Quantum®

WHITE PAPER

Backup and Recovery: The Benefits of Multiple De-duplication Policies

CONTENTS

Introduction

Data de-duplication is changing the way that IT departments protect their data. It allows users to expand their use of disk for backup and introduces the idea of replicating backup sets over a conventional WAN as a practical part of a disaster recovery program. There are several different approaches to data de-duplication—some technologies compare specific files to find repeated data between versions, others use a block-based approach that expands the data comparison to a more global scale. When it comes to effectiveness, several of the approaches seem to be able to reduce disk capacity and network bandwidth requirements by 90% or more - powerful savings that allow users to retain more data on disk, improve their restore performance, leverage networks for replication, reclaim floor space in the datacenter, and reduce power and cooling requirements.

There are also two different approaches to when de-duplication activity is carried out, and that choice can have a significant impact on how effectively de-duplication works with specific data protection tasks. The basic issue is whether data is de-duplicated 1) during the backup process (during ingest) or 2) after the backup (in a post process). Each methodology has advantages, but these have been largely obscured because until recently each vendor only offered one choice. That meant that the discussions, often heated, focused on which method was "the best" and why the other methodology was less effective. The reality is that there is no single best approach—different approaches have different results and the right way to think about the problem is to see which specific kinds of backup jobs are best matched to the different methodologies.

This white paper explains the differences between the de-duplication approaches, discusses the strengths and limitations of each, shows how different backup tasks are likely to benefit from them, and talks about the business benefits. There's an important caveat: there is no way to provide blanket guidance that will apply to all similar backup tasks. This paper can talk about general principles, discuss typical jobs, and show likely results, but it is important for you to talk to your backup system supplier and reseller to see what is likely to make most sense for your specific situation. What is close to being certain, however, is that any even moderately complex set of backup jobs will be likely to find different situations where different approaches make most sense.

It is important for the reader to know that some of Quantum's disk-based de-duplication products, the ones designed for larger and more complex backup environments, give users a choice of different de-duplication methods. While we believe that a choice between different de-duplication policies is likely to benefit most users, we have tried to describe the approaches as objectively as possible to allow the reader to decide whether or not a single approach might make more sense in their specific case.

The Basic Issue: Optimizing for Capacity or Performance

Data de-duplication systems differ in how they carry out the capacity optimization process, but all of them perform some kind of comparison of data segments to identify the ones that are repeated and that can be replaced with a reference or pointer. One method, file-by-file comparison systems, looks at two versions of the same file or fileset and looks for unique data. A second method, employed in block-based systems, divides data into segments and uses a mechanism for remembering which blocks in the data set of interest are unique and which have already been written to disk. In each case, the ultimate result of the process is that only unique data segments are stored while repeated ones are referenced instead of being stored again. When there are several versions of similar data

sets to be retained and accessed, either of these technologies can provide very powerful savings. Both systems can also support highly optimized replication in which only the unique segments need to be transmitted—again, assuming multiple similar data sets where similarity and uniqueness of elements can be identified.

The ability to locate and store only unique data segments is common to all de-duplication approaches, and an inevitable by-product of the process is some level of system overhead—it takes more processor cycles and more time to find the redundant segments and compare them to what has already been recorded than it would take to simply write all the data, redundancy and all, directly to the disk medium. De-duplication vendors have all worked hard to make this process occur as rapidly as possible, but when you compare any de-duplication system with simply writing data straight to disk, the de-duplication process always introduces some level of overhead which can introduce latency. The question that anyone using de-duplication has to address is "when do I want that overhead to take place?"

De-duplicating During Ingest Minimizes Capacity Requirements

One choice is to carry out de-duplication during ingest—while the backup is going and before data is written to the de-duplication storage location. This method has the advantage of using the least amount of disk capacity, and it can allow replication of unique segments to begin while the backup is going on—which means replication can finish faster. But de-duplication during ingest has the disadvantage of linking de-duplication rates to backup speeds. In some cases, where vendors have a strict in-line methodology and the backup speed is faster than the de-duplication rate, it can slow the backup down. This is normally a minor problem in small or mid-range sites where backup server speeds are usually limited, but when there is a high volume of new data to be processed, or where there are backup bursts, or very fast servers, this methodology may cause the backup to take longer to complete—lengthening the backup window. Adaptive buffering systems like the one that Quantum builds into its DXi-Series will minimize this effect but not eliminate it.

De-duplicating as a Post Process Maximizes Performance

The other method, also widely used on commercial systems, is to bring all the data onto the disk system first, and, after the backup ingest is complete, de-duplicate the data as a post data transfer process. This has the advantage of letting the backup window be as short as possible—there is no de-duplication overhead to slow down the backup. That means the application servers are back in service as rapidly as possible. But this method has downsides as well. 1) More disk is required since enough space has to be reserved to hold an entire backup job—some file-by-file approaches even require that room be saved for two full backup sets. 2) The replication of unique data is delayed until the de-duplication starts—so it will take longer. The other general issue to watch for with any post-process approach is the total, end-to-end time. It is very possible that the combination of ingest and de-duplication in a post process may take longer than ingesting and de-duplicating at the same time. This effect means that a post process system is likely to be preferred when the length of the initial backup window is the most important issue.

Remember to Ask Whether You Need De-duplication for All Your Data

Even though our discussion is focused on which kind of de-duplication, it also makes sense to look at backup jobs and think about whether they really benefit from the technology at all. Not all backup jobs do. Conventional VTL or NAS disk backup will make better sense for some data and we will discuss that option when we look at different typical backup tasks.

The Quantum Approach: Multiple De-duplication Policies, Multiple Benefits

Quantum's DXi-Series is a family of disk backup systems that feature de-duplication and replication, and provide a choice of different kinds of data de-duplication methodologies. The DXi3500 and DXi5500 models, which are designed for smaller environments, always de-duplicate during ingest. The DXi7500, designed for larger midrange and Enterprise environments, provides an intuitive GUI interface that allows users to schedule different de-duplication policies on a share-by-share basis to provide optimal results for a wide range of backup situations.

De-duplication During Ingest: Adaptive De-Duplication

Quantum's Adaptive policy, offered on all DXi-Series models, de-duplicates data as the data segments arrive at the system during the ingest process. This means that it minimizes the amount of disk used and that replication of unique data segments begins while the backup is still going on. The key difference between Quantum's adaptive system and conventional in-line systems is the use of an adaptive buffer. The DXi-Series approach begins by writing a small segment of data to a disk buffer and then de-duplicates it immediately. If backup ingest rates increase (during bursty operations, for example) or if de-duplication slows down (because a large volume of new data has to be analyzed, for example), the buffer helps keep transfer performance high. In the DXi7500, which supports usable capacities of up to 180TB, the buffer contents can also be retained as a cache to help accelerate restore operations as long as disk space is available.

The adaptive de-duplication approach makes most sense for:

- Backup jobs of any size where there are short to medium backup windows
- When immediate replication is required
- Where disk use needs to be minimized
- Where jobs have a high percentage of redundant data
- Where there are likely to be mixed or variable backup speeds

A few of the applications that meet these criteria are:

Email
In many companies, users have transformed their inbox into the "hold everything" data manager, so email has become a lynchpin. Backup at the mailbox level is vital, replicating the data rapidly may be important, there is often a very high rate of redundancy between mailboxes, and backup speeds are often relatively slow. It is usually a very good candidate for adaptive de-duplication.

User Shares/Unstructured data
This category includes individual user files (documents, spreadsheets, presentations, etc.) and shared assets (collaborative systems) both of which are critical to organizational operations. The loss of even one file can have a domino effect, halting work for any number of people, teams, or divisions so backup and prompt DR protection are important. The performance characteristics of these shares (medium performance) and need for replication make them likely candidates for the adaptive approach.

Virtual Servers
Virtual Machines, a salvation for datacenter overcrowding and wasted computing resources, have introduced a new challenge – managing their system backups. Backup that spans virtual and physical boundaries are especially good candidates for de-duplication and replication since the system files are highly redundant. Multiple jobs for different hosts are likely to create variable backup speeds—which, along with the replication requirements—makes the adaptive approach a likely choice.

De-duplication as a Post Process: Deferred De-duplication

Quantum's Deferred de-duplication approach, available on the DXi7500, lets users decide to de-duplicate in a post process. Data is first written to the DXi-Series disk in native, non-de-duplicated format, and then de-duplication is carried out at a later time, normally after the backup job has finished. Replication of unique data begins at the same time as the de-duplication operation. The Quantum deferred policy differs from some other post process de-duplication systems in three ways. 1) It requires landing space for only one backup job (some other systems require space for two). 2) It implements the post process by allowing the user to define when the de-duplication will start—so users can decide to begin de-duplication before a backup is completely finished if they wish. And 3) the same share can switch back and forth between the adaptive and deferred processing. A user might elect to use the deferred mode on a backup which contains a large amount of new data, but might decide that backups with small amounts of changed data would take place in the adaptive mode. A positive feature of the deferred process is that the native format data is always retained in cache to accelerate reads and tape creation. In the DXi-Series, the cache is maintained until the disk space is needed for other operations.

Deferred de-duplication makes most sense when:

- The Shortest possible backup window is required for de-duplicated data
- Deferred replication is acceptable
- Enough disk is available to provide a landing area for all of the backup
- Immediate reads from cache are desired

Applications that meet these criteria are:

OLTP database backup
This is often one of the most important and most challenging issues for business continuity. Backup and recovery can be very complicated and there is often a need to complete the backup and return the primary application to normal service as rapidly as possible. Deferred de-duplication provides very high-speed backup of data, shorter backup windows, and high performance from cache for most recent restores and for tape creation.

Any backup with large amounts of new data
The deferred policy can also be applied to any one of the data types already using the adaptive policy when jobs have an unusually high de-duplication workload. The situation may be an initial seeding of an exceptionally large backup or the periodic backup (monthly, quarterly, etc.) where the backup window limitations are more important than reserving additional workspace. Using deferred de-duplication will keep the initial backup window shorter, and the DXi-Series system allows users to switch to the adaptive mode when the amount of redundant data is higher.

Native VTL/NAS

It is important to mention that for some kinds of jobs, de-duplication does not provide an additional benefit and is likely to not be applied. These include data that has been pre-compressed, encrypted, or consisting of randomized data in which segment patterns do not recur. It also includes data that will be stored one time and that will not be retained.

Native VTL or NAS makes most sense when:

- Data does not de-duplicate well
- Data does not need to be retained on disk
- For non-backup archive applications

A few of the applications that meet these criteria are:

Archive copies of image files
Specialized image files, such as satellite images, have the unique characteristic of being almost impervious to de-duplication or additional compression. They are usually very large files and storing them on primary storage for long periods can be very resource intensive and expensive. Note, however, that if multiple copies of the files are being backed up, de-duplication may provide very high value. It makes sense to talk with your vendor about other users' experience with similar filesets.

Database log files
Files containing a description of changes made to a database allow administrators to take a previous backup and recreate the changes necessary to restore a more current state, so they are just as valuable as the databases they're associated with. The logs lose their value very rapidly so multiple versions of them are typically not retained. The DXi7500's native VTL mode gives them fast backup and restore on the same system where the database tables are backed up.

Summary

Data de-duplication systems, which in the past have required that users apply one de-duplication methodology to all of their jobs, now provide a choice. For larger backup environments where multiple jobs, different data types, and variable retention periods are the norm, users can deploy a single de-duplication system with flexible methodologies. Quantum's DXi7500 allows users to choose on a share-by-share basis to carry out de-duplication during ingest, in a fully deferred post process, or to operate in native VTL/NAS mode, allowing them to choose between optimizing capacity savings and optimizing length of the backup window. With a more flexible approach to backup and recovery using multiple policies, companies can enjoy new business benefits from data de-duplication:

- Performance that meets the most demanding backup windows and doesn't throttle the backup process
- The fastest possible ingest and restore rates
- Seamless integration with existing D2D2T architectures
- The option of more using frequent full backups that use less overall disk capacity
- Greater capacity and performance to handle larger datasets and more diverse data types

This policy-based de-duplication lets IT departments use a single system to properly match backup and recovery techniques to all of their data types and service level agreements, and its combination of adaptive buffering and caching techniques keep backup, restore, and tape creation performance high in all operating modes.

The Quantum Advantage

Quantum understands that different business segments require different business solutions. We are in the business of delivering superior backup, recovery and archive solutions that meet your specific needs, whether you are a start-up or Fortune 100 corporation.

Quantum's comprehensive data protection solutions include:

DXi-Series – Disk Backup Solutions with Data De-duplication and Replication
DXi-Series disk-based backup systems extend the benefits of data de-duplication across the Enterprise, integrating with tape, replication, and encryption forming a complete backup solution. Quantum's patented data de-duplication reduces disk and network bandwidth requirements by 90% or more, lowering disk backup costs and making WAN replication a practical DR tool. The DXi7500 offers policy-based de-duplication giving you the ability to choose the ideal method to optimize your data de-duplication, and integrated tape creation for a single solution that combines short term and long term data protection.

Scalar Series – Intelligent Tape Libraries
From entry-level to enterprise environments, Quantum's Scalar® tape libraries provide solutions that meet customers' most critical backup, recovery and archive demands. Scalar libraries are designed with Quantum's unique iPlatform architecture to provide rich features such as self-diagnosis, proactive customer alerting, drive and media integrity reporting, and monitoring and management. Scalar libraries also integrate functions such as partitioning, mixed-media, scalability and capacity-on-demand. Each system comes equipped with standards-based encryption to address growing security concerns and regulatory compliance.

StorNext – Data Management Software
Quantum StorNext® data management software enables customers to generate revenue faster and store more data at a lower cost. Combining high-speed data sharing with cost effective content retention, StorNext helps customers build an infrastructure for consolidating resources to ensure workflow operations run faster and enabling you to maximize storage resource utilization.

Quantum Vision – Consolidated System Management Tool
Quantum Vision™ is the consolidated management tool that offers "single pane of glass" administration for all Quantum disk and tape automation systems. Quantum Vision™ enables you to monitor and manage your entire Quantum data protection environment from one convenient window – giving you complete visibility into all the layers of the data protection infrastructure. It puts an end to juggling multiple management tools and interfaces when troubleshooting, generating reports, optimizing, and fine tuning your data protection environment for simplified systems configuration, monitoring, and reporting.

Quantum Global Services

Providing pre-sale and post-sale professional services, including backup architecture planning and assessment services, Quantum's Global Services division helps customer identify their critical requirements for a comprehensive data protection strategy and implement the right solutions. Quantum Global Services delivers customer support worldwide with Service Professionals in 80 countries and 300 stocking locations.

Talk to the Experts at Quantum

To learn more about this solution or any of Quantum's backup systems there is more information available online at http://www.quantum.com/solutions or call 1.800.677.6268. For background information about Quantum's patented de-duplication capability, please refer to the white paper "Data De-duplication Background: A Technical White Paper".

Quantum®

For contact and product information, visit quantum.com or call 800-677-6268

WP00125 Sep 2008