SCI StorInt™ Dispatch
EMC announces Celerra deduplication

Today, EMC announced their Celerra product line would supply deduplication with built-in compression functionality as a free upgrade to their Celerra customers to improve storage efficiencies of their primary file storage.

Disseminating Dedupe

EMC's intent seems to be to incorporate and integrate Avamar, RecoverPoint and other deduplication technology into the breadth of their product offerings. The latest chapter in this saga is their Celerra NAS offerings.

Celerra now supports compression and single instancing (deduplication) for file data on a file system by file system basis. This enables a fine-grained approach to saving storage.

Deduplication technology factors out common data across multiple files in a file system. EMC's Celerra deduplication is at the file level meaning that it can deduplicate or factor data across multiple files with the lowest levels of CPU or memory resources compared to other deduplication technologies like fixed block or variable block.

Historically, deduplication was first introduced for backup data by products such as EMC's Avamar, Data Domain, IBM's Dilligent, FalconStor, Quantum DXI, Sepaton, Symantec Veritas, and others. As such, for backup (full and incrementals) data deduplication can often reduce storage consumed by 20X or more. However, Celerra's announcement applies to improving storage efficiencies related to primary file storage. EMC and NetApp are the only two vendors that support deduplication for primary storage. Although Data Domain supports deduplication for NFS and CIFS, It's not sold as primary storage.

Within Celerra Data Deduplication EMC also announced support for software data compression. Data compression has been around for decades in gZIP and/or precursor file formats. EMC says they are not using gZIP technology but rather have used an internally developed, higher performing, software data compression capability. Historically compression has reduced storage requirements by around 2X, but this is entirely dependant on the file type.

Celerra Data Deduplication adds a new approach to improving storage efficiency by automatically targeting files that are the best candidates for compression and single instancing in terms of the level of frequency of file access and file size. It filters files that are too small, too big, and accessed very often from being processed and avoids compressing files if the space savings are minimal. All deduplication processing is accomplished as a background asynchronous operation that acts on file data after it is written into the file system, avoiding adding additional latency.

Silverton Consulting
Strategy, Storage & Systems
P.O. Box 492|Broomfield, CO 80038-0492|
720-221-7270| http://www.silvertonconsulting.com

Post processing normally must read the file data in, deduplicate and then compress it, and then write the final deduped/compressed data back out. While significant technology has reduced the CPU and IO required to deduplicate data, it cannot be eliminated altogether and as such it may impact IO throughput. In contrast, data compression is moderately CPU intensive but should need minimal I/O as long as data stays in memory.

Since Celerra supports file level deduplication, all the data is stored together for a file and thus, Celerra can retrieve the data without the traditional impact of reconstituting a deduplicated file. Also, by intelligently selecting data and avoiding active data Celerra minimizes the performance impact of both compression and deduplication. De-compression is only needed when files that are inactive are accessed. This decompression will add a moderate impact to file retrieval and is done only for the chunk being accessed minimizing the need to retrieve the whole file.

Both compression and deduplication can reduce the storage consumed by file data. NetApp currently "guarantees" that their product can save up to 50% of the storage capacity required for storage supporting VMware VMs. While EMC has targeted their feature towards primary storage deduplication used for generic file share data and has seen the additional gains to be upwards of 30-40% for this store where a majority of NAS unstructured data is stored.

Announcement significance

We can easily see how deduplication benefits backup and VM data, how well this helps other file data is open for debate. However, in EMC's and NetApp's defense, corporations duplicate lots of file data, mostly as attachments are sent from one user to another. Even if this data were modified, much common data would remain and could be factored out. For transaction data such as databases, deduplication makes little sense until one replicates this for data warehousing and development activities.

In contrast, we have always been proponents of data compression since our days with a former employer. Software compression, however, has significant CPU, mask-able on creation via post processing but visible on retrieval. As such, software data compression may be less useful for the average primary storage environment. However, since the majority of the data in primary storage is inactive. Celerra has targeted compression to only this inactive data and also intelligently decompresses only the chunk that is accessed for read to further minimize impact if the inactive data is accessed.

Nonetheless, one has to give EMC credit for making data compression available to primary storage. While deduplication and data compression may be relatively straightforward additions to file storage, it's another matter entirely to add to block storage. Stay tuned to see if and when they roll that out.

Silverton Consulting
Strategy, Storage & Systems
P.O. Box 492|Broomfield, CO 80038-0492|
720-221-7270|http://www.silvertonconsulting.com