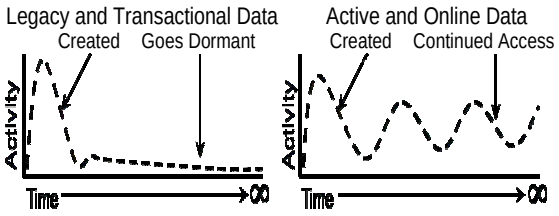
Real-Time Data Compression for active and on-line data

Author: Greg Schulz – Sr. Advisor **December 7, 2010**

This Industry Trends and Perspective Solution Brief is compliments of IBM Real-time Compression

Background and issues

There is no such thing as a data or information recession! IT organizations are looking for opportunities to stretch available resources (people, time, budgets, physical floor space, server, networking, storage along with power and cooling) to accommodate the ever-increasing amount of data being created.



Source: The Green and Virtual Data Center (CRC)

Figure 1 - Changing data access patterns

Many IT storage strategies are built around the premise that shortly after it is created, data is seldom, if ever, accessed again.

Figure 1 shows an example on the left of the traditional transactional data lifecycle with data being created and then going dormant. Once data goes dormant it is a candidate for archiving or other forms of data footprint reduction (DFR).

On the right hand side of Figure 1 is a different data lifecycle and access pattern. Unlike the transactional data lifecycle models where data can be removed after a period of time, unstructured primary data needs to remain online and readily accessible. This agrees with the idea that if data is online and accessible, it will be used and have increased value.

The result is that more data is finding and providing value by being on-line and accessible. The main challenge with this is cost.

What is needed

The question, then, is how to keep more data within a given set of constraints (power, cooling, floor space, capacity, management capabilities, software licensing, data protection windows) while enabling online access at a lower per unit cost for online active and primary data.

The solution is data footprint reduction (DFR) using real-time compression for online active and primary storage with changing data.

Applications and environments that benefit from real-time compression include:

- Databases on NAS based storage systems
- Home directories and general file sharing
- Financial, energy and telecommunications
- Business analytics and data mining
- Cloud, Web 2.0 and entertainment
- Pre-production video, CAD/CAM
- Backup/restore and active archive
- Server virtualization (e.g. VMware)
- Gaming security and video surveillance

Value proposition

Benefits of real-time compression include:

- Support on-line and primary applications
- Move more data in the same, or less, time
- Maximize usefulness of existing storage
  - Boost storage system capacity
  - Boost storage system performance
  - Off-load DFR processing overhead
  - Investment protection
- Complements other DFR solutions
- Reduce capital and operating costs
- Interoperability and application transparent
- Enhance energy efficiency
- Manage more data per staff time

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

Real-Time Data Compression for active and on-line data

The technology

A solution is IBM real-time compression for data storage systems. This technology is application and storage system transparent enabling it to plug into existing networks sitting between your application servers and network attached storage (NAS) systems from various vendors.

Instead of compression being performed on servers via software or on storage systems, taking away performance resources from other applications, data footprint reduction occurs in the IBM compression appliances. By performing data footprint reduction at the point of origin with less data being physically written to disk, storage systems can operate more efficiently.

At the heart of IBM real-time compression are three components:

- Random Access Compression Engine (RACE)
- Unified Protocol Manager (UPM)
- Monitoring and Reporting Manager (MRM)

RACE supports inline compression of random accessible active or changing data without performance compromise. For increased flexibility, compression can occur across file systems or on a selective basis per user configurable GUI screens. By supporting real-time and random access data compression, application transparency is enabled by default.

RACE leverages time tested and proven Lempel-Ziv (LZ) data compression algorithms providing lossless (no data loss) data footprint reduction. A further benefit of performing compressing on the fly within the IBM real-time appliances is the ability to off load storage system resources. For example, by reducing the size or footprint of data before it is written to a storage system's cache results in more cache for other functions. This allows storage resources to do more work as well as enhancing other downstream functions including snapshots, clones, replicas, and backups, including deduplicated backups.

UPM enables interoperability with various servers and storage systems via networking protocols including CIFS and NFS. MRM enables online storage compression trending, analysis and reporting.

How the solution works

RACE takes incoming data streams and compresses the data using an LZ based algorithm while preserving attached Meta data as it is written to a storage system. As the data is written and known to be intact on the storage system, an acknowledgement is sent back to the requesting application.

Since data is being compressed on the fly and has a smaller footprint as it is written to the storage system, the result is DFR without performance compromise. This approach of DFR on the fly and on a random basis makes for a suitable solution for online active file systems, home directories and even databases.

A byproduct of appliance based real-time compression is that subsequent downstream operations including storage system or array based snapshots, replication or other functionality perform without the need to re-inflate or uncompress the data. For example, data that is compressed and written to a storage system and then replicated will be already reduced enabling more efficient transmission on a local or wide area basis.

Read operations are the inverse of writes where data requests are made of the storage system and subsequently uncompressed on the fly. Since compressed data stored on the storage array system is smaller, it is also accessed faster enabling DFR without performance penalties.

About the solution

Currently there are two IBM Real-time Compression appliance models, the STN 6500 and STN 6800.

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

Real-Time Data Compression for active and on-line data

	STN 6500	STN 6800
Xeon 5600 Processors	Dual 2.4 Ghz	Dual 2.8 Ghz
DRAM Memory	72 GByte	72 GByte
1 Gb E Ports	16	
10 Gb E Ports		8
Mixed Ethernet[1]		8 x 1 GbE + 4 x 10 GbE
Power supplies	Dual	Dual

Figure 2 - IBM STN Appliance Source: IBM

Strategies and recommendations

Develop a holistic approach to managing to your growing data footprint. Leverage data compression as part of an overall data footprint reduction strategy.

Optimize and leverage your investment in existing storage across all types of applications. For on-line, active, primary applications and storage, real-time compression provides immediate DFR benefits with transparency and without introducing performance bottlenecks.

Avoid introducing problems such as performance bottlenecks or increased workload on existing storage systems in the course of implementing DFR. The solution to one problem should not introduce issues or challenges elsewhere. Look for DFR solutions that compliment your applications and storage environment providing investment protection without compromise.

Closing comments

IBM Real-time Compression appliances deliver real-time, random access, deterministic and lossless data compression while maintaining reliable and consistent performance and data integrity.

Where to learn more:

Learn more at www.storageioblog.com as well as at the IBM Real-time Compression landing page at www.ibm.com/storage/rtc where you can find additional information on the technologies and techniques discussed here.

[1] For the STN 6800, networking connectivity is either 8 x 10 GbE OR 8 x 1 GbE plus 4 x 10 Gb Ethernet Ports