FalconStor®

HOW DATA DEDUPLICATION WORKS

A WHITE PAPER

HOW DATA DEDUPLICATION WORKS

ABSTRACT

IT departments face explosive data growth, driving up costs of storage for backup and disaster recovery (DR). For this reason, data deduplication is regarded as the next evolutionary step in backup technology and a "must-have" for organizations that wish to remain competitive by operating as efficiently as possible. This paper explains in detail how data deduplication technology from FalconStor Software reduces backup storage requirements for backup and disaster recovery operations in a manner that exceeds other deduplication technologies.

TABLE OF CONTENTS

INTRODUCTION

Data deduplication technology has gained rapid acceptance in the IT industry over the past several years for its ability to dramatically reduce the amount of backup data stored by eliminating redundant data. In its simplest terms, data deduplication maximizes storage utilization while allowing organizations to retain more backup data on disk for longer periods of time. This tremendously improves the efficiency of disk-based backup, lowering storage costs and changing the way data is protected.

Although data deduplication solutions vary in terms of how deduplication is accomplished, in general, data deduplication works by comparing new data with existing data from previous backup or archiving jobs, and eliminating the redundancies. Because only unique blocks are transferred, replication bandwidth requirements are reduced.

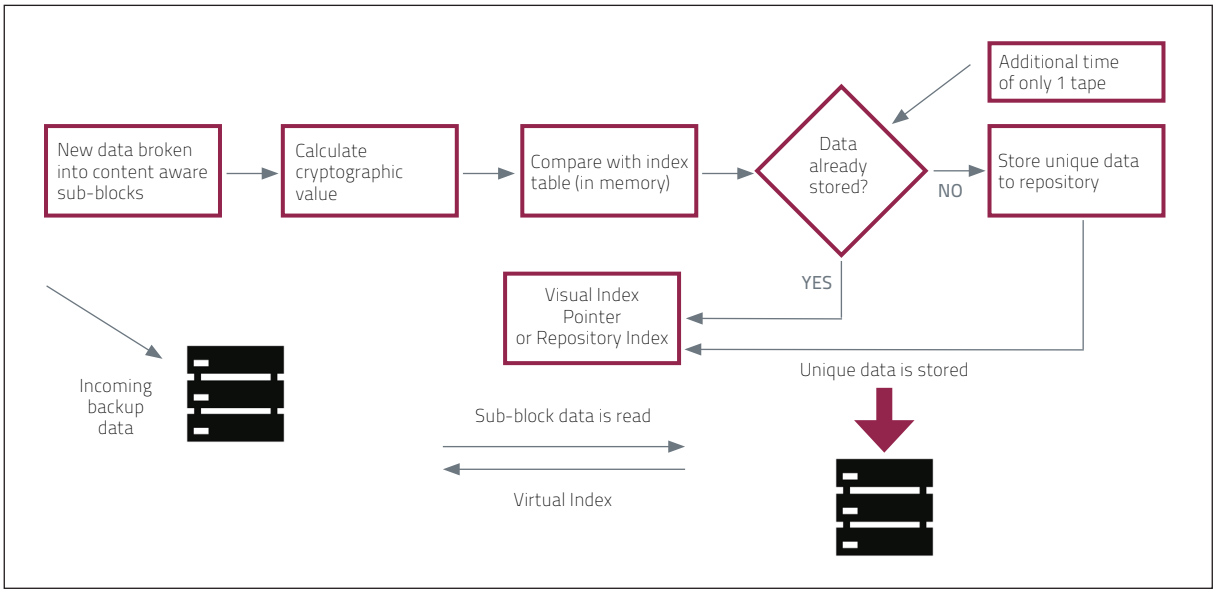
FalconStor Optimized Backup & Deduplication solution for Virtual Tape Library (VTL) provides high-performance backup and recovery, offering flexible, high availability, industry-leading performance-based deduplication to enable organizations to overcome backup challenges, optimize capacity, reduce storage costs, and minimize WAN requirements. With FalconStor global deduplication, only globally unique blocks of data are replicated. This means more cost-efficient and faster offsite backup.

HOW DOES IT WORK?

Data deduplication works by comparing blocks of data or objects (files) in order to detect duplicates. Deduplication can take place at two levels — file and sub-file level. In some systems, only complete files are compared. This is known as Single Instance Storage (SIS). This is not as efficient as sub-file deduplication, because entire files have to be stored again as a result of any minor modification to that file.

FalconStor data deduplication solutions provide sub-file or block-based deduplication. Block-based deduplication can either be fixed length block or sliding block. Using a patent-pending tape/file-format-aware parser, data blocks are broken into sub-blocks and assigned an identification key (index), calculated using a cryptographic hash function. If two identical hash keys are identified, it means that the related data blocks are identical.

OVERALL DEDUPLICATION DATA PROCESS FLOW



Once it is determined that a block of data already exists in the deduplication repository, the block is replaced with a Virtual Index Pointer linking the new sub-block to the existing block of data in the repository. If the sub-block of data is unique, it is stored in the deduplication repository and a virtual index is stored in memory for fast comparison with new data writes.

Because deduplication can occur independently from the backup process, it is transparent to the backup application. Similarly, when a file read request is initiated for data restore, the deduplication system can detect the links and read the blocks directly, in parallel from the deduplication repository, sending the right data blocks directly to the application.

FLEXIBLE DATA DEDUPLICATION

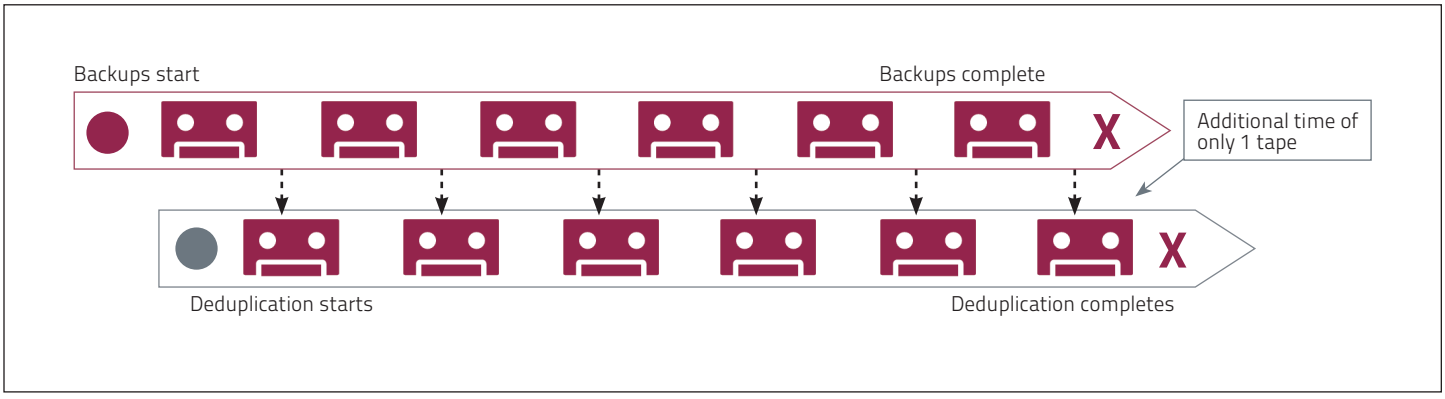
Today's experienced storage managers know the importance of data deduplication and understand that there are trade offs among the various deduplication options that are available in the market. It is important to be able to select the type of deduplication that best meets an organization's unique and changing requirements in order to best align capacity to business requirements.

FalconStor offers policy-based data deduplication technology, which gives users a choice of inline, concurrent, or post-process deduplication, configurable for maximum flexibility and performance. In addition, FalconStor offers Turbo Deduplication. Lastly, for those data sets whose characteristics do not yield the benefits of deduplication (such as compressed databases), FalconStor VTL offers the option of full tape integration without deduplication, so that this data can be written directly to high-speed high-density physical tape, or stored on disk and migrated to physical tape later.

Inline deduplication has the primary benefit of minimizing storage requirements, reducing them by as much as 40%. It is ideal for small storage configurations or environments where immediate replication is desired.

Along with full inline data duplication, FalconStor data deduplication can operate in concurrent mode on a file-by-file or backup job basis. Concurrent deduplication is more accurately described as a "concurrent overlap." Deduplication does not wait for all backup jobs to complete; rather, it begins as soon as the first virtual tape or file is completed. Meanwhile, other backup jobs continue to run concurrently with the deduplication process. One of the primary benefits of concurrent deduplication is more immediate replication. As soon as the deduplication process is complete, replication to the data center or to a disaster recovery site can be initiated, ensuring that the most critical data is available at all times.

CONCURRENT DEDUPLICATION PROCESSING



In post-process deduplication, also called off line deduplication, the deduplication process is performed independently from the backup process. The backup data is written to temporary disk space first, and then the deduplication process starts based on a user-defined schedule. Deduplicated data is copied to the repository disk for long-term retention. In this fashion, the backup speed is unaffected by deduplication workloads, and vice-versa. An administrator can apply deduplication policies, export data to physical tape, and schedule the deduplication to take place as a concurrent process or at a later point in time. This high-speed flexibility allows IT departments to maximize the efficiency of their operations while delivering reliable and predictable performance.

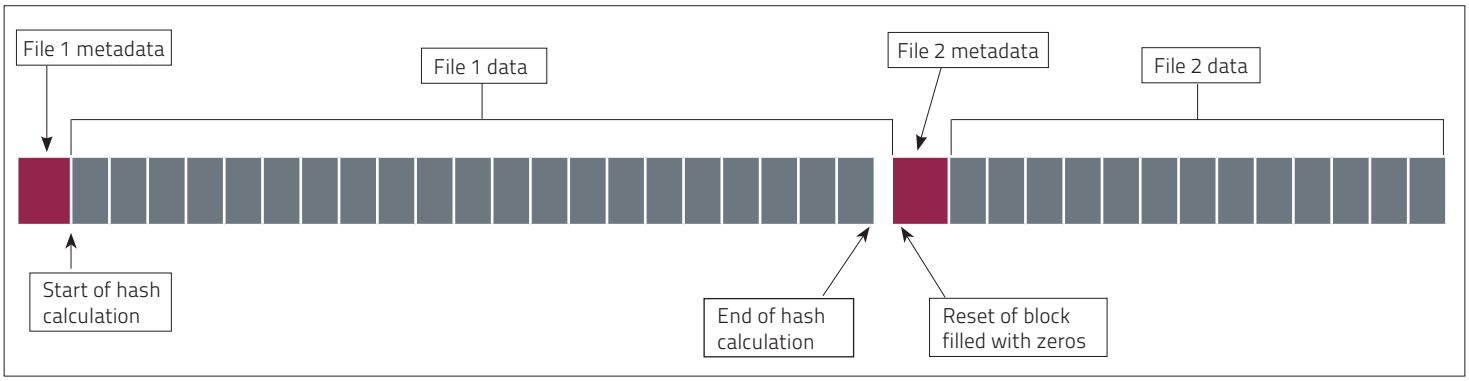
For example, an IT department can retain its most recent full backup on disk and automatically schedule deduplication on receipt of the next full backup. This enables the organization to quickly restore recently backed-up files. Data can also be exported to tape for archival and off-site protection.

Organizations can employ several deduplication methods using FalconStor deduplication technology. One use case for both inline and post-process deduplication might be a new, large file that has not been deduplicated before. The first deduplication always requires the most processing for any system, so it could be run using post-process deduplication for speed, to avoid affecting the backup window. Then, once that has been deduplicated, future deduplication could be performed inline to reduce ongoing storage requirements.

CONTENT-AWARE DATA DEDUPLICATION

The best deduplication ratios are obtained by segregating tape metadata from tape data and only deduplicating the tape data. For this reason, FalconStor data deduplication solutions are optimized to provide the best results based on the backup format awareness. The deduplication parser recognizes over 32 backup formats, as well as NDMP, OST, and multi-streamed backup sessions. This allows for proper alignment of the data prior to the start of the deduplication process. The metadata portion of the backup is identified and compressed. The parser then is aligned to the start of the data portion where the hash is calculated based on fixed block size, and duplicate blocks are discarded. This patent-pending, tape/file-format-aware deduplication model analyzes tape formats, confirms that the same files are aligned the same way each time, and achieves the most efficient deduplication ratio. Tape/file format-awareness allows the deduplication parser to align on different size blocks for different formats to ensure maximum detection of duplicate data, improving duplicate data detection by as much as 30% to 40% over generic raw fixed block deduplication analysis.

BACKUP FORMAT-AWARE PROCESSING



HASH COLLISION AVOIDANCE

FalconStor data deduplication technology is based on calculating a hash value to identify a block of data. FalconStor data deduplication solutions use the SHA-1 method. In this method, the hash algorithms take a sequence of input data and produce an output (often called a digest or simply a hash) of a much smaller size. SHA-1 produces a 160-bit hash.
Using SHA-1, the chance of hash collision for a 16 petabyte system has been shown to be less than 1 out of 10^{24}. FalconStor deduplication technology periodically checks the hash against the corresponding unique block to detect corruption of data for archiving.

Comparing SHA-1 cryptographic hash calculations to recognizable and acceptable error calculations, an IT administrator would need to store 432 zettabytes (432 x 10^{21} bytes) of data to reach the same odds of a single disk writing incorrect data and not knowing it (1:10^{15}), also known as an Undetectable Bit Error Rate (UBER). Similarly, an IT administrator would need to write 43 yottabytes (43 x 10^{24} bytes) to realize the same odds of a double-disk failure in RAID 5 (1:10^{5}). Based on these mathematical calculations and the amount of data typically retained in deduplication target systems, the probability of a hash collision is extremely rare.

SCALABILITY

In today's data-on-demand world, continuous uptime is mandatory. Unlike some deduplication providers who need to perform "forklift" upgrades to handle increased volumes of data, FalconStor provides a scalable, highly-available solution architecture that allows users to cope with growing and changing data protection needs. When more computing power is required, nodes can be easily and quickly added. When more capacity is needed, storage can be added easily and non-disruptively.

FalconStor's Flexible Repository Configuration storage can be configured and augmented with more flexibility, simplifying the configuration of a deduplication server. It is not necessary to organize data devices into fixed columns; any number of disks can be added to the system. This enables customers to be prepared for the future as well as for the present. Expanding a data repository from 1 node to 2 nodes or from 2 nodes to 4 nodes can be challenging with other vendors, but we have made this task easier, lowering the disruption, complexity, and cost of initial investments while allowing capacity to grow as the organization grows.

TAPE INTEGRATION

Many data centers opt to leverage both disk and tape for tiered backup and archive/compliance needs. FalconStor VTL seamlessly bridges physical and virtual tape operations through best-of-breed tape management capabilities. By integrating seamlessly with the tape environment, FalconStor VTL provides a full set of features for tape support, including tape caching and writing directly to physical tape. This comprehensive support makes FalconStor VTL a unique solution in the data deduplication industry.

CONCLUSION

FalconStor data deduplication technology tremendously improves the efficiency of disk-based backup, reduces the amount of stored data, and changes the way data is protected. Several key characteristics distinguish FalconStor data deduplication solutions from other deduplication solutions:

- The flexibility to choose from several deduplication methods: Inline, post-process, and concurrent
- The patent-pending tape and file format-aware deduplication model for parsing block data into properly aligned sub-blocks for maximum detection
- Industry-leading, enterprise-level deduplication performance
- High availability for greater backup and restore reliability
- WAN-optimized replication to reduce bandwidth costs
- Global deduplication ensures that redundant data from remote offices is eliminated prior to replicating to the central repository, minimizing space requirements
- Direct integration with physical tape and automated tape management simplifies operations, decreases media consumption, and reduces tape handling costs

In addition, clustering of data ingest and deduplication nodes allows customers to independently scale storage infrastructures to meet growing capacity and performance requirements while enhancing data retention and the overall deduplication efficiency.

Whether an organization backs up to disk or uses a combination of disk and tape, FalconStor data deduplication technology provides the right interface to fit its environment. Combined with multiple deployment models that enable scalability from the smallest business or branch office to the largest enterprise data center, FalconStor data deduplication solutions help today's organizations confidently move forward into the future to take the next evolutionary step in backup.

CONTACT US

Corporate Headquarters
2 Huntington Quadrangle
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