



**An EMC Perspective on Data  
De-Duplication for Backup**

## Abstract

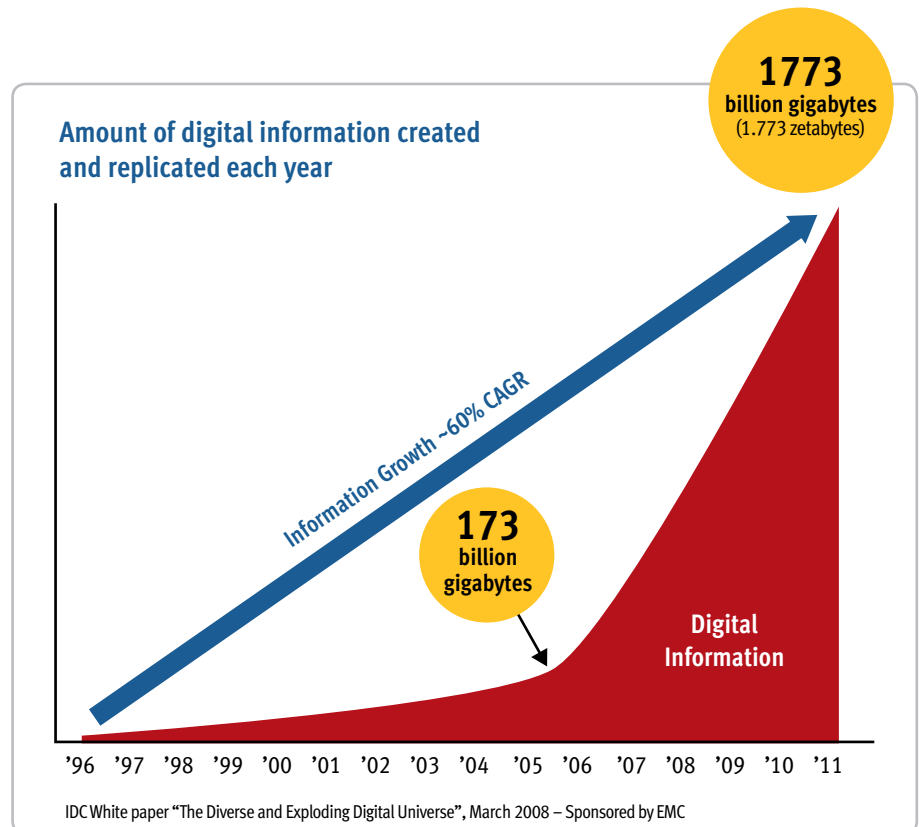
This paper explores the factors that are driving the need for de-duplication and the benefits of data de-duplication as a feature of an organization's backup strategy. Topics covered in this paper include:

- Aspects of information growth and the resulting strain on backup and replication
- Current data de-duplication technology and deployment
- Benefits of data de-duplication and its positive impact on backup, replication, and recovery
- Factors that determine the effectiveness of data de-duplication
- Specific consideration of backup-to-disk, remote office, and VMware backup environments
- Justifying a data de-duplication investment

Finally, the paper will briefly outline EMC's de-duplication-enabled backup solutions optimized to meet backup needs, recovery requirements, and application environments.

## The Data Growth Explosion

IDC predicts that the growth of the “digital universe”—information that is either created, captured, or replicated in digital form—will increase ten times in volume between 2006 and 2011 (IDC, The Diverse and Exploding Digital Universe, March 2008). Most organizations are struggling with the impact that information growth has on their information protection processes. More information, cost constraints, and increasingly stringent service levels are stressing traditional backup environments to the breaking point. Protecting critical data is a challenge for organizations of all sizes. Traditional backup tools and processes are just not able to keep up, leaving organizations struggling to manage backup windows, application uptime, and information recovery requirements.



### Backup Challenged by Multiplying Data Growth

Most traditional backup solutions protect the same data repeatedly, expanding total storage under management by five to thirty or more times. Consider the effect of a weekly full file server backup—with a five-week retention policy, the vast majority of the content is protected five times over. Now consider an e-mail server that is protected with a nightly full backup—much of the content in that mail server is backed up 35 times during those five weeks.

Organizations need solutions to help manage this information explosion. In addition, government regulations and requests for legal discovery can further strain the resources and capabilities of traditional data protection solutions. Failure to comply or provide information in a timely fashion can result in significant costs and penalties. Recent legislation has also exposed the risk of shipping physical tape cartridges as one of the greatest security concerns in today’s IT infrastructure.

As noted above, traditional backup solutions require a rotational schedule of full and incremental backups, which move a significant amount of redundant data week after week. And most organizations create a second copy of this information to be shipped to a secondary site for disaster recovery purposes. When taken in aggregate, the costs of traditional backup in terms of bandwidth, storage infrastructure, and time—let alone capital costs—are a significant burden to IT.

Backing up redundant files and data is a major reason that backup windows roll into production hours, over-utilize network resources, and require too much additional storage capacity to hold unnecessary backup data. This waste is all the more painful since replicating duplicate files and the repeated backup of redundant data add no additional value to the business.

## What is Data De-Duplication?

If it was possible, IT organizations would only protect the unique data from their backups. Instead of saving everything repeatedly, the ideal scenario is one where only the new or unique content is saved. Data de-duplication provides this basic capability. It offers the ability to discover and remove redundant data from within a dataset. A dataset can span a single application or span an entire organization. Redundant data elements can be entire files or sub-file data segments within a file. In all cases, the objective of the de-duplication process is to store unique data elements only once, yet be able to reconstitute all content in its original form on demand, with 100 percent reliability at disk speeds.

Data de-duplication is fundamental to improving information protection, streamlining backup operations, reducing backup infrastructure, shortening backup windows, and removing burden from information networks.

## Benefits of Data De-Duplication for Backup

Effective data de-duplication products help organizations cope with backing up information stores. Data de-duplication removes data that is redundant to economize the storage and disaster recovery requirements for data. As noted earlier, there is tremendous data redundancy in backup environments. Effective deployment of data de-duplication allows organizations to protect and restore information at a fraction of the footprint and operational expense of their current backup storage infrastructure.

There are significant business benefits derived from an investment in data de-duplication. Data de-duplication business benefits include:

**Lower infrastructure costs.** By eliminating redundant data from the backup, far less infrastructure is required to hold the backup images. Data de-duplication directly results in reduced storage capacities to hold backup images. Smaller capacity requirements means lower acquisition costs as well as reduced power and cooling costs.

**Longer retention.** Because data de-duplication reduces the amount of content in the daily backup, users can extend their retention policies. This can have a significant benefit to users who currently require longer retention, but are limited by current processes and policies.

**Improved data protection.** Data de-duplication enables many organizations to create daily full backup images. Many of these organizations had been forced to do weekly fulls and daily incrementals due to backup window constraints. De-duplication reduces storage capacity requirements, which permits more aggressive backup policies with improved restore times.

**Reduced quantity, better performance.** By reducing the total quantity of backup image size, companies are better able to afford the replacement of traditional tape storage with disk for backup. Backing up to disk enables high-speed, highly reliable backup images—which supports the need for both shorter backup windows and faster recovery times.

**Vast reductions in backup bandwidth.** Data de-duplication reduces the amount of content in a backup image, thereby reducing the expense of remote replication of that content—hence enabling remotely replicated backups. In addition, by utilizing data de-duplication at the client (source-based), redundant data is extracted from the backup process before any data is moved during the backup process. This means that backups are completed by moving only sub-file data elements not already in the backup database. Only a fraction of the network bandwidth used for traditional backup is required for source-based de-duplicated backups.

**Improved security and streamlined management.** By leveraging disk-based storage for backup images, organizations can improve information security by eliminating the risks, delays, and costs associated with physical tape handling and bunkering. Instead of traditional methods of shipping tapes between facilities, content can now be moved over the network.

**It is important to understand that data de-duplication is not a single solution.** There are many options as to how, when, and where de-duplication is performed. There are impacts to performance and efficiency that should be considered before finalizing your data de-duplication strategy.

## A Closer Look at Data De-duplication

There are many forms of data de-duplication. Typically, there is no one best way to implement data de-duplication across an entire organization. Instead, to maximize the benefits, organizations may deploy more than one de-duplication strategy. When selecting a de-duplication solution, it is important to clearly understand the environment and the backup challenges it represents.

There are three basic forms of data de-duplication. Although definitions vary, some forms of data de-duplication, such as **compression**, have been around for decades. Lately, **single-instance storage** has enabled the removal of redundant files from storage environments such as archives. Most recently, we have seen the introduction of **sub-file de-duplication**. These three types of data de-duplication are described further below.

### Data Compression

Data compression is a method of reducing the size of files. Data compression works within a file to identify and remove empty space that appears as repetitive patterns. This form of data de-duplication is local to the file and does not take into consideration other files and data segments within those files. Data compression has been available for many years, but being isolated to each particular file, the benefits are limited when comparing data compression to other forms of de-duplication. For example, data compression will not be effective in recognizing and eliminating duplicate files, but will independently compress each of the files.

### Single-Instance Storage

One form of de-duplication is the removal of multiple copies of any file. Single-instance storage (SIS) environments are able to detect and remove redundant copies of identical files. After a file has been stored in a single-instance storage system, all other references to the same file will refer to the original, single copy. Single-instance storage systems compare the content of files to determine if the incoming file is identical to an existing file in the storage system. Content-addressed storage is typically equipped with single-instance storage functionality.

While file-level de-duplication avoids storing files that are a duplicate of another file, many files that are considered unique by single-instance storage measurement may have a tremendous amount of redundancy within the files or between files. For example, it would only take one small element (e.g., a new date inserted into the title slide of a presentation) for single-instance storage to regard two large files as being different and requiring them to be stored without further de-duplication.

### Sub-file De-Duplication

Sub-file de-duplication detects redundant data within and across files as opposed to finding identical files as in SIS implementations. Using sub-file de-duplication, redundant copies of data are detected and eliminated—even when the duplicated data exists within separate files. This form of de-duplication discovers the unique data elements within an organization and detects when these elements are used within other files. As a result, sub-file de-duplication eliminates the storage of duplicate data across an organization. Sub-file data de-duplication has tremendous benefits even where files are not identical, but have data elements that are already recognized somewhere in the organization.

There are two forms of sub-file de-duplication implementation. **Fixed-length** sub-file de-duplication uses an arbitrary fixed length of data to search for duplicate data within files. Although simple in design, fixed-length segments miss many opportunities to discover redundant sub-file data. (Consider the case where an addition of a person's name is added to a document's title page—the whole content of the document will shift, causing the failure of the de-duplication tool to detect equivalencies). **Variable-length** implementations are not locked to any arbitrary segment length. Variable-length implementations match data segment sizes to the naturally occurring duplication within files, vastly increasing the overall de-duplication ratio (In the example above, variable-length de-duplication will catch all duplicate segments in the document, no matter where the changes occur).

## **Where Does Data De-Duplication Occur?**

There are some technology considerations when determining the optimal de-duplication solution for your organization. Considerations include whether the de-duplication occurs at the information source or at the backup target. Additionally, you should consider the appropriateness of immediate de-duplication or scheduled de-duplication architectures for your backup environment. This section further describes these features.

### **Source-based De-duplication**

Source-based data de-duplication provides for the elimination of redundant data at the source. This means that data de-duplication is performed at the start of the backup process—before the information is transmitted to the backup environment. Source-based de-duplication can radically reduce the amount of backup data sent over networks during backup processes. This is important if there are bottlenecks in the backup process related to networks, shared resources (as in VMware®), or backup windows. Furthermore, there is also a substantial reduction in the capacity requirements needed to store the backup images.

### **Target-based De-Duplication**

Target-based de-duplication is an alternative to source-based de-duplication. Target-based de-duplication happens at the backup storage device. This form of de-duplication does not require users to change their incumbent backup software. Target-based de-duplication does require that all backup images are copied to the backup appliance, so target-based backup is not a solution that reduces backup-client-to-target bandwidth requirements.

## **When Does Target Data De-Duplication Occur?**

### **Target-based Immediate De-Duplication**

There are two classifications of where the de-duplication occurs. Immediate de-duplication occurs as new data arrives at the target and the appliance recognizes duplicate data before writing the de-duplicated data to the target storage system. All de-duplication introduces overhead in the form of time required to identify and remove duplication. Due to the fact that immediate de-duplication implementations incur a time penalty, users must consider the effect of adding latency into the data path. This can be challenging for environments where backup window reduction is an imperative.

### **Target-based Scheduled De-Duplication**

An alternative to immediate de-duplication is scheduled de-duplication. De-duplication products that are classified as scheduled de-duplication perform the de-duplication outside of the data path. In order to remove overhead out of the data path, scheduled de-duplication implementations initially store data from the source, then perform de-duplication as a post-process. Scheduled de-duplication implementations have storage capacity to hold the intermediate data before de-duplication. The primary benefit of scheduled de-duplication is that it allows the data path from the source to operate with the least latency, resulting in faster data movement. Also, scheduled de-duplication permits the full, native backup image to be available, allowing for fast restore direct from the disk target. A consideration for users can be that scheduled de-duplication requires extra storage capacity to store the native backup images before they are de-duplicated.

## Factors that Determine Effectiveness of De-Duplication

The first decision to make is to decide whether de-duplication is warranted in a particular environment. Many application environments are well suited to data de-duplication and optimally suited to a particular implementation of de-duplication. Some environments will not maximize the benefits of de-duplication, and various de-duplication solutions have different de-duplication ratios. The de-duplication ratio is the size of the data backed up compared to the size of the backup images on disk. There are several factors that contribute to the final de-duplication ratio one will experience. Some of these factors are intrinsic to the de-duplication technology and others are environmental factors. Here is a summary of several common factors that affect de-duplication effectiveness.

- **Retention period.** The longer the data retention period in backup, the greater the frequency of duplicate data already in the backup image store.
- **Ratio of full backups to incremental backups.** The more frequently full backups are conducted, the greater the advantage of de-duplication reducing the size of the backup store.
- **Change rate.** The fewer changes to the content between backups, the greater the efficiency of de-duplication.
- **Data type.** The data from natural sources (audio, images, scans) is highly unique compared to application-generated data (documents, e-mail, presentations). The more unique the data, the less intrinsic duplication exists. File-level de-duplication can still occur, but sub-file level is less likely.
- **De-duplication method.** Variable-segment length, sub-file de-duplication will discover the highest amount of de-duplication across an organization

### Special File-type and Bandwidth Considerations

Certain files have low levels of redundancy within their content. Examples of these low-content redundancy include video, audio, compressed, and encrypted files. These files will not have a high frequency of duplication matches and, as a result, will not gain significant benefit from de-duplication. Environments that consist mainly of these low-redundancy files will fail to achieve high de-duplication ratios. Applying de-duplication to low-redundancy environments will resolve these files to single-instance storage in the backup images.

For many environments, backup bandwidth is also a special consideration. For environments with bandwidth or resource restrictions (e.g., remote offices, VMware), source-based de-duplication offers the most substantial advantages. Since source-based de-duplication identifies redundant data before the data is copied for backup, only a small fraction of the content will need to be moved over the network, reducing the bandwidth requirements for backup.

## De-Duplication by Environment Type

No one de-duplication strategy or offering is ideal for every application environment. Different application environments are best suited for either source-based or target-based de-duplication. In some environments immediate de-duplication is preferable over scheduled de-duplication. In this section we will look at several application environments and demonstrate how data de-duplication is ideally implemented.

### Data De-Duplication in LAN and SAN Environments

Data de-duplication can be leveraged in a variety of different formats—all the way from simple LAN-based backup-to-disk to virtual tape library infrastructures. Each of these solutions has its own advantages, and data de-duplication solutions have leveraged these architectures to extend their benefits for a particular environment.

LAN-based backup-to-disk allows users to easily build a solution leveraging their existing LAN infrastructure. For organizations that have not yet implemented a Fibre Channel SAN, LAN-based backup-to-disk is often the favored solution. For users who already own or are planning a Fibre Channel SAN, virtual tape libraries (VTLs) are a strong consideration. VTLs are specialized disk-based storage systems that present themselves as a tape library to the backup application, while providing improved speed and reliability. Either of these solutions is easy to implement within traditional backup application environments.

### Data De-duplication in VMware Environments

VMware is the industry's most widely deployed virtualization solution, and is increasingly used for more mission-critical production environments. Extending data protection of virtual machines is therefore an important function. VMware consolidates multiple virtual servers into a single physical server. In a traditional backup process, each of those virtual servers has a backup agent—which, when run, consumes significant CPU, NIC, cache, and memory resources. When multiple agents are initiated on a VMware machine, performance of the overall system can be impacted. By implementing de-duplication on the VMware infrastructure, significant resource savings can be attained—enabling a higher degree of consolidation and reduced strain on the server platform.

### Data De-Duplication in Remote Office Backup

Remote office backup is often a difficult challenge—many organizations have limited personnel with insufficient skills to provide adequate services levels. The key challenges of remote office data protection include not only trained staff, but also limited WAN bandwidth, the high cost of obtaining additional bandwidth, failure-prone equipment, manual processes, lack of centralized management, and high data growth rates. The risk of data loss or exposure from remote sites can be extremely high. As a result, it's no surprise that remote office data protection is top of mind for many IT executives.

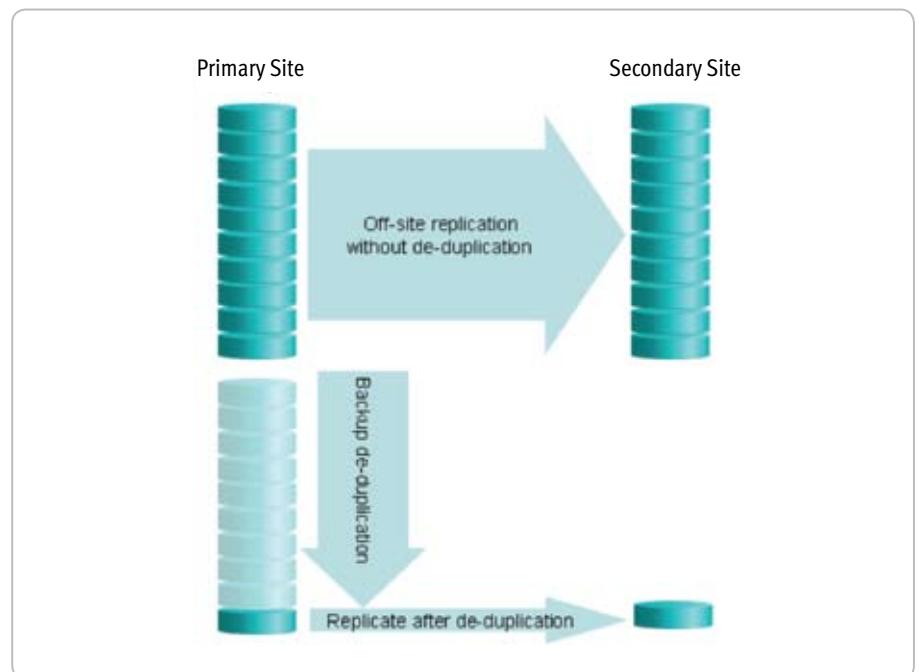
By implementing data de-duplication and its attendant reduction in bandwidth cost, users can centrally store all remote site backups. Centralizing backup relieves the operating expenses and reduces the risks associated with independent remote backup activities. By conducting the de-duplication at the source, before the data is shipped, the organization benefits from greatly reduced bandwidth requirements for backup while enabling centralized backup management.

## De-duplication Makes Offsite Replication Affordable and Reliable

De-duplication of backup provides considerable payoffs when considering disaster recovery. In order to have effective backup-based disaster recovery, backup images must be retained offsite at a sufficient distance from the data center to accommodate various site disaster scenarios. Unfortunately, backup replication (remote copy) over the wire doubles the storage requirements of backup, resulting in increased infrastructure, time, operational overhead, and bandwidth costs.

In order to ensure protection from site disasters, the backup image needs to be copied to the secondary site. Traditionally, organizations shipped backup tape to secure sites and retrieved those tapes to facilitate site recovery. Increasingly aggressive recovery service-level requirements have forced many companies to expedite the recovery process. As a result, backup replication over networks is emerging as a preferred method for protection from site disasters.

Performing data de-duplication, then replicating the de-duplicated data is a sound strategy for achieving disaster recovery more economically. The reduced backup image data resulting from de-duplication directly reduces the amount of content needed at the secondary site. Backup de-duplication directly lowers the infrastructure, time, operational overhead, and bandwidth costs needed to create backup images at the disaster recovery site. Additionally, disk-to-disk backup replication eliminates the risks associated with shipping physical tape cartridges.



## Why EMC for Backup De-Duplication

Today, EMC has the broadest portfolio of de-duplication offerings in the industry. Organizations of all types look to EMC for its experience and guidance in selecting the most optimal backup-to-disk solutions with advanced data de-duplication capabilities. EMC's de-duplication strategy is to offer solutions that:

- De-duplicate at the highest level of abstraction. EMC helps organizations preserve value by de-duping content as opposed to data and to enable pervasive re-use throughout the lifecycle.
- De-duplicate as close to the source as practical. EMC maximizes savings throughout the information lifecycle by maximizing the benefits of de-duplication while minimizing resource overhead and storage requirements.
- Find the most interoperable approach across multiple use cases. EMC technology and best practices enable combinations of storage and movement from primary to secondary to tertiary—for recovery and retention.

EMC backup products that feature de-duplication include:

**EMC® Avamar®.** Avamar provides source-based, global data de-duplication using variable segment length for the highest de-duplication efficiency. This industry-leading technology is well-suited for remote environments, VMware deployments, and bandwidth/backup-window-constrained environments in data centers.

**EMC Disk Library 3D 1500 and 3D 3000.** Highly scalable LAN backup-to-disk platforms with data de-duplication providing cost-effective storage, longer onsite retention and lower replication costs. These affordable platforms feature policy-based, sub-file data de-duplication to best match your various environments as well as hardware data compression and IP replication of de-duplicated data streamlines for offsite protection.

**EMC Disk Library DL4000.** This enterprise-scaleable SAN backup to disk platform now supports native data de-duplication. The EMC Disk Library 4000 grid architecture platform features policy-based, sub-file data de-duplication as well as separate backup, de-duplication and replication engines to ensure scaling. Additional savings are realized through disk spin-down technology and low-power 1TB SATA drives.

**EMC NetWorker®.** NetWorker is now available with data de-duplication capabilities. NetWorker offers the user a choice of service level—the opportunity to do traditional backups as well as non-traditional backups—to further optimize enterprise-wide backup and recovery from a central console.

**EMC Services.** These EMC and partner-delivered services include backup assessments, de-duplication estimation, and design/implementation services to clearly identify the benefits, architect a solution, and expedite the deployment of optimized backup leveraging data de-duplication.

## In Summary: Justifying De-Duplication

Many organizations are deploying backup-to-disk to augment or even eliminate their tape-based backup and recovery infrastructure. Why is the de-duplication of backup-to-disk content so compelling? De-duplication reduces backup costs and provides incremental benefits in the following areas:

- Reduces tape infrastructure costs
- Reduces disk capacity requirements
- Shortens backup windows
- Expedites data recovery compared to tape
- Reduces offsite replication costs
- Eliminates tape handling risks
- Reduces reliance on tape libraries for backup

Let EMC and EMC partners worldwide help your organization with world-class backup-to-disk and data de-duplication offerings. EMC's innovative product portfolio, extensive information management experience, and proven best practices ensure that you get the greatest value from your information infrastructure with the greatest efficiencies and lowest costs.



**EMC Corporation**  
Hopkinton  
Massachusetts  
01748-9103  
1-508-435-1000  
In North America 1-866-464-7381  
[www.EMC.com](http://www.EMC.com)