



EMC NetWorker and Avamar — An Integrated Pair for Traditional and Deduplication Backup

Analyst: Michael Fisch

Management Summary

Data deduplication has become mainstream technology for backup and restore. It solves a pressing problem enterprises face today with the growing amount of digital information and the need to protect it with consistent, timely backups and fast recoveries.

Client (source) deduplication and *storage (target) deduplication* are the two basic types of solutions. These two are often portrayed as competitive solutions but, in fact, they are complementary and offer different advantages. Enterprises may choose to employ one or both. In general, client deduplication provides faster backups, while storage deduplication is very easy to deploy in existing backup environments.

EMC has integrated its widely deployed and well-established backup application, *EMC NetWorker*, with next-generation client deduplication from *EMC Avamar*. This integration effectively creates a unified platform to deliver next-generation backup and recovery alongside traditional approaches, e.g., tape, for those customers who want a more evolutionary approach to adopting data deduplication for disk-based backup. Read on for details.

Data Deduplication – Mainstream Technology

Data deduplication has become mainstream technology for backup. In fact, the industry has nearly reached the point where the majority of large enterprises employ deduplication in their backup infrastructure. The capacity savings it delivers and the faster backups and restores it enables are simply too good to pass up.

Deduplication Facilitates Disk-Based Backup

One could say that deduplication is to backup what telephones are to interpersonal communication. People have always talked and communicated. This is an essential part of human relationships. With the development of telephones, talking became easier, faster, and less costly because it eliminated the barriers of distance. People no longer had to travel across town or across the country to have a conversation. They just picked up a phone and said “hello.”

In a similar way, enterprises always have backed up data. Business operations depend on continuous access to critical information. Backup is the standard way to protect information from corruption, system failures, and good, old human error. The development of data deduplication has made backup and recovery easier and faster because it facilitated the broad adoption of disk-based backup. Deduplication technology can reduce backup data by 20 to 50 times, or more. Such

IN THIS ISSUE

➤ Data Deduplication – Mainstream Technology	1
➤ EMC NetWorker and Avamar	3
➤ EMC NetWorker and Data Domain.....	4
➤ Conclusion.....	5

a dramatic reduction in capacity requirements makes backup to disk economically viable as a replacement for tape at least for the short and medium term, if not completely. In the new world of disk backup, enterprises no longer need to physically handle and manage a proliferation of backup tapes. Servers simply back up to disk over a network. And the data is ready and waiting, if needed, for a fast restore.

Disk-based backup is the most significant backup trend of the last decade. It all starts with the ever-rising tide of data with which enterprises have to contend. The norm is 30 to 100% annual data growth. Enterprises have to store and protect this data without growing the IT budget at a similar rate. In many cases, tape has become a weak link in the backup and recovery chain because backup jobs may not complete or they overlap into production hours. Recovery from tape may be too slow or unreliable to meet the uptime requirements of the business. Tapes can be misplaced or damaged during handling. The tape-centric backup approach is outmoded and may no longer be sufficient to meet current data protection requirements.

In contrast, disk is a faster storage medium, especially because it can process random reads and writes immediately. It is easier to manage than tape because disk does not require physical loading and unloading, offsite transport, and management. While tape is still the lowest-cost option for long-term data archiving, a large number of enterprises have adopted disk-based backup to overcome the aforementioned limitations. Data deduplication is the enabling technology because it shatters the cost barrier that stood in the way of widespread adoption of disk-based backup.

Since deduplication is an attractive technology – should an enterprise just plug it into their backup infrastructure? Well, yes, though the answer is nuanced. The best approach to integration will depend on your enterprise's current backup infrastructure and objectives.

Recall that there are two fundamental types of deduplication solutions for backup: client deduplication and storage deduplication – also called source and target deduplication, respectively. The two types are often portrayed as competitive or alternative solutions, but they are in fact complementary, and enterprises may choose to employ one or both since they possess different advantages.

Client Deduplication

Client deduplication identifies and eliminates redundant data at the client (typically a server) before it is sent over the network to the backup server or storage node. After an initial full backup, all subsequent backups transmit only changed, sub-file data. Well-designed client deduplication solutions can identify new and changed data, not just within a single client, but also across an entire data protection domain. Client deduplication is generally much faster than traditional backup because less data is backed up each time. This reduces network traffic and storage capacity requirements. Even with the additional deduplication processing at the client, backup jobs are faster as long as the server's incremental rate of data change is not too high (i.e., less than 10 to 20% daily).

Client deduplication is particularly effective for the following activities.

- **Remote and branch office backup consolidation** – Client deduplication makes it viable to send backup data over a WAN connection to a central data center. In this case, backup is more reliable because it is no longer dependent on tape and potentially inconsistent process execution by non-IT staff at remote sites.
- **Server virtualization** – Server virtualization solutions like *VMware*, *Hyper-V*, and *Xen* have dramatically increased physical server utilization, though the multiplication of virtual machines on a single server has increased the backup burden while leaving fewer system resources available for it. Client deduplication addresses this problem by providing faster and more efficient backup for virtualization environments.
- **Desktops and laptops** – Similar to remote and branch offices, data deduplication makes it viable to send backups over an Internet or LAN connection.
- **File servers** – File servers tend to be candidates for client deduplication backup because of the nature of their data, with relatively low rates of data change and the space and times savings from backing only altered file segments.

Storage Deduplication

Storage or target deduplication identifies and eliminates redundant data at the storage device during or after the backup. This tech

nology often is found in network-attached storage systems as well as virtual tape libraries (VTLs), which present themselves as tape media to the backup application while writing deduplicated data to disk.

Storage deduplication is particularly effective in the following situations.

- **Existing backup processes and policies** – Storage deduplication is virtually a plug-and-play solution for existing backup environments. This is a major reason why it is so popular for enterprises that want to enjoy the benefits of data deduplication without overhauling their backup infrastructure.
- **Transactional applications with a high rate of data change** – When a transactional application server is connected to a backup server via a high-speed SAN or LAN, storage deduplication can be faster when the database is heavily utilized and has a high rate of data change.
- **High-priority recovery** – Host servers with large volumes of data that require the fastest recoveries may be better served by storage deduplication.

Assessment Questions

Finally, here are some questions to help sort out which deduplication approach may work best for your backup environment:

- *What types of applications need to be backed up? Are they file-based or transactional? Is the data rate of change high or low?*
- *Is your organization geographically distributed or centralized? A distributed organization would benefit more from client deduplication.*
- *Do you want to transition to all-disk backup or will tape continue to play a significant role? Storage deduplication with a traditional backup application may be better, if tape will continue to play a major role.*

EMC NetWorker and Avamar

EMC has in its product portfolio the industry-leading solution for client deduplication, *EMC Avamar*, and has integrated it with *EMC NetWorker*, its unified backup software application that provides support for a wide range of data protection capabilities. The combined solu-

What Is Deduplication?

One could describe deduplication as a bouncer who stands at the door and checks data for redundancy. Only original (unique) data gets in, and the rest is asked to please step aside. This technology scans for repetitive data segments and replaces subsequent occurrences with pointers to the original. Since backup data is highly repetitive by nature, deduplication is especially effective in this context.

There are multiple deduplication techniques. File-level deduplication identifies and eliminates redundant files. This technique consumes relatively few system resources, as in CPU cycles and memory, though it also delivers the smallest space savings. Fixed-block, sub-file deduplication scans for redundancies in bit-level segments of equal size. It delivers high space savings and is more resource-intensive. Variable-length, sub-file deduplication operates on bit-level segments of variable size, which are optimized by intelligent algorithms to find natural breaks in a file and maximize data reduction. It delivers the highest space savings and is more resource-intensive. Additionally, data compression often is paired with deduplication to achieve greater results together. The difference is that compression transparently re-represents (i.e., “squeezes out”) repeating or positional data within a file, while deduplication operates more broadly across files.

The space savings that users will experience depend on three things: the nature of the data, the deduplication techniques employed, and the scope over which it is applied. The more redundant a data set is, the more amenable it is to reduction. Again, backup tends to be the most redundant because the same data is backed up repeatedly. Different techniques yield different degrees of space savings. Variable-length, sub-file deduplication is arguably the most effective. However, the optimal technique or combination of techniques for a given application should factor in space savings as well as resource consumption and effect on operations. Finally, the scope includes the amount of data deduplicated (e.g., single server or dozens) as well as the span of time over which deduplication is applied.

tion creates a single, integrated framework for introducing next-generation backup while at the same time allowing continued support for tape, snapshots, replication, and more. Users can manage, schedule, and initiate both deduplicated and non-deduplicated backups from a single user interface. It offers a convenient solution to the many customers who require an evolutionary approach to adopting data deduplication for disk-based backup.

EMC Avamar

Avamar is a deduplication backup application that reduces backup data at the client before transferring it across a network for storage on disk. It employs variable-length, sub-file deduplication to maximize savings in bandwidth and disk storage. Through a sophisticated identification and communication process, Avamar eliminates redundant backup data within and across protected systems, wherever they exist in the data protection domain. In fact, Avamar reduces the required daily network bandwidth consumption by up to 500 times and cumulative backend storage can be reduced by up to 50 times (depending on the nature of the data) across sites and servers.

Avamar uses a grid architecture that scales by adding storage nodes containing CPU, I/O, memory, and disk resources. Data is automatically and non-disruptively load-balanced across new nodes. For high availability, Avamar uses *RAIN (Redundant Array of Independent Nodes)* technology that provides redundancy and fail-over and eliminates single points of failure. Data within nodes is stored in a RAID 5 configuration. Avamar can also replicate deduplicated data asynchronously to a remote site for disaster recovery.

Avamar supports source deduplication of databases related to applications, such as Microsoft *Exchange*, *SQL* and *SharePoint*, *Oracle*, *IBM DB2*, and *Lotus Domino*. It is integrated with VMware for backing up virtual machines. Avamar automatically backs up servers in remote offices and even desktop and laptop PCs over a WAN connection to an enterprise's main data center, thereby consolidating the backup process. Avamar recently added the ability to export deduplicated backup data to tape for cost-effective long-term archiving.

EMC NetWorker

EMC NetWorker has progressed from a traditional tape-centric backup solution to one that

now embraces a range of next-generation capabilities for data protection. This is a product with a long history and substantial installed base. NetWorker provides centralized backup and recovery for a wide range of heterogeneous environments, both physical and virtual. Its features include support for SAN, NAS, and DAS storage, a variety of applications and operating systems, off-host "hot" backup while applications are in use, and 256-bit AES file encryption and user authentication for data security. Today, NetWorker also supports data deduplication through integration with Avamar and storage deduplication solutions, like Data Domain.

NetWorker and Avamar Integration

EMC's integration of NetWorker and Avamar has produced a tightly coupled solution. By combining the NetWorker and Avamar clients into a single backup agent and managing Avamar within NetWorker resources, administrators can manage the entire backup environment through a familiar user interface and common workflow. This reduces complexity by simplifying the day-to-day management and further consolidating the backup infrastructure. Backup procedures do not have to change to adopt client deduplication. You also have one go-to vendor for backup implementation and support.

To set up the integrated solution, first install Avamar and NetWorker in your backup environment. Then install the NetWorker client on the Avamar server and input the Avamar server name and login credentials into NetWorker. NetWorker will then represent the Avamar Data Store as a deduplication node on its console.

To configure a new client for deduplication backup or convert an existing one, just check the box in the client properties and choose the Avamar server to which data should be sent. NetWorker will track metadata about the deduplication backup locally while the protected data is sent to the Avamar server. Administrators can establish deduplication backup policies and schedules, monitor activities, and view backup summary results.

Administrators also can recover deduplicated data from the same user interface as traditional backup data. NetWorker automates the whole process, so there is no need for guesswork at recovery time, in terms of how data was protected or where it is located.

The integration of Avamar and NetWorker enables a flexibility and precision in how

backups are structured. For instance, it is possible to deduplicate backups operationally and occasionally write a full backup to tape for compliance and long-term archiving? This would leverage both Avamar and NetWorker backup facilities, respectively. It is also possible to parse out files within an application and deduplicate those with a low rate of change while not deduplicating others that consist mainly of unique data. So, this integration enables a finer balance of time, cost, and factors like compliance in structuring backups to meet business requirements. Furthermore, this flexibility extends to when to use the integrated approach, and perhaps when not to. If an enterprise uses Avamar for remote office backup and NetWorker in the main data center, and wants to implement deduplication backup for select host servers in the data center, it can choose to manage only those servers in the data center using the NetWorker console, while continuing to manage remote offices through Avamar. This use case might seem esoteric, though it illustrates the evolutionary way the Avamar and NetWorker integration can be applied as enterprises adopt next-generation backup technologies.

Deduplication reporting within NetWorker shows data moved versus protected and provides statistics on commonality ratios achieved. This information is useful for demonstrating the effectiveness of deduplication in reducing capacity and bandwidth requirements as well as making a case for return on investment.

An exception here is the Avamar remote replication feature. Enterprises that want to replicate Avamar storage nodes to a remote site will need to set that up apart from NetWorker.

EMC NetWorker and Data Domain

EMC also has in its portfolio the industry-leading solution for storage deduplication, EMC Data Domain, and has integrated it with NetWorker. Data Domain is an inline storage deduplication appliance that slips into a backup environment by connecting to the backup application as either a file server (CIFS, NFS) over an Ethernet network or as a VTL over a Fibre Channel network. It delivers high-performance throughput and typically 10-to-30-times data reduction. Data Domain also offers replication of deduplicated data to a disaster recovery site.

Data Domain is qualified with all major enterprise backup applications, including NetWorker. We anticipate EMC will announce some

exciting new capabilities this year for integrating Data Domain with NetWorker. Stay tuned.

Conclusion

Data deduplication for backup should be on your IT planning list, if it is not already in your data center. In many cases, this technology will be necessary to adequately protect your enterprise's growing data. Backup to and recovery from disk is simply faster, and client deduplication offers unique advantages for remote offices, desktops and laptops, and server virtualization.

So deduplication becomes a question of how, not if. With the integration of NetWorker and Avamar, EMC is offering an evolutionary and pragmatic approach to inserting deduplication into your backup process. For enterprises that do not have the luxury of starting from a blank slate with their backup infrastructure, the ability to deploy traditional and deduplication backup using a single client agent and manage both from a single console offers an attractive solution. Consider it as your enterprise looks at how to adopt deduplication and back up to disk.



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

Michael Fisch is a Senior Contributing Analyst for The Clipper Group. He brings over 14 years of experience in the computer industry working in marketing, sales, and engineering, the last nine of which he has been an analyst with Clipper. Before Clipper, Mr. Fisch worked at EMC Corporation as a marketing program manager focused on service providers and as a competitive market analyst. Before that, he worked in international channel development, manufacturing, and technical support at Extended Systems (since acquired by Sybase). Mr. Fisch earned an MBA from Babson College and a Bachelor's degree in electrical engineering from the University of Idaho.

- ***Reach Michael Fisch via e-mail at mike.fisch@clipper.com at 781-235-0085 Ext. 211. (Please dial "211" 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.

Disclosure

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.