

Deduplication as a Key Component of Data Management

Integrating deduplication as part of your storage expansion strategy

Obviously, we're aware of the booming growth in data and the challenges it presents. While storage solutions available through Network Appliance and EMC are getting more consideration, another prevailing strategy can be used to help curb increased storage requirements. Data deduplication is a process used to eliminate redundant data from the storage system. At the core, deduplication tools compare data to identify and eliminate redundancy.

Data deduplication can help lower your storage space requirements, increase space efficiency, allow for longer retention periods, save on storage expansion costs, and possibly eliminate tape. If it's not part of your backup and archive strategy, it's something you should think about to compliment your storage strategy. Understanding the basics of deduplication, and where the technology is deployed can provide insight for achieving the lowest total cost of ownership for storage resources.

So, what are the basics, and how does it work? The comparison phase, or fingerprinting, creates segments of blocks designed to give the highest match rate against other existing fingerprinted segments. Once data is fingerprinted, redundancy can be identified. By using hashing algorithms, bit stream comparison, custom content analysis, or a combination of these methods, the deduplication engine can decide whether to save the segment or create a reference to it, thus eliminating the storage of redundant segments.

When a match is made, the new segment becomes the master and the old segment is referenced by a pointer (forward referencing). The other architecture keeps the oldest segment and references all reoccurring segments (reverse referencing). The architecture selection probably isn't as important choosing and implementing a strategy. A definitive study proving one over the other isn't currently available, and probably shouldn't be a key to your purchase criteria.

Deduplication can be performed on the source, in the stream, or at the target. The source or host based schemes use a client/server program to communicate with the backup server to select matches in the data set, and thus avoid sending them. This can help reduce storage overhead and the network bandwidth impact of the backup process. Available host CPU cycles are used to generate the savings.

In-stream deduplication analyzes the data stream from a backup client and saves the stream directly to the storage system (VTL, D2D, etc.) as a de-duplicated backup. These in-band methods work well where expedient replication to a remote site is required, resulting in far less data having to be transmitted, and replication beginning upon backup completion.

Target based deduplication, or out-of-band, is performed by the storage appliance on the completed backup image. This method provides the best possible performance for a backup as data does not have to be manipulated during the backup process. The storage appliance can use CPU power to do the deduplication work.

How much could you reduce storage requirements? The truth is no one can answer this question without some in-depth analysis. Quoted numbers vary with a true return becoming evident over time. With more data for comparison, the increased hit rate probability results in higher deduplication rates. It's important to remember, deduplication rates will vary during the information lifecycle - starting low and growing over time. Additionally, the types of data sets being backed up will affect deduplication rates. Backing up servers with similar OS structures will carry a high deduplication rate. Monthly full backups will have a lower deduplication rate compared to weekly full rotations. I've read anecdotal

studies of deduplication processes providing a 20:1 reduction in backup storage requirements when using typical weekly full, daily incremental backup schemes in mixed database and file system environments.

If you're considering a hardware acquisition for your backup system, make sure you consider the deduplication options available on proposed systems. Implementing deduplication on a new system may be the easiest way to get started. Be sure to ask a trusted IT advisor about advanced data protection available. Use the highest protection level available, without making rebuilding arrays unreasonably long. Remember, many backups now depend on the current set of data segments.

Deduplication is one of the fastest growing technologies in the data management field, and can be more complex. You may find that just one method of deduplication cannot fulfill all of your deduplication requirements. With thorough evaluation, you can select the deduplication tools necessary to get the greatest return on your storage infrastructure investment.

#

Tom Stewart, Senior Storage Architect
(937) 384-0444 ext 2310
Email: tom.stewart@digitalcontrols.com

Tom Stewart has been a member the Digital Controls Corporation service team for 27 years. Tom's current position is Senior Practice Manager providing lead support in storage architecture.

Digital Controls Corporation is a local IT services and software company. Digital Controls Technology Services Group provides services to IT in the key areas of security, data storage management, networking, Microsoft infrastructure, and product sourcing. These services are delivered through consulting, professional services, on-site and remote managed services.