



“ I would not hesitate to recommend Coolspirit to any prospective customers. They are responsive, very easy to work with and tailor solutions to suit customer’s specific needs, resulting in very high customer satisfaction. ”

Giovanni Goduti
CA Technologies – Recovery Management
& Data Modeling Business Unit

Our address

24 The Bridge Business Centre
Beresford Way
Chesterfield
S41 9FG

Get in touch

Call us on: 01246 454222
Email us: web@coolspirit.co.uk
Find us: [View location map](#)
Web: www.coolspirit.co.uk

Office hours

mon - thurs 8:30am - 5:30pm
fri 8:30am - 5pm
sat - sun Closed

“

Boost your storage buying power...
use ours!

”

Buy with confidence from
Coolspirit your authorised
CA Partner

Data Deduplication: An Essential Component of your Data Protection Strategy

JULY 2010

Andy Brewerton

CA TECHNOLOGIES – RECOVERY MANAGEMENT AND DATA MODELLING

Table of Contents

DATA DEDUPLICATION – THE DRIVER	2
<i>Can Disk Based Backup Manage Exponential Data Growth?</i>	<i>2</i>
THE DATA DEDUPLICATION REVOLUTION	3
<i>How does Deduplication work?.....</i>	<i>3</i>
<i>When does Deduplication Occur?</i>	<i>3</i>
<i>Where does Data Deduplication Occur?</i>	<i>4</i>
<i>What can you expect from Data Deduplication?</i>	<i>4</i>
CONCLUSIONS.....	5
CA ARCSERVE BACKUP – INTEGRATED DEDUPLICATION	6

Copyright ©2010 CA TECHNOLOGIES. All rights reserved. All trademarks, trade names, service marks and logos referenced herein belong to their respective companies. This document is for your informational purposes only. CA TECHNOLOGIES assumes no responsibility for the accuracy or completeness of the information. To the extent permitted by applicable law, CA TECHNOLOGIES provides this document “as is” without warranty of any kind, including, without limitation, any implied warranties of merchantability, fitness for a particular purpose, or noninfringement. In no event will CA TECHNOLOGIES be liable for any loss or damage, direct or indirect, from the use of this document, including, without limitation, lost profits, business interruption, goodwill, or lost data, even if CA TECHNOLOGIES is expressly advised in advance of the possibility of such damages.

DATA DEDUPLICATION – THE DRIVER

Can Disk Based Backup Manage Exponential Data Growth?

Organisations are facing many challenges when it comes to managing and protecting their business data. According to various industry experts, companies will experience a number of changes to their environment that will cause them to reconsider how they deploy data protection and business continuity solutions.

Virtualisation of servers continues to be high on the agenda for many companies. More companies are looking to drive efficiencies by deploying new applications on virtual platforms, and they are beginning to migrate the bulk of existing applications away from physical servers. While server virtualisation promises to simplify the infrastructure, what many companies find is that it increases the complexity of management functions, especially with regards to Data Protection tasks.

Recent weak economic climates are driving many organisations to cut costs, and this means that increasingly organisations have to manage more data with less budget. These reductions in budget could range from reductions in staff - resulting in loss of expertise through to reductions in capital spending - meaning at best, that environments can only be enhanced instead of completely replaced. To address these, many organisations are finding new cost effective ways to protect and operate their infrastructure.

Another challenge facing a lot of organisations is Merger and Acquisition activity. This is resulting in either a distributed workforce, or a consolidation exercise. Either way, IT teams need to manage more users, with more data, spread over more locations.

Sitting behind these challenges is the acceleration in the acquisition and generation of data. This at best complicates the other business challenges, and at worst introduces risk to an organisations' ability to continue delivering services and generating income.

When we look at these challenges from a data protection viewpoint we can see that it is simply a case of:

- Addressing expanding data generated by more users in more locations.
- Keeping the organisation running by ensuring high availability of key applications to maintain business processes.
- Supporting the ever-evolving environment. This spans everything from a move to virtualised servers, thru hardware upgrades, to new versions of important business tools.

The need to protect these increasingly high volumes of data has led to many organizations adopting disk-to-disk techniques for their backup and recovery options. Disk offers a high performance destination for the protected data, and can enable an organizations' ability to quickly recover an application when disaster strikes.

However, storing vast amounts of long term backup data on disk is becoming more difficult, it has led companies to questioning the economics of sending all backups to disk, or using disk only as a short term staging area. This is reducing the benefits associated with the disk-to-disk solutions by driving data recovery of older data to revert to tape.

To help alleviate this problem, data reduction technologies are being deployed as part of the data protection strategy, allowing more data to be crammed onto the disk targets, and therefore cope better with the data growth. Over the last few years the data reduction technologies used to resolve the problem have evolved:

- **Data Compression Algorithms.** Most data protection products have included Compression Algorithms for many years. Either as part of the core product, or as an option, the compression features are used to reduce the consumption of resources, such as backup destination or network bandwidth. The data would often be compressed at the client before transmission to the destination media, incurring a processing cost which occasionally would be detrimental to the performance of the applications being protected. The algorithms and techniques' used in the design of the data compression functions were varied, and often involved trade-offs between the degree of compression and the processing resources required to compress and uncompress the data.
- **Single Instance Storage.** (SIS) This is perhaps best described as File Level Deduplication. It refers to the process of keeping one copy of content that multiple users or computers share. It is a means to eliminate duplicate data and to increase efficiency of storage systems. SIS is often found within file systems, e-mail server software, data backup and other storage-related solutions.

One of the more common implementations of single instance storage is within email servers. SIS is used to keep a single copy of a message within the database, even when multiple users have been sent the data. This was implemented within email products to help them handle the dramatic increases in email volume, and managed to resolve problems associated with both the architectural boundaries imposed on the size of the database, and the performance impact of delivering one email to multiple recipients. It's quicker to set the pointers than it is to write many copies to disk.

When used within a backup solution, single instance storage can help to reduce the amount of target media required since it avoids storing duplicate copies of the same file. When protecting multiple servers or environments with a lot of users of unstructured data, identical files are very common. For example, if an organization has not deployed a collaboration tool such as Microsoft SharePoint, then many users will save the same document in their home directories, resulting in many duplicates consuming space on the backup media, and causing longer backup processes.

- **Data Deduplication.** Increasingly the term Data Deduplication refers to the technique of data reduction by breaking streams of data down into very granular components, such as blocks or bytes, and then storing only the first instance of the item on the destination media, and then adding all other occurrences to an index. Because it works at a more granular level than single instance storage, the resulting savings in space are much higher, thus delivering more cost effective solutions. The savings in space translate directly to reduced acquisition, operation, and management costs.

As a result of this evolution in data reduction technologies, disk continues to be seen as the ideal destination of the bulk of a company's backup and recovery operations, delivering the speed and flexibility that businesses require in a cost effective, high performance solution.

THE DATA DEDUPLICATION REVOLUTION

Data Deduplication technologies are deployed in many forms and many places within the backup and recovery infrastructure. It has evolved from being delivered within specially designed disk appliances offering post processing deduplication through to today being a distributed technology found as an integrated part of backup and recovery software. Along the way some suppliers of solutions have identified the good and bad points of each evolution and developed what today are high performance efficient technologies.

How does Deduplication work?

As with many things in the world of IT, there are numerous techniques in use for deduplicating data, some are unique to specific vendors, who guard their technology behind patents and copyrights, others use more open methods. The goal of all is to identify the maximum amount of duplicate data using the minimum of resources.

The most common technique in use is that of "chunking" the data. Deduplication takes place by splitting the data stream into "chunks" and then comparing the chunks with each other. Some implementations use fixed chunk sizes, other use variable chunk sizes. The latter tends to offer a higher success rate in identifying duplicate data as it is able to adapt to different data types and environments. The smaller the chunk size then the more duplicates will be found, however, performance of the backup and more importantly the restore is affected. Therefore, vendors spend a lot of time identifying the optimal size for different data types and environments, and the use of variable chunk sizes often allow tuning to occur, sometimes automatically.

During Deduplication every chunk of data is processed using a hash algorithm and assigned a unique identifier, which is then compared an index. If that hash number is already in the index, the piece of data is considered a duplicate and does not need to be stored again, and instead a link is made to the original data. Otherwise the new hash number is added to the index and the new data is stored on the disk. When the data is read back, if a link is found, the system simply replaces that link with the referenced data chunk. The deduplication process is intended to be transparent to end users and applications.

When does Deduplication Occur?

Deduplication can occur in any of three places. "At the client"- where the source data sits, "in-line" – as the data travels to the target, or "on the target" – after the data has been written, the latter is often referred to as "post process". All three locations offer advantages and disadvantages, and one or more of these techniques will be found in the deduplication solutions available on the market today. The choice of which type of deduplication an organization deploys is governed by their infrastructure, budgets, and perhaps most importantly, their business process requirements.

Post Process Deduplication

This works by first capturing and storing all the data, and then processing it at a later time to look for the duplicate chunks. This requires a larger initial disk capacity than in-line solutions, however, because the processing of duplicate data happens after the backup is complete, there is no real performance hit on the data protection process. And CPU and memory requirements for use in the deduplication process are consumed on the target, away from the original application and therefore not interfering with business operations. As the target device may be the destination for data from many file and application servers, post process deduplication also offers the additional benefit of comparing data from all sources – this Global Deduplication increases the level of saved storage space even further.

In-Line Deduplication

The analysis of the data, the calculation of the hash value, and the comparison with the index all takes place as the data travels from source to target. The benefit being it requires less storage as data is first placed on the target disk, however, on the negative side because so much processing has to occur, the speed of moving the data can be slowed down. In reality, the efficiency of the in-line processing has increased to the point that the performance on the backup job is so small it is inconsequential. Historically, the main issue with in-line deduplication was that it was often focused only on the data stream being transported, and did not always take into account data from other sources. This could result in a less “global” deduplication occurring and therefore more disk space being consumed than is necessary.

Client Side Deduplication

Sometimes referred to as Source deduplication, this takes place where the data resides. The deduplication hash calculations are initially created on the client (source) machines. Files that have identical hashes to files already in the target device are not sent, the target device just creates appropriate internal links to reference the duplicated data and results in less data being transferred to the target. This efficiency does however, incur a cost. The CPU and memory resources required to analyze the data will also be needed by the application being protected, therefore, application performance will most likely be negatively affected during the backup process.

Where does Data Deduplication Occur?

As Data Deduplication solutions have evolved, they have been packaged into a variety of products. The first major deployment of deduplication technology came in disk storage arrays. The units consist of a processor to manage the deduplication process, a bunch of disk to store the data, and a number of data connections through which the source data will travel. Different vendors deployed different techniques, some delivered post-processing capabilities, and others did their deduplication in-line. Some tuned their boxes to store generic data, while others had intelligence built in that helped recognize specific data types to increase deduplication efficiency. Most vendors offered connections to servers over FibreChannel and iSCSI connections, and others included Network Attached Storage (NAS) options. The units either looked like standard disk, or emulated Tape Libraries. The latter allowed for seamless integration with backup and recovery solutions already in place at a customer’s site.

Deduplication technology has now evolved from within the hardware disk array and is built into a number of backup and recovery solutions. Putting deduplication within the backup application offers numerous advantages, not least the extra efficiencies that are often gained in performance, and the non-reliance on proprietary disk drives. The latter frequently leads to a more cost effective solution. These software solutions can be found in “standalone” software appliances, or fully integrated components of the backup product.

What can you expect from Data Deduplication?

The way in which an organization chooses to deploy data deduplication will depend on a number of factors, and many of these will be specific to the environment they operate in. In designing the deployment of deduplication careful thought should be put into:

- **The Types and the Location of the Data.** Depending on the technology in use, different types of data will affect the efficiency of the deduplication process. For example, backing up images of multiple virtual machines often leads to high disk saving figures, as there is a lot of duplication of operating system data across each of the images. In comparison, heavily compressed data, or encrypted data may not offer as much savings as the content is more unique. The use of different “chunk” sizes during the deduplication phase can alleviate some of these issues.

Highly Distributed Source Data can also prove difficult to deduplicate if the technology being used does not provide a “global” view.

- **Impact on Backup Performance.** The different technologies available for deduplication each have a different impact on the performance of the backup job. The post process techniques would on the face of it appear to offer the least impact, however, given the need for additional staging space on the target disk, and the subsequent additional I/O's to run the comparison jobs, the affect on the start to finish process can be extreme. In-band processing, although being done as part of the backup job, is beginning to prove more efficient as the deduplication vendors improve their algorithm and index lookup technology.
- **Restore Performance.** Often overlooked in the decision making process, but realistically the most important aspect. The restore speed of the deduplicated data has a direct impact on the availability of the business applications.
- **Deduplication Ratios.** This is the headline of many pieces of deduplication marketing, as this speaks directly to the space savings that the organization is going to achieve. The reality is that the deduplication ratio is affected by many factors; the data, the change rates of the data (few changes mean more data to deduplicate), the location of the data, the backup schedules (more fulls makes compression effect higher), and perhaps least by the deduplication algorithm. 95% reductions in disk usage are achievable, however, organizations should be pleased if the savings are over 50%, as these still represent big reductions in capital and operational spending.
- **Capacity and Scalability.** Even with highly efficient deduplication technology in place, an organization will eventually run out of capacity. Therefore, before choosing a technology, understand the implications of outgrowing your capacity. Will it mean maintaining numerous “silos of storage” or require a forklift upgrade to a new system? Will you be tied to a specific hardware vendor, or will you be able to flexibly add capacity as needed.

CONCLUSIONS

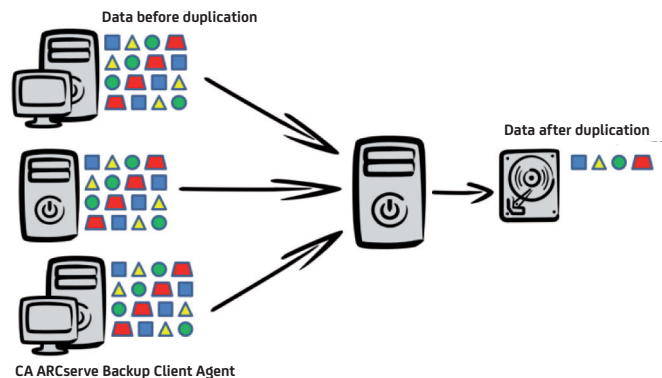
The data growth challenges that every organization is facing are pushing the implementation of new backup and recovery technologies to help meet service level agreements around availability and performance. Shrinking budgets are pulling IT departments in the opposite direction. Data Deduplication is a technology that helps organizations balance these opposing demands. Disk based backups can be rolled out in order to reduce the backup window and improve recovery time, and deduplication means the investment in those disk based targets is maximized.

Organizations should review the different deduplication technologies on the market and choose a solution that is able to integrate seamlessly into their backup environment, in a cost effective way. Making sure that any investment does not tie them into a hardware solution that is difficult to expand as the organization's data grows.

CA ARCSERVE BACKUP – INTEGRATED DEDUPLICATION

CA ARCserve Backup provides integrated data deduplication technology that enables you to

- Reduce your storage needs
- Analyze data at a block level to increase performance
- Improve restore times by using index files to identify data needed to reassemble the original data streams



CA ARCserve Backup includes a next generation patent-pending data deduplication technology built using highly efficient deduplication algorithms that help eliminate redundant data resulting in extended use of existing storage, or reducing the need for additional storage. CA ARCserve Backup Data Deduplication uses an in-line process occurring at the backup target that compares backups with previously stored data at the block-level resulting in considerable data reduction (up to 95% or more) it is highly efficient thus reducing the impact on the CPU during the backup process and does not affect the production servers being backed up.

The deduplication process works by working at the data block level rather than the files level, this approach provides greater economy and increases disk space savings when identifying data for deduplication. The first time a block is backed up, it will be copied to the backup media and a reference to this block added to an index. During later backup operations, for any files that contain the same data blocks, the block itself will not be copied to the media but references to the original unique block will added to the files index. In this way, CA ARCserve Backup can quickly recreate the file during restore operations.

In CA ARCserve Backup, data deduplication is a process that occurs in a single session at the backup server; it is an implementation of global deduplication, which enables you to identify redundancy between backup jobs across different computers in order to maximize disk space savings. CA ARCserve Backup delivers high levels of efficiency; for example, it compares Microsoft Exchange data to Microsoft Exchange data (per backup job) as opposed to comparing Exchange with Microsoft SQL data, resulting in a higher probability of finding duplicated data. The deduplication pass also eliminate any data that has not been changed since the last backup from being processed – further improving efficiency.

Data deduplication is included in CA ARCserve Backup at no extra cost.

CA Technologies is an IT management software and solutions company with expertise across all IT environments—from mainframe and physical to virtual and cloud. CA Technologies manages and secures IT environments and enables customers to deliver more flexible IT services. CA Technologies' innovative products and services provide the insight and control essential for IT organizations to power business agility. The majority of the Global Fortune 500 rely on CA Technologies to manage their evolving IT ecosystems. For additional information, visit CA Technologies at www.arcserve.com.