De-dupe: It’s not a question of if, rather where and when!

”What to Look for and What to Avoid“

By Greg Schulz

Founder and Senior Analyst, the StorageIO Group
Author “The Green and Virtual Data Center” (CRC)
“Resilient Storage Networks: Designing Flexible Scalable Data Infrastructures” (Elsevier)

StorageIO

July 29, 2010

Introduction

There is no such thing as a data or information recession! Data keeps growing in almost every economic climate. Fixed or shrinking budgets mean that IT departments have to store and protect more data in a denser, safer and more cost effective manner without reducing the quality of service they provide to their customers. An effective and enabling data storage technology able to help IT departments do more with less is data deduplication. This paper looks at the value of data de-duplication for optimizing data protection along with considerations about how to add it to your storage strategy most effectively.

Background and issues

Organizations are faced with the constant demand to store more data, including multiple copies of the same or similar data, for longer periods of time. The result is an expanding data footprint resulting in increased IT expenses, both capital and operational, due to additional Infrastructure Resource Management activities to support and sustain given levels of application Quality of Service (QoS) delivery (Figure 1).

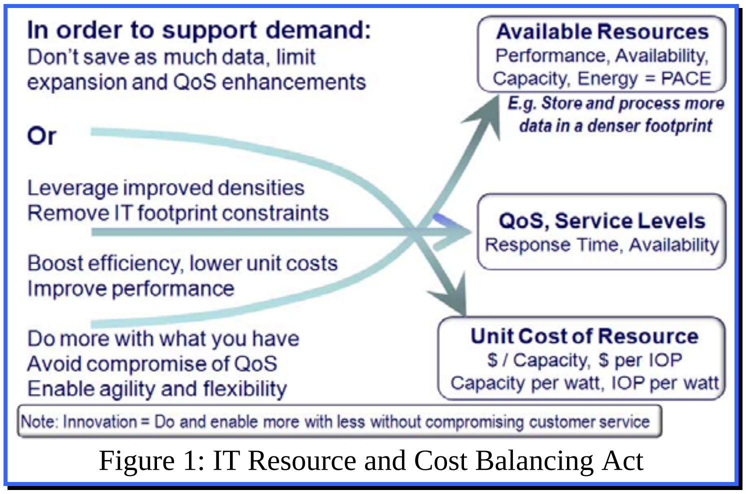

In order to support demand:
Don't save as much data, limit expansion and QoS enhancements
Or
Leverage improved densities
Remove IT footprint constraints
Boost efficiency, lower unit costs
Improve performance
Do more with what you have
Avoid compromise of QoS
Enable agility and flexibility
Available Resources
Performance, Availability, Capacity, Energy = PACE
E.g. Store and process more data in a denser footprint
QoS, Service Levels
Response Time, Availability
Unit Cost of Resource
$ / Capacity, $ per IOP
Capacity per watt, IOP per watt
Note: Innovation = Do and enable more with less without compromising customer service

Figure 1: IT Resource and Cost Balancing Act

Data de-duplication reduces data footprints

Data de-duplication is a set of techniques that recognizes redundant data and uses reference pointers to eliminate the need to store the same data multiple times. It reduces the impact of expanding data footprints by optimizing disk storage requirements, and it reduces network bandwidth capacity needed to transmit data sets. The result for data sets like back-up, with high levels of redundancy, is that less space and bandwidth are needed to store or transmit a given amount of information. An important attribute of some de-dupe solutions is the ability to coexist with existing environments, leveraging installed technologies and maximizing investment protection while minimizing IT operations disruptions.

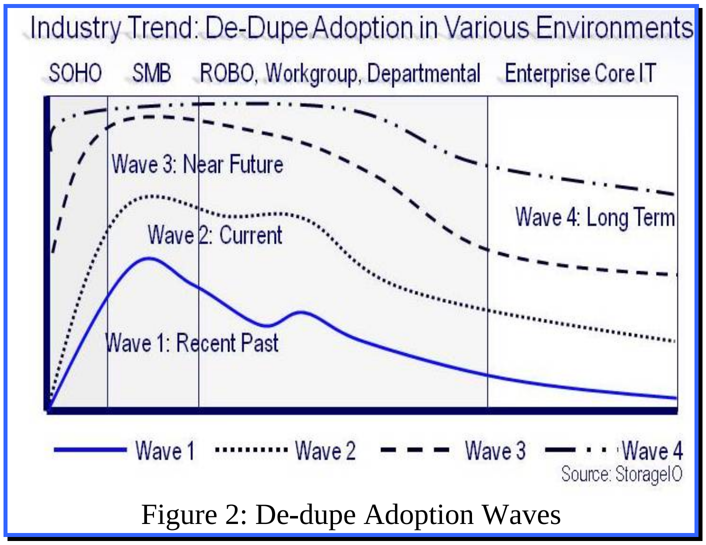
De-dupe adoption trends and directions

Data de-dupe continues to evolve as a technology, finding increased adoption with deployments in a variety of different customer environments (Figure 2). Data de-dupe in general provides a Green, efficient, effective and economically optimized IT data storage infrastructure, especially when it is combined with the use of removable media for long-term storage. In general, the business and IT benefits of de-dupe include:

- De-dupe can improve performance using less capacity than traditional disk based back-up
- Retention of more data in the same or smaller footprint enabling faster, fine grained restores
- Accelerate BC/DR moving more data over networks to meet RTO/RPO requirements
- Enhance data RPO and RTO by having more frequent copies of data for rapid restores
- Enable comprehensive remote office branch offices (ROBO) data protection using replication
- Streamline routing data protection management tasks including simplify media management

Data De-dupe: It’s not a question of if, rather where and when!

While de-dupe is a popular technology from a discussion standpoint and has good deployment traction, it is far from reaching mass customer adoption or even broad coverage in environments where it is being used. StorageIO research shows broadest adoption of de-dupe centered around back-up in smaller or SMB environments (Wave 1 in Figure 2) with some deployment in Remote Office Branch Office (ROBO) work groups as well as departmental environments.



Source: StorageIO

Figure 2: De-dupe Adoption Waves

StorageIO research also shows that complete adoption in many of those SMB, ROBO, work group or smaller environments has yet to reach 100%. This means that there remains a large population that has yet to deploy de-dupe as well as further opportunities to increase the level of de-dupe deployment by those already doing so.

There has also been some early adoption in larger core IT environments where de-dupe coexists with complimenting existing data protection and preservation practices. Another current deployment scenario for de-dupe has been for supporting core-edge deployments in larger environments that provide support for back-up and data protection of ROBO, work group and de-departmental systems.

Note that Figure 2 simply shows the general types of environments in which de-dupe is being adopted and not any sort of indicators as to the degree of deployment by a given customer or IT environment.

While some users do not yet understand the benefits, others want to make sure that the gains they can achieve by deploying de-dupe are not offset by the cost and complexity of integration or of potential performance impacts. There are many different approaches for deploying de-dupe technology and some can require changes to existing back-up and data protection environments. Consequently it is important to understand what to look for as well as what to look out for when evaluating candidate solutions.

Performance and Deployment: Two Key Issues

Look carefully at performance when you are evaluating de-duplication solutions. Some users of de-dupe have reported increases or enhancements for back-up performance in terms of shortening data protection windows or getting more data protected in the same amount of time when using de-dupe solutions. However, as with many things in IT and data storage in particular, the level of optimization will vary due.

For example, where the de-dupe operation is performed, such as software running on a media server, or whether the media or application servers being backed up can push data fast enough to the de-dupe target can have a bearing on performance optimization.

Keep in mind that data de-dupe technology in general trades processing time for reducing the amount of space or disk storage capacity needed for a given amount of data. Consequently de-dupe takes more processing or compute power to perform the calculations to implement the "thinking" needed to recognize what data has already been seen and reduce the data accordingly.

This means that more CPU or compute performance is needed than what is necessary to just write native data to a storage target device. As such, it makes sense that dedicated, special purpose appliances or platforms are often more efficient at carrying out de-duplication processing than general purpose application servers that are also doing other work.

When considering adding de-duplication software functionality to existing server hardware, it is important to think about how much additional load the server can handle without causing aggravation by creating bottlenecks with subsequent performance degradation. As is the case with any data footprint reduction technology or strategy, there needs to be a balancing act between optimizing for performance (time and rates), optimizing for capacity (space and ratios), and avoiding performance bottlenecks. This means that one of the things to look for in a de-duplication solution is the availability of tools to help you assess how effective de-duplication will be with your data sets.

Another issue to consider is how easily the de-dupe solution works within the existing IT back-up infrastructure, including hardware, software, and staff skill sets. Introducing new technologies that are not compatible or able to coexist with current policies or procedures can result in a step backward instead of forward.

What to watch out for when evaluating de-dupe solutions

De-dupe solutions should work for you, not the other way around, which means that they should not force or limit how data will be protected and preserved in your IT environment.

In general, watch out for de-dupe solutions that:

- Are limited to a small number of back-up applications or third party tools
- Require extensive hardware and software integration efforts
- Introduce point products with separate management and hidden licensing fees
- Force a compromise between performance and data reduction benefits
- Lack flexibility to be tailored to meet specific application and business needs
- Limit interoperability with existing storage devices and tiered mediums
- Add complexity for protecting ROBO or Workgroup with replication to Core IT
- Cannot adapt to different environment from SMB to ROBO to Core-Edge and Core IT

What to look for in a de-dupe solutions

The first and foremost question is whether the proposed solution is the right combination of size and performance for the specific data sets and environments. Ease of use and deployment is another important attribute along with the ability to merge seamlessly into the existing IT environment. This includes interoperability with existing hardware and software management tools as well as the ability for a solution to adapt to the specific needs at hand.

In general, look for de-dupe solutions that:

- Coexist with existing processes, technologies and tools to simplify management
- Are interoperable (non-intrusive) for seamless integration with the existing IT environment
- Avoid having to rip and replace to enable investment protection
- Provide flexibility to adapt to existing QoS, SLA, RTO and RPO needs
- Enable simplified management and resource situational awareness insight
- Support open initiatives and interoperability with approaches such as Symantec OST
- Facilitate protecting data from multiple locations to different locations

Putting it all together

Back-up and data protection environments change and evolve over time to reflect different business requirements, service objectives as well as technology shifts such as server virtualization[1]. Data de-duplication is an important innovation that is helping IT organizations keep up with the growing demands of more data to protect and preserve[2] in less time using less physical storage capacity in a cost and energy efficient manner.

Data de-dupe shrinks data footprint impact making it possible for IT organizations to see shorter back-up and restore windows while protecting more data. In addition to either reducing or protecting more data to sustain growth within available time window constraints, less physical disk capacity is needed to store a given amount of protected data. By using less disk storage capacity, more protected copies of data can reside on-line enabling faster restores and also more granular file restore capability.

Another by product of leveraging a denser protected data footprint using de-dupe is that the amount of data to be moved or migrated to tape is staged enabling more efficient streaming of data. The benefit is that existing tape devices and media are more fully utilized, further reducing overall data protection and storage costs, while maintaining or enhancing QoS and SLA requirements and supporting expanding data growth needs. This benefit can be especially powerful when it leverages deduplication to reduce the bandwidth needed to replicate back-up data between remote sites. A summary of a business benefits of de-dupe that seamlessly integrates with the existing IT infrastructure resources are summarized in Table 1 below.

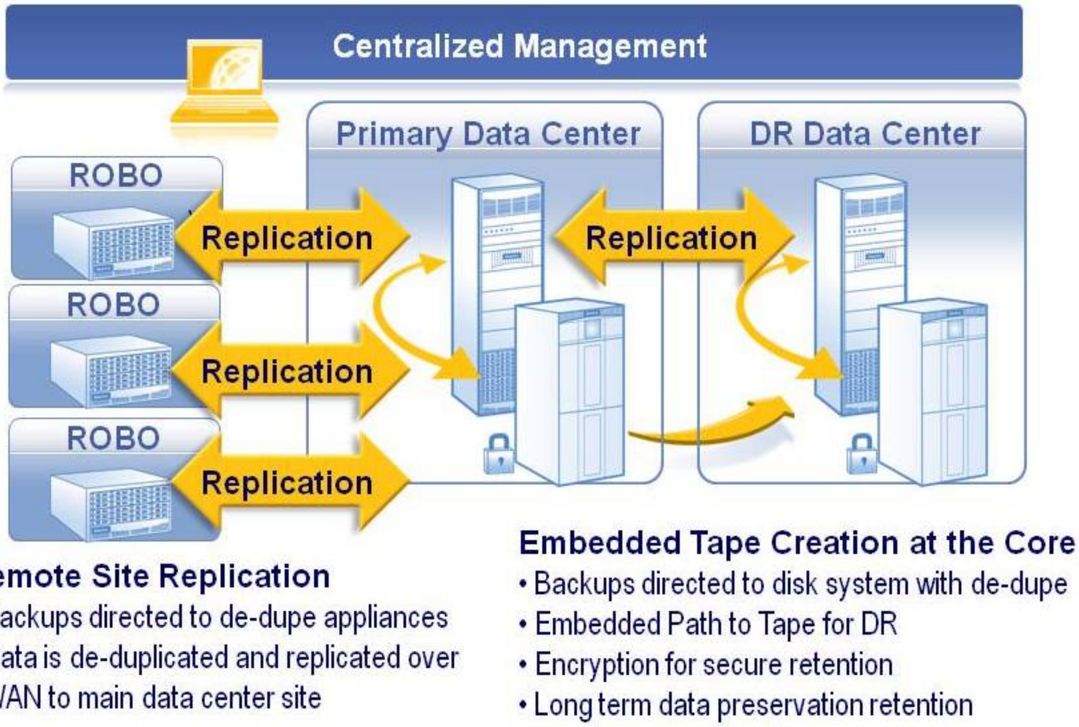
Value statement	Business and IT benefit
Simplified management and reduced data footprint	Reduces costs and enables more frequent and transparent back-ups or snapshots for BC/DR and faster fine grained file or data restoration
Storage and media optimization	Improved IT resource usage effectiveness, maximize investment or ROI, reduce TCO, meet QoS and SLA including RPO/RTO
Interoperability and technology co-existence	Simplify deployment, management and reduce or eliminate cost and complexity, maximize investment in people and resources
Support ROBO and workgroups	Leveraging remote data replication and de-dupe more data can be protected and preserved from ROBO and workgroups

Table 1: Summarizing value of data de-dupe as part of a data footprint impact reduction approach

[1] See "Demystifying Virtual Server Data Protection" white paper at www.storageio.com/reports

[2] See "The Changing and Evolving Role of Magnetic Tape" white paper at www.storageio.com/reports

Figure 3 shows how disk based de-dupe appliances or solutions combined with existing technologies can enable efficient and effective data protection for local and ROBO environments. Routine back-ups shift from tape to disk-based data protection solutions that include data de-duplication and replication. Replication of back-up data between sites provides first level data protection. The emphasis is on immediate back-up and file restore performance, improved RTO and RPO for speed of data restoration in addition to maximizing storage medium utilization.



Remote Site Replication
- Backups directed to de-dupe appliances
- Data is de-duplicated and replicated over WAN to main data center site

Embedded Tape Creation at the Core
- Backups directed to disk system with de-dupe
- Embedded Path to Tape for DR
- Encryption for secure retention
- Long term data preservation retention

Figure 3 Evolving data protection and retention paradigm leveraging multiple technologies

In the example shown, ROBO or work group sites are backed up locally to disk based de-dupe appliances leveraging existing back-up software and best practices. Data that is backed up locally to de-dupe enabled appliances can then be replicated to a primary data center using network resources more effectively. Data at the primary location can also be backed up to a local disk based de-dupe enabled appliance that also serves as a target for the ROBO or work group sites to replicate their back-up data to.

Also shown at the primary data center is a tape library where data is migrated to for long term data preservation or retention under centralized management. For BC and DR purposes, a secondary site is also shown where data can be replicated from the primary site using disk based de-dupe appliances as well as a remote tape device for archives or master "gold copies" of data.

From a data recovery standpoint, ROBO sites can rapidly restore individual files or entire volumes leveraging existing processes and tools. The added benefit for ROBO sites in addition to faster restores is flexibility to choose across a larger time frame of back-ups enabled by storing more data in a denser footprint.

Data De-dupe: It's not a question of if, rather where and when!

Summary

The questions around data de-dupe are shifting from if to when, where and how. De-dupe technology can be leveraged today to improve back-up, restore, and data protection operations with the resulting being streamlined routine day to day tasks including simplified or reduced media management handling. In addition, de-dupe technology helps to shrink data footprint impact enabling more data to be protected and preserved with less impact. The business benefit is thus to maximize IT environments investments in people, processes, hardware and software.

Look for systems that coexist with your existing best practices and technology investment. This means solutions that work with your existing back-up software and integrate easily into your day to day operations. Also look for solutions that provide tools to help you manage the data protection process across multiple sites and different tiers of technologies including different storage mediums. Avoid approaches that require high levels of integration that increase management complexity and cost both at ROBO as well as main data center locations.

Bottom line: Start thinking about how solutions will adapt to your requirements instead of how you will adapt your environment, processing, policies and procedures to meet a solution's capabilities. Learn more about data footprint reduction techniques and technologies along with other storage and data protection management related topics at www.storageio.com.

This Perspective was sponsored by Quantum Corporation. Quantum's DXi Series appliances comprise a complete family of high performance disk backup solutions with deduplication and replication capability designed to protect a full range of sites, from small and medium business to distributed Enterprises. DXi appliances support all leading backup software applications, including those using the Symantec OpenStorage API, and are designed to make deduplication easy and economical to add to users' existing backup environments. DXi systems operate effectively alongside existing tape backup devices, and Quantum offers an easy-to-use management interface that allows end users to manage all their disk and supported tape devices from a single console.

For more information about Quantum solutions, visit www.Quantum.com

About the author

Greg Schulz is founder of Server and StorageIO, an IT industry analyst consultancy firm, and author of the books *The Green and Virtual Data Center* (CRC) and *Resilient Storage Network* (Elsevier). Learn more at www.storageio.com or on twitter @storageio.

All trademarks are the property of their respective companies and owners. The Server and StorageIO (StorageIO) Group makes no expressed or implied warranties in this document relating to the use or operation of the products and techniques described herein. StorageIO in no event shall be liable for any indirect, inconsequential, special, incidental or other damages arising out of or associated with any aspect of this document, its use, reliance upon the information, recommendations, or inadvertent errors contained herein. Information, opinions and recommendations made by StorageIO are based upon public information believed to be accurate, reliable, and subject to change. This industry trends and perspective white paper is compliments of Quantum www.quantum.com