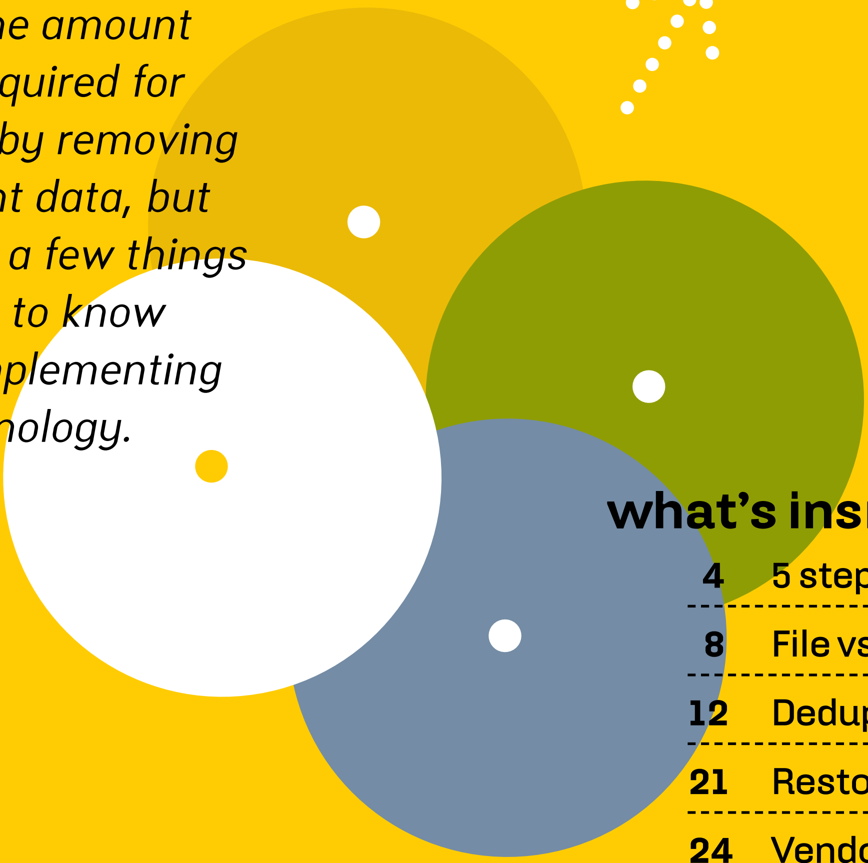
Managing the information that drives the enterprise

STORAGE

ESSENTIAL GUIDE TO data deduplication

Data dedupe can reduce the amount of disk required for backups by removing redundant data, but there are a few things you need to know before implementing this technology.



what's inside

TechTarget
The IT Media ROI Experts

Dedupe in detail

By Rich Castagna

deduplication is the hottest thing to hit backup since disk began popping up in backup configurations. With its ability to stretch usable capacity to accommodate ever-growing data stores, dedupe might just be the single technology that keeps disk in the picture as a viable alternative to tape for short-term retention of backup data.

Using disk as a backup target for short- or long-term retention is a no brainer at this point—backup windows aren't shattered, restores have never been faster or easier, and relatively cheap disk makes a perfect home for data before it gets spun off to tape. It's a pretty picture, but one marred by an ugly reality: There doesn't seem to be an end in sight to the growing amount of data that needs to be backed up.

The idea behind dedupe technology is simple—it trims the amount of data to be stored by eliminating the redundancies so typical of backups, allowing you to effectively cram far more data into the same physical space by factors ranging from 10 to 50 times or more.

The benefits are apparent and make pitching a dedupe deployment to upper management a relatively easy exercise. But dedupe does have its finer points, with capabilities, efficiencies and administrative issues varying from product to product, and from one environment to another.

> Dedupe does have its finer points, with capabilities, efficiencies and administrative issues varying from product to product, and from one environment to another.

To determine the best fit for your backup setup, you'll need to sort through the types of dedupe available and make decisions regarding file or block methods, hash-based systems vs. those that use byte-level comparison techniques and inline vs. post-processing implementations. But there are still numerous details to work out to get optimal dedupe performance and to restore deduped data in a timely manner.

A little homework up front will avoid some grief later, while helping you set reasonable expectations for deduplication. ⊙

Rich Castagna (rcastagna@storagemagazine.com) is Editorial Director of the Storage Media Group.

Can your backup software *do this*?

- ✓ Eliminate up to half your tape drives
- ✓ Eliminate the backup window
- ✓ Deduplicate enterprise data across all tiers, including tape
- ✓ Reclaim space on primary storage
- ✓ Archive, preserve, and search information for eDiscovery
- ✓ Reduce off-site tapes by up to 90%
- ✓ Reduce data management costs by up to 40%

introducing

SIMPANA® 8

Leading companies worldwide are flocking to CommVault and its #1 end-user-ranked enterprise backup product. But backup is just the beginning. With the industry's only truly unified single platform, Simpana® software provides a dramatically superior way for enterprises to handle data protection, eDiscovery, recovery, and information management requirements.

Now, the groundbreaking new features of Simpana 8 make it easier than ever to solve immediate backup and data management problems, improve operations, lower costs, eliminate disparate point products, and set your enterprise up with a scalable platform to meet your needs far into the future.

Sound too good to be true? We'll be happy to prove it to you. Call 888-667-2451. Or visit **www.commvault.com/simpana** to sign up for an introductory webcast.

commvault®
solving forward®

BACKUP & RECOVERY ► ARCHIVE ► REPLICATION ► RESOURCE MANAGEMENT ► SEARCH

CommVault Corporate Headquarters | 2 Crescent Place | Oceanport, NJ | 07757

Individual cost savings and performance improvements may vary.

All specifications are subject to change without notice.

5 steps to dedupe

Following these steps will get you the best results when deduplicating backup data.

By Jerome M. Wendt

The best way to select, implement and integrate data deduplication varies depending on how the deduplication is performed. Here are some general principles you can follow to select the right deduplicating approach and then integrate it into your environment.

Step 1

Assess your backup environment.

The deduplication ratio a company achieves will depend heavily on the following factors:

- Type of data
- Change rate of the data
- Amount of redundant data
- Type of backup performed (full, incremental or differential)
- Retention length of the archived or backup data

The challenge most companies have is quickly and effectively gathering this data. Agentless data gathering and information classification tools from Aptare Inc., Asigra Inc., Bocada Inc. and Kazeon Systems Inc. can assist in performing these assessments, while requiring minimal or no changes to your servers in the form of agent deployments.

Step 2

Establish how much you can change your backup environment.

Deploying backup software that uses software agents requires installing agents on each server or virtual machine, and doing server reboots after installation. This approach generally results in faster backup times and higher deduplication ratios than if you used a data deduplication appliance. However, it can take more time and require many changes to a company's backup environment. Using a data deduplication appliance typically requires no changes to servers, although a company will need to tune its backup software according to if the appliance is configured as a file server or a virtual tape library.

Step 3

Purchase a scalable storage architecture.

The amount of data a company initially plans to back up and what it actually ends up backing up are usually two very different numbers. Companies usually find deduplication so effective when using it in their backup processes that they quickly scale its use and deployment beyond initial intentions; you should therefore confirm that deduplicating hardware appliances can scale both performance and capacity. You should also verify that hardware and software deduplication products can provide global deduplication and replication features to maximize deduplication benefits throughout the enterprise, facilitate technology refreshes and/or capacity growth, and efficiently bring in deduplicated data from remote offices.

Step 4

Check the level of integration between backup software and hardware appliances.

The level of integration a hardware appliance has with backup software (or vice versa) can expedite backups and recoveries. For example, ExaGrid Systems Inc.'s ExaGrid appliances recognize backup streams from CA ARCserve Backup and can better deduplicate data from that backup software than streams from backup software it doesn't recognize. Enterprise backup software is also starting to better manage disk storage systems, so data can be placed on various disk storage systems with different tiers of disk so they can back up and recover data more quickly in the short term and then store it more cost-effectively in the long term.

Step 5

Perform the first backup.

The first backup using agent-based deduplication software can be a potentially harrowing experience. It can create a significant amount of overhead on the server and take much longer than normal to complete because it needs to deduplicate all of the data. However, once the first backup is complete, it only needs to back up and deduplicate changed data going forward. Using a hardware appliance, the experience tends to be the opposite. The first backup may occur quickly, but backups may slow over time depending on how scalable the hardware appliance is, how much data is changing and how much data growth a company is experiencing. ⊙

Jerome M. Wendt is lead analyst and president of DCIG Inc.

File vs. block dedupe

Deduplication can take place at the file or block level. Block-based dedupe gives you greater reduction, but requires more processing power to carry out.

By Lauren Whitehouse

DATA DEDUPLICATION has dramatically improved the value proposition of disk-based data protection, as well as WAN-based remote- and branch-office backup consolidation and disaster recovery (DR) strategies. It identifies duplicate data, removing redundancies and reducing the overall capacity of the data transferred and stored.

Some deduplication approaches operate at the file level, while others go deeper to examine data at a sub-file or block level. Determining uniqueness at the file or block level offers benefits, but the results will vary. The differences lie in the amount of reduction each approach produces, and the time each method takes to determine what is unique.

FILE-LEVEL DEDUPLICATION

Also referred to as single-instance storage (SIS), file-level data deduplication compares a file to be backed up or archived with those already stored by checking its attributes against an index. If the file is unique, it's stored and the index is updated; if the file isn't unique, only a pointer to the existing file is stored. The result is that only one instance of the file is saved and subsequent copies are replaced with a "stub" that points to the original file.

BLOCK-LEVEL DEDUPLICATION

Block-level data deduplication operates on the sub-file level. As its name implies, the file is typically broken down into segments—chunks or blocks—that are examined for redundancy vs. previously stored information.

The most popular way to determine duplicates is to assign an identifier to a chunk of data; for example, by using a hash algorithm that generates a unique ID or "fingerprint" for that block. The unique ID is then compared to a central index. If the ID exists, the data segment has been processed and stored before. Therefore, only a pointer to the previously stored data needs to be saved. If the ID is new, the block is unique. The unique ID is then added to the index and the unique chunk is stored.

The size of the chunk to be examined varies from vendor to vendor. Some have fixed block sizes, while others use variable block sizes (a few even allow users to vary the size of the fixed block). Fixed blocks could be 8KB or 64KB in size; the difference is that the smaller the chunk, the more likely the opportunity to identify it as redundant. This, in turn, means even greater reductions as even less data is stored. The only issue with fixed blocks is that if a file is modified and the deduplication product uses the same fixed blocks from the last inspection, it might not detect redundant segments. This is because as the blocks in the file are changed or moved, they shift downstream from the change, offsetting the rest of the comparisons.

Variable-sized blocks increase the odds that a common segment will be detected even after a file is modified.

Variable-sized blocks increase the odds that a common segment will be detected even after a file is modified. This approach finds natural patterns or break points that might occur in a file and then segments the data accordingly. Even if blocks shift when a file is changed, this approach is more likely to find repeated segments. The tradeoff? A variable-length approach may require a vendor to track and compare more than just one unique ID for a segment, which could affect index size and computational time.

The differences between file- and block-level deduplication go beyond just how they operate. There are advantages and disadvantages to each approach.

File-level approaches can be less efficient than block-based dedupe:

- A change within the file causes the whole file to be saved again. A file, such as a PowerPoint presentation, can have something as simple as the title page changed to reflect a new presenter or date; this will cause the entire file to be saved a second time. Block-based deduplication saves the changed blocks between one version of the file and the next. Reduction ratios may only be in the 5:1 or less range, whereas block-based deduplication has been shown to reduce capacity in the 20:1 to 50:1 range for stored data.

File-level approaches can be more efficient than block-based data deduplication:

- Indexes for file-level deduplication are significantly smaller, which takes less computational time when duplicates are determined. Backup performance is, therefore, less affected by the deduplication process. File-level processes require less processing power due to the smaller index and reduced number of comparisons. Therefore, the impact on the systems performing the inspection is less. The impact on recovery time is low. Block-based dedupe will require "reassembly" of the chunks based on the master index that maps the unique segments and pointers to unique segments. Because file-based approaches store unique files and pointers to existing unique files, there's less to reassemble. ⊙

Lauren Whitehouse is an analyst with Enterprise Strategy Group and covers data protection technologies.

When it comes to data de-duplication, most companies only offer one kind of solution. But with Quantum, you're in control. Our new DXi7500 offers policy-based de-duplication to let you choose the right de-duplication method for each of your backup jobs. We provide data de-duplication that scales from small sites to the enterprise, all based on a common technology so they can be linked by replication. And our de-duplication solutions integrate easily with tape and encryption to give you everything you need for secure backup and retention. It's this dedication to our customers' range of needs that makes us the smart choice for short-term and long-term data protection. After all, it's your data, and you should get to choose how you protect it.

Find out what Quantum can do for you.
For more information please go to www.quantum.com

Quantum.
BACKUP. RECOVERY. ARCHIVE.

dedupe myths

Data deduplication products can dramatically lower capacity requirements, but picking the best one for your needs can be tricky.

By Alan Radding

EXAGGERATED CLAIMS, rapidly changing technology and persistent myths make navigating the deduplication landscape treacherous. But the rewards of a successful dedupe installation are indisputable.

"We're seeing the growing popularity of secondary storage and archival systems with single-instance storage," says Lauren Whitehouse, analyst at Enterprise Strategy Group (ESG), Milford, MA. "A couple of deduplication products have even appeared for use with primary storage."

The technology is maturing rapidly. "We looked at deduplication two years ago and it wasn't ready," says John Wunder, director of IT at Milpitas, CA-based Magnum Semiconductor, which makes chips for media processing. Recently, Wunder pulled together a deduplication process by combining pieces from Diligent Technologies Corp. (dedupe engine), Symantec Corp. (Veritas NetBackup) and Quatrio (servers and storage).

Assembling the right pieces requires a clear understanding of the different dedupe technologies, a thorough testing of products prior to production, and keeping up with major product changes such as the introduction of hybrid deduplication (see "Dedupe alternatives," this page) and the emergence of global deduplication.

"Global deduplication is the process of fanning in multiple sources of data and performing deduplication across those sources," says ESG's Whitehouse. Currently, each appliance maintains its own index of duplicate data. Global deduplication requires a way to share those indexes across appliances (see "Global deduplication," p. 14).

Dedupe alternatives

Until recently, deduplication was performed either in-line or post-processing. Now vendors are blurring those boundaries.

- **FALCONSTOR SOFTWARE CORP.** offers what it calls a hybrid model, in which it begins the post-process deduping of a backup job on a series of tapes without waiting for the entire backup process to be completed, thereby speeding the post-processing effort.
- **QUANTUM CORP.** offers what it calls adaptive deduplication, which starts as in-line processing with the data being deduped as it's written. Then it adds a buffer that can increase dynamically as the data input volume outpaces the processing. It dedupes the data in the buffer in post-processing style.

STORAGE CAPACITY OPTIMIZATION

Deduplication reduces capacity requirements by analyzing the data for unique repetitive patterns that are then stored as shorter symbols, thereby reducing the amount of storage capacity required. This is a CPU-intensive process.

The key to the symbols is stored in an index. When the deduplication engine encounters a pattern, it checks the index to see if it has encountered it before. The more repetitive patterns the engine discovers, the more it can reduce the storage capacity required, although the index can still grow quite large.

The more granular the deduplication engine gets, the greater the likelihood it will find repetitive patterns, which saves more capacity. "True deduplication goes to the sub-file level, noticing blocks in common between different versions of the same file," explains W. Curtis Preston, a SearchStorage.com executive editor and independent backup expert. Single-instance storage, a form of deduplication, works at the file level.

DEDUPE MYTHS

Because deduplication products are relatively new, based on different technologies and algorithms, and are upgraded often, there are a number of myths about various forms of the technology.

In-line deduplication is better than post-processing. "If your backups aren't slowed down, and you don't run out of hours in the day, does it matter which method you chose? I don't think so," declares SearchStorage.com's Preston.

Magnum Semiconductor's Wunder says his in-line dedupe works just fine. "If there's a delay, it's very small; and since we're going directly to disk, any delay doesn't even register."

The realistic answer is that it depends on your specific data, your deduplication deployment environment and the power of the devices you choose. "The in-line approach with a single box only goes so far," says Preston. And without global dedupe, throwing more boxes at the problem won't help. Today, says Preston, "post-processing is ahead, but that will likely change. By the end of the year, Diligent [now an IBM company], Data Domain [Inc.] and others will have global dedupe. Then we'll see a true race."

Post-process dedupe happens only after all backups have been completed. Post-process systems typically wait until a given virtual tape isn't being used before deduping it, not all the tapes in the backup, says Preston. Deduping can start on the first tape as soon as the system starts backing up the second. "By the time it dedupes the first tape, the next tape will be ready for deduping," he says.

Global deduplication

"Global deduplication is the process of fanning in multiple sources of data and performing deduplication across those sources," says Lauren Whitehouse, analyst at Enterprise Strategy Group (ESG), Milford, MA. Global dedupe generally results in higher ratios and allows you to scale input/output. The global deduplication process differs when you're deduping on the target side or the source side, notes Whitehouse.

- **TARGET SIDE:** Replicate indexes of multiple silos to a central, larger silo to produce a consolidated index that ensures only unique files/segments are transported.
- **SOURCE SIDE:** Fan in indexes from remote offices/ branch offices (ROBOs) and dedupe to create a central, consolidated index repository.

Vendors' ultra-high deduplication ratio claims. Figuring out your ratio isn't simple, and ratios claimed by vendors are highly manipulated. "The extravagant ratios some vendors claim—up to 400:1—are really getting out of hand," says ESG's Whitehouse. The "best" ratio depends on the nature of the specific data and how frequently it changes over a period of time.

"Suppose you dedupe a data set consisting of 500 files, each 1GB in size, for the purpose of backup," says Dan Codd, CTO at EMC Corp.'s

Software Group. “The next day one file is changed. So you dedupe the data set and back up one file. What’s your backup ratio? You could claim a 500:1 ratio.”

Grey Healthcare Group, a New York City-based healthcare advertising agency, works with many media files, some exceeding 2GB in size. The company was storing its files on a 13TB EqualLogic (now owned by Dell Inc.) iSCSI SAN, and backing it up to a FalconStor Software Inc. VTL and eventually to LTO-2 tape. Using FalconStor’s post-processing deduplication, Grey Healthcare was able to reduce 175TB to 2TB of virtual disk over a period of four weeks, “which we calculate as better than a 75:1 ratio,” says Chris Watkis, IT director.

Watkis realizes that the same deduplication process results could be calculated differently using various time frames. “So maybe it was 40:1 or even 20:1. In aggregate, we got 175TB down to 2TB of actual disk,” he says.

Proprietary algorithms deliver the best results. Algorithms, whether proprietary or open, fall into two general categories: hash-based, which generates pointers to the original data in the index; and content-aware, which looks to the latest backup.

“The science of hash-based and content-aware algorithms is widely known,” says Neville Yates, CTO at Diligent. “Either way, you’ll get about the same performance.”

Yates, of course, claims Diligent uses yet a different approach. Its algorithm, he explains, uses small amounts of data that can be kept in memory, even when dealing with a petabyte of data, thereby speeding performance. Magnum Semiconductor’s Wunder, a Diligent customer, deals with files that typically run approximately 22KB and felt Diligent’s approach delivered good results. He didn’t find it necessary to dig any deeper into the algorithms.

Special data structures, unusual data formats, and other ways an application treats data and variable-length data can all fool a dedupe product.

“We talked to engineers from both Data Domain and ExaGrid Systems Inc. about their algorithms, but we really were more interested in how they stored data and how they did restores from old data,” says Michael Aubry, director of information systems for three central California hospitals in the 19-hospital Adventist Health Network. The specific algorithms each vendor used never came up.

FalconStor opted for public algorithms, like SHA-1 or MD5. “It’s a question of slightly better performance [with proprietary algorithms] or more-than-sufficient performance for the job [with public algorithms],”

says John Lallier, FalconStor's VP of technology. Even the best algorithms still remain at the mercy of the transmission links, which can lose bits, he adds.

Hash collisions increase data bit-error rates as the environment grows. Statistically this appears to be true, but don't lose sleep over it. Concerns about hash collisions apply only to deduplication systems that use a hash to identify redundant data. Vendors that use a secondary check to verify a match, or that don't use hashes at all, don't have to worry about hash collisions.

For example, SearchStorage.com's Preston did the math on his blog and found that with 95 exabytes of data there's a 0.00000000000001110223024625156540423631668090820313% chance your system will discard a block it should keep as a result of a hash collision. The chance the corrupted block will actually be needed in a restore is even more remote.

"And if you have something less than 95 exabytes of data, then your odds don't appear in 50 decimal places," says Preston. "I think I'm OK with these odds."

DEDUPE TIPS

Sorting out the deduplication myths is just the first part of a storage manager's job. The following tips will help managers deploy deduplication while avoiding common pitfalls.

1

Know your data. "People don't have accurate data on their daily changes and retention periods," says Magnum Semiconductor's Wunder. That data, however, is critical in estimating what kind of dedupe ratio you'll get and planning how much disk capacity you'll need. "We planned for a 60-day retention period to keep the cost down," he says.

"The vendors will do capacity estimates and they're pretty good," says ESG's Whitehouse. Adventist Health's Aubry, for example, asked Data Domain and ExaGrid to size a deduplication solution. "We told them what we knew about the data and asked them to look at our data and what we were doing. They each came back with estimates that were comparable," says Aubry. Almost two years later the estimates have still proven pretty accurate.

2

Know your applications. Not all deduplication products handle all applications equally. Special data structures, unusual data formats, and other ways an application treats data and variable-length data can all fool a dedupe product.

When Philadelphia law firm Duane Morris LLP finally got around to using Avamar Technologies' Axiom (now EMC Avamar) for deduplication, the company had a surprise: "It worked for some apps, but it didn't work with Microsoft Exchange," says John Sroka, CIO at Duane Morris LLP.

Avamar had no problem deduping the company's 6 million Word documents, but when it hit Exchange data "it saw the Exchange data as completely new each time, no duplication," reports Sroka. (The latest version of Avamar dedupes Exchange data.) Duane Morris, however, won't bother to upgrade Avamar. "We're moving to Double-Take [from Double-Take Software Inc.] to get real-time replication," says Sroka, which is what the firm wanted all along.

3

Avoid deduping compressed data. As a corollary to the above tip, "it's a waste of time to try to dedupe compressed files. We tried and ended up with some horrible ratios," says Kevin Fiore, CIO at Thomas Weisel Partners LLC, a San Francisco investment bank. A Data Domain user for more than two years, the company gets ratios as high as 35:1 with uncompressed file data. With database applications and others that compress files, the ratios fell into the single digits.

When deduping a mix of applications, Thomas Weisel Partners experiences acceptable ratios ranging from 12:1 to 16:1. Similarly, data the company doesn't keep very long isn't worth deduping at all. Unless the data is kept long enough to be backed up multiple times, there's little to gain from deduplication for that data.

4

Avoid the easy fix. "There's a point early in the process where companies go for a quick fix, an appliance. Then they find themselves plopping in more boxes when they have to scale. At some point, they can't get the situation under control," says ESG's Whitehouse. Appliances certainly present an easy solution, but until the selected appliance supports some form of global dedupe, a company will find itself managing islands of deduplication. In the process, it will miss opportunities to remove data identified by multiple appliances.

Magnum Semiconductor's Wunder quickly spotted this trap. "We looked at Data Domain, but we realized it wouldn't scale. At some point we would need multiple appliances at $80,000 apiece," he says.

5

Test dedupe products with a large quantity of your real data. "This kind of testing is time consuming, so many companies avoid it. Usually a company will try the product with little bits of data, and the results won't compare with large data sets," says SearchStorage.com's Preston.

Ideally, you should demo the product onsite by having it do real work for a month or so before opting to buy it. However, most vendors won't go along with this unless they believe they're on the verge of losing the sale.

Adventist Health got lucky. It made a decision based on lengthy onsite meetings with engineers from Data Domain and ExaGrid. Based on those meetings and internal analysis, it opted for ExaGrid. Once the decision was made, Adventist Health's Aubry called Data Domain as a courtesy. Data Domain wouldn't give up and offered to send an appliance.

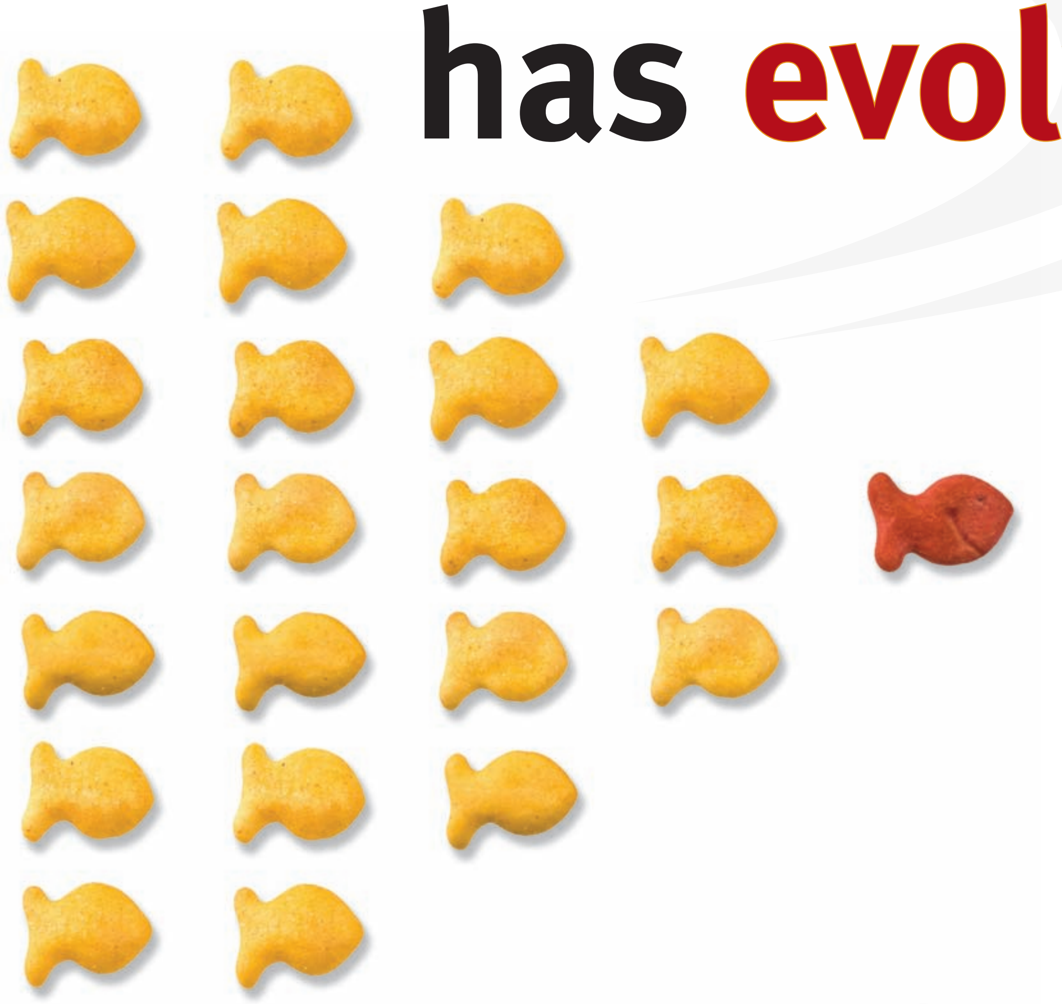
"I was a little nervous I might have made a wrong decision. We put in both and ran a bake off," says Aubry. ExaGrid was already installed on Adventist Health's routed network. It put the Data Domain appliance on a private network connected to its media server.

"I was expecting Data Domain to outperform because of the private network," he says. Measuring the time it took to complete the end-to-end process, ExaGrid performed 20% faster, much to Aubry's relief as he was already committed to buying the ExaGrid.

Just about every consumer cliché applies to deduplication today: buyer beware, try before you buy, your mileage may vary, past performance is no indicator of future performance, one size doesn't fit all and so on. Fortunately, the market is competitive and price is negotiable. With the technology-industry analyst firm The 451 Group projecting the market to surpass $1 billion by 2009, up from $100 million just three years earlier, dedupe is hot. Shop around. Informed storage managers should be able to get a deduplication product that fits their needs at a competitive price. ⊙

Alan Radding is a frequent contributor to *Storage*.

DEDUPLICATION has evolved.



You need more than just simple storage reduction to truly reap the benefits of deduplication. Spectra's nTier Deduplication sets the standard for the next generation of dedupe technology by giving you a robust, preconfigured and easy-to-install appliance that does more than reduce data. nTier doesn't affect network performance, seamlessly integrates with tape and gives you an affordable remote replication option via global deduplication. Saving you space AND much more—now that's dedupe that you can really use.

WHICH MODEL IS RIGHT FOR YOU?	
v80	protects 2.5 TB for 12 weeks
vX	protects 20 TB for 12 weeks

SPECTRA®

DEDUPLICATION
Demand it.

WWW.SPECTRALOGIC.COM/DEDUPE

restore deduped data

When restoring deduped data, performance depends on the backup software, network bandwidth and disk type.

By Lauren Whitehouse

THE BENEFITS OF data deduplication are significant. It allows you to retain backup data on disk for longer periods of time or extend disk-based backup strategies to other tiers of applications in your environment. Either strategy implies that recovery times can be greatly improved (over tape-based backup) for a larger share of the data in your environment.

The capacity reduction achieved through data deduplication also reduces network traffic. Depending on where the deduplication occurs, this can impact the volume of data transferred over a LAN, SAN or WAN, and make it more practical for organizations to implement backup consolidation for remote/branch offices and offsite replication for disaster recovery protection. Both scenarios introduce significant improvements over tape-based strategies where media has to be physically handled and transferred between sites.

There's been a lot more focus (and vendor marketing) on the data deduplication process for backup; specifically, when, where, how and to what degree deduplication impacts the process of writing data.

However, that focus isn't accompanied by increased enlightenment on how deduplication affects the recovery process (specifically, how quickly you can recall data for restoration).

During the recovery process, the requested data may not reside in contiguous blocks on disk even in non-deduplicated backup. As backup data is expired and storage space is freed, fragmentation can occur, which may increase recovery time. The same concept applies to deduped data as unique data and pointers to the unique data may be stored non-sequentially, slowing recovery performance.

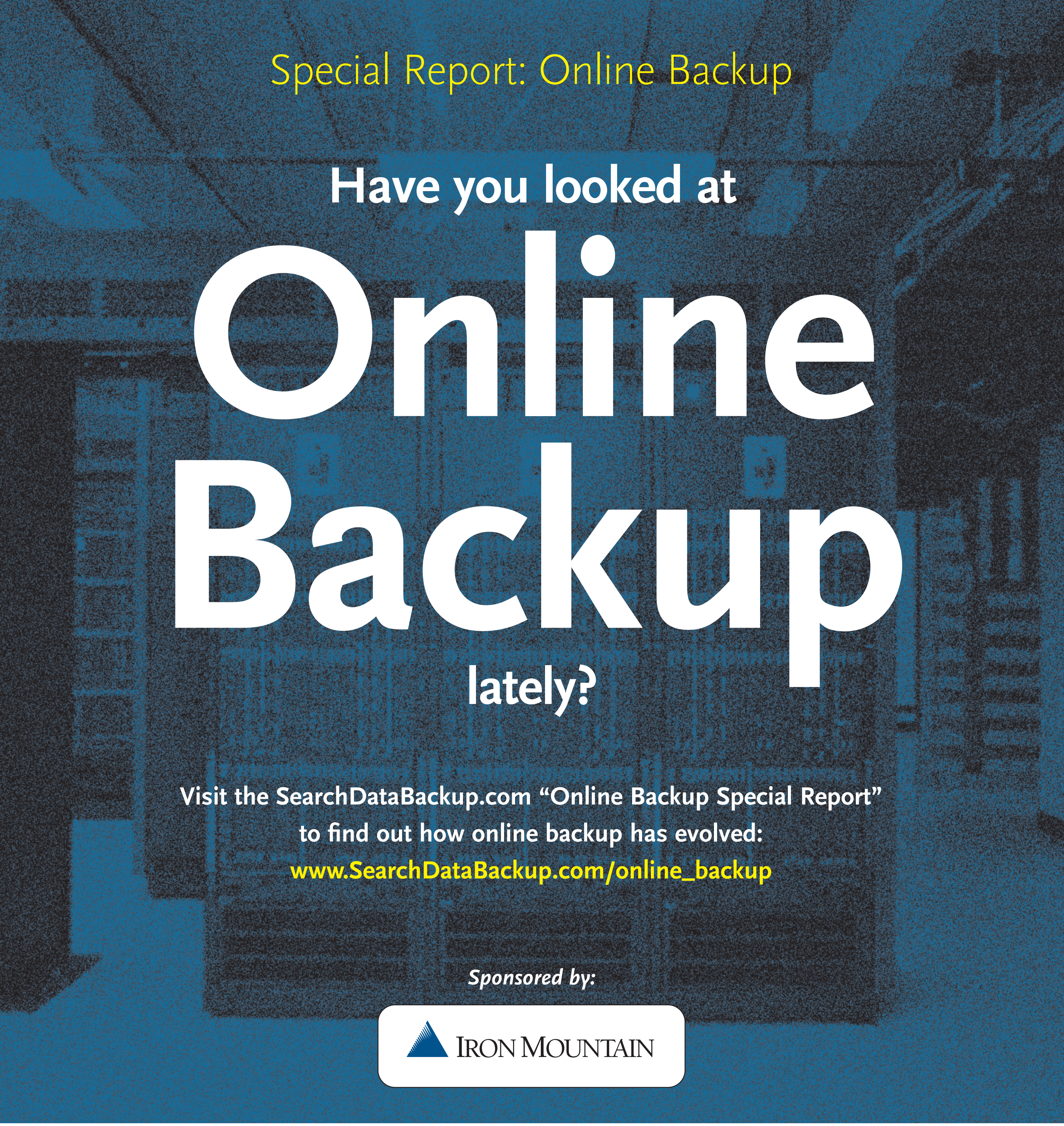
Some backup and storage systems vendors that offer deduplication features anticipated these recovery performance issues and optimized their products to mask the disk fragmentation problem. A few vendor solutions, such as those from ExaGrid Systems Inc. and Sepaton Inc., may keep a copy of the most recent backup in its whole form, enabling more rapid restore of the most recently protected data vs. other solutions that have to reconstitute data based on days, weeks or months of pointers.

The capacity reduction achieved through data deduplication also reduces network traffic.

Other solutions are architected to distribute the data deduplication workload during backup and reassembly activity during recovery across multiple deduplication engines to speed processing. This is the case with software- and hardware-based approaches. Vendors that spread deduplication activities across multiple nodes and, more importantly, allow additional nodes to be added, may provide better performance scalability over those that have a single ingest/processing point.

Performance is dependent on several factors, including the backup software, network bandwidth and disk type. The time it takes for a single file restore will differ greatly from a full restore. It's therefore important to test how a deduplication engine performs in several recovery scenarios (especially for data stored over a longer period of time) to judge the potential impact of deduplication in your environment. ⊙

Lauren Whitehouse is an analyst with Enterprise Strategy Group and covers data protection technologies.



Check out the following resources from our sponsors:

commvault
solving forward

CommVault Webcast: Introducing Simpana® 8
CommVault Webcast: The Simpana Singular Architecture
Visit the CV Deduplication Solutions Page for a Capabilities Overview and Downloadable Content

FalconStor
Software

Book Chapter: SAN for Dummies - Using Data De-duplication to Lighten the Load
White Paper: Demystifying Data De-Duplication: Choosing the Best Solution
Webinar: Enhancing Disk-to-Disk Backup with Data Deduplication

Quantum

Quantum Deduplication Solutions
Quantum's DXi7500 Cuts Backup Times
Northeast Delta Dental Achieves Data Reduction Rates of More than 95% with Quantum DXi3500

SPECTRA
SPECTRALOGIC.COM

Whitepaper: Deduplication and Tape: Friends or Foes?
Whitepaper: Is Deduplication Right for You?
Video: nTier With Deduplication

Sun
microsystems

White Paper: Blending Tape Virtualization and Data Deduplication to Optimize Data Protection Performance and Costs
Combining Storage Capacity Optimization and Replication to Optimize Disaster Recovery Capabilities