



“ For over a decade Coolspirit have been supplying the UK’s top organisations with storage products and solutions so be assured we will meet your requirements head on.

It’s all about getting things right first time, quickly and simply! ”

**Damon Robertson**  
Coolspirit Ltd

#### 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](mailto:web@coolspirit.co.uk)  
Find us: [View location map](#)  
Web: [www.coolspirit.co.uk](http://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  
FalconStor Partner

**FalconStor®**  
Software

## How Data Deduplication Works

*Abstract: IT managers and executives 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 DR operations in a manner that exceeds other deduplication technologies.*

## The backup challenge: Duplicate data

As IT managers and executives are looking at increasing business productivity and continuity, system and data availability becomes a priority. Traditional tape backup has been replaced by disk-to-disk (D2D) backup to accelerate the backup and recovery process and in turn to improve operational efficiency. However, managers moving to D2D backup solutions face explosive data growth, driving up costs of storage for backup and disaster recovery (DR). Each time a full backup is performed, it contains many of the same files and data as prior full backups, leading to multiple copies of the same data, taking up valuable disk capacity. The same applies to duplicate data within a backup job, across servers, and across backup jobs (full and incremental) over time.

## Data deduplication

Data deduplication is an advanced technology that can dramatically reduce the amount of backup data stored by eliminating redundant data. Data deduplication maximizes storage utilization while allowing IT to retain more nearline backup data for a longer time. This tremendously improves the efficiency of disk-based backup, changing the way data is protected.

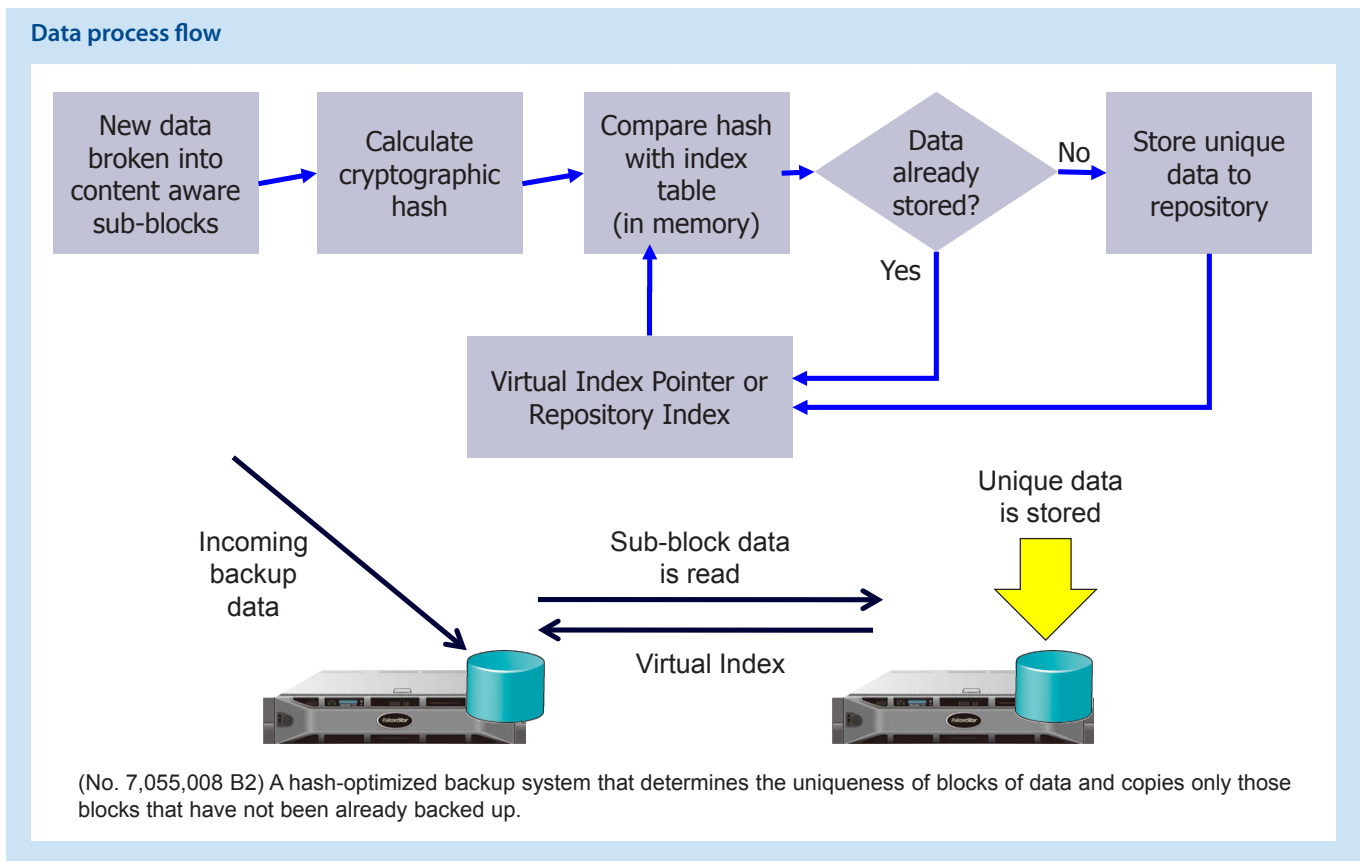
In general, data deduplication compares new data with existing data from previous backup or archiving jobs, and eliminates the redundancies. Advantages include improved storage efficiency and cost savings, as well as bandwidth minimization for less-expensive and faster offsite replication of backup data.

FalconStor data deduplication solutions, FalconStor® Virtual Tape Library (VTL) and FalconStor® File-interface Deduplication System (FDS), offer flexible, high-performance data deduplication to enable organizations to overcome backup challenges, optimize capacity, reduce storage costs, and minimize WAN requirements.

## 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, which is called Single Instance Storage (SIS). This is not as efficient as sub-file deduplication, as 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. 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 most likely the related data blocks are identical. An additional check can be performed to make sure the data is, in fact, identical. 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 reads.



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.

## High-performance data deduplication

FalconStor data deduplication technology is policy-based, offering flexible post-process or concurrent deduplication, configurable for maximum flexibility and performance.

In post-process deduplication, also called offline deduplication, the deduplication process is performed independently from the backup process. The backup data is written to temporary disk space first, then the deduplication process starts, based on a user-defined schedule, and 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 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 and data if needed, and/or export data to tape for archival.

On a file-by-file or backup job basis, FalconStor data deduplication can operate in a concurrent mode. 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.

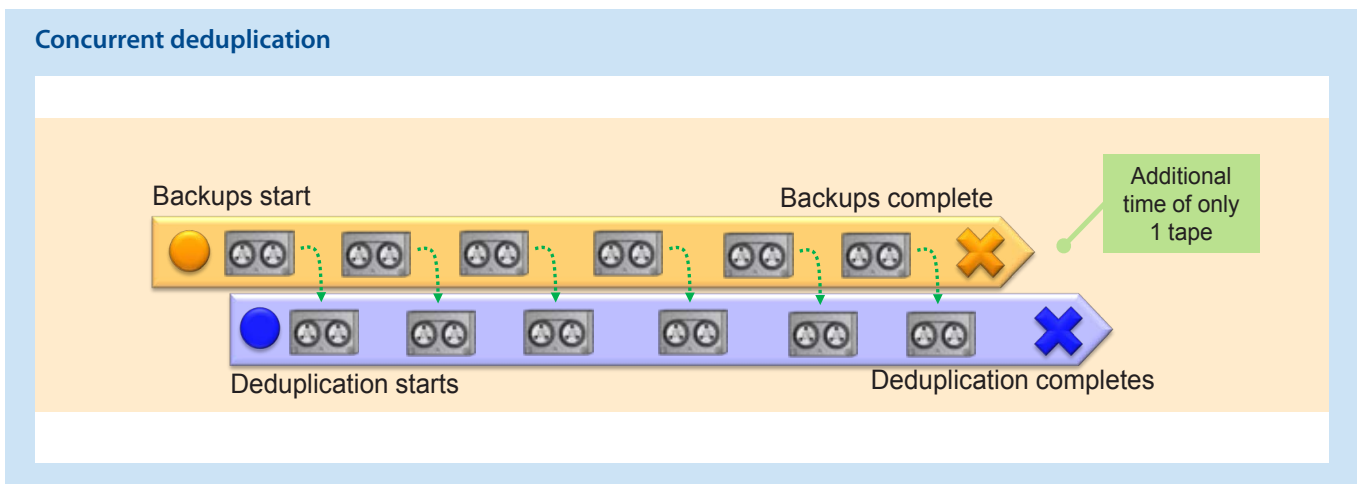
One of the primary benefits of concurrent deduplication is faster replication. As soon as the deduplication process is complete, replication to the data center or to a DR site can be initiated, ensuring that the most critical data available at all times.

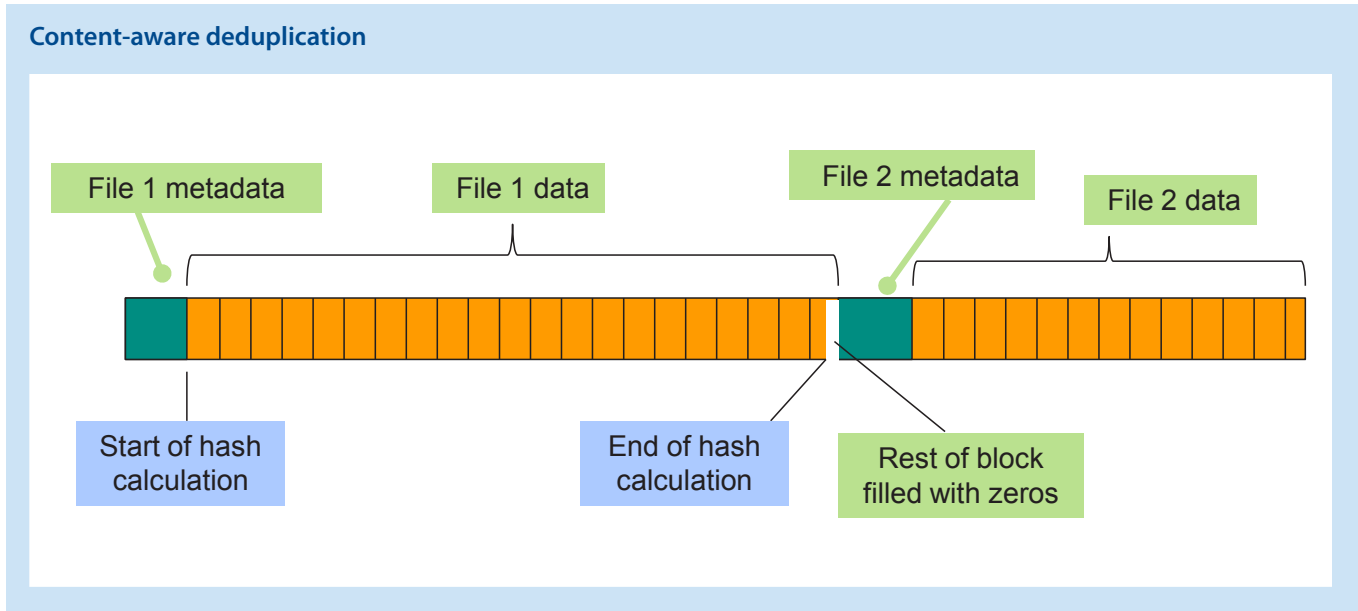
## Content-aware data deduplication

Because backups write data sequentially to either tape or files, FalconStor data deduplication solutions are optimized to provide the best results based on the backup format awareness. A deduplication parser recognizes over 32 backup formats, as well as NDMP 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, guarantees 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.

## Hash collision avoidance

FalconStor data deduplication technology is based on calculating a hash value to identify a block of data. There are various methods to create a hash value; the most commonly used methods are SHA-1 and MD5. 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  $10^{24}$ . Even though the possibility of hash collision is almost zero, FalconStor provides an additional verification mechanism that allows for full data block validation prior to deleting a new block with a matching hash of an existing block. Although this method can consume additional deduplication CPU cycles, it does not impact backup performance during post-process or concurrent deduplication.

Comparing SHA-1 cryptographic hash calculation to recognizable and acceptable error calculations, an IT administrator would need to store 432 zettabytes ( $432 \times 10^{21}$  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 ( $43 \times 10^{24}$  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.

## 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 and benefits of

the FalconStor post-process and concurrent deduplication models; the patent-pending tape and file format-aware deduplication model for parsing block data into properly aligned sub-blocks for maximum detection; and the cryptographic technology behind the almost-negligible likelihood of hash collisions. In addition, clustering of data ingest nodes and deduplication nodes allows organizations to independently scale to meet growing capacity and performance requirements while enhancing data retention and the overall deduplication efficiency.

Whether an organization backs up to disk or 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.