symantec *NetBackup*, and Backup Exec, ProtecTIER operates effectively with these applications as well as with *Tivoli Storage Manager (TSM)* and all other major backup applications. It should be noted that all of EMC's benchmarks were run with NetBackup using the OST option.

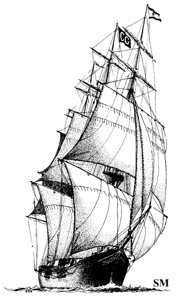
[7] The typical production environment might expect at least a 10-25% change rate.

Conclusion

When it comes to data deduplication, the two big elephants in the room are EMC with their Data Domain product set and IBM with their ProtecTIER family. In addition to data deduplication, IBM has a full portfolio of data protection and data retention products including both disk and tape, enabling them to provide a wider range of solutions, for both long-term and short-term data protection problems. EMC only offers disk. No matter what the question, EMC's answer seems to be "use more disk," which may not always be the best response when the optimum TCO solution calls for tape.

With regard to the specific issue of throughput performance in the recovery mode, we know that, based upon IBM's own projections, ProtecTIER's restore performance is even faster than their backup throughput. We cannot really be sure about Data Domain, as EMC does not disclose restore performance figures. IBM claims to be quoting real user's backup/recovery results in a sustainable mode, while EMC's figures may be citing benchmark results for peak mode, not in a production environment. While backup performance may be comparable, restore performance appears to tilt the scales toward IBM; however, you need to benchmark this with your own data!

Clearly, IBM has the edge in capacity, scalability, performance, and also in data integrity with their unique (non-hashing) deduplication technology. IBM also has superior flexibility with a wider collection of back-end solutions, including their own disk storage families and supporting storage offerings from other vendors, including HP, EMC, HDS, and others. With software compatibility across a wide range of data protection applications, ProtecTIER enables the data center to deploy deduplication without disrupting existing data center operations. If you are looking to maximize your restore performance, take a good look at IBM's ProtecTIER deduplication solutions for both your short-term and long-term requirements.



About The Clipper Group, Inc.

The Clipper Group, Inc., is an independent 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.