WHITE PAPER

Real-time Compression Advances Storage Optimization

Sponsored by: IBM

Laura DuBois

May 2012

EXECUTIVE SUMMARY

It should come as no surprise that storage budgets are constantly under pressure from opposing forces: On one hand, economic forces are pushing budgets to either stay flat or, in many cases, shrink as a percentage of a company's revenue. On the other hand, the infrastructure struggles to keep up with the pace of data growth, pressured by many variables, both social and economic. Businesses have no choice but to acclimate their storage infrastructure to the unprecedented levels at which data is growing.

For the most part, the inefficiency of compute infrastructure has been tackled via server virtualization. Racks and racks of idle servers have been slowly but surely replaced with ultradense virtualized compute environments that occupy a fraction of the space, putting the onus on storage vendors to deal with the challenge of making data storage efficient and economically sustainable.

Storage vendors have responded in kind by developing a suite of solutions aimed at slowing down the consumption of storage. Storage optimization technologies, as they are called, seek to reduce the storage footprint by removing redundancy and wastage and by optimizing data placement. Traditionally, storage optimization suites have consisted of thin provisioning, deduplication, and postprocess compression. Thin provisioning targets wasted storage that is allocated but unused because of overprovisioning. Similarly, deduplication targets redundant or duplicate data, creating single instances of data. For data that can be neither "thinned" nor "deduplicated," compression targets the reduction of the overall footprint by squeezing the data to make it smaller. The promise has been that by deploying one or more — or all — of these technologies, businesses can seek to reduce their overall storage spend, delay future purchases, and realize a better return on their investments.

In reality, however, every optimization puts some overhead on the system, in terms of either performance or efficiency. Historically, the worst offender in this mix has been compression because of its inherent compute-intensive postprocess nature. This has often led businesses to deploy thin provisioning and/or deduplication but leave compression out of the mix when it comes to active primary storage.

IN THIS PAPER

In this paper we examine how IBM Real-time Compression is changing the game by offering a robust, efficient, and cost-effective optimization solution. Moreover, IBM is demonstrating that compression does have a place in storage optimization and nicely complements other optimization techniques in play today. IBM estimates that by

Global Headquarters: 5 Speen Street Framingham, MA 01701 USA P.508.872.8200 F.508.935.4015 www.idc.com

deploying Real-time Compression–enabled solutions such as IBM Storwize V7000, businesses can save 50% or more in the amount of physical space required and reduce overall storage growth by nearly 30%. According to IBM, this typically results in a 30–40% reduction in costs per gigabyte (GB) for primary storage configurations without compromising any capacity or performance.

IBM Real-time Compression is a core component of IBM Smarter Storage, which in turn is part of IBM's Smarter Computing strategy, an approach built on years of storage industry leadership and innovative and forward-looking technology that help guide the design and deployment of storage systems. Smarter Storage enables organizations to take control of their storage so that they can focus on gaining more valuable insights from their data and delivering more value to the business.

Market Situation

We live in a post-PC world. Mobile computing devices continue to proliferate everywhere, including in businesses. By 2015, the mobile computing market will consist of over 243 million tablets and over 1.5 billion smartphones. According to IDC, around 9.57 zettabytes (ZB) of information is consumed each year, of which 1.2ZB is unstructured media data that is growing at a 62% compound annual growth rate (CAGR). As the enterprise, prosumer, and consumer demographics shift toward data creation and access on the go, so does the demand for more storage. IDC research shows that worldwide external disk storage systems revenue posted year-over-year growth of 7.7%, totaling just under $6.6 billion, in the fourth quarter of 2011. The total disk storage systems capacity shipped in 2011 reached 6,279 petabytes, growing 22.4% year over year.

With enterprise demand for storage capacity worldwide projected to grow at a CAGR of over 43% from 2008 to 2013, the demand for data vastly outstrips the supply of storage. Additionally, the lingering sluggishness in the economy means that businesses can no longer afford to accommodate this data growth using traditional approaches to data storage. Gone are the days of treating storage as a dumb tier with adequate protection and performance mechanisms such as RAID groups and volume managers. Costs per GB for disk may be going down, but infrastructure costs are shooting up, forcing businesses to discard traditional approaches to data storage in favor of newer and better approaches.

Storage vendors have largely followed the example set by server virtualization vendors by creating an ecosystem of technologies that improve the utilization of the storage infrastructure, making it more efficient in the process. The evidence for the increasing market adoption of such technologies is in the fact that today, almost all storage solutions are prebuilt with one or more storage optimization technologies that can be deployed straight out of the box. While it may make it a bit challenging to measure the size of the storage optimization market as a fraction of the overall storage market, the rate of adoption of these technologies means that eventually all installed storage will be optimized in some way or other.

Why Is Storage Optimization Not Optional Anymore?

While the primary driver for storage optimization may be rampant data growth, storage optimization is also influenced in large part by inefficiencies introduced by how data is created, accessed, and/or stored. Some of these inefficiencies have a human element to them, but some of them are by-products of technology sprawl.

- The amount of static data (i.e., data that is hardly accessed on an ongoing basis) continues to grow at a much faster pace than the amount of active hot data (i.e., data that is accessed all the time).
- There is rampant duplication in data that is created thanks to newer technologies. For example, server virtualization may create images of guest operating systems that are nearly identical to each other. Another example is users creating duplicate copies of images, files, and other data.
- Some of the formats used for creating static content are natively inefficient (i.e., the format creates a lot of empty space). This results in extraneous space on disk.
- Structured data storage such as databases may result in inefficiencies on the storage side because of the metadata elements that support various attributes for this data.
- Businesses may have disparate (and, at times, multivendor) storage assets that have created silos in the infrastructure. Because of varying interoperability guidelines or performance considerations, businesses may find it challenging to deploy such assets in a homogeneous manner, resulting in severely underutilized assets.
- Storage tiering sounds like an ideal plan to rightsize the storage infrastructure, but in practice, the lack of automated tiering mechanisms makes it challenging for storage administrators to move data from one tier to another. This results in the inefficient use of storage tiers and fewer benefits from tiering.

Storage optimization technologies are aimed squarely at improving asset utilization by changing data placement and organization. As an additional layer of technology, storage optimization is ultimately designed to overcome the inefficiencies introduced by humans or technology itself:

- **Automated tiering** is the ability to move only active portions of data to higher (and more costly) tiers while the less active majority of the data remains on lower tiers of storage. The underlying principle is that while data may grow, an ever-increasing percentage of this data is static or cold data that does not need to reside on high-performance tiers at all times.
- **Storage virtualization** is the ability to pool together disparate and often multivendor storage resources under a common management and presentation layer.
- **Thin provisioning** is the ability to define a storage unit (full system, storage pool, or volume) with a logical capacity size that is larger than the physical capacity assigned to that storage unit. The host or device accessing this storage unit sees the logical capacity and not the physical capacity.

- **Deduplication** is the ability to examine a single data block or file or a series of data blocks or files for common patterns and replace them by pointing them to a single instance of that pattern, thereby reducing the duplication of such patterns in the storage frame. Because of the nature of block access, deduplication is offered mostly in file-based storage.

- **Compression** is the ability to squeeze data so that the blocks become smaller. The use of compression allows this data to consume much less storage compared with other optimization technologies, such as deduplication or thin provisioning.

As the drivers and responding storage optimization techniques demonstrate, storage optimization is not a one-size-fits-all approach, unlike some other storage technologies such as replication, in-array clones, and so on. The diverse nature of data itself makes some storage optimization techniques better suited for certain data types than others. However, by deploying storage optimization technologies as a bundle in a shared storage infrastructure, businesses can derive significant tangible benefits without compromising service quality.

Real-time Compression

Before we highlight the benefits of Real-time Compression, let us first examine why compression is often excluded from the list when it comes to storage optimization. Many vendors are quick to point out that unlike thin provisioning or deduplication, compression can have mixed and sometimes adverse results. That is partly because the traditional approach to compression forces vendors to strike a delicate balance between compressibility and performance penalty, resulting in suboptimal results. This is why it typically is deployed for data that is used infrequently.

Challenges with Using Compression

The traditional and commonly adopted approach is to compress data at rest. This type of compression, which is also known as postprocess compression, kicks in after the data has already been written to disk. This is very similar to several host-based utilities that compress files and folders once they have been created. This method is inherently inefficient because its algorithms consume significant compute cycles, are disk intensive, and require additional "jump" space to store the uncompressed data. On a host, under most circumstances, one can get away with this added penalty, but on a purpose-built storage array, an added overhead of this sort can easily produce noticeable degradation in service quality, which is automatically noticed on the server and applications. As a result, vendors are quick to downplay compression in most scenarios unless the data set is insignificant or the storage system as a whole is underutilized.

In traditional compression, when applications make multiple updates to data, such changes are written to disk in an uncompressed manner. Subsequently, a separate operation has to be scheduled that samples and compresses this data based on its physical location on a volume, irrespective of its relationship with other data blocks that the application may be accessing.

Results produced by traditional compression engines also vary greatly. Such compression engines ingest a fixed preprogrammed chunk of data and produce a variable output depending on the compressibility of that data. Furthermore, compression ratios are based on chunk sizes: The larger the chunk, the greater the compression ratio and the greater the performance overhead. Smaller chunk sizes result in poorer compression ratios. The happy medium between chunk sizes and performance overhead is not so happy after all.

A side effect of compression is that it results in fragmentation. Due to the postprocess and variable nature of traditional compression engines, compression "degenerates" the continuity of data over time. The compressed data is spread out in chunks across the volume and requires frequent garbage collection. The impact of this effect is performance impact over time.

Real-time Compression Reinvigorates the Role of Compression in Storage Optimization

Today, most vendors offer compression as an optional licensable feature but strongly recommend that administrators enable it only on select data sets. In most situations, the adoption of compression for active primary data in the storage infrastructure at large remains limited. Because every performance penalty on the storage system can have a ripple effect on the rest of the application stack, most businesses choose to leave compression disabled.

A newer approach known as Real-time Compression may be bringing the role of compression in the primary storage optimization picture to the fore again. Real-time Compression makes the use of compression in general-purpose configurations feasible without operational penalties. It promises compression multiples of up to five times for primary online data. Furthermore, by shrinking primary data, it also compresses all derived copies of that data, such as backups, archives, snapshots, and replicas. IBM has ported the technology to the Storwize V7000 storage system, eliminating the need for external appliances.

IDC expects other vendors to follow IBM's lead in moving away from traditional at-rest compression and in developing similar but competing compression technologies. This will validate and reinforce the role of compression as a core component of efficiency. In fact, as storage processors become faster and more powerful, storage vendors could offload compression/rehydration cycles to dedicated cores or processors in the storage controller in addition to deduplication cycles.

What Is Real-time Compression, and What Are Its Benefits?

Unlike traditional compression technologies that compress data as a postprocess operation, Real-time Compression operates on active primary data as it is being accessed. This expands the realm of compression to a much wider set of workloads with predictable and measurable results. Moreover, this compression is "always on," which means it can be enabled on active workloads and does not require scheduled periods for postprocessing, unlike its predecessors.

The primary differentiator is that in Real-time Compression, the compression engine processes a variable data stream based on the patterns of data that are actually written. Real-time Compression takes advantage of "temporal locality" and not physical location: Data that is accessed together is compressed together independent of its physical location. So when applications make related updates to different parts of the volume, such updates are handled together in a contiguous manner. This is akin to real system operation because it takes advantage of the structure of the data, the size of the data, and the data's relationship with other data. This awareness of workload minimizes the number of compression and/or decompression operations, resulting in fewer disk IOPS and less overhead on the storage controller. This increases compression ratios and efficiency without compromising performance.

IBM Storwize V7000 Implementation

IBM acquired Real-time Compression technology via its Storwize acquisition in 2010. Initially, IBM offered this technology via purpose-built in-band appliances for transparently compressing primary NAS data. IBM still sells these appliances today but has also taken the bold step of porting the technology into its Storwize V7000 platform. The Storwize V7000 platform will use the same Random Access Compression Engine (RACE) technology as the appliances.

IBM plans to introduce compression into the Storwize V7000 as an optional licensable component via a firmware update. Thus, existing users of Storwize V7000 will be able to select compression as an attribute (or preset) when creating new volumes. Administrators will also have the option to convert existing volumes using volume mirroring to compress thinly provisioned volumes and eliminate unused space in the process. IBM initially plans to support up to 200 compressed volumes but may increase this limit with later releases. The Storwize V7000 GUI will present compression-related performance information and allow administrators to manage and monitor compression from a single console. Real-time Compression enhances the functionality of Storwize V7000 without creating additional administrative overhead.

How Does Compression in Storwize V7000 Compare with the Traditional Approach?

As noted previously, traditional approaches to compression have received a bad rap for their unpredictable performance, overhead, and cumbersome postprocess nature. Real-time Compression on the other hand changes the equation by creating an optimization layer that is inline and efficient. The immediate impact of implementing Real-time Compression is that the storage system has to deliver with fewer IOPS during the compression routine. Active workloads such as databases and email systems often perform small updates to existing data. IBM's analysis suggests a significant improvement compared with traditional approaches for these workloads (see Table 1).

TABLE 1

Comparison of Traditional Compression and IBM Real-time Compression

Traditional Compression	IBM Real-time Compression
1 MB "chunk"	100 Byte update
1 MB read	0 MB read
1 MB decompress	0 MB decompress
100 Byte update	0 Byte update
1 MB compress	100 Byte compress
1 MB write	< 100 Byte write
Total IO per compression operation: 2 MB	Total IO per compression operation: < 100 Bytes

Source: IBM, 2012

Storwize V7000 Real-time Compression achieves compression ratios similar to those of IBM Real-time Compression appliances (see Table 2). Because Real-time Compression can be deployed with a wider range of data than traditional compression, the potential benefits can be greater. Predictable compression ratios make it easier for businesses to plan budgets accordingly.

TABLE 2

Compression Ratios Observed by IBM in Client Environments

Application		Observed Compression
Databases		Up to 80%
Virtual Servers (VMware)	Linux virtual OS	Up to 70%
	Windows virtual OS	Up to 50%
Office	2003	Up to 75%
	2007 or later	Up to 20%
CAD/CAM		Up to 70%

Source: IBM, 2012

By enabling compression on new or existing volumes in Storwize V7000, businesses can extract greater usable capacity from their existing investments. This allows many businesses to flatten or reduce their storage spend for most common configurations.

An additional benefit of using Real-time Compression with Storwize V7000 is that even volumes carved out from external virtualized storage can be compressed. In many cases, this presents nearly double the amount of usable capacity for modest capex investments and lower opex costs.

Challenges and Opportunities for Storwize V7000

In current times, when storage budgets are either capped or reduced, IT organizations continue to look for ways to efficiently store rapidly growing data at lower costs. Real-time Compression, while attractive, is still in its early adoption phase. Many organizations still regard compression as a compute-intensive postprocess technology.

One of the principal challenges that organizations will face today while enabling Real-time Compression is that of risk versus reward. Organizations would certainly demand to know answers to questions such as the following: How much am I paying to compress my data, and what savings will I see as a result? What are my risks with compressing my data? What are the performance penalties? Will I be able to maintain the compression ratios, or will they degenerate over time? Being able to quantify savings will be a huge factor when deciding if the software is worth acquiring or not. Organizations can avail themselves of IBM's Compresstimator tool to model expected compression benefits for specific environments. Using the results and expected growth rates, organizations could derive potential savings by deploying Storwize V7000 with compression or enabling compression on existing volumes. IBM could proactively leverage this tool to provide both customers and prospects with a view of potential savings.

The task ahead for IBM is to help organizations shed their perception of compression as an inherently postprocess feature. IBM can continue to gain mindshare among its customers by educating them about best practices and use cases for deploying Real-time Compression, creating documents that describe the use of compression alongside other storage optimization technologies such as thin provisioning, storage tiering, and deduplication.

CONCLUSION AND ESSENTIAL GUIDANCE

Storage optimization technologies are here to stay. Thanks to Real-time Compression, compression has a place in the storage optimization portfolio. IBM has taken a step in the right direction by offering this technology in one of its flagship storage products.

The predictable, reliable, and linearly scalable nature of Real-time Compression will fuel its adoption in environments with diverse workloads. This in turn will lead to businesses moving out of their comfort zones and making Real-time Compression one of the "must have" technologies in their environments. Being able to quantify, predict, and measure the savings brought about by storage optimization technologies such as Real-time Compression will better equip businesses to tackle explosive data growth.