TANEJA GROUP
TECHNOLOGY ANALYSTS

T E C H N O L O G Y B R I E F

Introduction to Capacity Optimization

May, 2005

THE NET NET

The past 3 years have seen rapid and significant shifts in how storage solutions approach capacity management. New archiving and backup solutions are now achieving massive levels of capacity efficiencies, far beyond any technology in history. To date, the industry has not collectively agreed on a common descriptor for this extremely critical enabling technology, and therein rests the purpose of this report. We call this new technology Capacity Optimization (CO), and the various storage solutions stemming from this technology are termed Capacity Optimized Storage (COS.)

Unlike traditional compression techniques that may provide 2x benefits, CO techniques can reduce the storage requirements of content by several magnitudes, in some cases, delivering in excess of 20x compression. CO technologies are seeing adoption today primarily in disk-based data storage and enterprise networking industries where the benefits of reduced capacity and network utilization are most beneficial. Capacity Optimization will play a foundational role in storage system architecture throughout this decade, and its principles therefore merit detailed analysis.

What is Capacity Optimization?

Capacity Optimization (CO) is a new technology designed to massively reduce data down to its raw essentials. Unlike traditional compression which typically reduces data to half of its previous size, capacity optimization can reduce standard business data down to a twentieth or less of the original size. This is achieved by breaking the data into a small number of fundamental parts which when replicated can be used to rebuild the original data. This technique is being used both in storage devices and also networking devices to build much more cost-effective systems. For example, a storage device that utilizes CO can store 20x as much data as a standard one at the same cost. A capacity optimized network can transmit 20x as much data as a non-optimized one again all for the same price. This order of magnitude cost reduction in capacity optimized technologies vs. standard ones is ensuring that all future system designs will at some point implement this technology. Around ten companies today are shipping products that use capacity optimizing technologies and much of the continued development being done on the fundamental algorithms are tuning them for performance and scalability now that the basic principles of operation are well understood.

87 Elm Street, Suite 900 Hopkinton, MA 01748 Tel: 508-435-5040 Fax: 508-435-1530 www.tanejagroup.com

A Lexicon for Capacity Optimization

In order to discuss compression and capacity optimization it's useful to define some simple terms that will be used throughout this report:

Object – A data container such as a file or network transmission

Part – A piece of an object containing a fixed or variable amount of data

Plan – A plan that shows how to assemble ***parts*** into an object

Optimized Store – A location for storing ***parts*** and ***plans***

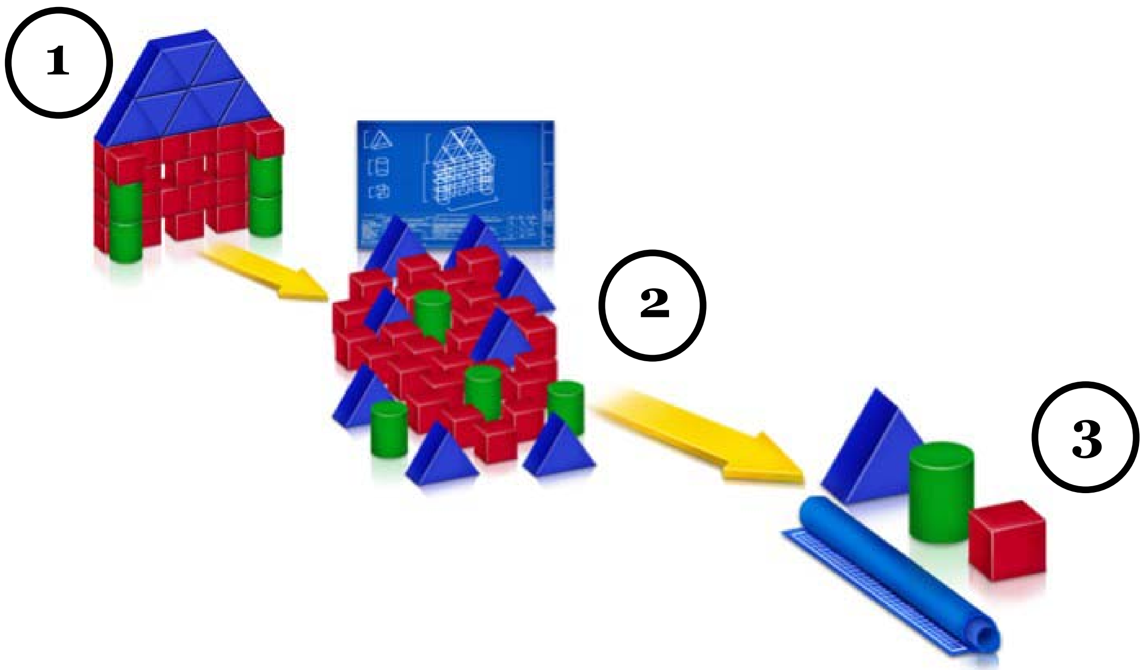
CO Versus Compression

Regular compression examines the data in an object for repeating patterns or extended runs of the same data. For example, an object which contains a thousand zero's one after the other could be represented by two numbers **1000 0**, rather than a thousand separate elements. These two numbers could then be stored or transmitted over a network much more efficiently than the one thousand individual zeros. Likewise, the pattern **123412341234123412341234** could be stored as the five numbers **6 1 2 3 4** (six times one two three four) rather than the twenty-four original digits (almost a five fold compression).

Capacity Optimization Explained

At a basic level, capacity optimization works in a very similar manner to regular compression technologies. ***Objects*** are broken down into ***parts*** of either fixed or variable sizes. A ***plan*** is constructed to show how to build the ***parts*** back into the original ***object***. The ***parts*** are then compared to see which are unique. Any non-unique ***part*** is discarded resulting in immediate compression. The unique ***parts*** and the ***plan*** showing how to assemble them back into the original ***object*** are then stored.

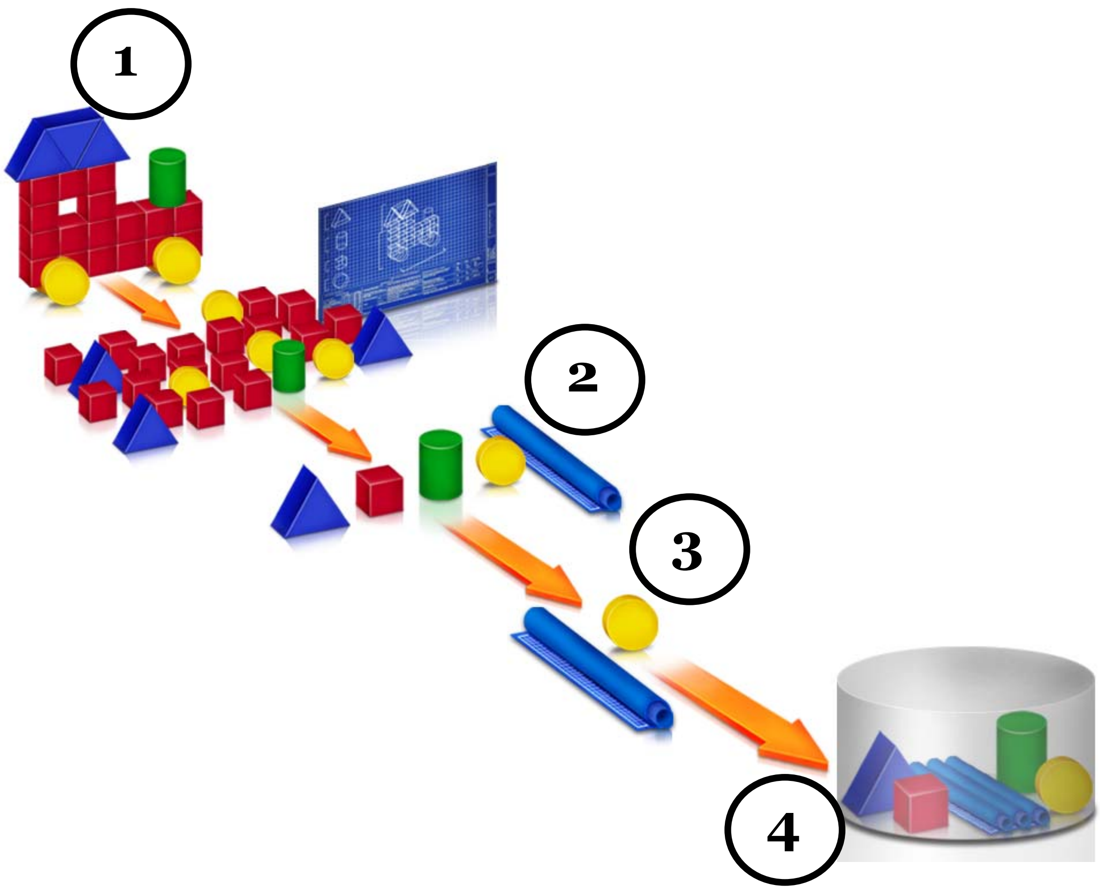
By analogy, consider a house made out of children's building blocks, as illustrated below.



(1) *We begin with a house created from blocks and then we break the house down, ending up with many blocks that take up quite a lot of space. Importantly, we maintain a blueprint, or plan, showing how to rebuild the house. This "house plan" takes up very little space (**2**). Because many of the blocks are exact copies of each other we can always make more as long as we keep at least one of them. So we discard any block that isn't unique and keep the plan and the unique blocks. By doing this, we have everything we need to rebuild the house at any future time (**3**).*

This analogy demonstrates the first principle by which CO technologies approach the storing of blocks of data. Because they maintain maximally granular plans for all objects, they only need one instance of any given object. Most importantly, CO technologies only increase their efficiency the more they are used for storing content. This is true because over time, a capacity optimizing system is introduced to new ***objects***, breaking those ***objects*** down into ***parts*** while maintaining all new ***plans*** in an ***optimized store***. Before the new ***parts*** are added to an ***optimized store*** those parts are compared with *all* of the other ***parts*** already located in the store. Any ***parts*** already stored, even if from other ***objects*** stored earlier are discarded. Then, any remaining unique ***parts*** are inserted into the ***optimized store*** along with their corresponding ***plan.*** Over time the ***optimized store*** contains a plan for literally each ***object*** stored, along with all of the necessary ***parts*** to rebuild them. The concept of the ***optimized store*** finally achieves compression efficiencies of over twenty times in real world environments for typical data objects.

For a clarifying example, we can return to our illustration with a new scenario:



*(1) After breaking the house from the example above down to its blocks, we decided to build a train out of those same blocks, but we have to introduce one new type of yellow block to be included for the train wheels. (2) Along with our block train, we have a"train plan" for how to build it. When we broke the train down, we keep the plan and a single copy of each type of block. (3) We then discard all the blocks except the new yellow train wheel block and our plan. Recall that we can do this because we already had a copy of the others stored from our earlier breaking down of the house, and the plan will let us get access to however many of them we need. (**4**) Now we only have to add the new "train wheel" block and one new "train plan" to the optimized store in order to be able to recreate the whole train (**4**).*

Through this analogy, we can see how CO technologies achieve increasingly higher levels of efficiencty as more ***objects*** are added. This is possible because we have an ***optimized store*** of the minimum number of unique ***parts*** and ***plans*** required to assemble those ***parts*** back into their original ***objects***.

This is precisely the technological approach being utilized by various companies in the data storage and networking industries today to achieve maximal efficiencies in how content is stored or transmitted on a network.

Capacity Optimized Storage (COS)

The primary emerging market utilizing CO today is the data protection marketplace. In this market, corporate data has traditionally been stored on tape rather than disk. Although disk is widely recognized to keep data in a manner that is safer and more accessible than tape, until now it had been too expensive to use for longer-term data protection, due to cost differentials making tape as tape about 1/20th the cost of disk storage solutions.   By utilizing CO, a new generation of solutions are able to provide a massive savings over regular disk based storage, brigning them into price equivalence with more traditional tape solutions.

This savings occurs because when data is migrated to the capacity optimized solutions, only the changed and unique new data (typically only 5% of the data stored) needs to be stored on the disks.

Capacity Optimized Transport (COT)

New solutions in networking are also utilizing capacity optimization. Traditionally transferring data over the wide area is a slow and costly business. The network links connecting geographically disparate sites have only a tenth or less of the bandwidth of their local counterparts. Also the latency inherent in the dominant network protocols also ensures the transfers speeds are often limited well below even this lower level of bandwidth. As a result, there is a need for optimizing traffic flowing over the wide-area network links.

In this scenario, capacity optimized network transport devices sit at either end of a wide-area network link each maintaining an ***optimized store***. When new network traffic (the ***object***) arrives destined to be sent over the link the device examines the traffic for ***parts*** already transmitted (these are kept in its local ***optimized store***). Only the new unique ***parts*** of the traffic are sent over the link along with a ***plan*** of how to reassemble the ***parts*** back into the original traffic. At the other end of the link the second capacity optimized device receives and stores the new unique ***parts*** in its ***optimized store*** and then reassembles the original traffic using the transmitted ***parts*** and other ***parts*** sent earlier which it pulls from its store are directed by the received ***plan***. This reassembled traffic is identical to the original that would have been transmitted over the link and this is sent onto its final destination. As only the new ***parts*** from the traffic were transmitted, the bandwidth used on the network link is dramatically reduced overall. As with all capacity optimized solutions, as more data is sent over the link the level of efficiency will increase and more ***parts*** are accumulated in the ***optimized stores*** at either end of the link.

CO Players Today and Tomorrow

Today, we see CO technologies in a range of emerging companies. They all leverage unique intellectual property, but they achieve single-instance storage or transport through the principles of Capacity Optimization. In the disk backup and archiving realm, this list includes Archivas, Avamar, Data Domain, DataCenterTechnologies, Rocksoft, and Permabit. Amongst established players, HP's RISS solution also utilizes CO technology for disk archiving. In the areas of network transport, we see this technology in use by companies in the WAN optimization space, including Peribit and Riverbed. Over the course of the next 2 years, CO technologies will become a ubiquitous feature in the arsenals of established players for both storage and networking solutions, either through acquisition of smaller companies or internal development efforts. In short, CO is going to happen. Quite simply, there is no other way for enterprises or individuals to deal with the massive data deluge than to pursue these technologies to such ends, and all vendors will have to respond.

Brad O'Neill *is a Senior Analyst and Consultant with Taneja Group Inc.*

NOTICE: The information and product recommendations made by the TANEJA GROUP are based upon public information and sources and may also include personal opinions both of the TANEJA GROUP and others, all of which we believe to be accurate and reliable. However, as market conditions change and not within our control, the information and recommendations are made without warranty of any kind. All product names used and mentioned herein are the trademarks of their respective owners. The TANEJA GROUP, Inc. assumes no responsibility or liability for any damages whatsoever (including incidental, consequential or otherwise), caused by your use of, or reliance upon, the information and recommendations presented herein, nor for any inadvertent errors which may appear in this document.