StorageIO Industry Trends and Perspective Solutions Brief
Performance Considerations for Real-Time Data Compression

Author: Greg Schulz – Sr. Analyst   Compliments of Storwize   **July 23, 2008**

This industry trends and perspective solution brief looks at performance considerations and impacts of real-time data compression for active data on-line primary NAS storage as part of a organization wide, application independent data footprint reduction strategy.

Background and Issues

IT organizations of all sizes are looking at data footprint reduction techniques (archiving, compression, and de-duplication) to sustain business growth. Archiving is great for taming the virtual "paper tiger" known as digital data by moving inactive data off-line that must be kept for compliance purposes. Archiving is also good for removing data no longer needed or seldom accessed by an organization. Meanwhile, data de-duplication continues to gain popularity for reducing data footprints of recurring backup or static data, with real-time compress being a good fit for on-line primary NAS storage where data is changing.

Most organizations rely on on-line accessible data for time sensitive applications. Real-time data compression can provide performance enhancements on both reads and writes for existing storage systems while preserving data integrity. The result is a reduction of data footprint for active on-line data.

There are many myths about data footprint reduction of active, on-line data (including databases and other applications) with changing data that cannot incur performance delays on reads or writes. Some myths are based on older technology or dated experiences, some are based on current technology that can be performance limited.

Value Proposition

Compression of active on-line data can speed up backups, snapshots, replication for BC/DR along with general access of data, including reading and writing of individual data blocks.

Applications that can benefit from real-time data compression that support performance optimization for reads and writes include:

- Databases on NAS based storage systems
- Home directories and general file sharing
- Financial, energy and telecommunications
- Web 2.0, entertainment, social networking
- Cloud based data and storage services

Benefits of real-time data compression for active and changing on-line data include:

- Address IT power and green issues
- Move more data in the same, or less, time
  o Enhance HA, BC and DR capabilities
  o Improve productivity of IT users
- Maximize usefulness of existing storage
  o Boost storage system capacity
  o Boost storage system performance
  o Off-load post processing overhead
  o Investment protection
- Compliment de-dupe and archive

The Technology

A common misconception is that all real-time compression solutions in the data path will degrade performance.

The reality is that some solutions can sustain effective compression rates, enabling more data to be written or read form a storage device. The result is improved performance on both reads and writes.

Compressed data has a smaller footprint, which means that available caches on storage systems can be used more effectively to boost both read and write performance. For example, a NAS storage system processing compressed data can move more effective in a shorter period of time data maximizing cache and boosting storage system performance.

Likewise, since data is smaller, less data needs to be moved over the network, the effective performance of storage systems is also boosted. The net benefit is that more data can be written or read while reducing the data footprint using existing NAS storage.

Not all NFS file access operations require that an entire file be sequentially read or written. With the advent of NFS V3 several years ago, NAS based solutions and applications that support direct I/O (DIO) operations can see marked performance improvements using DIO.

DIO enables applications such as Oracle databases to be stored on NFS storage with direct or random block type access being performed for reads and writes. This differs from traditional NFS file operations which tend to be sequential reads and sequential writes.

What this means is that applications, such as Oracle, that support DIO when working with real-time compression solutions that also support NFS DIO, can see benefits of a smaller data footprint and improved effective performance.

Strategies and Recommendations

Compression ratios and wire speed performance rates can be interesting; however, they may not be relevant. Look at the effective compression rates and effective compression impact of a solution. For example, at a 20 to 1 ratio on 20TB of backup data, reducing it to 1TB is impressive (e.g. elimination of 19 1TB disk drives).

What is even more significant is with as little as a 3 to 1 compression ratio, for 1,000TB of on-line primary data, the data footprint can be reduced to about 333TB. Real-time compression that provides an effective performance benefit also enables the equivalent savings of 666 1TB disk drives. The impact is much larger if smaller high capacity disk drives are being used.

The net result is the ability to delay installing more storage, or, to increase the effective capacity of existing and new storage while enhancing performance. This approach, even on a smaller scale, can be used to maximize available power and floor space during technology upgrades. For example, by compressing active data, data from other storage systems can be consolidated to free up floor space and power during instillation and migration to newer technologies.

A smaller data footprint also means that more data can fit into existing storage system cache to boost performance and maximize cache effectiveness. With a smaller on disk data footprint, for subsequent data access, more data can be pre-fetched as part of read-ahead algorithms common on many storage systems. Consequently, frequently accessed data is more likely to be readily available in a storage system protected cache to boost application performance and user productivity.

Not all on-line solutions provide real-time compression of both read and write data. Some solutions will un-compress static data that results in read only performance enhancements with no benefit to write or changed data optimization.

For on-line solutions that do not reduce the data footprint of changed data, there is a lost

performance benefit of not being able to store more data in a shorter amount of time. There can also be added overhead for the required post processing to go back and re-read data and re-write it back to a storage system.

Closing Comment

Real-time data compression for on-line primary NAS storage that can both reduce data footprint for changing data on the fly as well as compliment both read and write performance of various applications should be part of a holistic data footprint strategy.

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.