Optimizing backup using deduplication: select the right method to fit your environment

The promise of data deduplication

Storage budget and backup time windows have come into sharp focus as a result of growing data volumes, longer data retention requirements and stringent Recovery Time and Recovery Point (RTO/RPO) Objectives. While many IT administrators have turned to disk storage platforms for backup because of the significant advantages versus traditional tape-based practices, the promise of adding deduplication into the mix is compelling. Simply put, retain what data you need but do so at a lower cost. Typical backup data stores contain massive amounts of redundant data—incremental and daily backups, unstructured content, file archives and virtual images. Although compression and Single Instance Store (SIS) can help reduce media requirements, deduplication brings this benefit plus it improves recovery time since the entire image of the file (original data and pointers) is available from the same storage device.

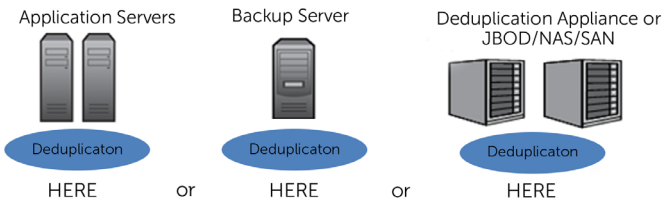
Business goals and IT practices help determine source or target dedupe

Deduplication software has been shown to reduce the amount of backup data stored in typical environments by a ratio of 20:1*—significantly lowering the amount of backup storage media required. Operational impacts include the ability to retain data for longer periods of time, decreased bandwidth for remote backups, and more comprehensive disaster recovery policies. However, deciding how to implement deduplication requires a better understanding of the alternatives available.

Choosing the right deduplication architectures depends upon several factors

Several considerations come into play when determining the deduplication architecture that's right for your environment. These include where and when the deduplication process takes place—on the host, i.e., the source server or at target backup environments.

Deduplication Technology can be applied at different points in the Backup process



In source-based dedupe environments, processing occurs on the host server. This dramatically reduces the amount of data sent over the network to the backup server and thus demand for shared resources such as network bandwidth and virtual machine backup agents. Source-based deduplication can be deployed as dedicated server-based software agents or as integrated functionality in backup software. While typically less expensive than dedicated appliances, this architecture can be disruptive if it requires administrators to recreate backup job configurations, schedules and alerts.

For some environments, target-based deduplication is easier because it can be added to existing backup environments as a dedicated backup-to-disk, VTL or a storage appliance supporting all applications in the network. With in-line target dedupe, data arrives at the target site and files are scanned for duplicate data segments before writing to storage. However, because in-line dedupe adds time and latency, an alternative is post-process deduplication which performs processing outside the data path after the files have been stored. Target-based dedupe requires greater overall capacity, but results in faster data movement and permits full, native backup images for fast file recovery.

Another approach (a.k.a hybrid dedupe) combines these methods—backup software handles data compression on client servers, sends files to backup media servers for deduplication processing, and then directs the deduplicated data stream to storage systems. This creates minimal impact because it leverages existing legacy equipment.

Reduction ratios and performance

Data deduplication processing uses software algorithms to scan files and detect and remove repeated block-level patterns of data in a file and replace them with pointers in a catalog. Unique files or data chunks are preserved while duplicate files or data chunks may be optionally rechecked using bit-level comparison or secondary hash algorithms to ensure data is truly a duplicate and not a collision.

There are currently three types of dedupe processing in use today—File, Block and Sub-Block. An example of file-level deduplication is the storage of a PowerPoint presentation in a file share. The first backup is a full copy written to storage. Subsequent full or incremental backups recognize that the file is already stored and won't write it to storage again. However, if just one slide changes, file-level dedupe considers this a new file and writes the entire thing again. In contrast, block or sub-block dedupe would only back up the changed portions. The overall impact of deduplication on the amount of changed data stored and the processor workload also varies upon how files are scanned during deduplication—as fixed blocks or variable (sub-block) increments. The rule of thumb is that the more efficient the deduplication technology, the more processor intensive it is.

Other performance factors include the ability to scale the backup architecture with additional nodes, frequency of the dedupe process, secondary processing, and the types of applications or files the technology is applied to. Frequently changing transactional data stores aren't good candidates for deduplication. Static files, large file servers, archives, unstructured repetitive content such as virtual desktop images are all good candidates for deduplication processing.

Comprehensive data protection solutions from Dell

Dell provides a comprehensive portfolio of data protection solutions to enhance your IT environment. With world class storage platforms, a rich suite of software offerings and partnerships Dell can help you plan and deploy backup and recovery solutions customized to meet your needs and budget. Supporting the portfolio is a comprehensive set of services designed to optimize the storage environment. Dell's IT Consulting** Services can provide the expertise to assess the application of deduplication to your environment with automated tools that gather workload, backup performance and CPU statistics to establish data-driven recommendations designed to optimize and protect your data.

*Source: "Introducing New Tools for Better Backups—an eZine from Storage magazine and SearchStorage.com, Chapter 1: "Catch up with dedupe" Jerome M. Wendt, April 2008
**Availability and terms of Dell Services vary by region. For more information, visit www.dell.com/servicedescriptions

Simplify your storage at www.dell.com/PSseries

CB131_USA_120909_090310