HP StoreOnce D2D

Understanding the challenges associated with NetApp's deduplication

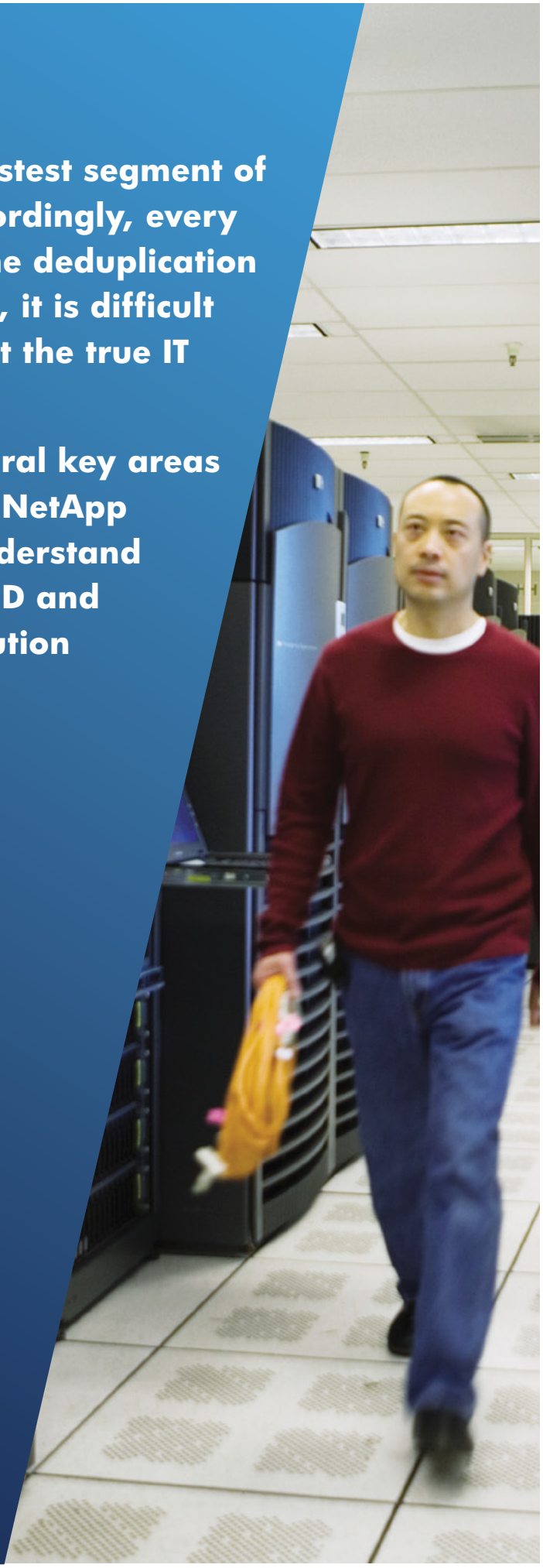
Business white paper

Table of contents

Challenge #1: Primary deduplication: Understanding the tradeoffs 4
Not all data is right for primary storage.....4
Challenge #2: Fixed vs. variable chunking 5
Challenge #3: Performance issues and high deduplication ratios 6
Challenge #4: One size fits all 6
Challenge #5: Backup applications and presentation....7
Challenge #6: Snapshots vs. backup.....7
Summary: HP StoreOnce—building storage for tomorrow 8

The deduplication market is the fastest segment of the storage industry, per IDC. Accordingly, every vendor is telling their version of the deduplication story. With emerging technologies, it is difficult for IT managers to determine what the true IT benefits vs. marketing spin are.

This brief document discusses several key areas of deduplication and uncovers six NetApp "challenges" customers should understand when evaluating HP StoreOnce D2D and NetApp as their deduplication solution of choice.



HP StoreOnce D2D vs. NetApp

Challenge #1: Primary deduplication: Understanding the tradeoffs

Deduplication is emerging as the primary solution to address customers' data growth problems, and is the top initiative for large enterprise customers. Deduplication's optimal benefits are for sequential I/O data patterns, like backup data. NetApp promotes its ability to deduplicate primary data. On the surface, this sounds logical: Less data stored equals less storage investment.

While some may find this surprising given the continuing interest in optimization technologies, primary deduplication can impose some potentially significant performance penalties on the network.[1] Primary data is random in nature. Deduplicating data leads to various data blocks being written to multiple places. NetApp's WAFL file system exasperates the problem by writing to the free space nearest to the disk head. Reading the data involves recompiling these blocks into a format presentable to the application. This data reassembly overhead mandates a performance impact, commonly 20–50 percent.[2]

Not all data is right for primary storage

Can customers accept this level of performance degradation for primary data applications? When NetApp customers have performance issues, NetApp's first recommendation is not to implement its deduplication. Data that requires frequent access with optimum write performance, like production and primary data, is not the correct fit for data deduplication.

According to NetApp's website, primary deduplication can reduce data by 35%. Deduplication vendors claim and deliver data reductions of 20x. There is an enormous difference in efficacy between the two.

Reducing storage capacity growth makes sense and saves customers money. However, for primary data, other techniques such as Thin Provisioning deliver similar storage reduction benefits without the accompanying performance impact; capabilities found on HP P4000 and InServ Storage systems.

Takeaway

Deduplication is often the wrong technology for data reduction of primary storage.

[1] End Users Hesitate on Primary Deduplication; TheInfoPro TIP Insight, October 21, 2010

[2] Evaluator Group, August, 2010

Figure 1: Fixed vs. variable chunking
Example: Sixteen different versions of a PowerPoint presentation (41.5 MB total)

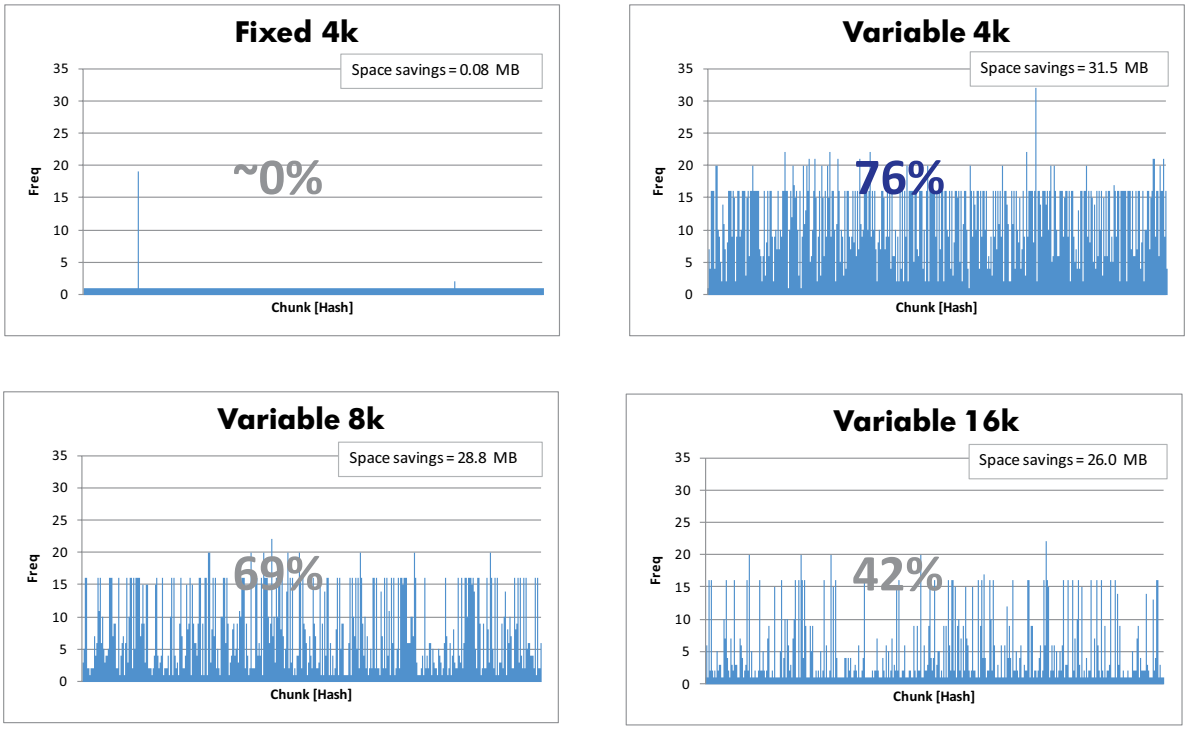
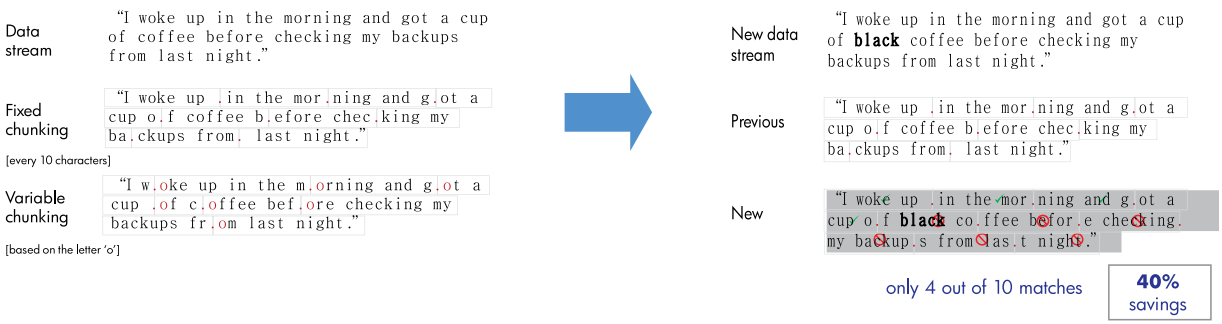


Figure 2:
Chunking: Image backup as a sequence of sentences and every word is a file.
Fixed chunking does not match after the change in the data stream.



Challenge #2: Fixed vs. variable chunking

With deduplication, customers maximize their TCO benefit when the deduplication ratio is high. Accordingly, backup applications, with their repetitive data patterns and sequential throughput are the ideal target environment (figure 1). There are types of deduplication implementations: fixed chunking and variable chunking.

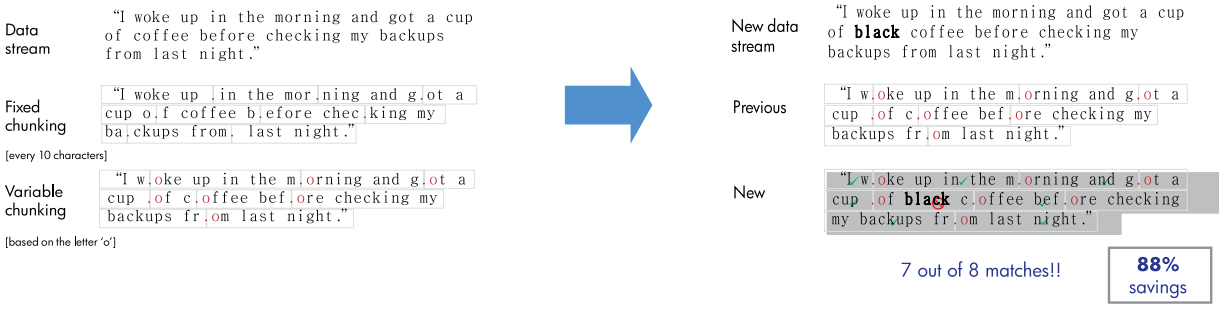
NetApp uses a fixed 4k chunking algorithm, meaning that it "chunks" the data into fixed block size segments. If a subsequent update adds new header information and any other block changes to the file or backup stream, the result is a "shift" in the data pattern and all subsequent data will be shifted as well, and will not match existing data. Thus, everything following the initial change needs to be written to the disk. The fixed chunking methodology means changes in the data can lead to a very poor deduplication ratio (figure 2). Notice that there are no matches after the change in the data stream.

Variable chunking is the technique deployed by many deduplication vendors including HP. Variable chunking groups the data into chunks based on patterns in the data itself. In the above example, the new data "black" or block change causes the same shift. With variable chunking, the new data blocks are written to disk as is the case with fixed deduplication. All other information, following the new data block, is then resynchronized and deduplicated, accordingly, because the variable chunking identifies the same data that was already written. This drives orders-of-magnitude higher deduplication ratios and thus, lowers TCO (figure 3).

Takeaway

Using variable chunking allows HP StoreOnce D2D solutions to provide a more intelligent and effective approach for deduplication.

Figure 3:
Chunking: Image backup as a sequence of sentences and every word is a file.
Fixed chunking does not match after the change in the data stream.



Challenge #3: Performance issues and high deduplication ratios

NetApp suffers performance issues with high deduplication ratios; something NetApp engineers said on a post to the **NetApp technical forum**.[3]

"If we look at a typical scenario (impossible, I know, but bear with me)—let's say we have a FAS3070, a 1 TB volume with 5 percent duplicate data, and the system is fairly quiet. This would be a typical setting for running dedupe overnight on a regular basis. I would expect this system to complete dedupe in less than an hour and have no impact on workloads (since there aren't any running).

On the other hand, if we have a FAS2050, 90 percent duplicate data, and the system is running at peak load—the dedupe process will take many hours and you will likely see some performance degradation resulting from dedupe.

The problem is that there are too many variables for us to give an exact number. Instead, we recommend two things:

- If your application or system is extremely performance-sensitive, don't run dedupe
- If you are concerned that dedupe will create an excessive performance penalty, run a POC first

Also, remember that you can easily turn off dedupe, and/or "undo" dedupe if you don't like the results you get."

NetApp is so concerned about the performance of their deduplication technology that Chris Cummings, senior director of data protection solutions for NetApp told CRN customers must acknowledge the "**chance of performance degradation when implementing the technology**" should they turn on the technology.[4]

"In order to activate the dedupe license, which is free, customers do need to spend about 10 minutes filling out a form that states that they know there is a chance of performance degradation when implementing the technology, depending on data type and other factors."

Takeaway

HP typically finds 95 percent duplicate data in backup and deduplicates the data without impacting performance on the primary array.

Challenge #4: One size fits all

NetApp is a NAS company, by heritage. To address the SAN market, NetApp engineered their NAS filers to mimic block level environments to present SAN style storage. NetApp then added deduplication, compression and other features to the same architecture. Any product line required to provide high performance storage, NAS, SAN storage, data compression, and deduplication while managing overhead cannot be optimized for any one specific purpose.

HP Converged Infrastructure uses industry-standard components to build proven SANs, deduplication and compression appliances, NAS filers, Tape, Archive, and the tools and software to make them to work in harmony (along with servers, networking, print, and client devices). When you buy an HP solution it is a symphony orchestra; each section specialized in purpose but standardized in components, optimized but in step with the rest of the organization. This is not "one box fits all". This is HP Converged Infrastructure.

Takeaway

Backup solutions are optimized for sequential data patterns and are purpose built. HP Converged Infrastructure delivers proven solutions. NetApp's one-size-fits-all approach is ineffective in the backup and deduplication market.

[3] **Deduplication Performance Impact?** NetApp Community Public Forum

[4] **NetApp Allows Dedupe Of Competing Primary Storage**; CRN

Challenge #5: Backup applications and presentation

Changing backup environments can be difficult. Every change has benefits and costs. The best deduplication vendors provide the flexibility to match the existing customer environment and for their solution to present itself to the backup application as tape, with a virtual tape library implementation or as disk, with a NAS implementation. Customers should have the choice to determine the right solution for their environment.

NetApp does not provide this choice. It only offers its disk presentation. For some customers, this is acceptable; for others it will not be. Customers want the flexibility to adapt their deduplication solutions as their requirements evolve.

NetApp frequently recommends eliminating the backup application altogether and just using their snapshot technology for recovery. This solution can be effective in pure NetApp environments, but unfortunately most customers have more than one type of primary storage. Multiple applications and storage arrays require customers to manage and protect their data with a centralized backup application. Finally, many enterprise customers use multiple backup applications due to mergers and acquisitions.

Takeaway

NetApp does not provide enough flexibility for today's complex backup environments.

Challenge #6: Snapshots vs. backup

Taking a snapshot is not the same as backup. NetApp supports up to 255 snapshots. If a customer were to space these an hour apart, this will allow 10 days of recovery. HP solution of snapshots, disk-based deduplication and tape provides years of recovery (up to 30 years) with very fine recovery with snapshots, fast backup recovery from disk on StoreOnce D2D, and then long-term archive with tape.

To move the data between filers NetApp has to re-inflate it for replication. This requires sufficient available space to re-inflate the data on both the destination and target filers. It then requires the destination filer to re-deduplicate the data again. This processing overhead again impacts overall performance. It will consume large amounts of the bandwidth between the sites and slow replication performance. HP StoreOnce D2D moves deduplicated data for replication which reduces bandwidth, disk space, cost, and the need for re-inflation and re-deduplication of data.

By ignoring tape, NetApp increases the long-term archival costs of the customer and does not provide a "green" solution. For example, LTO-5 tape cost around $50/TB—a price per TB SAN disks cannot match. Since tape does not draw power once placed in archive (onsite or off), no running costs are incurred and the overall solution is greener.

Takeaway

Snapshots are part of a data protection solution, but are incomplete by themselves. Long-term storage requirements are not addressed effectively by snapshots alone. HP Converged Infrastructure provides industry-leading solutions, including StoreOnce for disk-based deduplication for a complete data protection strategy.

Summary: HP StoreOnce—building storage for tomorrow

Customers are showing their interest in deduplication with the investments they are making in deduplication solutions. The goal of this document is to inform customers of some of the tradeoffs they will be making when investing in NetApp for deduplication.

As an IT manager, you want simple, cost-effective solutions for your complex IT challenges. HP kept this in mind while designing StoreOnce. New technologies, such as deduplication can be difficult to understand but StoreOnce makes its simple.

StoreOnce is the common deduplication technology across all HP products; you need to learn it just once. StoreOnce D2D appliances plug into your existing HP software frameworks so they can be managed alongside the rest of your storage infrastructure.

With its flexible deployment and compelling price/performance advantage, HP StoreOnce is a key emerging technology for HP and your enterprise.

HP Services

HP Storage Services offers a comprehensive portfolio of onsite and remote services that help your business handle all phases of storage—from preliminary planning and equipment delivery to installation, configuration, integration, testing, and every level of ongoing support. We take a holistic approach to your entire environment, bridging storage, servers, blades and software, and network infrastructure. Our storage services portfolio is designed to help you accelerate growth, mitigate risks, and lower costs throughout the technology life-cycle.

For more information visit **www.hp.com/go/StoreOnce**

4AA3-2649ENW, Created January 2011