

WHITE PAPER

Data Deduplication for Backup: Accelerating Efficiency and Driving Down IT Costs

Sponsored by: EMC Corporation

Laura DuBois
May 2009

EXECUTIVE SUMMARY

Data deduplication is dramatically improving IT economics by optimizing network bandwidth, backup windows, and storage footprint requirements in distributed and datacenter locations alike. In real-world environments, deduplication is accelerating backup 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 backup portfolio of deduplication offerings and 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 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, the virtual machines process redundant data, which 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. Standard approaches such as deploying a traditional backup agent in a guest virtual machine or using a VCB proxy backup do nothing to reduce the volume of virtual machine data that needs to be backed up or the network bandwidth requirements. Deduplication offers significant backup storage capacity savings. In addition, some forms of deduplication also reduce the amount of data to be backed up, resulting in faster backups and less impact on the network. 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.
- ☒ **Reducing carbon footprint.** Deduplication reduces the power, cooling, and space requirements for storage, thus reducing carbon footprint and enabling environmental responsibility.
- ☒ **Improving backup and recovery service levels.** Deduplication significantly improves 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 has had a role in enterprise datacenters due to its economics and archival properties. However, cost/GB declines for disk when used with deduplication could make disk costs equal to or less than tape costs.

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. It also identifies the deduplication approach best suited to address each challenge.

TABLE 1

Backup Challenges and Deduplication Impact

Backup Challenge(s)	Deduplication Impact	Best Fit Deduplication
Recovery time frames 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.	Source- or target-side deduplication
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.	Source- or target-side deduplication
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 reduces the amount of data that needs to be backed up, thus enabling more data to be backed up in an available window.	Source-side deduplication
Increased server virtualization means fewer resources are available for backup, which can increase backup times and stress backup windows.	Source-side deduplication means duplicate data never requires shared resource processing, reducing contention and speeding virtual machine backups.	Source-side deduplication
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.	Source-side deduplication
Secure offsite copy using traditional tape methods leaves data at risk due to loss or theft.	Reliance on removable tape media for offsite storage in the event of a disaster introduces risk in compromise to the physical media. 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 improving security.	Source- or target-side deduplication
Backup infrastructure costs increase 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. Tape drives and automation may address current performance bottlenecks and perform the backups more quickly but with cost and management overhead. Deduplication reduces the ongoing spend on tape infrastructure to keep pace with these trends.	Source- or target-side deduplication

Source: IDC, 2009

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 objects (e.g., blocks, chunks, files) 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 operates on only unique data and eliminates the costs associated with keeping multiple copies of the same data object.

Different from single-instance storage (SIS), which deduplicates data at the file or object level, data deduplication is most often associated with subfile deduplication processes. Subfile deduplication examines a file and breaks it up into "chunks." These smaller chunks 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. However, deduplicated data can also be compressed for further space savings.

Where Deduplication Occurs

Backup data deduplication can occur at the source or target. An example of source-side 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 network during the backup process. An example of target-side deduplication would be reducing the size of backup data after it crosses the network when it reaches a deduplication appliance. Deduplication at the source provides network bandwidth, backup window, and storage savings. Deduplication at the target provides storage savings, works with existing backup software, and can reduce the network impact, although it requires a hardware appliance at every location. 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 (refer back to Table 1).

Source-side Deduplication

Performing deduplication at the source (client) provides an extended set of benefits beyond capacity optimization. It also means significantly less data is sent, thus relieving congested virtual/physical infrastructure and LAN/WAN links. Because only new or changed subfile data segments are sent, 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. The overall impact of source deduplication is actually much less than that of traditional agents over a seven-day period. Environments with very large databases or databases with high daily change rates may want to consider a target-side solution instead. Fortunately,

vendors typically have data assessment tools to help customers make the best choice. Source-side deduplication also offers deployment flexibility, since smaller remote offices can simply deploy just the software backup agent, with no additional local hardware required.

Target-side Deduplication

Performing deduplication at the target optimizes backup disk storage capacity since only new, unique subfile data is stored to disk. However, redundant backup data is still sent to the deduplication target using traditional backup software. Thus, it does not offer relief to an available backup window. A critical factor to consider when using a target-side approach is the ability to keep pace with backup window performance, and whether or not inline or post-process deduplication is warranted given a particular workload. (Refer to the following section for more on inline versus post-process deduplication).

Additionally, a target-side approach requires the purchase of an incremental deduplication appliance, which needs to be budgeted for and managed like any other system. When the appliance runs out of capacity, another deduplication appliance needs to be deployed. Some solutions offer clustering of a number of appliances to mitigate this issue. Target-side deduplication can be used both in central datacenters for large volumes of data and in remote locations. However, this does mean implementation of a deduplication appliance in each remote location with remote replication from many branches to a central, larger appliance in the datacenter. Source-side deduplication can eliminate this remote branch hardware investment.

When Deduplication Happens

There are two different approaches available today for determining *when* the deduplication process occurs: inline or post-process. Some suppliers are also working on a third approach called hybrid or adaptive deduplication. 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.

An inline process is more capacity efficient, and there is no lag time for a deduplication process to begin. For large-capacity environments with backup window considerations, post-process deduplication gives precedence to completing the backup but requires greater initial storage capacity. These approaches can mean a trade-off in performance and capacity requirements. A third approach, still in the developmental stages, is called hybrid or adaptive deduplication. This method of deduplication gives precedence to an inline approach until a performance threshold is reached and then automatically switches to a post-process approach, tuning the method for the current workload in the environment. Some leading solutions offer policy-based deduplication that allows for customer configuration of deduplication to occur either immediately or on a schedule or to be disabled based on characteristics of a data set. For example, smaller data sets and unstructured data can be set for immediate deduplication and large backup jobs configured for post-processing, while

deduplication can be turned off for backups of rich media or encrypted data. Policy-based deduplication gives users the greatest level of flexibility in setting deduplication according to their environmental conditions.

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 chunk. This process calculates a unique number for each chunk, 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 chunk size is fixed or variable in length. A variable-length approach can dynamically, based on content type, adjust a chunk size to accommodate redundant data chunks 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 chunks again, even though they are already in the backup repository. A potential issue with a hash-based approach is performance and disk I/O. The hash index is kept in memory, but as a hash index grows, it may spill over from memory into disk, requiring disk I/O for the lookup and chunk retrieval. Vendors have varying ways of dealing with these practical technology challenges.

An alternative approach is delta-based data deduplication, 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. The approach used, hash or delta based, becomes a trade-off between a deduplication ratio yielded versus performance. Larger chunk sizes tend to decrease the data deduplication ratio, but smaller chunk sizes result in more indexing overhead.

Another factor impacting the deduplication ratio is whether or not the deduplication engine can recognize a particular data format (particular backup application, Microsoft Exchange data, etc.). The ability to detect the format of the data requires understanding where application-specific metadata is injected into a stream. The deduplication engine can then tune the chunk size so that it's ideal for the data format according to natural application, resulting in potentially greater deduplication results.

CONSIDERATIONS IN EVALUATING DEDUPLICATION TECHNOLOGY

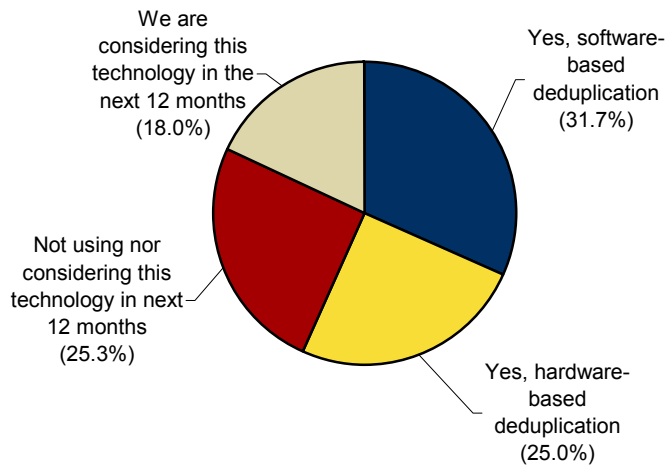
A number of different types of products with deduplication capabilities are available on the market today. Backup applications, appliances, virtual tape libraries, WAN optimization solutions, and primary disk storage subsystems all may have some form of deduplication functionality. It's important for 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-side deduplication solutions can reduce required daily network bandwidth by an order of magnitude compared with traditional daily full backup methods, with several leading vendors claiming up to 500:1. 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 premise 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. Firms using encryption or compression at the software layer and employing target-based deduplication will likely need to disable these functions to gain deduplication benefits or employ hardware-based compression. 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 needs to be disabled if firms want to benefit from deduplication.
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 an agent in each guest, a VCB proxy server backup, or an agent in the Service console. 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 decreased server consolidation. Deduplication can address these limitations. Source-side 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 systems. Another factor to consider is that a VMDK file, when updated, will be offset. Only variable-length deduplication can account for this offset, finding and moving only the unique changes within the VMDK.
4. **Deduplication for remote branches.** Like datacenter operations, remote branches require both local and disaster (remote) recovery. However, the characteristics of remote branches introduce challenges. Remote branch 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 global deduplication eliminates redundant data across branch locations and the datacenter. With limited IT staff at remote branch locations, many firms are

looking to reduce storage hardware footprint in distributed locations. Source-side deduplication can be deployed through software, which mitigates this concern. A recent IDC study looked at the use of deduplication in remote branch 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.** For datacenter environments, data volumes scale significantly but with the benefit of LAN connectivity and faster connections to disaster recovery sites. 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 policy-based 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. Global deduplication.** The definition of what constitutes global deduplication can vary. Some users view global deduplication across sites, while others view it within a single storage frame. However, ideally global deduplication should be across both sites and frames to gain the greatest benefit. Target-side deduplication can offer global deduplication of many remote offices to a single frame and its replica pairs. However, when a capacity or performance ceiling is

reached, a new appliance must be set up, which introduces another standalone deduplication appliance. Source-side deduplication also can offer global deduplication of many remote offices datacenter servers and its pairs. Firms should keep in mind that global deduplication is a term that has different meaning depending on the vendor and its approach.

7. **Deduplication and replication.** Replication is really the next battleground for deduplication technology. Established and innovative suppliers alike have proven that it works, and user reactions in evaluating 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. As more firms look to more strategically (i.e., use cases such as compliance) rely on tape in datacenters and minimize its use in remote locations, the role of remote replication becomes paramount. 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 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.

8. **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, 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.

9. **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 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.

10. **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 network-attached storage (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-ENABLED SOLUTIONS

EMC offers a broad range of deduplication-enabled products to assist customers with driving down IT costs and accelerating backup efficiency. Backup deduplication solutions include EMC Avamar, which provides a source-side approach to deduplication; EMC Disk Library, which offers a target-side approach to deduplication; and EMC NetWorker, which can be deployed with either a source-side or target-side approach, or both. 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 backup and recovery solutions include 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, and datacenter LAN and NAS servers. Avamar also deduplicates backup data globally 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 local 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 communicate 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.

Avamar backup and recovery solutions provide source-side and global 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

EMC Avamar can be deployed in four types of configurations:

- ☒ **Avamar software.** For smaller remote offices, the Avamar software agent can be deployed on the systems to be protected (clients) with no additional local hardware required.
- ☒ **Third-party Avamar server.** 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-side deduplication approaches on the market. For example, Avamar's deduplication uses subfile variable-length data segments, which deliver superior efficiency and performance compared with solutions that use fixed-length segments. 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 backup.** An Avamar agent running on the VCB proxy server backs up only unique data and offloads the processing for the guest machines. Deduplication occurs within and across VMDK files and supports VCB 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 Disk Library

The EMC Disk Library (DL) family offers policy-based deduplication with its 1500, 3000, and 4000 series systems. The EMC Disk Library 1500 and 3000 provide LAN-based backup to disk with deduplication included. The DL1500 is designed for midsize customers that want improved performance, longer onsite retention, and lower replication costs. The DL1500 begins at 4TB of usable capacity and expands to 36TB, with a sustained backup ingest rate of 0.72TB/hour with immediate data deduplication — or up to 0.84TB/hour when the deduplication process is deferred.

The DL3000 begins at 8TB of usable capacity and expands to 148TB, with a sustained backup ingest rate of 1.44TB/hour with immediate data deduplication. With both the DL1500 and DL3000, policy-based deduplication is included with the system. Unlike the DL1500 and DL3000 models, the DL4000 deduplication is via an add-on hardware option for new and installed DL4000 Virtual Tape Library systems. Firms can deploy it to reduce capacity requirements for backup to disk and reduce network traffic for replication between datacenters.

EMC Disk Library deduplication is ideal for datacenter, large storage volume, and highly changing database environments looking to introduce disk for backup. Firms using Disk Library deduplication are:

- ☒ Seeking to curb **large volume data** growth for backup to existing EMC Virtual Tape Library environments

- ☒ Deploying a new **backup to disk strategy** for improved recovery and/or reliability while minimizing the storage costs
- ☒ Introducing both disk and deduplication into an **existing EMC Disk Library** environment
- ☒ Making use of **electronic vaulting of backups** to a disaster recovery datacenter and minimizing the use of physical tape
- ☒ Seeking to **replace tape with disk for backup** with little disruption to current backup operations

The Disk Library systems use a target-side deduplication method. The same deduplication capability works across the entire Disk Library family, providing block-level, variable-length hash-based deduplication at the target. The Disk Library deduplication uses "application sensing filters" that can detect the format of the data stream and understands where application-specific metadata is injected into a stream. The filter will place markers around this metadata and sift it out for greater deduplication impact.

The systems use a policy-based, customer-configurable deduplication process. Deduplication can be set to occur immediately or on a schedule, or it can be disabled entirely. The policy is set at a file share or virtual library level. Deduplication in "immediate mode" or inline is ideal for smaller data sets and unstructured data. Deduplication in "scheduled mode" allows the deduplication process to give precedence to the backup, which completes before deduplication starts. Ideal for larger data sets, this allows for backups to complete as much as 150–200% faster (according to EMC) than when done in an immediate mode.

For data types that do not deduplicate well, the capability can be disabled. For deduplication-enabled remote replication for disaster recovery purposes, replication can be configured by system, application, directory level, or virtual tape cartridge. The Disk Library deduplication index is clustered into branches, and similar objects are grouped into buckets for efficient index lookups while minimizing disk I/O. Hardware compression provides another level of storage optimization.

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 the EMC Disk Library, 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
- ☒ Introducing both disk and deduplication into an **existing EMC Disk Library** environment
- ☒ 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 or the EMC Disk Library 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 benefit of using deduplication within the backup application is the correct provisioning and sequencing of encryption and compression. 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 what EMC deduplication-enabled product is right for them.

EMC has a myriad of different products with deduplication functionality. While deduplication is a feature or technology, rather than 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 increase customer confidence in the application of the technology with a given product.

TABLE 2

Selecting an EMC Deduplication-Enabled Product

	EMC NetWorker	EMC Disk Library	EMC Avamar
Deduplication for backup	<ul style="list-style-type: none"> • Source side • Inline deduplication 	<ul style="list-style-type: none"> • Target side • Policy based and configurable for immediate, scheduled, or disabled deduplication 	<ul style="list-style-type: none"> • Source side • Inline deduplication
Ideal for environments with:	<ul style="list-style-type: none"> • NetWorker environments • Need for physical tape support • Large, heterogeneous environments 	<ul style="list-style-type: none"> • High-speed backup and recovery requirements • Replication for offsite backup • Support for current backup environment — no operational changes • Support for datacenter and remote sites 	<ul style="list-style-type: none"> • Virtual environments • Remote branch offices • LAN/NAS servers
Deployment options	<ul style="list-style-type: none"> • Single NetWorker agent • Agent for smaller remote offices • For deduplication node, any of the following: <ul style="list-style-type: none"> • 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 	<ul style="list-style-type: none"> • Appliance hardware 	<ul style="list-style-type: none"> • 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, 2009

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

External Publication of IDC Information and Data — Any IDC information that is to be used in advertising, press releases, or promotional materials requires prior written approval from the appropriate IDC Vice President or Country Manager. A draft of the proposed document should accompany any such request. IDC reserves the right to deny approval of external usage for any reason.

Copyright 2009 IDC. Reproduction without written permission is completely forbidden.