IDC
Analyze the Future

WHITE PAPER

Backup and Recovery: Accelerating Efficiency and Driving Down IT Costs Using Data Deduplication

Sponsored by: EMC Corporation

Laura DuBois Robert Amatruda

February 2010

Global Headquarters: 5 Speen Street Framingham, MA 01701 USA P.508.872.8200 F.508.935.4015 www.idc.com

EXECUTIVE SUMMARY

Data deduplication is dramatically improving IT economics by minimizing storage footprint requirements, backup windows, and network bandwidth consumption in distributed enterprises and datacenter locations alike. In real-world environments, deduplication is accelerating backup and recovery efficiency and driving down IT costs. This white paper looks at the various approaches to deduplication for backup data and outlines the considerations in selecting a solution. It also highlights EMC's portfolio of deduplication offerings for backup and recovery and explores specific use cases for optimal backup efficiency and cost reduction.

Deduplication Adoption

The demand for data deduplication in both midsize and enterprise environments is escalating as firms look for ways to keep pace with the near doubling of storage growth annually. This growth is fueled by new applications, the proliferation of virtualization, creation of electronic document stores and document sharing, use of Web 2.0 technologies, and the retention or preservation of digital records. With constrained IT budgets, the need to curb data growth is heightened as firms look to reduce capital and operating costs. From a physical perspective, many datacenter managers are also dealing with limited infrastructure in terms of power, cooling, and floor space. Deduplication is a technology that not only aids in accelerating storage efficiency by reducing cost but also alleviates physically constrained datacenters.

Deduplication also addresses challenges associated with management, backup, and network inefficiency. As data grows, there is an increasingly disproportionate relationship between the number of IT personnel and the amount of storage requiring management. Deduplication reduces the data footprint, keeping this ratio in balance. Similarly, as the gap between server processing power and disk continues to widen, firms are looking for ways to improve performance throughout their environment over a WAN, within disk storage subsystems, and across limited backup windows. Data deduplication technology can optimize available physical and virtual infrastructure by sending less data over local or remote network links. It can also improve service-level response times and help meet shrinking backup windows. Deduplication also makes use of random access media (disk), improving recovery times, data security, and reliability.

More recent challenges have come as a result of virtualization. As firms continue to deploy virtual machine technology to aid in server consolidation and disaster recovery (DR), the virtual machines process data that can be highly redundant but still needs to be protected. To account for different failure scenarios or to recover an image, a physical server and discrete files are typically required within a single backup solution and backup process. Deduplication offers significant backup storage capacity savings because it can eliminate the redundancy typically found in VMDK files. However, standard approaches such as deploying a traditional backup agent in a guest virtual machine and using a VCB proxy to create an image-level backup do not reduce the volume of virtual machine data that needs to be backed up or the local network bandwidth required to move the data. Deduplication in concert with backup software addresses the need for complete, efficient, and cost-effective protection of virtual machine environments.

The Benefits of Deduplication

Firms are deploying data deduplication in a number of places in the infrastructure stack to address these practical, real-world challenges. The benefits of deduplication include the following:

- **Driving down cost.** Deduplication offers resource efficiency and cost savings that include a reduction in datacenter power, cooling, and floor tile demands as well as storage capacity, network bandwidth, and IT staff.

- **Improving backup and recovery service levels.** Deduplication can significantly improve backup performance to meet limited backup windows. Deduplication technology also leverages random access disk storage for improved recovery performance compared with sequential access (tape) methods.

- **Changing the economics of disk versus tape.** Deduplication makes disk-based backup feasible for a wider set of applications. Tape still has a role in enterprise datacenters due to its economics and archival properties. However, cost/GB declines for disk when used with deduplication are leading to disk costs equal to or less than tape costs.

- **Reducing carbon footprint.** Deduplication reduces the power, cooling, and space requirements for storage, thus reducing carbon footprint and enabling environmental responsibility.

Deduplication technology addresses many of the long-standing backup challenges that firms large and small have been dealing with for over a decade. These challenges have included keeping up with doubling of data growth, meeting shorter backup windows, enabling faster recovery from operational and disaster-related failures, and the like.

Table 1 outlines the myriad of backup challenges that exist and how deduplication can address them.

TABLE 1

Backup Challenges and Deduplication Impact

Backup Challenges	Deduplication Impact
Backup windows are shortening as operations run 24 x 7 to meet global customer demands.	Traditional backups mean the transfer of vast quantities of redundant data, which can overrun tight or nonexistent backup windows. Deduplication in backup software can reduce the amount of data that needs to be backed up or fast inline deduplication storage systems can speed the performance of the backup target, each enabling more data to be backed up in an available window.
Recovery time requirements are becoming shorter to minimize the cost of downtime.	Deduplication reduces the cost of storing more backup data on disk. Keeping backups on disk rather than tape significantly improves recovery times for a broad set of applications.
Reliability of backups leaves data recovery at risk.	Reliance on tape media for backup introduces risk of media errors (bad media, contaminated heads, etc.), running out of available media, or hardware failures. Deduplication uses disk in the data protection process, eliminating or reducing these failure scenarios. Leveraging disk also facilitates health checks and other self-healing or failure prevention measures.
Increased server virtualization means fewer resources are available for backup, which can increase backup times and stress backup windows.	Deduplication can be used to eliminate shared resource processing of redundant data, reducing contention for physical resources and speeding virtual machine backups. Deduplication also allows longer retention of virtual machine backup data on a much smaller storage footprint, ensuring that operational recovery can happen quickly from disk, never from tape.
Data growth means not all data can be backed up in available backup windows.	Firms face on average 50% annual growth in the amount of data requiring protection. This growth is at odds with limited nightly backup windows and traditional methods. Deduplication addresses this growth challenge and enables efficient backup of growing data sets.
Secure offsite copy using traditional tape methods leaves data at risk due to loss or theft.	Reliance on removable tape media for offsite storage to use in the event of a disaster introduces risk that the physical media may become compromised. Deduplication, in concert with secure replication processes, enables an electronic copy to be kept offsite, eliminating the need for manual handling of tape media and therefore improving security.
Distributed data in remote branch offices needs centralized protection and recovery.	Remote branch locations are replacing standalone tape backup processes with a centralized edge to core backup approach for improved backup, recovery, and management. Deduplication makes the process of sending large volumes of backup data over congested WAN links to a centralized datacenter feasible.
Backup infrastructure costs are increasing to keep pace with capacity growth and backup windows.	Most firms deal with data growth and backup window challenges by putting more tape infrastructure in place. Adding tape drives and automation may address current performance bottlenecks and perform the backups more quickly but with cost and management overhead. Deduplication solves the problem at root cause to reduce the ongoing spend on tape infrastructure while keeping pace with capacity growth and backup window trends.

Source: IDC, 2010

DEDUPLICATION: WHAT, WHERE, WHEN, AND HOW

What Deduplication Is

IDC defines data deduplication as a technology that normalizes duplicate data to a single shared data object to achieve storage capacity efficiency. More specifically, data deduplication refers to any algorithm that searches for duplicate data (e.g., blocks, chunks, segments) and discards duplicate data when located. When duplicate data is detected, it is not retained; instead, a "data pointer" is modified so that the storage system references an exact copy of the data object already stored on disk. Furthermore, data deduplication alleviates the costs associated with keeping multiple copies of the same data object.

Data deduplication is most often associated with subfile comparison processes. This is different from single-instance storage (SIS), which compares data at the file or object level. Subfile deduplication examines a file and breaks it up into "segments." These smaller segments are then evaluated for the occurrence of redundant data content across multiple systems and locations. Deduplication is also different from compression, which reduces the footprint of a single object rather than across files or pieces of a file. Additionally, deduplicated data can also be compressed for further space savings.

Where Deduplication Occurs

Data deduplication for backup can occur at the source or target. An example of source deduplication would be reducing the size of backup data at the client (e.g., Exchange or file server) so that only unique subfile data is sent across the local or wide area network during the backup process. With source deduplication, the backup application has deduplication technology embedded in its architecture. An example of target deduplication would be reducing the size of backup data after it crosses the local network when it reaches a deduplication storage system. Deduplication at the source provides local and wide area network bandwidth, backup window, and storage savings. With target deduplication, the storage system itself has deduplication technology embedded in the storage controller. Deduplication at the target provides storage savings, works with existing backup software, and can reduce the wide area network impact of replication. Where deduplication is implemented not only yields different benefits but also affects implementation times and cost. Firms should evaluate their current backup problems and map these challenges to the different deduplication approaches.

Source Deduplication

Performing deduplication at the source (or client backup software) provides an extended set of benefits beyond storage capacity optimization. It means significantly less data is transferred from the source device to the storage repository, thus relieving congested virtual/physical infrastructure and LAN/WAN links. Because only new or changed subfile data segments are transferred from the source device to the storage repository, the amount of data moved is significantly reduced, enabling extremely fast daily full backups. The incremental overhead on the client CPU to perform source deduplication can be up to 15%, but the backup completes much faster than traditional methods — and some architectures provide throttling mechanisms to manage any short-term overhead increases. The overall impact of source deduplication is actually much less than that of traditional agents over a seven-day period. Source deduplication also offers deployment flexibility, since smaller remote offices can simply deploy a software backup agent. Environments with very large databases or databases with high daily change rates may want to consider a target solution instead. Fortunately, vendors typically have data assessment tools to help customers make the best choice.

Target Deduplication

Target deduplication optimizes backup disk storage capacity since only new, unique subfile data is stored to disk. All backup data is still sent to the deduplication target using traditional backup software, thus providing seamless integration into existing IT infrastructure. It only offers relief to available backup windows if the prior backup target, usually tape, was the bottleneck to backup solution performance. With target deduplication, the storage system itself (also called a deduplication storage system) is performing the deduplication to optimize data protection and disaster recovery performance while offloading the application servers from the deduplication process. Target deduplication is easy to implement by creating a fast, application-independent storage system (attachable as network-attached storage [NAS] over Ethernet or a virtual tape library [VTL] over Fibre Channel). No client software or other configuration is required. Deduplication storage systems are often used with larger data sets and databases. Target deduplication can also be used both in central datacenters for large volumes of data and in remote locations for local backup followed by replication to a central datacenter.

When **Deduplication Happens**

There are two different approaches available today for determining *when* the deduplication process occurs: inline or post-process. Inline deduplication eliminates redundant data before it is written to disk so that a disk staging area is not needed. Post-process deduplication analyzes and reduces data after it has been stored to disk, so it needs a full-capacity staging area upon which to start a deduplication process. In selecting an approach, organizations need to make considerations with regard to backup speed and disk capacity.

Inline deduplication is a more immediately efficient and economic method of deduplication. It reduces the raw disk capacity needed in the system since the full, not-yet-deduplicated data set is never written to disk. If replication is supported as part of the inline deduplication process, inline also optimizes time to recovery as the system does not need to wait to absorb the entire data set and then deduplicate it before it can begin replicating to the remote site.

Post-process deduplication implies waiting for the data to land on disk before initiating the deduplication process. This approach requires a greater initial capacity than inline solutions. Also, a post-process approach introduces lag time before deduplication is complete and also when replication will complete. There is also the risk of inconsistency between a local system and a remote system since there are two storage zones, each with policies and behaviors to manage.

How Deduplication Occurs

How the process of deduplication occurs depends on the implementation. A hash-based method of deduplication breaks up a file or backup stream into fixed or variable-length chunks of subfile data. A hash value is calculated for each segment. This process calculates a unique number for each segment, which is then stored in an index. If a file is updated, only the changed subfile data is saved; the changes don't require an entirely new file to be saved. An important distinction to be made with hash-based implementations is whether the segment size is fixed or variable in length. A variable-length approach can dynamically, based on content type, adjust a segment size to accommodate redundant data segments whose position has been shifted or offset in a byte stream during a file change. A fixed-length approach will not recognize redundant data that has been repositioned or offset, so it will inefficiently back up the segments again as they appear to be unique, even though they are already in the backup repository. The hash index is typically kept in memory, but as a hash index grows, it may spill over from memory into disk, requiring disk I/O for the lookup of hash values. Vendors have varying ways of dealing with these practical technology challenges, and consequently, the results range from eliminating the problem to significant performance degradation.

An alternative approach is delta-based data deduplication (also known as delta differencing or delta encoding), which stores or transmits data in the form of differences from a baseline copy. The baseline is a complete copy of the data used to recreate other versions of the data. Delta-based data deduplication may be performed at the block or byte level. Rather than using a hash to determine net-new data, a delta differencing method will scan and index an incoming data stream, looking for data that's similar to data already stored. Benefits of a delta-based approach include lower CPU utilization since there is no need to compute a strong hash. However, a delta differencing process requires significant disk I/O for comparing old data with the incoming new data. Therefore, the long-term advantage of each approach may depend on the relative performance improvements in CPU versus disk technology.

Another factor that may impact the deduplication ratio is whether or not the deduplication engine can recognize markers inserted into the data stream by the backup application or a particular data format (such as a backup application, Microsoft Exchange data, etc.). The ability to detect markers and the format of the data requires an understanding of where application-specific metadata is injected into a stream. If the deduplication engine understands the marker offsets or the format of the data, it can then tune the segment size so that it's ideal for the data format of the natural application, resulting in potentially greater deduplication results. However, leveraging this approach requires developing and maintaining an understanding of each backup application's (NetWorker, NBU, TSM, etc.) and user application's (Oracle, Exchange, etc.) changing formats.

CONSIDERATIONS IN EVALUATING DEDUPLICATION TECHNOLOGY

A number of different types of products with deduplication capabilities are available on the market today. Backup applications, appliances, virtual tape libraries, WAN optimization solutions, and primary disk storage subsystems all may have some form of deduplication functionality. It's important for a firm to agree upon what problems it is trying to address per application or data type before selecting the type of deduplication. Different deduplication approaches yield different capacity, performance, and network efficiency benefits.

1. **Deduplication ratios.** The deduplication ratio obtained will vary based on a myriad of factors, including data type, rate of data change, retention periods, variable versus fixed-length segments, backup policies, file format awareness, and the like. IDC research points to real-world total back-end disk storage deduplication ratios of between 8:1 and 22:1 based on the previously mentioned factors. Source deduplication solutions can reduce required daily network bandwidth by an order of magnitude compared with traditional full backup methods. However, as with all performance metrics, mileage will vary based on the environment. Firms must beware of throughput, scale, or performance guarantees offered and test the deduplication on premises with their own data sets.

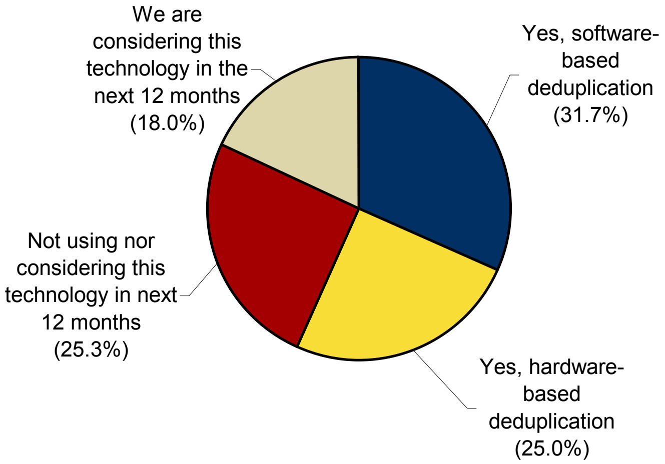
2. **Role of compression, encryption, and multiplexing.** Compression, the encoding of data to reduce its storage, can be a complementary technology to deduplication. Compression is optimized for a single object and reduces its footprint, while deduplication works across objects. However, compression can be applied to data that has already been deduplicated to provide further space savings. However, if deduplication is applied to a file already compressed (or encrypted), the benefit of deduplication will be negligible or nonexistent unless that compressed file is backed up again. Firms using deduplication along with compression may gain additional benefits if the deduplication of the data occurs first. Also to be considered is current usage of multiplexing for backups that interleaves data from multiple clients into a single stream sent to a tape drive. However, this process makes it difficult to detect segments of data that already exist. Multiplexing, a feature in most backup applications used to avoid shoeshining and improve performance when writing data to tape media, needs to be disabled if firms want to benefit from deduplication. However, in this scenario, disabling multiplexing does not erase any backup performance gains that it had enabled.

3. **Deduplication for virtual machines.** The use of virtual machines in production has heightened the need to protect and recover the virtual machine, the physical host, and files. Options for backup of virtual machines include image- and/or file-level backup using a backup agent running in a guest or on a service console or leveraging a virtualization vendor's backup API. Traditional backup solutions are inefficient for backup of virtual machines because they move large amounts of redundant data and require lots of CPU cycles to run a backup, resulting in poor backup performance and limited server consolidation. Deduplication can address these limitations. Source deduplication means duplicate data never travels across the shared underlying physical infrastructure, so daily full backups are fast and efficient. Deduplication can also occur globally, across VMDKs, to eliminate the backup of redundant data across virtual systems. This vastly improves the user's ability to recover virtual machines without tape and provides effective DR using the replication capabilities of deduplication systems.

4. **Deduplication for remote branch offices.** Like datacenter operations, remote offices require both local and disaster (remote) recovery. However, the characteristics of remote offices introduce challenges. Remote office locations typically have limited WAN bandwidth, no dedicated IT staff, and a disproportionate number of branch offices to regional or main datacenters. Deduplication can minimize data movement over the WAN and can eliminate redundant data across office locations and the datacenter. With limited IT staff at remote office locations, many firms are looking to reduce storage hardware footprint in distributed locations. Source deduplication can be deployed through software, which mitigates this concern. Smaller deduplication systems can also be deployed at remote sites when fast, local recovery is needed, and some vendors support replicating from these systems to a central datacenter. A recent IDC study looked at the use of deduplication in remote branch office locations and the type deployed (see Figure 1).

FIGURE 1

Use of Deduplication Technology in Remote Branch Data Protection



n = 300

Source: IDC's Remote Branch Special Study, 2009

5. **Deduplication for production/disaster recovery datacenters.** Large datacenters still struggle to meet their backup window for at least some of their applications and cannot afford to compromise backup performance. This reality may warrant a deduplication approach that includes source and target deduplication depending upon the application and environment. Optimizing network bandwidth within a datacenter may be less of a priority than remote replication to a disaster recovery site. But as backup windows continue to shrink, network bandwidth will become an issue over time.

6. **Deduplication and replication.** Replication is really the next battleground for deduplication technology. Established suppliers have proven that it works, when properly designed and implemented, and user reactions to the technology result in excitement and demand. Deduplication is being deployed in enterprise environments, in both edge and core locations, driving up efficiency while reducing infrastructure cost. The use of remote replication becomes vitally important as more firms minimize tape usage in remote locations but still need to support tape in a centralized location for archive and compliance use cases. User requirements for replication are becoming increasingly more sophisticated and include the following:

   - ❑ **Deduplication-aware replication that replicates a deduplicated data set and not a full volume.** Some vendors offer replication services with a deduplication-enabled product. However, firms must make sure the replication feature is deduplication aware.
   - ❑ **All-or-nothing and directory/tape-level replication.** Some use cases warrant a full system replication, while others may require flexibility to determine which shares or virtual tapes to replicate.
   - ❑ **Replication monitoring, performance tuning, and troubleshooting.** Despite deduplication, most large enterprises still have a lot of data to replicate. This is managed using a scheduled or asynchronous replication process, monitoring the replication process and the bandwidth used. Tuning and troubleshooting tools help to ensure the replication process stays on track within an available replication window.
   - ❑ **Scheduled and real-time replication for higher and lower latency links.** Some links/sites warrant real-time replication, whereas others may be fine with a scheduled replication process. Remote branch office characteristics vary significantly and may have lower latency links, while links between two datacenters may not face the same issue.

7. **Seeding and migration.** While deduplication is great for reducing the storage and/or transmission of redundant data, it does require an initial baseline or first backup to be established. For edge to core deduplication and replication, users need to consider how to create this baseline over bandwidth-constrained links. Most vendors offer some form of seeding service to quickly create this baseline, either through a bulk deduplication-aware replication process with systems side by side or by using a series of tapes from a last full backup and restoring them locally into a deduplication system. With storage refresh cycles on a three- to five-year rotation cycle, other considerations include how a migration is done and how disruptive it will be to an existing environment.

8. **Vendor selection.** Vendors make many claims and statements with regard to their deduplication approach. IDC research shows that not all deduplication products generally available actually work as advertised. Firms should consider how long a particular deduplication-enabled product has been shipping, how many customers are using the product in production, and how mature the product is in real-world environments. Firms should fully investigate the scalability of a product: Ask for user references and an application and/or system support matrix. Firms that choose not to conduct a proof of concept (POC) run the risk of finding surprises in performance and reliability.

9. **Use cases for deduplication.** Deduplication is a technology that promises to move further up the storage infrastructure stack. To date, the technology has largely been deployed in the backup arena where a large amount of redundant data already exists. This same data is backed up every week — calling on unnecessary server, network, and storage resources. Some firms are starting to look at or test existing deduplication in primary storage environments within a NAS approach. However, this implementation requires improved performance to avoid latency and response time implications. Today, deduplication technology is well-positioned for backup of virtual machines, remote and branch offices, and datacenter environments.

EMC'S PORTFOLIO OF DEDUPLICATION SOLUTIONS

EMC offers a broad range of backup and recovery products and services to assist customers with driving down IT costs and accelerating backup efficiency. Backup solutions enabled by deduplication include EMC Avamar deduplication backup software; EMC Data Domain deduplication storage systems; and EMC NetWorker, which can be deployed with Avamar, Data Domain, and other third-party target systems. Additionally, although not included in the scope of this paper, EMC offers a deduplication solution for primary storage and backup data with its network-attached storage EMC Celerra system and a disk archive deduplication solution with its Centera product line.

EMC Avamar

EMC Avamar deduplication backup software includes integrated deduplication technology to identify redundant data at the source, minimizing backup data before it is sent over the LAN/WAN. With Avamar, a firm gains data reduction and fast, daily full backups for VMware environments, remote offices, desktops/laptops, and datacenter LAN and NAS servers. Avamar also deduplicates backup data across sites and servers and over time. Unlike products that utilize traditional recovery methods, Avamar can quickly restore data in a single step — eliminating the hassle of recovering the last good full backup and subsequent incrementals to reach the desired recovery point. Avamar capabilities are a fundamental departure from traditional backup applications.

The Avamar agent keeps track of files that are new or have changed. The agent does not need to walk the entire file system tree to identify new or changed data and will check the local file cache for those files first. Upon identification, the agent will break the new or changed files into subfile, variable-length data segments and assign a hash value (unique ID) to each segment. The agent will then check the local hash cache to see if the hash already exists and has been previously backed up. If it has, it does not back it up again. Finally, the agent communicates with the Avamar server to determine if the hash is unique or already exists. If the data segment is new, it will be sent across the LAN/WAN during the daily full backup.

These processes will increase the CPU utilization on the host compared with a traditional backup agent. However, because the backup is efficiently protecting only net-new data segments, Avamar backups complete significantly faster than traditional full and incremental backups. For example, an incremental backup that typically took 10 hours might take closer to 1 hour to complete with Avamar, thus cutting down the weekly impact of backup from 50 hours to 5 hours for Monday through Friday incrementals. And Avamar's daily full backups are an order of magnitude faster than traditional full backups.

Additionally, Avamar now offers desktop/laptop protection that operates in the background using existing network links and does not throttle valuable CPU cycles. User data is automatically backed up at log-in during normal backup windows, or it can be user initiated, enabling users to recover their data at will.

Avamar backup and recovery solutions provide source and source deduplication, making it ideal for firms with the following environments:

- Deploying virtual machines and evaluating a new protection strategy to recover physical servers, virtual servers, and discrete objects

- Improving their remote branch office backups to gain fast, daily full backups; centralized management; improved reliability; secure replication; and reduced backup traffic over congested WAN links

- Seeking to curb data growth, backup windows, and network traffic for backup of local NAS and file server environments

- Protecting valuable desktop/laptop data, including offices and mobile professionals

EMC Avamar deployment configurations are as follows:

- Client options

  - ❑ Avamar software agents are deployed on the systems to be protected (clients) with no additional local hardware (i.e., media servers) required. Agents exist for most operating systems and for leading applications and databases. Unlike many backup applications, Avamar does not charge for its client software; instead, it relies on deduplicated capacity-based licensing. This can result in extremely cost-effective implementations and subsequent expansion of clients.

- Server options (backup repository)

  - ❑ Third-party Avamar servers. Avamar software can be purchased and deployed on a range of certified industry-standard servers with internal disk storage.

  - ❑ Avamar Data Store. This scalable, all-in-one solution includes Avamar software preinstalled and preconfigured on EMC hardware for simplified ordering, deployment, and service.

- ❑ Avamar Virtual Edition for VMware. An industry first, this configuration enables an Avamar server to be deployed as a virtual appliance on an existing ESX Server, leveraging the attached resources and disk storage.

Avamar is different from other source deduplication approaches on the market. For example, Avamar's deduplication uses subfile variable-length data segments, which deliver superior efficiency and performance. Avamar uses grid architecture for scaling performance and capacity, where each incremental node increases CPU, memory, I/O, and storage for the entire system.

The Avamar grid uses a redundant array of independent nodes (RAIN) configuration for built-in fault tolerance and high availability across the grid and eliminates single points of failure. Avamar distributes its internal index across Avamar nodes for reliability, load balancing, and scalability. Also, every day and automatically, Avamar verifies that backup data is fully recoverable, and the Avamar server checks itself twice daily to ensure server integrity. Lastly, Avamar offers a broad range of application and client support, including Exchange, SQL, Oracle, DB2, SharePoint, Lotus Notes, and NDMP support.

Avamar offers a number of ways to protect virtual as well as physical machines. Options for Avamar backup of VMware virtual machine environments include the following:

- **Avamar agent in guest OS.** An Avamar agent inside each guest OS provides a backup approach that is an order of magnitude more efficient than traditional agent backup approaches. Lightweight Avamar agents reduce backup data at the guest, reducing network requirements and contention for shared CPU, NIC, disk, and memory resources. Because only new or unique subfile data is backed up, Avamar enables fast daily full backups.

- **Avamar for VCB or vStorage API backup.** An Avamar agent running on the proxy server backs up only unique data and offloads the processing for the guest machines. Deduplication occurs within and across VMDK files and supports both file- and image-level backup. Avamar's efficient replication enables VMDK files to be quickly transferred across the WAN in support of disaster recovery objectives.

- **Avamar agent on ESX console.** An Avamar agent on the ESX console can deduplicate within and across VMDK files. This method provides an image-level backup and restore option, without a dependency on VMware VCB or shared storage. However, it does not provide for file-level restore.

EMC Data Domain

Data Domain deduplication storage systems dramatically reduce the amount of disk storage needed to retain and protect enterprise data. By identifying redundant files and data as they are being stored, Data Domain systems provide a storage footprint that is 10x–30x smaller, on average, than that of the original data set. Backup data can then be efficiently replicated and retrieved over existing networks for streamlined disaster recovery and consolidated tape operations.

Data Domain systems are available in a range of configurations that vary by performance and capacity. Several software options enhance the functionality and value of its systems. For example:

- Data Domain Appliance Series provides high-throughput, cost-effective, and scalable deduplication storage systems with integrated storage.

- Data Domain DDX Array Series offers a scalable, high-performance disk array storage system with up to 56.7 petabytes of logical capacity and up to 86.4 terabytes/hour of throughput. It includes management and support for up to 2,880 remote locations when using DD880 controllers.

- Data Domain Gateway Series provides enterprise gateways offering deduplication and compression benefits to datacenters that want to use certain third-party external storage systems.

- Data Domain Replicator Software is a network-efficient, automated replication software solution available for disaster recovery, remote office data protection, and multisite tape consolidation.

- Data Domain Virtual Tape Library Software emulates multiple tape libraries over a Fibre Channel interface, providing deduplication storage for SAN environments.

- Data Domain OpenStorage Software provides seamless integration between Data Domain deduplication storage systems and Symantec Veritas NetBackup.

- Data Domain Retention Lock Software enables users to easily implement deduplication with file locking to satisfy IT governance and compliance policies. IT administrators are given the operational flexibility required to efficiently run the enterprise on a day-to-day basis, all at the right cost. The software includes secure data shredding capabilities.

Data Domain deduplication solutions provide a consolidated storage tier for backup, nearline, and archive data. In addition, Data Domain solutions integrate with existing infrastructure and provide a good measure of investment protection, thus making them ideal for firms with the following environments:

- Datacenters showing explosive data growth for backup and fixed content data while facing increasingly difficult operational and disaster recovery requirements

- Larger firms needing to support remote office to datacenter consolidation for data protection and disaster recovery

- Backup and archive infrastructure requiring a consolidated deduplication target

- Infrastructures built on tape systems and media that need to redesign backup and recovery to eliminate burdensome cost and management issues

The Data Domain Data Invulnerability Architecture ensures the integrity of all backup data by providing high levels of data protection, data verification, and self-healing capabilities. Key areas of data integrity protection include the following:

- Continuous recovery verification, which ensures the backup data is correct and recoverable from every level of the system throughout the data's entire life cycle
- Uniqueness verification, which protects against random and malicious hash collisions to ensure successful data recoveries
- Dual disk parity RAID (RAID 6), which protects against up to two simultaneous disk faults

The Data Invulnerability Architecture has extra levels of data integrity protection built in to detect faults and repair them to ensure that recovery of backup data is risk-free.

Data Domain Replicator software transfers only the deduplicated and compressed unique changes across any IP network, requiring only a fraction of the bandwidth, time, and cost, compared with traditional replication methods. If multiple Data Domain systems replicate to the same destination system, the deduplication effect becomes more efficient, as the destination system will store each unique segment across all inbound replication streams only once, further minimizing bandwidth.

Key benefits of Data Domain replication technology include the following:

- WAN vaulting for disaster recovery, providing network-based data protection by securely and automatically replicating backup data to a secure offsite location
- Remote office data protection, allowing vaulting of backup data from many branch offices to a central hub or datacenter
- Cascaded replication, enabling deduplicated data to be replicated from a remote office to a central datacenter and then on to additional sites such as a DR facility for enhanced recoverability and disaster protection
- Tape consolidation, eliminating the need to duplicate backup data at each remote office and leaving a significantly reduced tape infrastructure only at a central hub with IT staff

EMC NetWorker

EMC NetWorker is an enterprise backup application that centralizes backup and recovery operations. NetWorker provides a common platform that supports a wide range of data protection options, including backup to disk, replication, continuous data protection, and deduplication across physical and virtual environments. NetWorker's versatility makes it the ideal backup software for customers choosing to simplify their management across environments, from large datacenters to remote offices. The core NetWorker application provides deduplication at the source through integration with EMC Avamar's deduplication technology and can also leverage target deduplication solutions, such as Data Domain systems, within the scope of its operations. Firms using NetWorker deduplication are:

- Seeking to curb large volume data growth for existing NetWorker environments
- Deploying a new backup to disk strategy for improved recovery that still requires the use of physical tape for archival or long-term needs

- Meeting a mix of requirements — some ideally suited for source deduplication and some better suited to target deduplication
- Driving down cost and complexity by consolidating multiple data protection strategies under one application

The deduplication approach within the NetWorker application has advanced the market in terms of its integration of deduplication with a traditional backup application. The NetWorker client software for both nondeduplicating and deduplication-aware backups is a single agent. Source deduplication capabilities have been fully integrated, minimizing deployment and maintenance. The NetWorker console can manage and monitor both types of backups — traditional and deduplication. For NetWorker customers that want the benefits of deduplication, there is no additional client-side cost.

Unlike other offerings, NetWorker has no incremental software SKUs or pricing for deduplication integration. NetWorker customers can add the appropriate deduplication engine into the backup environment, either the Avamar software or the Data Domain back-end solution. One of the benefits of using NetWorker-enabled deduplication is the support for physical tape, ensuring that users who continue to have a tape requirement can meet the need within the same application. Another is the ability to leverage NetWorker's rich application support, allowing for disaster and granular recovery as well as snapshot management for off-host backup. NetWorker gives firms the strong features of deduplication without disrupting their current backup environment.

CHALLENGES: WHICH APPROACH?

As shown in this paper, different deduplication technologies and approaches provide distinct advantages per use case, so it is important to have an easy way to map each EMC product to the environment to which it provides maximum efficiency. Table 2 outlines how firms can think about which EMC backup and recovery product is right for them.

EMC has a broad portfolio of products with deduplication functionality. Since deduplication is not a standalone product, EMC needs to accelerate customer education on the most appropriate place to leverage the capability, given customers' specific environmental challenges. Education, in concert with documented case studies and scale and performance testing benchmarks, will further increase customer confidence in the application of the technology within a given product.

TABLE 2

Selecting an EMC Deduplication Product

	EMC NetWorker	EMC Data Domain	EMC Avamar
Deduplication for backup	• Source • Inline deduplication	• Target • Inline deduplication	• Source • Inline deduplication
Ideal for environments with:	• NetWorker environments • Need for physical tape support • Large, heterogeneous environments	• High-speed backup and recovery requirements • Replication for offsite backup • Support for current backup environment — no operational changes • Support for datacenter and remote sites	• Virtual environments • Remote branch offices • LAN/NAS servers • Desktop/laptops
Deployment options	• Single NetWorker agent • See Avamar deployment options	• Appliance hardware or gateway	• Agent only — for smaller remote offices • Avamar Data Store — turnkey all-in-one solution (hardware and software) • Third-party server — create own Avamar server • Avamar Virtual Edition — virtual appliance leveraging existing ESX Server and disk

Source: IDC, 2010

CONCLUSION

Deduplication technology can accelerate backup efficiency and drive down IT costs. Firms are deploying different types of deduplication-enabled solutions to address a myriad of cost and operational challenges with the growing volume of backup data. IDC finds that deduplication is a core, must-have feature for a variety of storage solutions to address these challenges. EMC as a vendor is well-positioned to address these long-standing problems, offering a range of solutions for a variety of environments and use cases to meet customer demand for the technology over the next five years.

Copyright Notice