

White Paper

DeltaStor Software Next Generation Deduplication

DeltaStor™ software is a next-generation data deduplication application for the SEPATON® S2100®-ES2 virtual tape library that enables enterprises to store more data online at a cost that is comparable to physical tape systems. DeltaStor leverages SEPATON's patented ContentAware™ architecture to enable wire speed backup and restore performance, modular scalability, and industry-leading capacity reduction efficiency. By changing the economics of data storage, DeltaStor software allows enterprises to reduce risk, handle exponential data growth, cut backup times, and significantly lengthen online retention periods.

Table of Contents

Overview 1

Next-Generation Technology 2

Benefits of SEPATON Deduplication Software 3

Technology Fundamentals 4

Designed to Meet Enterprise Storage Needs..... 8

Overview

Exponential data growth and the increasing need for data availability are creating a variety of challenges for enterprise IT departments. These challenges include backing up and restoring rapidly growing data volumes, maintaining compliance with stringent regulatory requirements, meeting increasingly aggressive recovery point objectives and staying within tight budgets. DeltaStor software is a next-generation data deduplication application for the SEPATON S2100-ES2 virtual tape library that enables enterprises to store more data online at a cost that is comparable to physical tape systems. By changing the economics of data storage, DeltaStor software allows enterprises to reduce risk, handle exponential data growth, cut backup times, and significantly lengthen online retention periods for faster restores. DeltaStor software draws on SEPATON's unique ContentAware™ architecture, which contains built-in intelligence about file content and the backup data relationships of leading backup applications, to deliver unparalleled speed, simplicity, scalability, and data integrity. This whitepaper will discuss the technology fundamentals of the product and how it can be used to save money, save time, and increase efficiency of backup and restore operations.

Next-Generation Technology

DeltaStor's fundamental design leapfrogs the capabilities of existing data compression software to provide deduplication that is many times more efficient. Its unique approach to deduplication enables it to identify duplicate data at the byte-level for optimal capacity reduction.

Common compression technologies, such as Lempel-Ziv, use a stream-based approach to minimize data storage. These technologies create a dictionary of repeated data patterns and refer to it within a small (typically 8 kB) data window. They use a compression algorithm that eliminates any repeated data pattern within that window and replaces it with a reference to a dictionary item. At the end of an 8 kB window, the dictionary is cleared and the process restarts. Because these technologies only use the previous 8 kB window for their data reference, they miss a large volume of repeated data. As a result, they only provide compression ratios ranging from 1.6:1 to 3:1 depending on data types.

In contrast, DeltaStor software searches any number of versions of a given data object for repeated data sequences and then replaces the redundancies with pointers to a single baseline version. Only a single instance of a data sequence is actually stored onto disk. With this technique, DeltaStor can deduplicate a typical mix of application data at a ratio of as much as 25:1.

DeltaStor software can be used in conjunction with stream-based software for significant additional reduction. For example, a DeltaStor deduplication data ratio of 25:1 can be compressed yielding another 2:1 reduction for an overall deduplication ratio of as much as **50:1**.

Benefits of SEPATON Deduplication Software

DeltaStor software allows enterprise storage managers to take advantage of the speed, flexibility and efficiency of disk storage at a cost that is comparable to physical tape. In addition, by storing more data on disk in less physical space than tape, DeltaStor software significantly reduces power, cooling, security and other operating and infrastructure costs.

- **Instant data restores.** Data is online and instantly accessible through the random access power of disk.
- **Dramatically faster backups.** DeltaStor software performs data deduplication outside of the primary data path, enabling the S2100-ES2 to deliver the industry's fastest backups—as much as 30 times faster than tape.
- **Scalability to handle exponential data growth.** The S2100-ES2 has a powerful grid architecture that can handle any size backup set. In addition, the S2100-ES2 allows simple, non-disruptive scaling of capacity and performance to let you purchase what you need, when you need it. Scale physical capacity of a single appliance from 10 TB to more than 1.6 PB – as much as twice that capacity with the built-in hardware compression enabled. With DeltaStor enabled, you can store as much as 60 PB of data on a single S2100-ES2 VTL.
- **Reduce time-consuming tape management tasks.** Keeping more data on disk reduces the labor needed to handle tapes, address tape failure issues and manage capacity provisioning.
- **Eliminate physical threats to data.** Unlike physical tapes that can be lost, stolen or damaged, data on disk is maintained in a secure, highly available environment.
- **Simplify Data Management.** Adding DeltaStor software is as easy as checking a box in the S2100-ES2 management console. As deduplication reduces data volume, capacity is automatically made available and managed through its built-in self-functions.
- **Retain more data on disk to meet compliance/recovery time requirements.** For example, the maximum retention time for 5 TB of full daily backups on a 50 TB system is only ten days. With an S2100-ES2 with DeltaStor software, you can store that data on line for 450 days in the same space while providing the performance and other benefits of disk-based data protection.

Technology Fundamentals

SEPATON ContentAware architecture was designed from the ground up to serve as a comprehensive data protection platform. With this architecture, the SEPATON S2100-ES2 virtual tape library appliance contains powerful software, such as the Dynamic Disk File System (DDFS) and the SEPATON I/O Subsystem (SiS) that work with the DeltaStor software to build an intelligent grid-based data protection platform.

At the heart of DeltaStor software is the SEPATON database. During backup sessions, as data is stored on the disk arrays, software modules in DeltaStor called data readers populate this database with metadata content. The data readers capture metadata related to individual backup datasets, as well as the metadata for each object within each backup dataset. The DeltaStor software uses this metadata to identify relationships between data objects to make intelligent decisions about how to handle them for maximum performance and deduplication efficiency. For example, when the metadata indicates that a relationship exists between objects in two backup sessions, the software performs a more careful comparison of the objects at the byte level.

The database is stored on the storage arrays within SEPATON's DDFS in a fully redundant configuration. Therefore, as the size of primary data grows, the database automatically scales to accommodate it. In keeping with SEPATON's high availability design model, the software can perform a complete rebuild of the metadata database with a simple scan of virtual cartridges in the S2100-ES2.

Process Overview

For every backup job, the SEPATON grid engine software goes through the five phases of the DeltaStor deduplication process. This process is conducted concurrently with the backup process. The grid engine software manages the scheduling and execution of jobs using all computing power in the appliance in a way that load balances the work. This capability yields virtually infinite scalability of the solution to meet the data processing requirements because jobs can be allocated to any available computing resource. Additionally, customers can optionally add DeltaStor nodes that provide additional computing power to accelerate deduplication activities. The five phases of the deduplication process (see figure 1) are as follows: data grooming, data discrimination and/or data comparison, reassembly, and integrity checking.

Space Reclamation

1. Data Grooming

In the data grooming phase, the software narrows the scope of data it needs to analyze at the byte level by comparing the in-coming backup data to the previous backup. It uses the SEPATON database to identify relationships between that indicate a high likelihood of redundant data. For example, if the filename “\root\documents\abc.txt” exists twice in backups from the same client, the DeltaStor software determines what type of action is

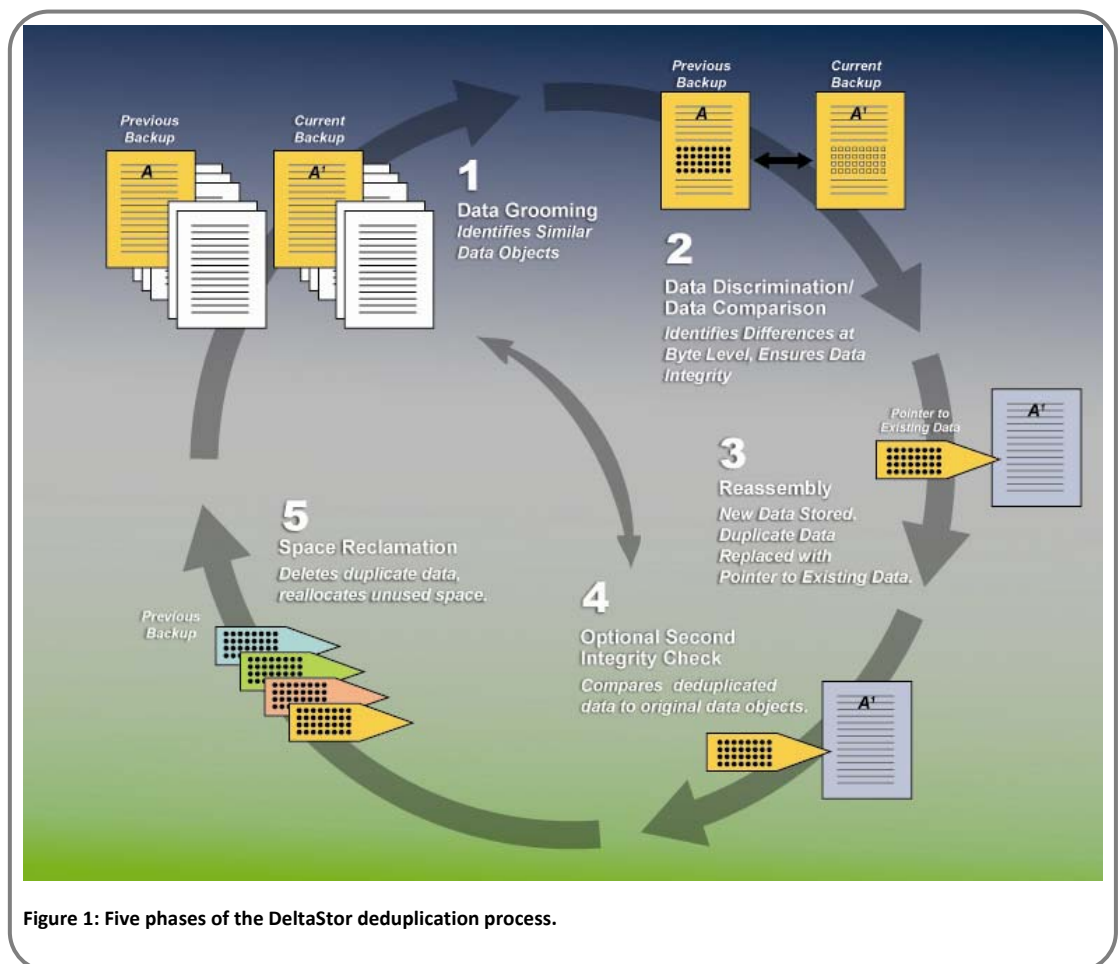


Figure 1: Five phases of the DeltaStor deduplication process.

needed for deduplication. If the data is a modified version of existing data, it is sent to the next phase (Data Discrimination, Data Comparison) where specific changes will be identified. If it is an identical duplicate of existing data, then DeltaStor sends it to the next

phase where the data duplication will be verified. Additional data grooming operations, include identifying copies of the same object kept in different places (i.e. different clients, directories, etc.). The software creates a work list of pairs of redundant data that were found in the Data Grooming stage, which is sent to the Data Discrimination/Data Comparison phase for further analysis.

2. Data Discrimination, Data Comparison

In the Data Discrimination/Data Comparison phase, the software performs a byte-level analysis of the objects identified as containing duplicate data in the Data Grooming phase. If the work list created in the Data Grooming phase identifies the need for data discrimination, then the software uses a technique called delta differencing to determine which data in the backup set is unique and which data is redundant.

The DeltaStor software efficiently maps changed data to byte-level granularity, and is insensitive to offset or positional changes within the data objects. This makes it capable of locating the redundant data even in related objects that have undergone significant structural changes.

If the Data Grooming phase determines that data in the backup set is identical to data in the previous backup at the metadata level, then a data comparison is performed at the byte level in the Data Discrimination, Data Comparison phase. In this step, the software identifies non-identical files in which the data has changed even though its metadata stayed the same.

3. Reassembly

In the Reassembly process, new data is stored and duplicate data that was identified in the previous phases is replaced with pointers to the stored data. Data is stored in a temporary holding cartridge that appears to backup software applications as identical to a native cartridge. However, the volume of data actually stored on this new cartridge is much smaller. The data appears to the backup software as if it is laid out sequentially and not deduplicated. The SEPATON software follows pointers embedded in the file system to read the single-instance copy of the redundant data along with the unique data stored separately. The result is a deduplicated view of a backup set.

4. Optional Integrity Check

Before any redundant data is removed, the software can be set to perform an additional data integrity analysis. In this phase, the software validates the structure and entirety of data content of the “holding cartridge” (which represents the DeltaStor deduplicated data) by comparing it to the original representation.

5. Space Reclamation

In the Space Reclamation phase, the software removes the redundant data from the file system, freeing the previously-occupied disk for other uses. The holding cartridge exchanges its place (assuming barcode, slot position, properties, etc) with the original, non-deduplicated version. The software then intelligently frees the duplicate extents and moves them back to the free space pool. Once there, any other ongoing data process that requires storage can re-use the space previously occupied by redundant data.

Designed to Meet Enterprise Storage Needs

SEPATON's DeltaStor software is leading the evolution to tapeless data centers by delivering a simple, cost-effective, highly secure way to protect data. DeltaStor software leapfrogs other data reduction technologies to provide a solution that was designed specifically to meet enterprise storage needs.