



Deduplication –  
InTechnology  
White Paper

**inTechnology**

# Deduplication – Hype or Reality

## InTechnology White Paper

June 2008

### Introduction

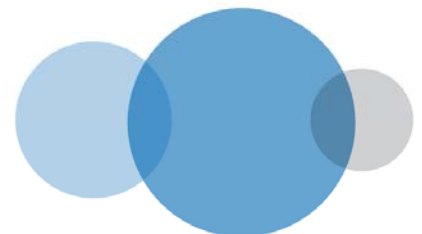
Deduplication is undoubtedly one of the hottest topics in data storage. The rationale behind deduplication is simple: Eliminate your duplicate data and reduce the capacity needed for primary storage as well as during backups and other data copy activities. Unfortunately, the many different deduplication approaches from various vendors, with much hype about their unique benefits, can leave users confused. As they consider the variety of deduplication offerings, they often fail to understand the basic differences between them. This paper looks beyond the hype, focuses on the meaning and application of deduplication and seeks to clarify what is rightly or wrongly encapsulated in the term deduplication.

### What Is Deduplication?

#### Definition 1 -Whatis.com

Data deduplication (often called "intelligent compression" or "single-instance storage") is a method of reducing storage needs by eliminating redundant data. Only one unique instance of the data is actually retained on storage media, such as disk or tape. Redundant data is replaced with a pointer to the unique data copy. For example, a typical email system might contain 100 instances of the same one megabyte (MB) file attachment. If the email platform is backed up or archived, all 100 instances are saved, requiring 100 MB storage space. With data deduplication, only one instance of the attachment is actually stored; each subsequent instance is just referenced back to the one saved copy. In this example, a 100 MB storage demand could be reduced to only one MB.

Data deduplication can generally operate at the file, block, and even the bit level. File deduplication eliminates duplicate files (as in the example above), but this is not a very efficient means of deduplication. Block and bit deduplication looks within a file and saves unique iterations of each block or bit. Each chunk of data is processed using a hash algorithm such as MD5 or SHA-1. This process generates a unique number for each piece which is then stored in an index. If a file is updated, only the changed data is saved. That is, if only a few bytes of a document or presentation are changed, only the changed blocks or bytes are saved,



the changes don't constitute an entirely new file. This behavior makes block and bit deduplication far more efficient. However, block and bit deduplication take more processing power and uses a much larger index to track the individual pieces.

## **Definition 2 -Wikipedia**

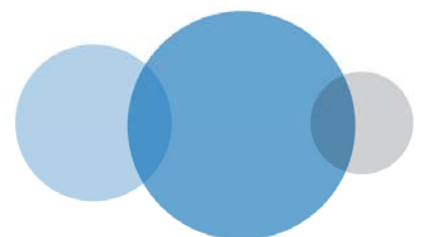
In computer data storage it is also known as *capacity optimisation* or *single-instance storage*.

*Capacity optimisation* technologies are similar to data compression technologies, but they look for redundancy of very large sequences of bytes across very large comparison windows. Typically using cryptographic hash functions as identifiers of unique sequences, long (8KB+) sequences are compared to the history of other such sequences, and where possible, the first uniquely stored version of a sequence is referenced rather than stored again. Capacity optimisation generally refers to the use of this kind of technology in a storage system. There are also implementations in networking (especially Wide Area networking), where they are sometimes called bandwidth optimisation technologies.

Commercial implementations of capacity optimisation are most often found in backup/recovery storage, where storage of iterating versions of backups day to day creates an opportunity for reduction in space using this approach. The term was first used widely in 2005.

*Single instance storage* is a system's ability to keep one copy of content that multiple users or computers share. It is a means to eliminate data duplication and to increase efficiency, SIS is frequently implemented in file systems, e-mail server products, data backup and other storage-related solutions.

In the case of an e-mail server, single-instance storage would mean that a single copy of a message is held within its database whilst individual mailboxes access the content through a reference pointer. However there is a common misconception that the primary benefit of single instance storage in mail server solutions is a reduction in disk space requirements. The truth is that its primary benefit is to greatly enhance delivery efficiency of messages sent to large distribution lists.



## InTechnology's View on Deduplication

In the context of disk storage, deduplication refers to any algorithm that searches for duplicate data objects, such as blocks, chunks, or files, and discards these duplicates. When a duplicate object is detected, its reference pointers are modified so that the object can still be located and retrieved, but it "shares" its physical location with other identical objects. This data sharing is the foundation of all types of data deduplication.

Data deduplication offers other benefits. Lower storage space requirements will save money on disk expenditures. The more efficient use of disk space also allows for longer disk utilisation periods, better recovery time objectives (RTO) and reduction of the data that must be sent across a WAN for remote backups, replication, and disaster recovery.

### How Does Deduplication Work?

Regardless of operating system, application, or file system type, all data objects are written to a storage system using a data reference pointer, without which the data could not be referenced or retrieved. In traditional (non-deduplicated) file systems, data objects are stored without regard to any similarity with other objects in the same file system.

In a deduplicated file system, two new and important concepts are introduced:

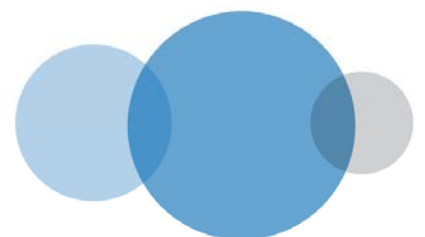
- A catalogue of all data objects is maintained. This catalogue contains a record of all data objects using a "hash" that identifies the unique contents of each object.
- The file system is capable of allowing many data pointers to reference the same physical data object.

Referencing data objects, comparing the objects, and redirecting reference pointers forms the basis of the deduplication algorithm.

### Inline or Post-processing

#### Inline deduplication

Deduplication is performed as the data is written to the storage system. With inline deduplication, the entire hash catalog is usually placed into system memory to facilitate fast object comparisons. The advantage of inline deduplication is that it does not require the duplicate data to actually be written to disk. The duplicate object is hashed, compared, and re-referenced on the fly.



## Post-processing deduplication

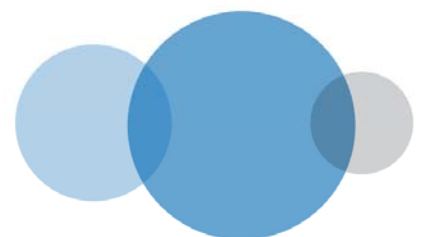
Deduplication is performed after the data is written to the storage system. With post-processing, deduplication can be performed at a more leisurely pace, and it typically does not require heavy utilisation of system resources. The disadvantage of post-processing is that all duplicate data must first be written to the storage system, requiring additional, although temporary, physical space on the system.

## Hype or Reality? Important Aspects of Inline and Post-processing

The decision regarding inline versus post-processing deduplication has more to do with the application than with any technical advantages or disadvantages. When performing data backups, the user's objective is completion of backups within an allowed time window. When adding deduplication to the backups, the user's objective is to free up the redundant storage space required for these backups.

These two objectives should not compete—the additional time required for deduplication should not drive backups beyond their allotted time window. An assessment should be made to determine if the time penalty of deduplication is offset by the space savings realised after deduplication, regardless of whether the deduplication is performed inline or post-processing. Other applications, such as primary storage and data archiving, do not lend themselves well to inline deduplication. In the case of primary storage, systems rarely have the performance headroom to facilitate inline deduplication. In the case of archival data, users may simply want to “clean” their file systems of duplicate data during periods of low activity, similar to other occasional storage housekeeping chores. In these environments, post-processing deduplication is the preferred method.

The bottom line is that organisations should evaluate their application environment and the cost of deduplication. If the priority is high-speed data backups with optimal space conservation, choose inline deduplication. If the organisation is interested in deduplicating primary storage or archival data, post-processing deduplication is probably the best bet.



## Source or Destination Deduplication

### Source Deduplication

Source deduplication refers to the comparison of data objects at the source, before they are sent to a destination (usually a data backup destination). The advantage of source deduplication is that less data is required to be transmitted and stored at the destination point. The disadvantage is that the deduplication catalogue and indexing components are dispersed over the network so that deduplication potentially becomes more difficult to administer.

### Destination Deduplication

Destination deduplication refers to the comparison of data objects after they arrive at the destination point. The advantage of destination deduplication is that all the deduplication management components are centralised. The disadvantage is that the entire data object must be transmitted over the network before deduplicating.

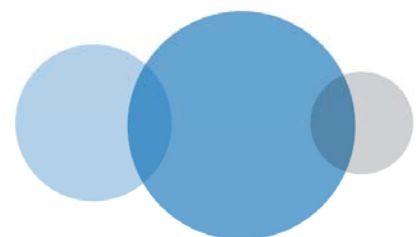
### Hype or Reality? Important Aspects of Source and Destination Deduplication

The decision to select source or destination deduplication is determined by the organisation's objectives. If the main objective is to reduce the amount of network traffic when copying files, source deduplication is the only option. If the goal is to simplify the management of deduplication and the amount of storage required at the destination, destination deduplication is the preferred choice.

### Deduplication Space Savings

Deduplication vendors often claim that their products offer 20:1, 50:1, or even greater data reduction ratios. These claims actually refer to the "time-based" space savings effect of deduplication on repetitive data backups, i.e. it refers to incremental backup by which only new and changed will be transmitted during the backup. Because the backups contain mostly unchanged data, once the first full backup has been stored, all subsequent full backups see a very high occurrence of deduplication.

But what if the business doesn't retain 64 backup copies? What if the backups have a higher change rate? Realising that space savings numbers from a vendor's marketing department often don't represent a real-life environment, what should be expected for space savings on backup data sets.



## Hype or Reality? Important Aspects of Space Savings

When evaluating deduplication space savings, the business should have two goals:

1. Examine the backup data. It is reasonable to expect 5:1 to 20:1 space savings (over time) with a backup change rate of 2%. If the business stores more than 20 backup copies on disk, or if the change rate is less than 2%, the deduplication ratio will increase. Conversely, if the business retains a lower number of backup copies, or if the data change rate is higher than 2%, the deduplication space savings ratio will be reduced.
2. Examine the non-backup data. Are there any opportunities to eliminate duplicate data on those volumes? Generally speaking, if this data can be reduced by 1.25 to 1.75:1, deduplication would be economically feasible. Think of it as receiving a “storage rebate” by reducing the storage capacity of these volumes by 20% to 40%. Enterprise storage systems today can easily exceed £5,000 per TB. Saving just a few TBs across your enterprise could justify the implementation of volume-based deduplication.

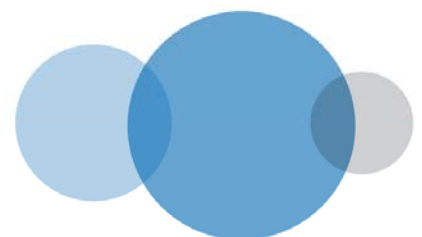
## InTechnology and How To Embrace Deduplication

### Data Analysis Audit

The Data Analysis Audit provides customers with an overview of data across their estate. InTechnology will assess the data storage on the current file and application servers to identify data usage, data type, age of data, data access date, storage growth, storage trends and – importantly- duplicate data. This will enable InTechnology to advise customers on the appropriate storage and data management policies for the customer’s primary data, backup data, potential archive data and what duplicate data could be deleted from primary storage.

### Main Features

- Display of storage resources and how to improve storage management
- View file-level detail from either server or NAS views to ensure consistent management
- Reduce or eliminate server outages and application downtime due to out-of-control disk space consumption
- Reclaim wasted capacity, identify age and out-of compliance files to save money and lower your total cost of ownership (TCO)
- Shorten backup windows so more work can be done in less time
- Increase storage utilization, further reducing the storage TCO



## Managed Archiving Service

InTechnology's Managed Archiving Service addresses soaring Microsoft Exchange server data volumes and mailbox archive files (pst files) on Windows file and print servers as these are creating unsustainable levels of cost and corporate risk in many organisations today.

This ever-growing volume of data - where typically 70% of data becomes inactive within 30 days of creation - is typically stored on expensive server attached disk systems. The impact of this never-ending capacity growth includes:

- o Additional capital expenditure for storage and server upgrades
- o Increased operational costs for the associated space, power and cooling requirements
- o Increased support staff costs in managing the growth in systems and storage
- o Increased backup hardware, software and media costs
- o Longer backup windows impacting application availability
- o Extended disaster recovery times

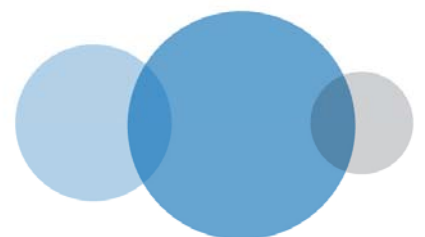
Traditionally the impact of these has been minimised by implementing operational processes to constrain data growth through techniques such as applying storage quotas for end-users, Exchange mailbox quotas, asking end-users to locally delete or archive aged email and mailbox archive files (pst files) on Windows file and print servers and using administrators to manage data growth - typically by backing it up and then deleting it.

These methods are typically unpopular with end-users, are labour intensive and involve unacceptable business risks that could conflict with the businesses data retention policies.

Intermixed with these operational challenges are growing legislative and business compliance requirements that demand businesses pay due diligence to data protection and availability.

This combination of cost, operational and legislative pressures are driving businesses to look at new data storage technologies and management solutions. To address these growing issues the InTechnology Managed Archiving Service provides data archiving for Microsoft Exchange email and mailbox archive files (pst files) on Windows file and print server data, based on the industry-leading Symantec Enterprise Vault software.

The Managed Archiving Service provides a variety of operational, commercial and business compliance benefits for Exchange email and file system environments and utilises destination deduplication to reduce the amount of data stored.



## Managed Backup Service

The Managed Backup Service (MBS) is a unique alternative to traditional backup and restore methods, replacing conventional tape based systems with a fully automated online solution. It provides centralised and automated backups of PC's, file servers and application/database servers with secure offsite storage and immediate online restoration.

More than ever, organisations of all sizes must strategically leverage their brand as well as manage costs to foster growth and innovation. A company's information, whether it be intellectual property or in the form of historical records and files, are their competitive assets. Access, or conversely, a lack of access to that information, can render its network- and PC-tethered workforce completely ineffective.

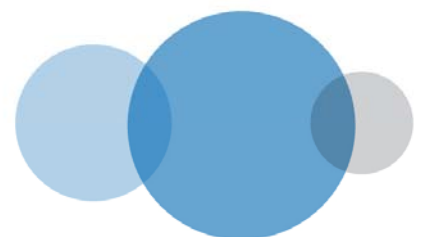
Those employees responsible for managing information access and protecting its integrity, from the server room to the board room, face increasing pressure focused around the following issues:

- Heightened awareness of business continuity and risk assessment
- Exploding data growth and the ability to manage it
- Dispersed environment fuelled by an increase in mergers and acquisitions
- Constant OS and application changes
- Increased regulatory requirements

Enterprise-level companies, with tens of terabytes under management, combat these issues with a team of experienced, well-compensated IT professionals armed with comparatively larger budgets than small- and mid-sized organisations. Small and medium businesses also deal with the same competitive pressures, but must alleviate them despite having small or no dedicated IT staff and tight budgets.

Because of these factors, mid-sized companies requiring data protection and rapid recovery want simplified management through one vendor, cost-effectiveness, more operational control, reliability and secure and fast recovery. Where mid-tier organisations dramatically differ and where MBS has a technology advantage is by helping them overcome their smaller budgets and IT staff.

MBS is a data protection and recovery service that enables a server or a group of heterogeneous servers to backup their data to a remote storage device over common telecommunication connections. The MBS Service allows for online restores of backup data transmitted over the same or alternate telecommunication connections, as well as facilitating the migration of data to lower-cost media for long-term storage.



Not only does the Managed Backup Service back up incremental data only – once the initial seed backup has been taken- but it also compresses and deduplicates data stored at the backend through destination deduplication. This functionality works on a backup set basis and allows for an average of 20-30% of data to be deduplicated. Thus even for backup set retentions for up to 3 months InTechnology will store no more than a ratio of 1:1 compared to primary source data. However this represents up to 20 x the stored data volume in recoverable data.

## **Managed Replication Service**

The InTechnology Managed Replication Service provides Customers with an affordable way to cope with the demands of storing, managing and recovering business-critical data from IT infrastructure failures, outages or site disasters within a timescale that meets business requirements.

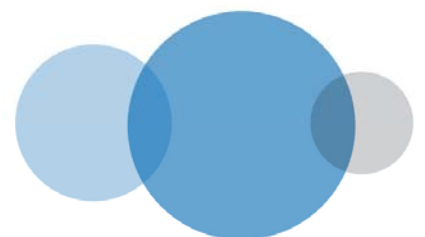
The traditional way to increase the overall resilience of your IT infrastructure is to address the issue on a local level, e.g. with the introduction of clustered servers and a highly resilient local storage.

However, this method does not give protection against site outages or a full site disaster. The next logical step should be to replicate your primary storage to InTechnology's secure off-site data centre using the Managed Replication Service. This will offer extra resilience and ensure your business will recover quickly.

InTechnology offers:

- A fully resilient on-site storage solution, by providing a choice of clustered storage head units, local snapshots and application specific recovery options.
- A network attached storage solution at one of InTechnology's data centres.
- A 24x7 managed, operated and maintained data replication service.
- Fast recovery time for your replicated data over our network or delivery of a portable storage system to your DR site.
- Provision of optional standby servers at InTechnology's data centres.
- Management information to enable you to track performance and trends.
- The capability to scale the solutions in line with your business growth.

InTechnology's Managed Replication Service uses backend storage from NetApp, a leader in data storage efficiency since 1992, who has established A-SIS deduplication as the first deduplication product to be used broadly across many applications, including data backup, data archival and primary data. A-SIS (Advanced Single Instance Storage) deduplication combines the benefits of granularity, performance, and resiliency to provide users with significant data deduplication benefits.



## Summary

Data deduplication is an important new technology that is quickly being embraced by users as they struggle to control data growth and distribution. By eliminating redundant data objects, an immediate benefit is obtained through space efficiencies. When choosing a deduplication technology or service, however, it is important to consider all aspects, including the type of deduplication, indexing, inline or post-processing, source or destination and of course space savings efficiency.

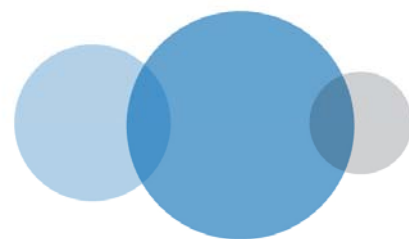
Many vendors currently offer deduplication, with more sure to follow, all with various approaches and techniques. It is clear that data deduplication will someday be a requirement for every storage vendor, much as snapshots became a requirement years ago.

Well-designed deduplication must perform without compromising data integrity and reliability. Deduplication magnifies the effect of data corruption. If a deduplicated data object becomes corrupt, it has far-reaching implications, because it is referenced by many other files and applications. Vendors will be required to provide 100% assurance that their design will prevent any such data inconsistencies

Deduplication must operate seamlessly in existing user environments. Users will not build a storage infrastructure around deduplication; rather, deduplication must fit into their existing environment with minimal disruption. Ultimately, deduplication must be a transparent background process.

Finally, deduplication will be required to have minimal impact on system performance. Users will not implement deduplication if it has a negative impact on their system workloads. This is particularly true as deduplication makes its way from backup applications to more performance-sensitive primary storage environments.

InTechnology's Managed Data Service not only embraces the concept of deduplication and the best practices listed above, but also uses best of breed storage hardware and storage software as well as industry-leading data management applications, which include data deduplication functionality as default.



InTechnology designs and supports the best IP solutions for business with a range of applications seamlessly integrating clients' communications needs through the delivery of secure voice, data and mobile solutions.

InTechnology employs 200 people and has data centres in Harrogate, London and Reading.

#### Head Office

Central House  
Beckwith Knowle  
Harrogate  
HG3 1UG  
Tel: 01423 850 000

#### London Office

17 St Helens Place  
Bishopsgate  
London  
EC3A 6DG  
Tel: 0203 040 5000

#### Reading Office

Commensus House  
3 – 5 Worton Drive  
Reading  
RG2 0TG  
Tel: 0870 777 7778

Enquiries: 0800 528 2522  
[www.intechnology.co.uk](http://www.intechnology.co.uk)

