Acronis®
COMPUTE WITH CONFIDENCE
WWW.ACRONIS.COM

How Deduplication Benefits Companies of All Sizes

An Acronis® White Paper

Table of contents

Executive Summary

Primary storage in small and large companies alike is growing at 50% - 100% a year. And, according to IDC research conducted in the second half of 2008, the amount of global digital data created and stored on a worldwide basis has increased over 3,000% in just three years. In addition, many multiple-site organizations are working to consolidate data assets along with system consolidations (including virtualization) to create a less energy-intensive collection of assets that fit in a reduced physical space.

Carrying costs associated with storing and managing all that data on disk or tape can be cut dramatically by deduplication

The benefits of data deduplication have been well publicized; but in the most basic sense they enable an organization to:

- store far more backup data for a given expenditure or
- substantially lengthen disk purchase intervals,
- store to disk cost efficiently, taking advantage of its speed and eliminating the need for tape
- effectively reduce its backup window.

If deduplication is such a cost effective data reduction technique, why doesn't every IT organization use it? Until recently, the cost of proprietary hardware deduplication products has priced large and small organizations out of consideration. That same cost concern forced the relatively small percentage of organizations who *could* afford it to reserve it only for server data, despite the fact that workstation data frequently represents half of the entire data owned by an organization. However, the advent of software-only deduplication has substantially lowered the threshold for purchase, making it attractive to organizations of all sizes, and allowing workstation data to be deduplicated as well.

In this white paper, we'll define deduplication, detail its benefits and make a business case for using it in Windows® and Linux® environments

What is Deduplication?

Deduplication is designed to eliminate redundant data in a storage system, and it is designed to reduce the amount of data that must be stored as a backup. It can act at the file or block level.

How do these levels differ?

File-level Deduplication

File-level deduplication searches for any files that are exactly alike and stores only one copy, placing 'pointers' in place of the other copies. While file deduplication is more efficient than no deduplication whatsoever, even a single minor change to the file will result in an additional copy being stored.

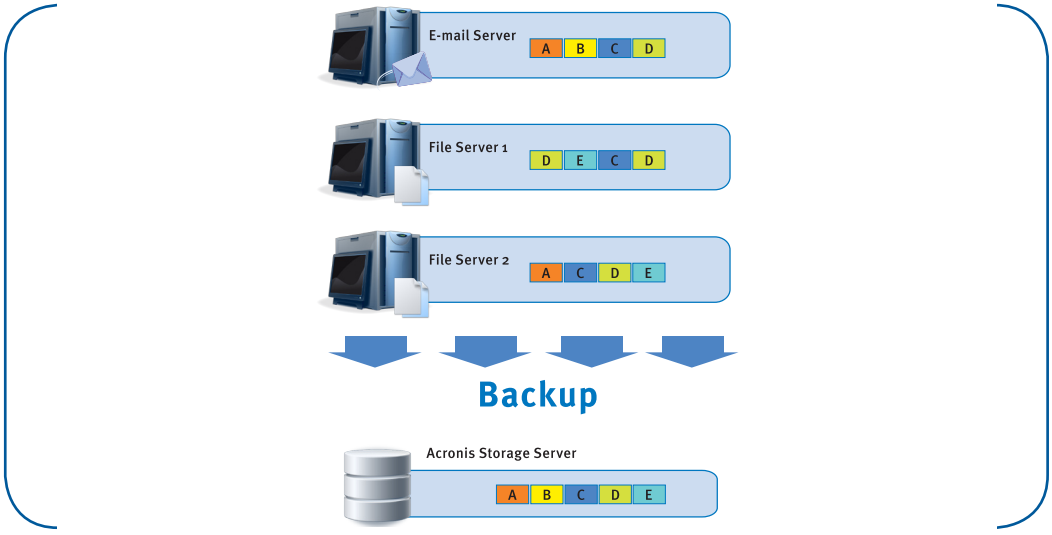
Block-level Deduplication

Block-level deduplication promises much greater overall storage efficiency. It works by searching for instances of redundant information by looking at chunks of data sized 4KB and larger and stores only one copy, regardless of how many copies there are. The copies are replaced by pointers which reference the original block of data in a way that is seamless to the user, who continues to use a file as if all of the blocks of data it contains are his or hers alone.

Deduplication cuts data storage volume by as much as 90%

To illustrate the power of deduplication, consider the impact on your back up system when you email a Microsoft® Powerpoint® presentation, full of graphics and eating up 9 MB of space, to 10 colleagues in your company. When you push the 'send' button, you clone 10 copies of that 9 MB file. When each recipient's data is backed up using traditional techniques, each instance of the presentation is backed up and stored. Suddenly a 9 MB file now occupies 99 MB of backup storage. Multiply this by hundreds of other instances of data cloning that occur throughout each day and you start to understand why disk storage requirements, and resulting costs, have climbed so steeply.

Deduplication is a proven way to reduce initial storage acquisition costs while saving network bandwidth. It makes it possible to either increase the data storage capacity per storage unit (stretching the time between additional data storage purchases), or to retain online data for longer periods of time.



Users might start to invest more—not just in raw capacity, but on tools that would help to maximize storage utilization [e.g., thin provisioning, data deduplication and storage virtualization].
Natalya Yezhkova, Research Manager, storage systems, IDC. January 2009

The whole process (left) can be carried out as a:

- Source function where duplicates are eliminated before they are written to a target disk
- *Target* function that identifies duplicate data already written to disk and removes them

Why is it important?

For many companies, deduplication will reduce data volumes so substantially that all backups can be kept on disk, obviating the need for tape and offline tape storage except perhaps for long-term archives. With this transition, administrators can accomplish faster backups and recoveries inherent with disk-based data protection solutions. Deduplication also makes it easier to meet government and financial reporting requirements, having to store all the copies that would be generated over a several years.

Addressing Data Integrity Concerns

While deduplication can save vast amounts of disk space, its very concentration makes it critical to store it properly. If a data block found on several sources (as in our earlier example of the Powerpoint presentation) is deduplicated and then lost, all of the backups associated with it will be damaged, since the source data is now non-existent. It applies to full system backups as well. If disaster strikes, a single damaged data block corresponding to a vital part of a Windows OS will make all the backups unusable for a system recovery. Please consider using RAID array to store deduplication data to provide an extra level of protection.

How Can Deduplication Benefit Your Organization?

General Benefits

Deduplication promises that organizations can store many times as much data per storage unit than before. Alternatively, for the same expenditure, they can chose to retain online data for longer periods of time. Either way, it translates into several business benefits:

- **effectively increased network bandwidth** - no copies need to be transmitted over the network if deduplication takes place at the source
- **a "greener" environment** - less electricity, fewer cubic feet of space required to house the data in both primary and remote locations
- **faster recoveries** ensure that line-of-business processes continue unimpeded
- preserves **the ability to respond to legal and corporate data storage compliance requirements** without adding storage bloat
- **fast return on investment** - because you're buying and maintaining less storage
- **smaller backup window** Backing up pointers to the data rather than the *data copy* itself takes only a tiny fraction as much space
- **lower overall cost of storage -** because you're storing less

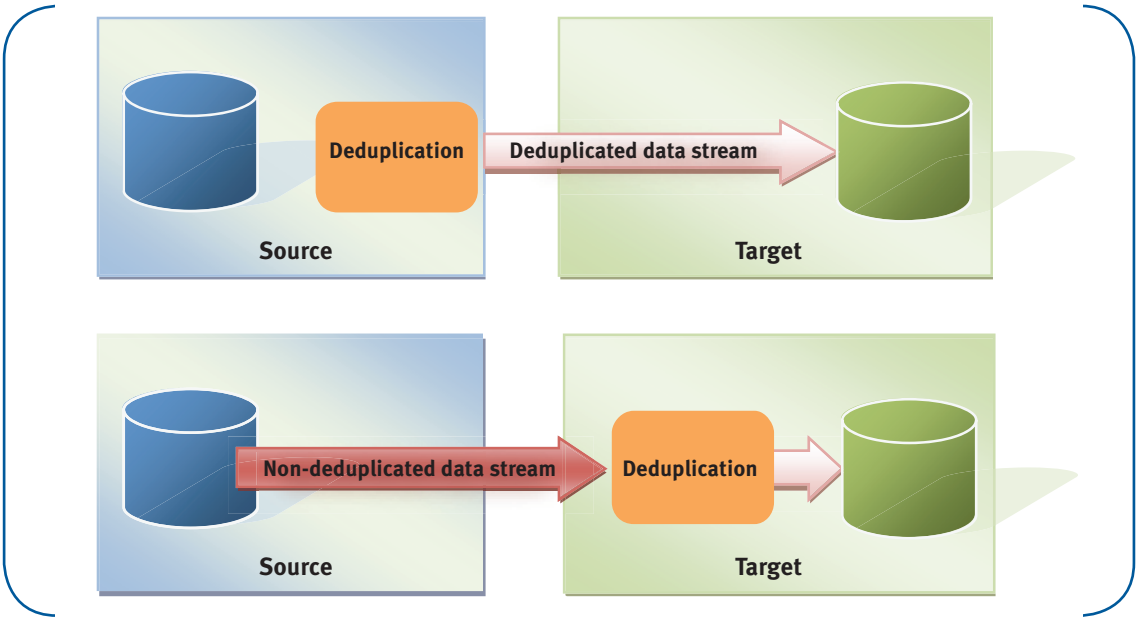
Specific Advantages of Source Deduplication

Source (or server-side) deduplication (shown in the top part of the following graphic) can:

- reduce the amount of data transferred over a network to a target storage location by 10 to 20 times
- eliminate a potential transmission bottleneck, particularly in scenarios where existing networks are already running at near capacity or where you are carrying out remote office backups over limited-bandwidth communications lines
- be effective for all types of stored data whether they are application-aware or not
- be easier to implement, as it doesn't require additional hardware or clients on the target side

Its main disadvantage?

Backups can take longer and use a lot of CPU cycles in the process of deduplicating data, possibly introducing performance issues on production machines. However, as we'll discuss later in the white paper, a new technology called *performance-optimized source deduplication* can eliminate most of source deduplication's performance tradeoffs.



Source vs. Target Deduplication

Specific advantages of target deduplication (bottom part of above graphic).

Target deduplication takes place after the source has been backed up, at the target storage location, typically on an attached storage node (ASN).

Its main advantage?

The initial backup at the source can be completed more quickly by moving CPU-intensive deduplication off the source machine, shortening the backup window. Target deduplication is often preferred in situations where administrators are supporting deduplication-unaware clients and data sources, or when the processing overhead associated with source deduplication will lengthen backup windows beyond the time limits set by administrators.

Its main disadvantage?

All copies that existed prior to deduplication must be sent over the network, potentially causing a bandwidth bottleneck. The choice of source versus target deduplication will depend on which constraint – *client CPU processing overhead or bandwidth considerations* – is most important to your organization.

Summary

Deduplication used to be an exclusive tool of the large enterprise, with an imposing cost, a daunting learning curve, and – with file-only deduplication – a limited ability to use deduplicated data to restore a failed machine. Until now, deduplication has been too expensive to implement in any but the largest organizations. Moreover, it could be applied only in support of servers, despite the fact that enormous data stores are contained at the workstation level within most IT infrastructures.

Most deduplication products have been designed and sold as combined software/hardware solutions. In most cases the hardware alone has been difficult to justify because of its high cost. To illustrate the latter, consider the fact that one well-known vendor reduced the cost of one of its high-end data deduplication appliances in March 2009 by more than one third. But at a reported $130,000 for 12 TB of storage capacity, it is still an expensive proposition. Roadblocks like these have limited the promises of deduplication to the largest of organizations. However, such limitations are finally being swept aside, and deduplication can be specified more broadly:

- not only by enterprises, but also by many smaller organizations which have very significant data storage challenges
- not only on servers, but on workstations as well

Introducing Acronis Backup & Recovery 10 Deduplication

Acronis deduplication has several advantages that distinguish it from the offerings of many other vendors:

- **Your choice of source or target deduplication**
  Many organizations need to be able to implement both in different parts of the organization.
- **File-and block-level backups**
  File-only backups severely limit the potential savings possible with deduplication.
- **Fast source deduplication**
  Often eliminates the need to deduplicate at the target. Acronis has made great strides in reducing CPU overhead in source-side deduplication, eliminating the need for many organizations to offload the job to target deduplication servers or appliances.Affordable. More organizations can now cost-justify deduplication, not only for their servers, but for their workstations, too.
- **Integrated with Acronis backup and recovery products**
  Works seamlessly with Acronis Backup & Recovery 10 software, so deduplicated data is just as well protected as unduplicated data.

Acronis Backup & Recovery 10 Deduplication is delivered as an optional, fully integrated module for our just-introduced Acronis Backup & Recovery 10 software products. A software-only solution, the Acronis deduplication offering may be purchased with these *advanced editions* of Acronis Backup & Recovery 10:

- Advanced Server Edition
- Advanced Workstation Edition
- Advanced Server SBS Edition
- Advanced Server Virtual Edition

Acronis' Deduplication Advantages

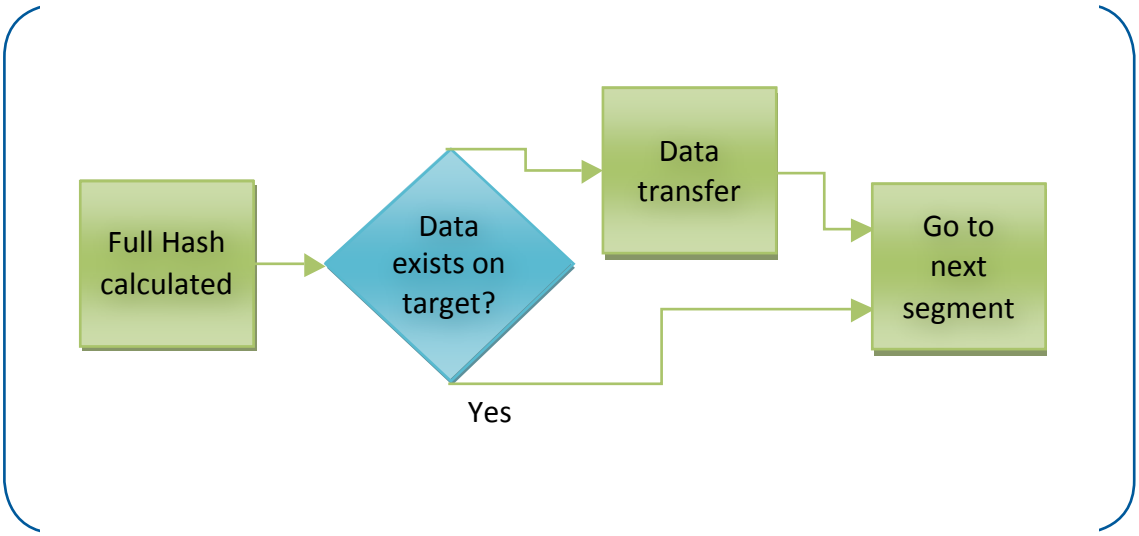
Unlike many other deduplication solutions, Acronis supports both source and target deduplication. But it also distinguishes itself in several other ways:

- **Image-based backup**
  Data can be deduplicated, providing either network or storage savings.
- **Fully integrated with Acronis disaster recovery software**
  Recovers files and systems – deduplicated and otherwise – in minutes rather than hours or days. Eliminates storing multiple copies of large sources of data, like multi-gigabyte operating systems, in the images.
- **Multi-type backups**
  Deduplication can be applied to full, incremental and differential backups.
- **Sensible cost**
  Software-only deduplication from Acronis is affordable.
- **Uses commodity storage hardware.** Acronis deduplication bypasses the need for costly proprietary hardware.
- **Installs fast**
  Typically it is up and running in about an hour, rather than in the several days required with hardware/software systems.
- **Easy to use**
  The same ease of use and reduced training requirements that distinguish Acronis products make Acronis deduplication a pleasure to set up and use.

Unparalleled storage efficiency is a reality, especially when combined with powerful Acronis compression algorithms (and other efficiency-oriented features) available in Acronis Backup & Recovery 10. Used in conjunction with Acronis' powerful data compression utility, IT administrators can cut overall data store size further, *after deduplication*, by an average of *50% - 60%* depending on file types, creating substantial additional disk storage savings. An attached storage node can be used to compress the data itself during its repack procedure, lifting the processing burden from production line servers. Both the Acronis .tib file and the deduplication data storage blocks will be compressed.

Quick Hash Algorithm: *key to optimized source deduplication*

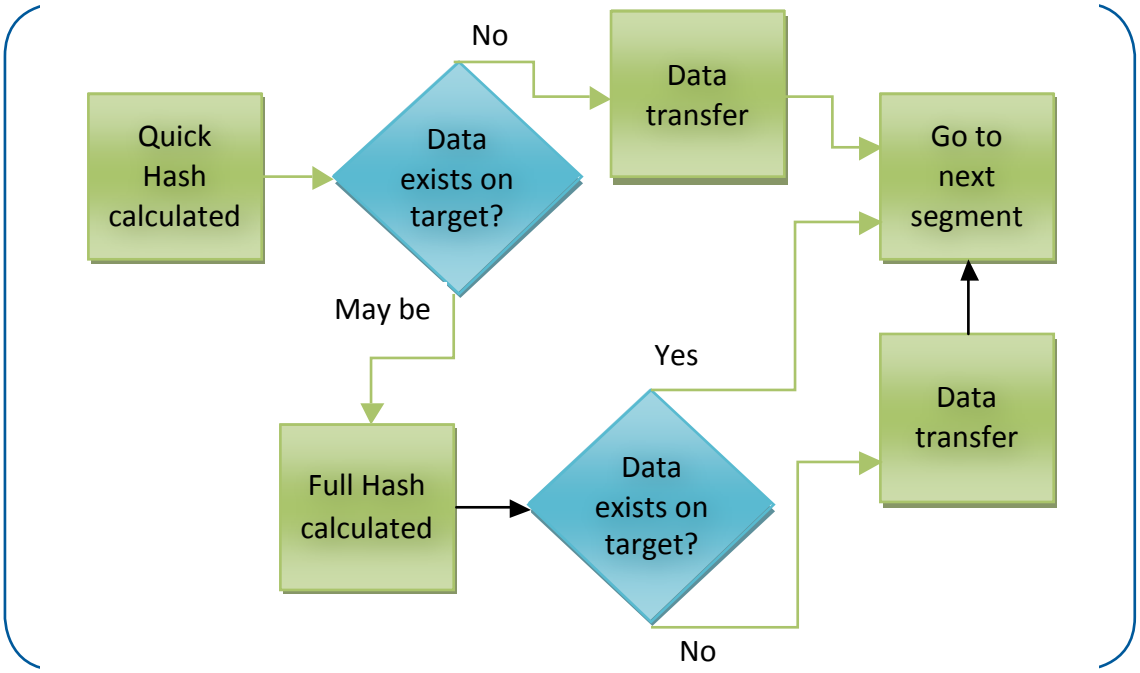
Acronis delivers a more efficient approach to source-side deduplication. To explain what we've done, let's look at a standard source deduplication algorithm (below). Here, the client software first calculates the data checksum of the data to be backed up – called *hash*. This hash is then sent to the target – which responds with either "I do not have the data" or "I do have the data." In the first case, the client will send the actual data to the target before proceeding with the next portion. In the second case, no further action from the client software is required, and the next portion of data may be processed, as shown here.



Standard source deduplication

Unfortunately, standard source deduplication creates significant overhead by *always* calculating the hash, regardless of whether the target does or does not have the data. This is required because the target cannot tell if the data is already available before the hash is provided by the source. On heavily loaded systems, standard source deduplication can create a system slowdown that might turn an IT manager against using it at all.

Acronis offers a much-less CPU-intensive approach that makes source deduplication a viable option for most companies. It's called **performance-optimized source deduplication.** This powerful algorithm *eliminates* most full hash calculations for data which has yet to be written on the target.



Acronis performance-optimized source deduplication

In this approach Acronis first creates *quick hash* by selecting a small amount of data that is statistically most likely to change when the data is modified. Quick hash is very fast, responding either to "I do not have the data" or "I may have the data." In the first case, the actual data is sent by the client. In the second case, full hash is calculated, which ensures that the target will respond reliably.

For security, we encrypt deduplicated data. One can specify a vault encryption password –protected in Windows secure storage – during vault creation. The encrypted data is accessible only through that password, and any attempt to retrieve data from the deduplication data-storage vault will fail without it.

Taking the next step

While Acronis is not the first company to offer deduplication, our image-based technology, with its fast backups and nearly immediate restores, brings deduplication to a new level, applicable to both file and system backup data, and to servers and workstations. Acronis makes duplication more accessible – both financially and from an ease-of-use perspective – to more users. When used in conjunction with Acronis Backup & Recovery 10, it redefines data protection. Here is what you can do to bring deduplication into your organization:

1: Try our Deduplication Calculator on our website. You can quickly determine just how much you can save using Acronis Backup & Recovery 10 Deduplication software.

2: Try it for yourself with a trial download. You'll need to also download Acronis Backup & Recovery 10 in order to use it.

3: Learn more at our website, www.acronis.com, or call us at one of the numbers listed at the end of this document for more details.

Acronis®
COMPUTE WITH CONFIDENCE
WWW.ACRONIS.COM

For additional information, please visit http://www.acronis.com

Enterprise/SMB sales:
Email: sales@acronis.com
Call +1 877 669-9749

OEM inquiries:
Email: oem@acronis.com
Call +1 650 875-7593