FalconStor®

DEMYSTIFYING DATA DEDUPLICATION

A WHITE PAPER

DEMYSTIFYING DATA DEDUPLICATION

ABSTRACT

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 choices on key initiatives. There are many providers of data deduplication solutions today, and each vendor lays claim to offering the best approach. However, some vendors set unrealistic expectations by predicting huge reductions in data volume, ultimately disappointing customers. In order to make the right decision, multiple factors must be considered in order to select a data deduplication solution which suits the needs of the organization while providing significant value and minimal disruption to infrastructures and processes.

TABLE OF CONTENTS

INTRODUCTION

Ongoing data growth and industry-standard backup practices have resulted in large amounts of duplicate data. In order to protect data, the traditional backup paradigm is to make copies of the data on secondary storage every night, creating an overload of backed-up information. Under this scenario, every backup exacerbates the problem. Incremental and differential backups help to decrease the amount of data required compared with full backups. However, even within incremental backups, there is significant duplication of data when protection is based on file-level changes.

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. When considered across multiple servers and multiple sites, the opportunity for storage reduction by implementing a deduplication solution becomes huge. As a result, data deduplication has become a required technology for any company trying to optimize the performance, efficiency, and cost-effectiveness of storing data. 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 information resources.

KEY CRITERIA FOR A ROBUST DEDUPLICATION SOLUTION

There are some important criteria to consider when evaluating a deduplication solution:

1. INTEGRATION WITH CURRENT ENVIRONMENT

An effective deduplication solution should be as non-disruptive as possible. Solutions requiring proprietary appliances tend to be more costly 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.

Many companies turn to virtual tape library (VTL) technology as a method of implementing a deduplication solution and improving the quality of their backups without having to make significant changes to policies, procedures, or software. VTL-based data deduplication is one of the least disruptive ways to implement a deduplication solution in a tape-centric environment. In such cases, the capabilities of the VTL itself must be considered as part of the evaluation process. 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. It is very important that the vendor emulate a large number of tape libraries and tape drives in order to provide administrators with maximum flexibility regarding any future decisions they may make.

2. 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. There are trade-offs for the various options, so you will want to have maximum flexibility. For example, inline deduplication processes the backup stream as it comes into the deduplication appliance. This reduces the storage requirement because inline does not require a staging area, but it also tends to be slower than a post-processing approach and can impact the backup window.

Conversely, any solution that runs concurrently with tape backup processes or after the backup job is complete will provide higher speeds and minimal impact on the backup window, but will require extra staging space. That is because the backup data is read from the backup repository after backups have been cached to disk, so the production system is free to resume. An enterprise-class solution that offers various deduplication options is ideal because this flexibility gives storage administrators the ability to select a method based on the characteristics and/or business value of the data being protected. For example, data that is being deduplicated for the first time might start by using post-processing to handle the initial load for speed during the demanding initial deduplication set up and then switch to inline for subsequent space savings.

Most deduplication solutions only provide a single deduplication method in a forced "one size fits all" scenario. For maximum manageability, the solution should allow granular (tape- or group-level) policy-based deduplication that can accommodate variety of factors: resource utilization, production schedules, time since creation, and so on. Certain types of data, such as pre-compressed data and large databases, are not good deduplication candidates. In those cases, given the remarkable increases in physical tape recording densities and performance, moving those specific files to tape offers the best level of protection in the event that the data needs to be restored. 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. Compare backup and restore performance metrics for the vendors you are considering.

3. SCALABILITY

A deduplication solution is usually chosen for rapidly growing longer-term data storage, so 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? A deduplication solution should have 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 a deduplication solution 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 (HA) environment, protecting the backup repository interface (VTL or file interface) and deduplication nodes by offering failover support. In addition to clustering, the deduplication vendor should be able to easily add nodes or storage as needed.

4. DISTRIBUTED TOPOLOGY SUPPORT

Data deduplication should occur throughout a distributed enterprise, not just in the data center. A deduplication solution that includes replication and global deduplication provides maximum benefits to its customers. For example, a company with a corporate headquarters, regional offices, and a secure disaster recovery (DR) facility should be able to implement deduplication in its regional offices to facilitate efficient local storage and replication to the central site. Only unique data across all sites should be replicated to the central site and subsequently to the disaster recovery site, to avoid excessive bandwidth usage.

5. HIGHLY AVAILABLE DEDUPLICATION REPOSITORY

It is extremely important to create a highly available deduplication repository. Since a very large amount of vital data is 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 deduplication solution 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.

6. 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 sentence in a 4MB Powerpoint presentation. With a file level based dedupe solution, the entire file may have to be re-stored, doubling the storage requirements. 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 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 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 is intelligent and 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 deduplication solution.

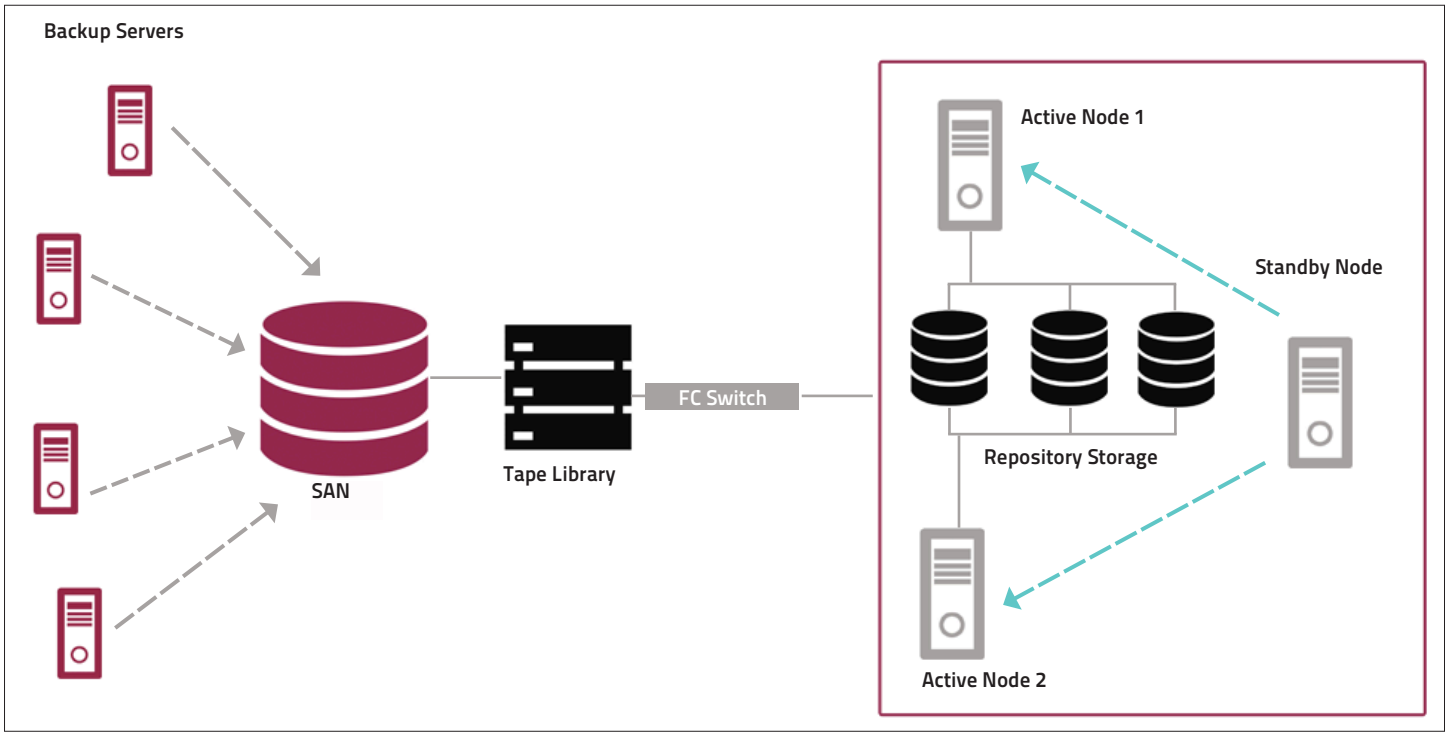

Figure 1: An example of a clustered deduplication architecture.

7. END-TO-END BACKUP AND RECOVERY PROCESS

By focusing only on deduplication, you may find yourself with a solution that breaks down somewhere within the larger process. Deduplication is part of a tiered storage protection hierarchy that includes storage virtualization, continuous data protection, and physical tape support on the back end. The larger data protection and recovery process includes backing within a specified window, copying data to tape, deduplicating data, replicating data, restoring data, and managing all of these processes. Before deciding on a deduplication solution, you should ensure that you understand the entire data protection 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 deduplication solutions based on the full spectrum of their IT and business needs.

FalconStor technology is unique in that it allows users to select the deduplication method best suited to their data recovery and protection criteria. The user has the choice of inline, post process, concurrent deduplication, or no deduplication at all. If deduplication is not an option for the specific data-type (IE: encrypted, compressed, or image files) then the data can be written to our high performance disk cache, and then directly to an attached tape library. Some data (for example, large compressed databases) are poor deduplication candidates and are best protected on high-performance, high-density tape. In non-intelligent dedupe solutions, the process is always on, even if the data is known to NOT be dedupable. Storing this type of data within a dedupe repository is wasteful, an may require up to twice the amount of storage in the long run. On a per node basis, FalconStor provides the fastest and most efficient data deduplication solution in the world today. When integrated with physical tape for data archives, we provide a holistic approach to data protection, and are able to move the data to the storage media best suited to the business policy you create.

REFERENCES

1. Ryan Murphy, "Where In The World Is Your Next Data Center," ReadWrite, May 2, 2013
2. CA Technologies Survey, "Cost of Downtime, November 2010
3. Jacques Bughin, Michael Chui, and James Manyika, "Ten IT-Enabled Business Trends For The Decade Ahead," McKinsey Quarterly, May 2013
4. "Data Data Everywhere," The Economist, February 25, 2010
5. Steve LaValle, Michael S. Hopkins, Eric Lesser, Rebecca Shockley, and Nina Kruschwitz, "Analytics: The New Path to Value, MIT Sloan Management Review, October 24, 2010
6. Brian Proffitt, "Why Business Still Needs Help Managing Explosive Data Growth," ReadWrite, November 20, 2013

Corporate Headquarters
2 Huntington Quadrangle, Suite 2S01
Melville, NY 11747
Tel: +1.631.777.5188
salesinfo@falconstor.com

Europe Headquarters
Landsberger Str. 312
80687 Munich, Germany
Tel: +49 (0) 89.41615321.10
salesemea@falconstor.com

Asia Headquarters
PICC Office Tower No. 2
Room 1901
2 Jian Guo Men Wai Avenue
Chaoyang District
Beijing 100022 China
Tel: +86.10.6530.9505
salesasia@falconstor.com