Quantum

WHITE PAPER

Effectiveness of Variable-block vs Fixed-block Deduplication on Data Reduction: A Technical Analysis

CONTENTS

Executive Summary 3
Fixed vs. Variable-block Deduplication 3
Test Configuration 4
Data 5
Methodology 5
Observations, Results, and Commentary 5
Unstructured Data 6
Semi-structured Data 6
Structured Data 7
Retention Matters 8
Summary 8

EXECUTIVE SUMMARY

The Problem: Deduplication has become a 'checkbox' technology for purchasers, so every vendor of data storage products has incorporated something they can call "deduplication." But deduplication is a technique, not a discrete technology. There are a number of ways to deduplicate data, and individual implementations vary greatly in their effectiveness. The value of deduplication is in reducing costs; and the more effective the data reduction, the lower your TCO will be. It doesn't help that some vendors are overly creative with their measurement techniques in their attempt to generate the biggest datasheet numbers, and industry analysts shy away from direct head-to-head product comparisons. The avalanche of vendor hype and FUD and the lack of real information make it very difficult for IT buyers to find the best deduplication solution for their needs and budget.

The Objective: While data reduction effectiveness isn't the only criteria to use when evaluating products that leverage deduplication, it is the single biggest factor affecting TCO. It directly affects storage capacity, replication network bandwidth, power, cooling, floor space requirements, and even whether DR SLAs are met or exceeded. Deduplication approaches may be divided into two broad categories: fixed-block and variable-block. The goal of this document is to demonstrate the fundamental differences between these two approaches by performing simple lab tests with real data. Out of necessity, specific hardware and software products were used in these tests, but the results are generally applicable to any product that uses fixed- or variable-block deduplication. The body of this paper will provide additional background, define terms, describe methods, and share and interpret the results of the tests.

The Bottom Line: The results will show that variable-block deduplication provides greater data reduction than fixed-block deduplication. Variable-block deduplication is two to three times more efficient, at minimum. With longer retention times and certain types of data, the difference will be far greater. 3X may not sound very large, but it means that a fixed-block solution will require three times the storage, three times the network bandwidth for replication, and more power, cooling, and rack space compared to a variable-block solution. It could even be the difference between meeting the business SLA for DR protection, or not.

FIXED VS. VARIABLE-BLOCK DEDUPLICATION

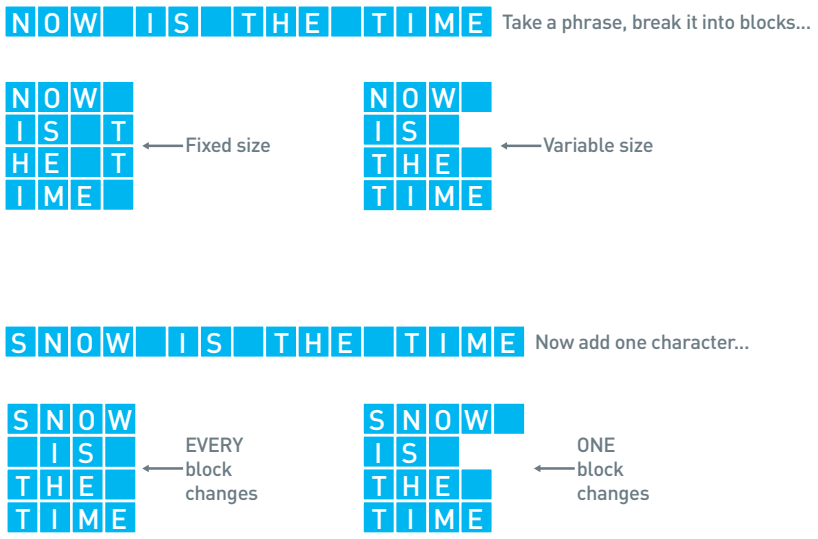
Fixed-block deduplication breaks the incoming data stream into pieces or 'blocks' that are all the same size. The blocks are compared, and only new unique blocks are stored on disk after being compressed. Duplicates are discarded. A system of pointers is used to map the ingested data to the pool of unique blocks. Some approaches let the administrator choose the block size and refer to this as 'variable'. But once a block size is selected, that size is used for all data—the ability to change the block size manually does not qualify a solution as 'variable-block'.

With variable-block deduplication, the block size is not fixed. Instead, the algorithm divides the data into blocks of varying sizes based on "natural boundaries" in the data. The block size is varied automatically, in real time, in response to the incoming data. New unique blocks are compressed and stored on disk, and pointers are used to map the ingested data to the unique blocks.

What happens when the data changes? With fixed-block methods, unless the changes made to the file are exactly a multiple of the fixed-block size, all of the data past the first change is shifted. This shift changes subsequent blocks in the file with respect to the fixed-block boundaries, so to the fixed-block algorithm they look 'new'.

Variable-block algorithms divide data into blocks based on the characteristics of the data itself, not an arbitrary block size. This makes them flexible when data changes. Only the new or changed data is stored and the uniqueness of the remainder of the file is not impacted.

Figure 1. Fixed vs. Variable Deduplication Example



TEST CONFIGURATION

Deduplication capability of three products was compared:

- Quantum's DXi6900 appliance, utilizing Quantum's patented variable-block deduplication
- A Symantec NetBackup 5200 appliance with fixed-block deduplication
- CommVault Simpana 10, also using fixed-block deduplication

The same hardware, software, and data was used for all tests. Source data was hosted on the data mover server directly, and backups were performed to a DXi® system or the NetBackup 5200 appliance. For the Simpana deduplication test, the data was sent to an NFS disk share. The NFS share was configured on a DXi, but with the DXi's deduplication and compression disabled. Any disk could have been used as the target, but by using the DXi, the DXi's standard Advanced Reporting capability could be leveraged to examine the results. For the NBU 5200, statistics were recorded from the device's interface. For the DXi and NetBackup 5200 tests, Symantec NetBackup 7.6 was used as the data mover.

DATA

It's easy to make any deduplication system look fabulous if you feed it artificially generated, highly redundant data, but it's not a real-world test. Unlike most vendors, Quantum uses a corpus of real data (supplied by Quantum's IT department) for DXi testing. This corpus consists of a chronological sequence of images of three data sets. Unstructured (user home directories), semi-structured (e-mail), and structured (database) data types present different challenges to a deduplication system. Testing with all three provides a more complete picture of how a system will perform in a typical corporate IT backup environment. Deduplication results vary depending on the data used, but only up to a point. The differences we identify between fixed- and variable-block deduplication are generally applicable.

METHODOLOGY

For each of the three data types, the following procedure was followed for each deduplication system in turn:

- Back up the oldest version of the data, record the amount of unique data stored to disk
- Back up the next oldest version of the data, record the amount of new unique data stored to disk
- Continue with each newer version of the data in sequence and graph the results

This methodology provides an accurate simulation of the original sequence of backups for each data type. Changes are captured in order at a series of points in time. Each deduplication system was exposed to the same data in the same order.

There are a number of ways to measure deduplication—for example, as a ratio or a percentage of disk savings—but they all boil down to the same thing: the amount of disk and bandwidth required. The more effective the deduplication, the less disk capacity is needed for storage, with a corresponding reduction in WAN bandwidth for DR replication. Keep in mind when examining the graphed results that lower is better, as it means less disk capacity was consumed.

OBSERVATIONS, RESULTS, AND COMMENTARY

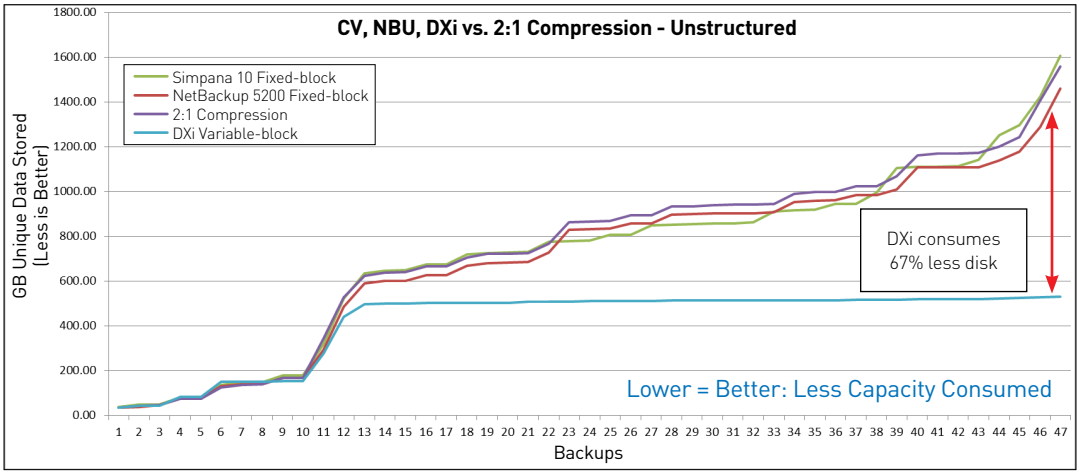
Deduplication, Compression, and Total Data Reduction: When a deduplication system identifies a unique block of data, that block is compressed before being stored to disk. Thus the total data reduction reported is the product of both deduplication and compression. Reported "10:1 reduction" could be 5:1 deduplication combined with 2:1 compression.

Compression techniques used in data storage commonly provide ~2:1 compression, more or less depending on the data. The graphs below include a hypothetical "2:1 Compression" line. This line represents how much data would be stored if the source data were simply compressed 2:1, as might be achieved with LTO tape. The position of the 2:1 reference line provides an indication of the overall effectiveness of the deduplication systems vs. simple compression. The shape of the lines can provide a clue to how much of the total reduction for each system is provided by compression vs. deduplication. If the shape of the line for a system closely tracks the shape of the 2:1 line, it indicates a significant proportion of the reduction benefit is from simple compression, not deduplication. Of the three systems tested, only the DXi provides reporting that explicitly displays how much reduction is due to compression vs. deduplication.

UNSTRUCTURED DATA

In the typical corporate IT environment, a large proportion—often the majority—of the data is unstructured. According to industry data, unstructured data is by far the fastest growing data type in most organizations.

Figure 2: Unstructured Data, 48 Sequential Backups

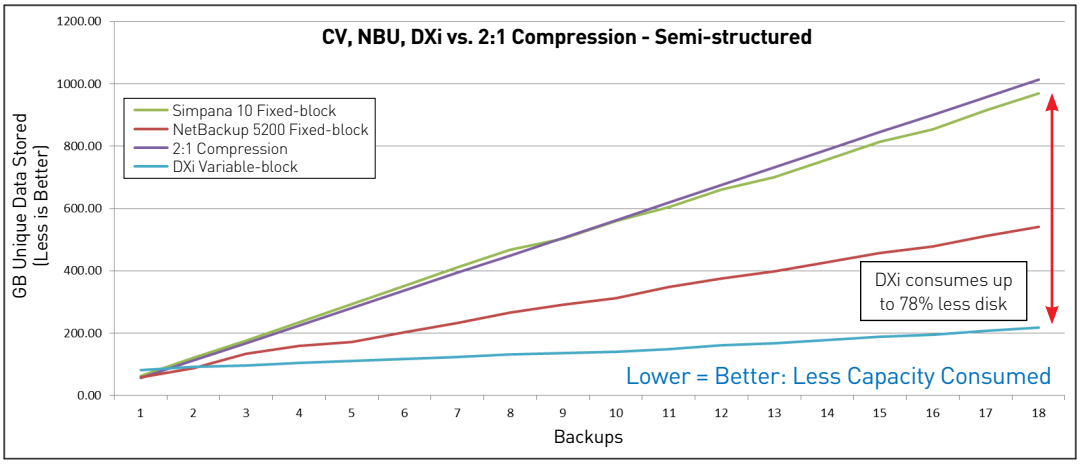


- **Key Finding:** DXi variable-block deduplication finds significantly more redundancy in the data vs. the fixed-block systems.
  - *Simpana consumes 3x* the capacity required by DXi
  - *NetBackup 5200 consumes 2.75x* the capacity required by DXi
- **Key Finding:** Simpana and NetBackup lines closely track the shape of the 2:1 line, indicating that most of the data reduction leverage with these fixed-block products is due to compression, not deduplication. The DXi variable-block deduplication is clearly operating in a fundamentally different way, as the shape of the DXi line does not closely track the 2:1 line.

SEMI-STRUCTURED DATA

E-mail is the classic example of semi-structured data, and every organization relies on e-mail and other forms of messaging. Unless messaging services are farmed out to a third party, IT must protect and retain messaging data to serve a variety of backup, disaster recovery, and compliance needs. Messaging grows constantly, but specialized archive solutions are often deployed that minimize the fraction of messaging data that must be protected.

Figure 3: Semi-structured Data, 18 Sequential Backups - see Retention Matters section below for a 60-day retention extrapolation.



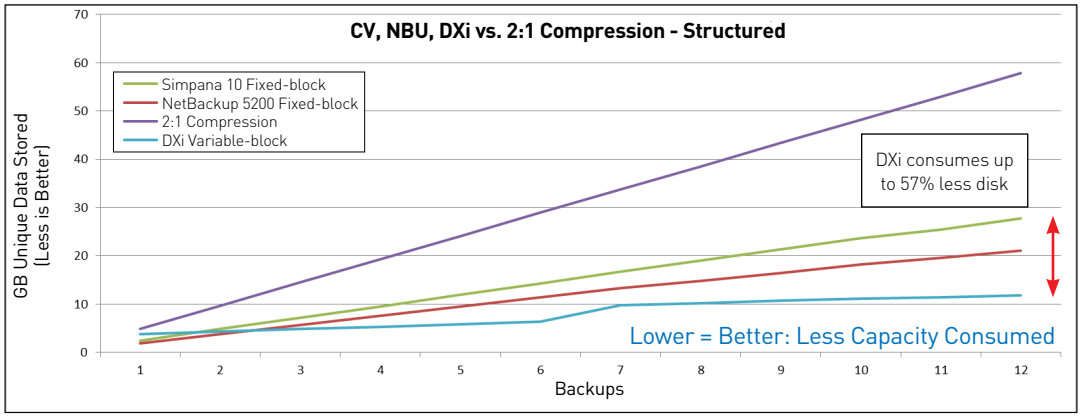
- **Key Finding:** DXi variable-block deduplication finds significantly more redundancy in the data vs. the fixed-block systems.
  - *Simpana consumes 4.5x* the capacity required by DXi
  - *NetBackup 5200 consumes 2.5x* the capacity required by DXi
- **Key Finding:** The data type matters. Unlike the previous example, there is a significant difference in the ability of the two fixed-block products to reduce this data. Simpana struggles to beat 2:1 compression; the NetBackup 5200 is much more effective. Neither is as efficient as DXi's variable-block approach.

STRUCTURED DATA

Structured data is just another term for traditional databases such as Oracle, MS SQL Server, and DB2. In most organizations, the quantity of structured data is dwarfed by the mountain of unstructured data. Whether large or small, however, the structured data usually grows much more slowly than unstructured data.

There are many options for protecting databases, and normally a combination of approaches is leveraged to meet the recovery time objective (RTO) and recovery point objective (RPO) demanded by the business. In many cases, deduplication of structured data is difficult and does not produce dramatic reduction results.

Figure 4: Structured Data, 12 Sequential Backups

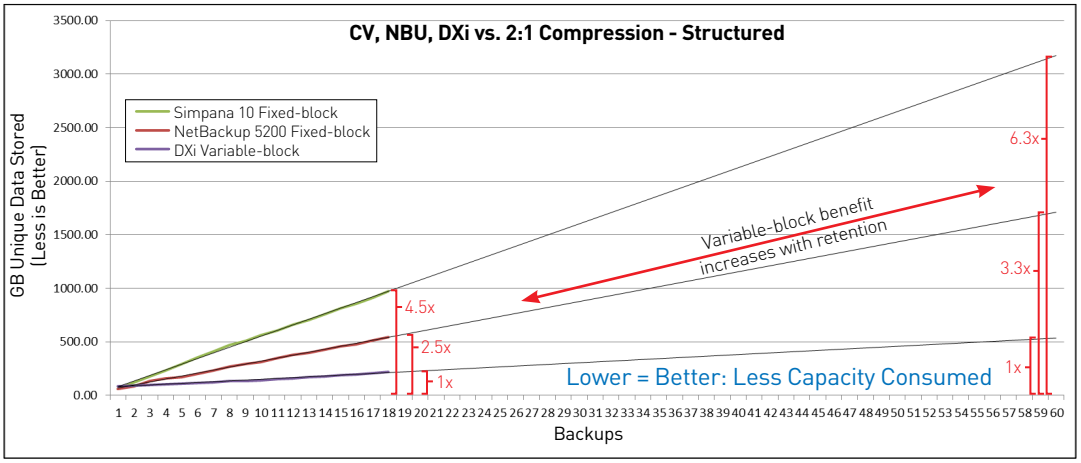


- **Key Finding:** All products tested performed better than the 2:1 compression reference, but DXi variable-block deduplication is again the most efficient.
  - *Simpana consumes 2.3x* the capacity required by DXi
  - *NetBackup 5200 consumes 1.8x* the capacity required by DXi

RETENTION MATTERS

It's critical to understand that the differences in deduplication leverage highlighted above are not fixed ratios—their impact is compounded over time as the curves diverge. A key factor to consider is the retention requirement of the backup data. More efficient deduplication provides greater benefit as more backups are retained. The lab testing was limited by the number of backup images in the corpus, but the effect of greater retention is illustrated below by using trend lines.

Figure 5: Semi-structured Data, With Trend Lines



- **Key Finding:** The longer backups are retained, the greater the benefit.
  - *Simpana consumes 4.5x* the capacity required by DXi after 18 backups
    - *This difference grows to 6.3x* if 60 backups are retained
  - *NetBackup 5200 consumes 2.5x* the capacity required by DXi after 18 backups
    - *This difference grows to 3.3x* if 60 backups are retained

SUMMARY

Deduplication enables new options for data protection, and makes previously expensive options viable. The benefits and TCO scale directly with the effectiveness of the deduplication approach chosen.

Fixed-block vs. Variable-block: Side-by-side testing with real corporate data shows that variable-block deduplication reduces data by 2x to >6x more effectively than fixed-block approaches.

Compression vs. Deduplication: Deduplication systems use two methods of data reduction: compression and deduplication. In some cases, fixed-block deduplication systems add little to no value beyond what could be achieved with traditional compression alone.

Infrastructure Impact: Storage efficiency is not just about storage. Data reduction maps directly to the amount of WAN bandwidth required for replication; a system that reduces data 2x better will require half the bandwidth. WAN connections are expensive—it is vital to meet DR SLAs with the smallest links possible. Better data reduction also equals lower power and cooling expense and fewer rack U consumed. Over time, infrastructure costs will dwarf the acquisition cost of a solution, so it pays to pay attention to them. When considering software-based approaches, remember they require significant server resources to implement; make certain the benefit justifies the cost.

Retention: The efficiency benefit of variable-block deduplication is not a simple multiple. It compounds over time as additional backups are retained. The greater the retention requirement for the data, the greater the benefits realized by using a system that employs a variable-block approach to deduplication. When evaluating products, you must test with enough cycles of data to understand the impact at your maximum on-disk retention period.

Don't believe it? Quantum approached this testing with as much neutrality as possible. If you don't believe the results, let us show you the difference in your environment. For qualified prospects we support proof-of-concept testing to prove the strength of our variable-block deduplication.

Why Choose Quantum DXi?

Inline, Variable-block Deduplication: All data is deduplicated and compressed prior to being stored on disk.

Simplified Portfolio: 3 scalable models provide 1TB-510TB of customer-usable capacity.

Scalable Architecture: DXi provides seamless capacity growth and performance scaling

- Pay-as-you-Grow – Capacity-on-demand enables capacity expansion with a license key
- Capacity & Density – 4TB disk drives
- Self-Encrypting Disks (SEDs) encrypt data-at-rest with no performance penalty

Quantum StorNext® 5: Fast, flexible file system enables high performance

- DXi uses dedicated disk pools for storing metadata and deduplicated content on disk tuned to that specific data type. This allows for full performance across all capacity points in a given model; no other appliance vendor has adopted this architecture

Increased IT staff Productivity

- User Interface – Home page with all operational information; installation wizards ensure easy and efficient setup and deployment
- System Scheduler –Primary system operations (replication, reclamation, CLI execution) can be scheduled to allocate DXi resources to priority tasks
- Advanced Reporting – DXi system retains 6-year history of all activity to enable optimized resource utilization and proactive system planning for future growth
- Vision – Consolidated Management and Analytics for all Quantum products
- Includes policy-based warnings for low-capacity situations

Additional Background and Reference Material

Data Deduplication Background: A Technical White Paper – White Paper containing a more general discussion of data deduplication [WP00126A]

ESG Lab Validation of Quantum's DXi6900 Deduplication System – Independent lab validation report from ESG of Quantum's DXi6900 [WP00201A]

ESG Lab Validation VMware vSphere Data Protection Feb 2013 – Independent report from ESG highlighting the fact that variable-block deduplication is superior to fixed-block approaches for VM backups

The Growth and Management of Unstructured Data

Your Unstructured Data is Sexy – You Just Don't Know It

Structured vs. Unstructured Data

ABOUT QUANTUM

Quantum is a leading expert in scale-out storage, archive and data protection, providing solutions for sharing, preserving and accessing digital assets over the entire data lifecycle. From small businesses to major enterprises, more than 100,000 customers have trusted Quantum to address their most demanding data workflow challenges. With Quantum, customers can Be Certain™ they have the end-to-end storage foundation to maximize the value of their data by making it accessible whenever and wherever needed, retaining it indefinitely and reducing total cost and complexity. See how at **www.quantum.com/customerstories.**

www.quantum.com • 800-677-6268

Quantum

WP00200A-v02 Apr 2016