

Business Benefits of Policy Based Data De-Duplication

Data Footprint Reduction with Quality of Service (QoS) for Data Protection

By Greg Schulz

Founder and Senior Analyst, the StorageIO Group
Author “The Green and Virtual Data Center” (Auerbach)

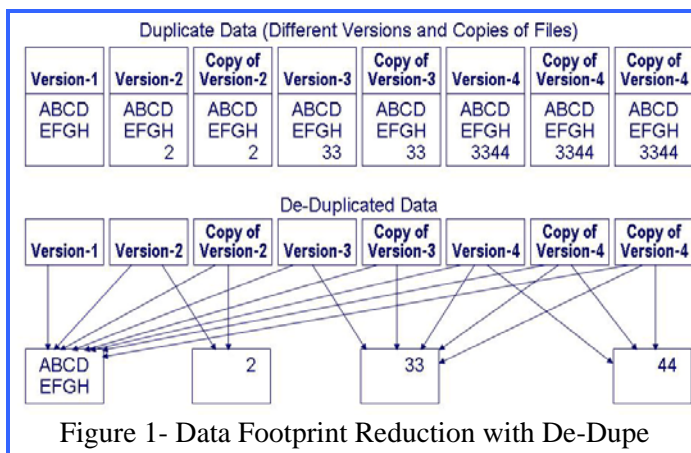


November 3rd, 2008

Introduction

Organizations of all sizes are faced with the demands of storing more data, including multiple copies of the same or similar data, for longer periods of time. The result is an expanding data footprint that results in increased IT resource management costs to support and sustain given levels of application and information Quality of Service (QoS) delivery.

An approach to providing relief from the pressures, costs and complexities associated with managing data protection for an expanding data footprint including reducing the amount of data transmitted over wide area network (WAN) links is data de-duplication (de-dupe).



Data de-dupe (Figure 1) reduces the impact of expanding data footprints¹ by optimizing storage and network bandwidth capacity used to store protected data, enabling improved utilization and more information to be processed in a given time frame.

Background and Issues

No two IT organizations are alike. All have different recovery point objectives (RPO) and recovery time objectives (RTO) to meet different application and business QoS, service level agreement (SLA) and cost objectives. For some organizations the emphasis will be on performance optimization to reduce or eliminate any negative impact on application availability as a result of data protection operations. For other organizations, or even different applications in a given organization, there will be a priority of optimizing storage space capacity instead of optimizing for performance.

Backup and data protection environments change and evolve over time to reflect different business requirements and service objectives. In order to keep up with the growing demands of more data to protect in less time using less physical storage capacity in a cost and energy efficient manner, technology evolution and innovations have been leveraged. In order to meet the diverse needs of different environments of various sizes and focus, new products and techniques need to be scalable in both performance and capacity, as well as resilient while being adaptable, to co-exist and preserve backwards compatibility for investment protection while leveraging new technologies to move forward.

Data Footprint Reduction Using Data De-Deduplication

The impact of an expanding data footprint extends further when a document or file gets backed up to some other medium, perhaps even multiple copies, some of which are sent to alternative sites as part of a business continuance (BC) or disaster recovery (DR) strategy. As new versions

¹ See "Business Benefits of Data Footprint Reduction" white paper at www.storageio.com/xreports.html

of documents and files are updated, circulated for review and revised, similar versions of the document are stored with similar and duplicated data. In addition to operating, capital and management costs, other aspects of the footprint to consider are the power, cooling, floor-space and environmental (PCFE) or green impact.

Data de-duplication eliminates the impact and overhead associated with protecting and managing expanding data footprints by preserving an image or snapshot of the original document or file while only saving the changes or differences unique to each version or copy. The result is that significantly less information or, specifically, duplicate copies of data are stored thus addressing cost and complexities associated with storing expanding data footprints.

Fundamentally, data footprint reduction and space optimization using de-dupe can be accomplished either at the source of where the data is being protected from (e.g. host or application server) or at the target destination where data is being backed up. The balance of this paper looks at target based data de-dupe.

A popular approach for performing de-dupe, given the relative ease of deployment and co-existence with existing installed server software and data protection and data management utilities, is target based data de-duplication. Target based data de-duplication has a relatively low barrier to entry and ease of deployment due to the ability to co-exist with currently installed backup and data protection software, procedures and policies on a local as well as remote or LAN basis.

As is the case with any data footprint reduction technology or strategy, there is a balancing act between optimizing performance and optimizing capacity. With data de-duplication, the balancing act exists between the quest to reduce data footprint by boosting storage capacity utilization and to meet QoS and SLA requirements and avoid performance gaps or bottlenecks².

These balancing acts and requirements lead to industry and vendor debates on the pros and cons of different approaches while often missing the key point of shifting the focus away from the architecture and looking at the QoS requirements for different application needs across different sized deployments.

Most industry discussions or debates have centered around in-line, immediate, or in-band de-dupe architectures and more recently around these versus scheduled, deferred or post-processing forms of de-dupe. Vendor driven de-dupe debates typically center on the architecture and capabilities of different solutions. These debates tend to force a decision to perform the data reduction in-line with a focus on algorithms and architecture and de-dupe data reduction ratios to optimize space or on post-processing performance with secondary focus on de-dupe ratios.

As a result, most de-dupe discussions driven by vendors involve detailed architectural comparisons and data reduction ratios as opposed to a focus on effective performance for different customer applications, needs and requirements. Policy based data de-dupe gives IT organizations the flexibility to select the applicable policy, for example immediate data footprint

² See "Data Center I/O and Performance Impacts" at www.storageio.com/xreports.html

reduction to optimize storage capacity or scheduled data footprint reduction to meet various QoS and SLA requirements.

In the quest to reduce data footprint impacts and optimize storage capacity, investment in existing backup infrastructure, including hardware and software as well as skill set and experience, can be compromised by introducing new technologies that are not compatible or able to co-exist with current policies or procedures or QoS and SLA requirements, This can result in a step backwards instead of going forward.

Consequently, when looking at space optimization solutions such as de-dupe, IT organizations need to assess and determine on an application by application basis what tradeoffs in QoS can be made to balance performance and capacity optimization. Put another way, can you afford to degrade backup performance for key applications in the effort to boost de-dupe ratios and storage capacity optimization at the risk of missed SLA objectives? Or would taking the time to align the most applicable de-dupe or space optimization approach to the specific QoS needs of different applications in your environment be the better approach?

De-Dupe Architecture Debates

As data de-dupe continues to evolve as a technology and feature function, it will become more common in storage solutions as opposed to current separate standalone first generation solutions. With this and the advent of adaptive and flexible policy based de-dupe with the ability to support both immediate de-dupe along with deferred de-dupe, the focus will shift from debates over one architecture versus another to that of where and what approach to use for which application.

Granted solutions that can only perform de-dupe using an immediate architecture or, solutions whose architectures can only support deferred de-dupe will continue to debate the merits of their approach. However in general, as has been the case with other technology features including WAN optimization, RAID levels, synchronous and asynchronous data replication among others, with technology maturity and evolution, conversations shift from how architected to when and where to deploy. For example, shifting conversations pertaining to de-dupe are occurring around which policy based mode or configuration to use when and where to meet different QoS and SLA objectives.

Different Data De-dupe Modes

Real-time or In-line de-dupe:

Aka: On-the-fly, in-band, immediate
De-dupe occurs as data is ingested
Optimizes storage capacity
Use when performance not top priority

Deferred or post-processing de-dupe:

Aka: Scheduled or time delayed
De-dupe occurs at a later point in time
Optimizes performance
Use when performance is top priority

Many discussions about data de-duplication focus on reduction ratios as opposed to de-duplication rates. While effective ratios need to be considered, so does the rate at which data can be de-duplicated, i.e. the de-duplication rate. The importance of this is the ability to gauge if data is being processed within available timeframes, such as backup windows. There is also continued growing awareness of how much effective data is reduced over a broader or larger capacity basis as well as increased focus on effective read and write or backup and restore performance.

The benefit of immediate de-duplication is that data is reduced on the fly, resulting in a smaller initial footprint on disk. The caveat with immediate is that the de-duplication process (e.g. ingestion rate) must keep up with the backup or data protection process in order to be able to ingest the data without causing a bottleneck and performance degradation to occur. The inverse of immediate data de-duplication is deferred where no performance degradation occurs and there is no negative impact to server based applications or data protection processes.

Deferred data de-duplication solutions should be able to ingest or receive data as fast as it is received; assuming no solution based bottlenecks such as a slow or non-intelligent disk storage system exists. For environments where performance and time sensitive processing are important, also look at de-duplication or data compression rates as an indicator of how much data can be processed in a given timeframe such as a constrained backup window. The trade off is some extra disk space to maintain performance for time sensitive applications or processing.

In-line vs. Immediate

Both in-line and immediate are real-time modes with the data de-duplication occurring while the data is being ingested. With EMC DL3D immediate mode, data de-dupe starts as soon as the first 256MB of data is ingested; the de-dupe process is continuous until the backup is completed. Look at the overall performance required for meeting backup windows when comparing in-line or immediate based de-dupe solutions.

The Changing De-dupe Landscape

In general, data de-dupe is a function and not a product although it may be presented as such. Data de-dupe, like RAID and other common storage technologies, is a feature that can be brought together with other storage features to enable IT organizations to select the modes and configuration options needed to meet specific QoS and availability requirements. A flexible data de-duplication solution in the form of policy based de-dupe, is now appearing as a selectable option that can be tailored to meet service requirements, similar to how different RAID levels or types of snapshots and replication or even types of disk drives can be used in intelligent storage systems to meet various needs and requirements.

Similar to aligning the appropriate WAN optimization, replication mode or RAID level to meet performance, availability, capacity and energy needs or requirements, policy-based de-dupe (Table 1) enables IT organizations to align and adapt the technology to their specific application QoS and SLA requirements.

	Immediate or In-line	Deferred or Scheduled	Policy-Based Adaptive
De-dupe occurs as data is ingested Optimizes storage capacity Use when performance not top priority	Primary Focus		Selectable
De-dupe occurs at a later point in time Optimizes performance Use when performance is top priority		Primary Focus	Selectable

Table 1 – Immediate vs. Deferred vs. Policy-based De-dupe

With policy-based de-dupe, IT organizations can shift their time and focus to learning how and where to deploy the most applicable mode or approach to meet different application QoS needs in

their environments. For example, with a policy-based de-dupe solution that enables both immediate and deferred (e.g. scheduled) modes of operations (Figure 2), IT organizations assign the applicable approach to specific application requirement.



Figure 2 - Selectable Policy and QoS Based Data De-Dupe

Benefits of Policy-Based De-duplication

- Flexibility to select different de-dupe modes to meet various application QoS needs
- Immediate mode for space savings combined with replication for BC/DR
- Deferred or scheduled mode to meet performance needs with deferred space savings
- Disable for data and applications where de-dupe does not provide a benefit
- Support more performance when needed to meet backup windows or move more data
- Enable effective BC and DR capabilities by moving more data with less resources
- Enable enterprise scalability with stability (performance, availability, capacity)

By selecting performance as a priority, de-dupe of data is deferred in order to meet time sensitive processing requirements such as copying or ingesting all data within a given backup or data protection window. Another policy option is to optimize storage capacity space as a primary objective when performance is not a priority by enabling immediate de-dupe of data. The result is that less space is initially needed to store de-duped data; however, this comes at the expense of performance, for example extending a backup window during data ingestion.

EMC Policy Based De-duplication

EMC with the DL3D, offers flexibility in implementing target-based de-deuplication. Rather than forcing IT customers into one model or another, for example, selecting a target-based solution based on if it is in-line or post-processing, EMC DL3D provides flexibility. The flexibility of DL3D policy based de-dupe provides the ability to tailor the solution to particular needs and adjust to changing workloads, service level objects and data protection requirements.

Benefits of the EMC DL3D Policy-Based De-dupe include:

- Provides approximately the same amount of performance as other in-line architectures
- Scheduled or deferred mode maximizes performance for time sensitive backup operations
- Provides best of both worlds – immediate and deferred deduplication
- Adaptability and flexibility to co-exist with existing technologies for investment protection
- Scalable in terms of performance, availability and capacity to meet diverse needs
- Policy based de-dupe is included in the purchase price of the EMC DL3D eliminating the need to decide between immediate or deferred mode based architectures

Summary

It is important to keep performance and subsequent application impacts in perspective when looking at data footprint reduction solutions, particularly when looking at heavy thinking approaches, including de-dupe, that add latency to time sensitive processing such as backup and restore.

Look beyond data de-dupe ratios. Instead, consider the overall effective amount of data footprint capacity that is reduced - for example, how applicable and achievable are high ratios for your environment, applications and data as well as data protection processes. Most environments will need a mix and balance of performance and capacity optimized capabilities to meet different applications needs that are addressed by policy-based de-dupe.

Bottom line, get solution providers to start talking about how their solution will adapt to your requirements instead of how you will adapt your environment, processing, polices and procedures to meet their solution capabilities. If solution providers want to continue the de-dupe debate, then make sure that you include policy-based de-dupe in the discussions.

About the author

Greg Schulz is founder of the StorageIO Group, an IT industry analyst and consultancy firm as well as author of the books *Resilient Storage Network* (Elsevier) and *The Green and Virtual Data Center* (Auerbach) at www.thegreenandvirtualdatacenter.com and www.storageio.com .

All trademarks are the property of their respective companies and owners. The StorageIO Group makes no expressed or implied warranties in this document relating to the use or operation of the products and techniques described herein. The StorageIO Group 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 the StorageIO Group are based upon public information believed to be accurate, reliable, and subject to change.