StorageIO Group Information Technology Industry Analysts and Consultancy

StorageIO Industry Trends and Perspective Solutions Brief
Real-time Data Compression Integrity and Reliability

Author: Greg Schulz – Sr. Analyst

Compliments of Storwize

July 23, 2008

This piece looks at how lossless Lempel-Ziv (LZ) type real-time data compression provides data integrity while reducing data footprint for on-line active NAS primary storage. LZ compression is a proven technique for data footprint reduction that scales with stability as more data is ingested without introducing performance bottlenecks during writes (compressing) or reads (un-compress).

Background and Issues

Preserving data integrity, which is insuring that data remains in its original state without loss of data quality or data corruption due to missing bits, bytes or blocks of data is an important consideration when looking at data footprint reducing techniques.

Not all data footprint reduction techniques are the same; some solutions are able to achieve very high ratios or effective amounts of reduction only at the expense of lost or changed data.

With lossless real-time data compression, compressed data is preserved and un-compressed exactly as it was originally saved with no loss of data. Some algorithms can achieve higher levels of compression by introducing some amount of data loss ("lossy" data compression).

Generally, lossless data compression is required for digital data requiring an exact match or 100% data integrity of stored data. Some audio and video data can tolerate distortion in order to reduce the data footprint of the stored information, however digital data and in particularly documents, files and databases have zero tolerance for lost or missing data.

Value Proposition

The value proposition of using real-time data compression for on-line storage is the ability to store more data in a reduced footprint for active changing data.

With active data including databases, unstructured files and other documents, caution needs to be exercised so as to not cause performance bottlenecks when introducing data footprint reduction techniques in addition to maintaining data integrity.

The Technology

Real-time compression techniques using time proven algorithms, such as Lempel-Ziv (LZ) as opposed to MD5 or other compute, "heavy thinking" hashing techniques, provide a scalable balance of uncompromised performance and effective data footprint reduction. This means that changed data is compressed on the fly with no performance penalty while maintaining data integrity and the same for read operations.

LZ is variable length for a wide range of uses and, thus, a popular lossless compression algorithm. LZ for compression in general involves a dictionary or map of how a file is compressed that will be used later for restoring a file to its original form.

The size of the dictionary can vary depending on the specific LZ based algorithm implementation. The larger the file or data stream combined with recurring data, including white spaces or blanks, results in a larger effective compression ratio and subsequent reduced data footprint benefit.

www.storageio.com P.O. Box 2026 Stillwater, MN 55082 651-275-1563 info@storageio.com

StorageIO Industry Trends and Perspective Solutions Brief
Real-time Data Compression Integrity and Reliability

Strategies and Recommendations

Data footprint reduction techniques for on-line data and primary NAS storage should be evaluated on a combination of effective compression rate (how much data is compressed in a given timeframe), the effective reduction ratio (amount or percent data reduced) and data integrity.

The data compression rate is how fast data can be compressed while maintaining data integrity and how fast the same data can be restored from a reduced footprint. Key to enabling data footprint reduction techniques, including deduplication and compression, are dictionaries or tables that describe how the recurring data is removed to enable data to be reassembled to its original form.

Where the meta data dictionary is stored, how it is maintained and accessed, along with how scalable it is impact data footprint reduction solutions. In the case of data deduplication, to be effective, a vast amount of data must be seen and remembered with a global dictionary. For example, a "zipped" file has its dictionary stored in the compressed file to speed de-compression.

Closing Comment

Look beyond the effective data compression ratio and consider the data compression rate and data integrity capabilities of a candidate solution. What good is data that has been significantly reduced in size in terms of its resulting data footprint that results in loss of data integrity making a file or document un-readable, inaccurate or corrupted for use?

Ask technology vendors or solution providers what data integrity features exist for their solution to insure that all data remains intact. Look for solutions that enable compressed data to be expanded or returned to its original form exactly as it appeared before compression.

Another question to ask is where and how the data reduction dictionary is safely stored. Because the dictionary is the key to reconstructing a compressed or de-duped file to its original state, it needs to be preserved and secured.

Where to learn more:

An example of a solution that enables real-time compression of active data on NAS based primary on-line storage without performance compromise is the Storwize STN-6000 appliance. Learn more about the STN-6000 and its capabilities at www.storwize.com.

Additional material pertaining to data footprint reduction including the StorageIO Industry Trends and Perspective report "***Business Benefits of Data Footprint Reduction***", companion solutions brief for real-time data compression and other topics can be found at www.storageio.com.