A White Paper

Demystifying Data Deduplication: Choosing the Best Solution

Abstract: *Data deduplication has become a primary requirement for backup implementations. Many vendors lay claim to offering the best data deduplication approach, leaving customers to face the difficulty of separating hype from reality. A number of key factors must be considered in order to select a data deduplication solution that actually delivers cost-effective, high-performance, and scalable long-term data storage. This document provides the background information required to make an informed data deduplication purchasing decision.*

Introduction

While data redundancy was once an acceptable operational part of the backup process, the rapid growth of digital content in the data center has pushed organizations to rethink how they approach this issue and to look for ways to optimize storage capacity utilization across the enterprise.

With the advent of different data deduplication technologies and methods to optimize storage capacity utilization, IT directors and storage administrators are left to make a difficult choice on key initiatives. There are many providers of data deduplication solutions today, and each vendor lays claim to offering the best approach. What's more, some vendors set unrealistic expectations by predicting huge reductions in data volume, ultimately disappointing customers. Companies must consider a number of factors in order to select a data deduplication solution that suits their needs and provides significant value with minimal disruption to infrastructures and processes.

In general, we will refer to a data deduplication solution as an Intelligent Disk Target (IDT) throughout this document.

Data Deduplication is an Operational Requirement

One might wonder what caused the proliferation of duplicated data in the first place. Ironically, the current industry-standard backup practice is the number-one cause of duplication. In the interest of data protection, the traditional backup paradigm copies data to a safe secondary storage repository over and over again, creating a monstrous overload of backed-up information. Under this scenario, every backup exacerbates the problem.

Because secondary storage volumes are growing exponentially, companies need a way to dramatically reduce these data volumes. Regulatory requirements magnify the challenge, forcing businesses to change the way they look at data protection. By eliminating duplicate data and ensuring that data archives are as compact as possible, companies can keep more data online longer – at significantly lower costs. As a result, data deduplication is now a required technology for any company wanting to optimize the performance, efficiency, and cost-effectiveness of its data storage environment.

Data deduplication can minimize the bandwidth needed to transfer backup data to offsite archives. With the hazards of physically transporting tapes being well-established (damage, theft, loss, etc.), electronic transfer is fast becoming the offsite storage modality of choice for companies concerned about minimizing risks and protecting essential resources.

Although compression technology can deliver an average 2:1 data volume reduction, this is only a fraction of what is required to deal with the data deluge most companies now face. Only data deduplication technology can provide the reductions in data volumes that customers need.

Key Criteria for a Robust IDT

There are ten important criteria to consider when evaluating an IDT:

1. Focus on the largest problem

The first consideration is whether the IDT can attack the area where the largest problem exists: backup data in secondary storage. Duplication in backup data can cause its storage requirement to be many times that which would be required if the duplicate data could be eliminated.

Figure 1, courtesy of the Enterprise Strategy Group (ESG), illustrates why a new technology evolution in backup is necessary. Incremental and differential backups were introduced to decrease the amount of data required compared to a full backup. Even within incremental backups, there is significant duplication of data when protection is based on file-level changes. When considered across multiple servers at multiple sites, the opportunity for storage reduction by implementing an IDT becomes huge.

2. Integration with current environment

An effective IDT should be as non-disruptive as possible. Many companies turn to virtual tape library (VTL) technology as a method of implementing an IDT and improving the quality of their backups without significant changes to policies, procedures, or software. It also focuses on the largest pool of duplicated data: backups.

Other companies leverage a disk-to-disk (D2D) backup paradigm, which requires that an IDT presents a network interface to the backup application. This process simplifies and enhances D2D backups, performing deduplication without disruption to ongoing operations.

Solutions requiring proprietary appliances tend to be far less cost-effective than those providing more openness and deployment flexibility. An ideal solution is one that integrates with an organization's existing backup environment and is available in flexible deployment options to provide global coverage across the data center as well as branch and remote offices.

3. VTL capability

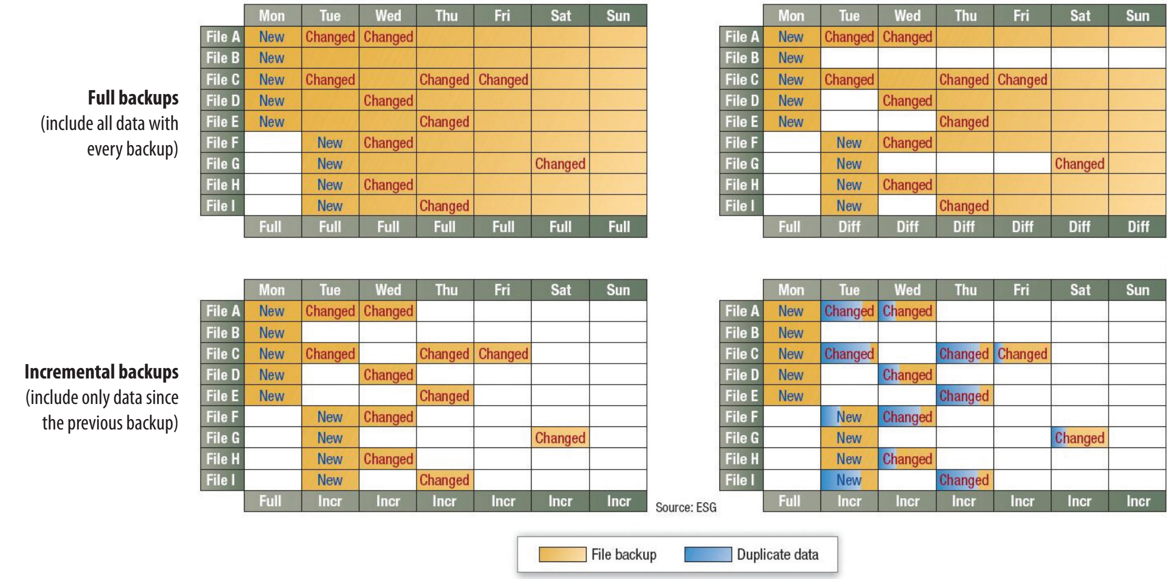
As noted, VTL-based data deduplication is one of the least disruptive ways to implement an IDT in a tape-centric environment. In such cases, the capabilities of the VTL itself must be considered as part of the evaluation process. It is unlikely that the savings from data deduplication will override the difficulties caused by using a sub-standard VTL. Consider the functionality, performance, stability, and support of the VTL as well as its deduplication extension. Keep in mind how well the VTL can emulate your existing tape environment (e.g. same libraries, same tape formats) and communicate with your physical tape infrastructure if required.

4. Impact of deduplication on backup and restore performance

It is important to consider where and when data deduplication takes place in relation to the backup process. Although some IDTs attempt deduplication while data is being backed up, this inline method processes the backup stream as it comes into the deduplication appliance, making performance dependant on the single node's strength. Such an approach can slow down backups, jeopardize backup windows, and degrade the IDT performance over time.

By comparison, IDTs that run after backup jobs complete, or concurrently with backup processes, avoid this problem and have no adverse impact on backup performance. The backup data is read from the backup repository after backups have been cached to disk. This ensures that backups are not throttled by any storage limitations. An enterprise-class solution that offers this level of flexibility is ideal for organizations looking for a choice of deduplication methods.

Figure 1: How various backup methods measure up



Full backups
(include all data with every backup)

	Mon	Tue	Wed	Thu	Fri	Sat	Sun
File A	New	Changed	Changed				
File B	New						
File C	New	Changed		Changed	Changed		
File D	New		Changed				
File E	New			Changed			
File F		New	Changed				
File G		New				Changed	
File H		New	Changed				
File I		New		Changed			
	Full	Full	Full	Full	Full	Full	Full

Differential backups
(include all data modified since the last full backup)

	Mon	Tue	Wed	Thu	Fri	Sat	Sun
File A	New	Changed	Changed				
File B	New						
File C	New	Changed		Changed	Changed		
File D	New		Changed				
File E	New			Changed			
File F		New	Changed				
File G		New				Changed	
File H		New	Changed				
File I		New		Changed			
	Full	Diff	Diff	Diff	Diff	Diff	Diff

Incremental backups
(include only data since the previous backup)

	Mon	Tue	Wed	Thu	Fri	Sat	Sun
File A	New	Changed	Changed				
File B	New						
File C	New	Changed		Changed	Changed		
File D	New		Changed				
File E	New			Changed			
File F		New	Changed				
File G		New				Changed	
File H		New	Changed				
File I		New		Changed			
	Full	Incr	Incr	Incr	Incr	Incr	Incr

Source: ESG

Deduplicated backups
(includes files modified and unique data created since the previous backup)

	Mon	Tue	Wed	Thu	Fri	Sat	Sun
File A	New	Changed	Changed				
File B	New						
File C	New	Changed		Changed	Changed		
File D	New		Changed				
File E	New			Changed			
File F		New	Changed				
File G		New				Changed	
File H		New	Changed				
File I		New		Changed			
	Full	Incr	Incr	Incr	Incr	Incr	Incr

🟧 File backup 🟦 Duplicate data

For maximum manageability, the solution should allow for granular (tape- or group-level) policy-based deduplication based on a variety of factors: resource utilization, production schedules, time since creation, and so on. In this way, storage efficiencies can be achieved while optimizing the use of system resources.

Restore performance is also crucial. Some technologies are good at deduplicating data but perform much slower when it comes to rebuilding data (often referred to as "re-inflating" data). If you are testing systems, you need to know how long it will take to restore a large database or full system. Ask the solution provider to explain how they can ensure reasonable restore speeds.

5. Scalability

Because an IDT is often chosen for longer-term data storage, scalability is an important consideration, particularly in terms of capacity and performance. Consider growth expectations over five years or more. How much data will you want to keep on disk for fast access? How will the data index system scale to your requirements?

An IDT should provide an architecture that allows economic "right-sizing" for both the initial implementation and the long-term growth of the system. For example, a clustering approach allows organizations to scale to meet growing capacity requirements – even for environments with many petabytes of data – without compromising deduplication efficiency or system performance. Clustering enables an IDT to be managed and used logically as a single data repository, supporting even the largest of tape libraries. Clustering also inherently provides a high-availability environment, protecting the backup repository interface (VTL or file interface) and deduplication nodes by offering failover support.

6. Beyond tape backup

Although backup operations have primarily depended on a tape backup paradigm, D2D backup operations can be more suitable for organizations that don't have long-term data retention requirements and are willing to eliminate or lessen the use of tape. Data deduplication can reduce storage consumption in other data management processes beyond backup, including data archiving or database dumps. An ideal IDT should be able to effectively support these processes.

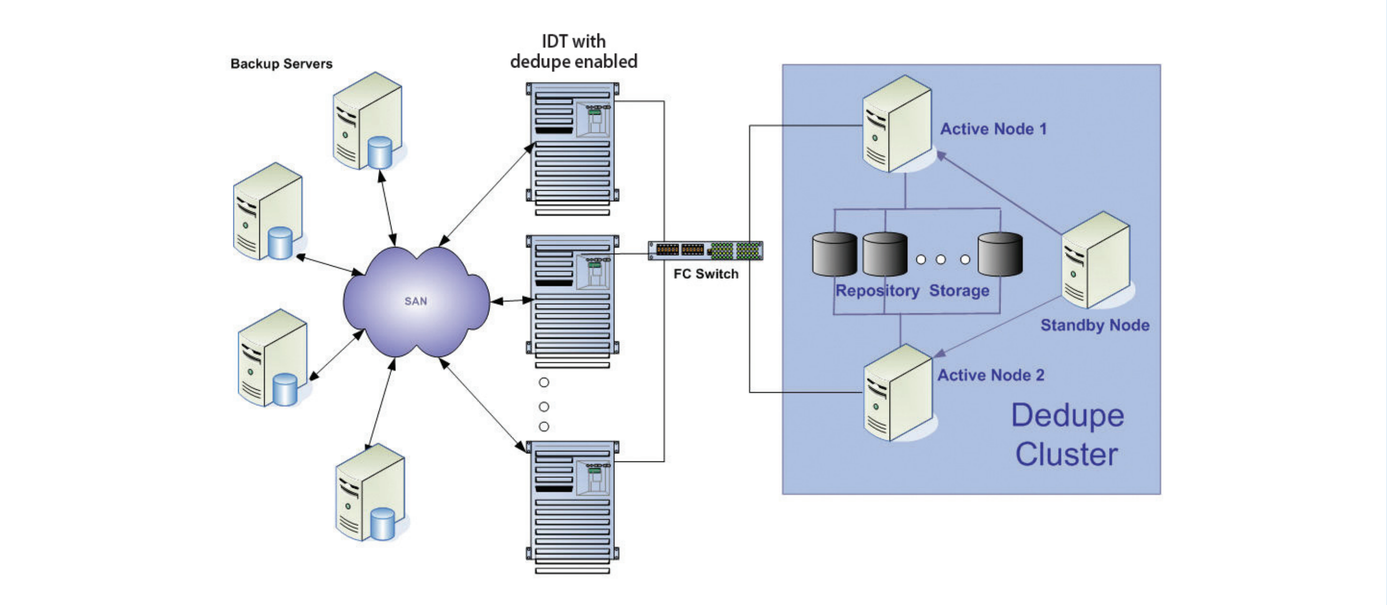
7. Distributed topology support

Data deduplication should occur throughout a distributed enterprise, not just in the data center. An IDT that includes replication and multiple levels of deduplication provides maximum benefits to customers. For example, a company with a corporate headquarters, three regional offices, and a secure DR facility should be able to implement deduplication in the regional offices to facilitate efficient local storage and replication to the central site. The IDT should only require minimal bandwidth for the central site to determine whether the remote data is contained in the central repository. Only unique data across all sites should be replicated to the central site and subsequently to the DR site, to avoid excessive bandwidth usage.

8. Highly available deduplication repository

It is extremely important to create a highly available deduplication repository. Since a very large amount of data has been consolidated in one location, risk tolerance for data loss is very low. Access to the deduplicated data repository is critical and should not be vulnerable to a single point of failure. A robust IDT will include mirroring to protect against local storage failure as well as replication to protect against disaster. The solution should have failover capabilities in the event of a node failure. Even if multiple nodes in a cluster fail, the company must be able to continue to recover its data and maintain ongoing business operations.

Figure 2: An example of a clustered deduplication architecture



9. Efficiency and effectiveness

File-based deduplication approaches do not reduce storage capacity requirements as much as those that analyze data at a sub-file or block level. Consider, for example, changing a single line in a 4-megabyte presentation. In a file-based solution, the entire file must be stored, doubling the storage required. If the presentation is sent to multiple people, as presentations often are, the negative effects multiply. Most sub-file deduplication processes use some sort of "chunking" method to break up a large amount of data, such as a virtual tape cartridge, into smaller-sized pieces to search for duplicate data. Larger chunks of data can be processed at a faster rate, but less duplication is detected. It is easier to detect more duplication in smaller chunks, but the overhead to scan the data is much higher.

If the "chunking" begins at the beginning of a tape (or data stream in other implementations), the deduplication process can be fooled by the metadata created by the backup software, even if the file is unchanged. However, if the solution can segregate the metadata and look for duplication in chunks within actual data files, the duplication detection will be much higher. Some solutions even adjust chunk size based on information gleaned from the data formats. The combination of these techniques can lead to a 30 to 40% increase in the amount of duplicate data detected. This can have a major impact on the cost-effectiveness of the IDT.

10. The end-to-end backup and recovery process

It is important to keep in mind that deduplication is only one part of a larger data protection and recovery process that likely includes some or all of the following: backup within a specified window, copying data to tape, deduplicating data, replicating data, restoring data, and managing all of these processes. By focusing only on deduplication, you may find yourself with a solution that breaks down somewhere within the larger process. Before deciding on an IDT, you should ensure that you understand the entire process, from backup to restore, and know how to manage it.

Summary: Focus on the total solution

As stored data volumes continually increase while IT spending decreases, data deduplication is fast becoming a vital technology. Data deduplication is the best way to dramatically reduce data volumes, slash storage requirements, and minimize data protection costs and risks.

Although the benefits of data deduplication are dramatic, organizations should not be seduced by the hype sometimes attributed to the technology. No matter the approach, the amount of data deduplication that can occur is driven by the nature of the data and the policies used to protect it. In order to achieve maximum benefits from deduplication technology, organizations should choose IDTs based on a comprehensive set of quantitative and qualitative factors rather than relying solely on statistics and hype.

For more information, visit www.falconstor.com or contact your local FalconStor representative.

Corporate Headquarters
USA
+1 631 777 5188
sales@falconstor.com

European Headquarters
France
+33 1 39 23 95 50
infoeurope@falconstor.com

Asia-Pacific Headquarters
Taiwan
+866 4 2259 1868
infoasia@falconstor.com

FalconStor® Software