Managing the information that drives the enterprise

STORAGE

ESSENTIAL GUIDE TO Data Deduplication Technology

Tips on implementing data deduplication, one of the hottest technologies in data backup and primary storage, are highlighted in this guide.

INSIDE

TechTarget

TOTALLY Open™ solutions form the roots of green IT

"**FalconStor VTL with dedupe offers us the ability to optimize storage capacity and minimize not only our physical storage footprint but also our overall carbon footprint.**

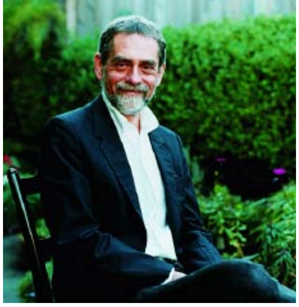
In addition, the solution will improve data security to meet our future compliance requirements via its encryption capabilities."

- Chris Watkis, IT Director, Grey Healthcare Group

GREY HEALTHCARE GROUP

Grey Healthcare Group is one of the world's top five healthcare communications companies, with a global network that includes 43 companies in 22 countries. It has an extensive array of integrated multichannel digital and traditional marketing services in support of brand acceleration and sales.

editorial | rich castagna



Data deduplication—what are you waiting for?

Data deduplication may be the best thing that ever happened to backups—it drastically reduces the size of stored backup data, streamlines the backup process and slashes media costs.

STORAGE is a 24/7 thing—it runs around the clock and keeps growing by the minute. Coping with unrestrained capacity growth is no small matter, but protecting all that data effectively is an even bigger challenge. One way to get an upper hand on data backup is to cut the data down to size—literally—using data deduplication technology.

Six or seven years ago, adding disk to the backup mix was the "big" thing that brought fairly radical changes—and benefits—to backup operations for most companies. Data deduplication takes disk-based backup a step or two further by making it a much more practical data protection solution. Take the redundancy out of backups and backup data that used to require hundreds of terabytes of disk or tape capacity can now be wedged onto media a tenth—or even smaller—that size.

Backup data that used to require hundreds of terabytes of disk or tape capacity can now be wedged onto media a tenth—or even smaller—of that size.

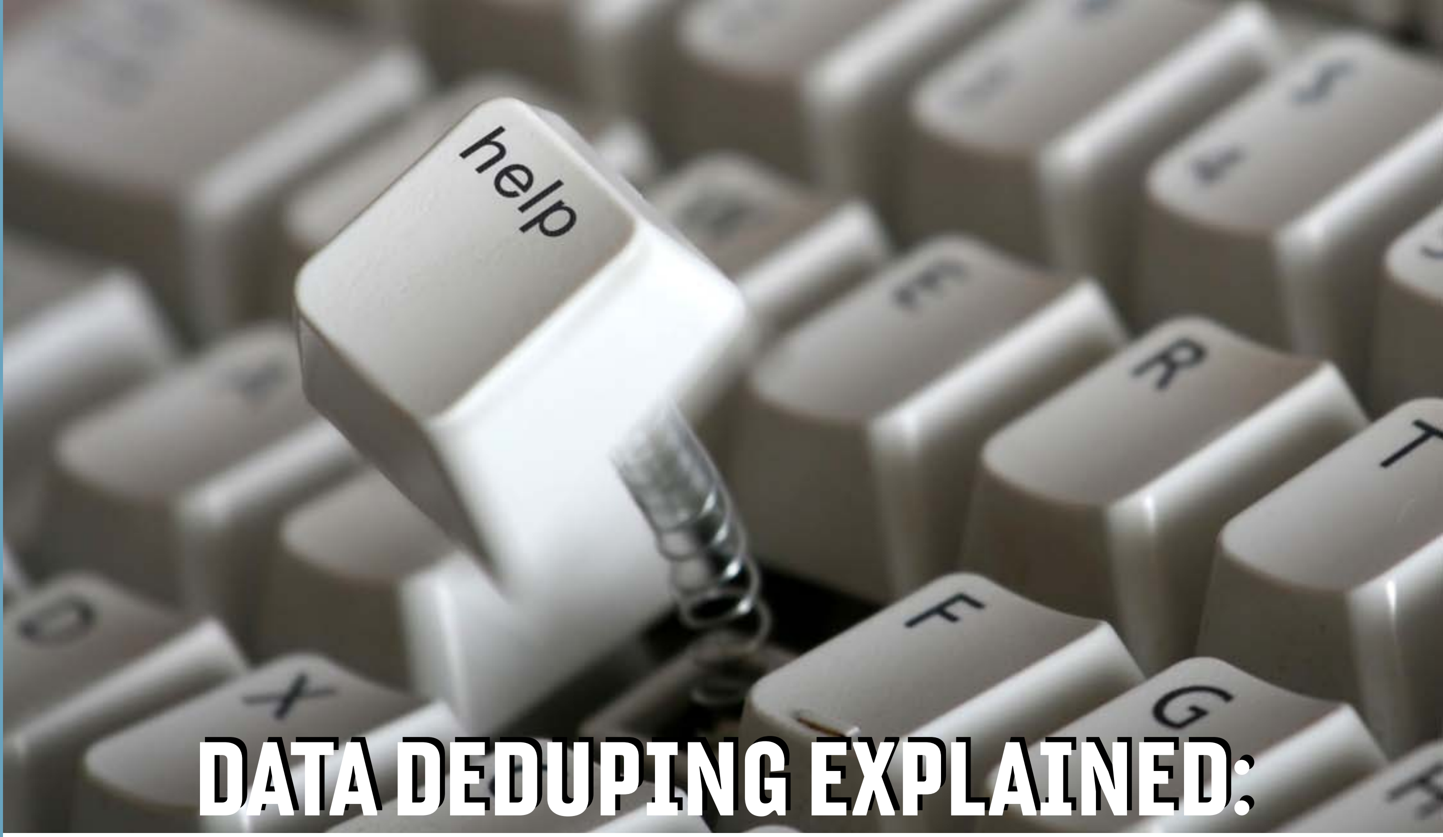
All backups have duplicate data, but how much air a dedupe appliance or app can squeeze out of your backups depends on the types of files, how often you back up and a few other variables. There's a wide enough selection of data deduplication products to meet the needs of most organizations, large and small. You can implement dedupe through software (it might even be an option you can turn on in your current backup application), with an appliance or with a virtual tape library. Deduplication can be performed at the data source, as the data travels to its disk target or after the data has landed on the target. These are just some of the basic options available; there are more subtle differences among dedupe products that can help you find the best fit for your backup environment.

For large organizations, perhaps the most compelling development in deduplication technology is the ability for many dedupe to scale to ever-growing and often dispersed backup environments. Global deduplication effectively pools backup data from disparate dedupe installations and further deduplicates across the backup sets.

Still a relatively young technology, data deduplication is already branching out of its backup role and showing up in products designed to reduce redundancies in primary and nearline storage. These primary storage data reduction products cut data down to size where it lives and before it even hits the backup system. For storage managers swimming against the tide of unchecked capacity growth, primary dedupe could be a lifesaver.

If you don't already know that data deduplication is the biggest thing to hit backup in a long, long time, either your inbox's "out of office" sign has been hanging out there for too long or you're so buried by your own backup woes that you've barely had time to survey the scene. Odds are it's the latter case—so we put together this *Storage* magazine Essential Guide to give you an in-depth dose of all things dedupe. Isn't it time you dug yourself out of that backup hole? ⊙

Rich Castagna (rcastagna@storagemagazine.com) is editorial director of the Storage Media Group.



DATA DEDUPING EXPLAINED: Deduplication in data backup environments tutorial

IT shops must consider a variety of approaches when selecting the optimal data deduplication option.

By W. Curtis Preston

DATA DEDUPLICATION is one of the biggest game-changers in data backup and data storage in the past several years, and it is important to have a firm understanding of its basics if you're considering using it in your storage environment. The purpose of this tutorial is to help you gain that understanding.

When the term "deduplication," also referred to as data dedupe or data deduping, is used without any qualifiers (e.g., file-level dedupe), we are typically referring to subfile-level deduplication. This means that individual files are broken down into segments and those segments are examined for commonality. If two segments are deemed to be the same (even if they are in different files), one of the segments is deleted and replaced with a pointer to the other segment. Segments that are deemed to be new or unique are, of course, stored as well.

Different files—even if they are in different file systems or on different servers—may have segments in common for a number of reasons. In backups, duplicate segments between files might indicate that the same exact file exists in multiple places. Duplicates are also created when performing repeated full backups of the same servers. Finally, duplicate segments are created when performing incremental backups of files. Even if only a few bytes of a file have changed, the entire file is usually backed up by the backup system. If you break that file down into segments, most of the segments between different versions of the same file will be the same, and only the new, unique segments need to be stored.

INLINE VS. POST-PROCESSING DEDUPLICATION

The two primary approaches (inline deduplication and post-processing deduplication) are roughly analogous to synchronous replication and asynchronous replication. Inline deduplication is roughly analogous to synchronous replication, as it does not acknowledge a write until a segment has been determined to be unique or redundant; the original native data is never written to disk. In an inline system, only new, unique segments are written to disk. Post-process deduplication is roughly analogous to asynchronous replication as it allows the original data to be written to disk and deduplicated at a later time. “Later” can be in seconds, minutes, or hours later depending on which system we are talking about and how it is has been configured.

The two primary approaches (inline deduplication and post-processing deduplication) are roughly analogous to synchronous replication and asynchronous replication.

There is not sufficient space here to explain the merits of these two approaches, but the following few statements will give you an overview of their claims. Inline vendors claim to be more efficient and require less disk. Post-process vendors claim to allow for faster initial writes and faster read performance for more recent data—mainly because it is left stored in its native format. Both approaches have merits and limitations, and one should not select a product based on its position in this argument alone. One should select a product based on its price/performance numbers, which may or may not be affected by their choice to do inline or post-process.

Another consideration is whether the deduplication product is hash-based, modified hash-based or delta differential. Hash-based vendors take segments of files and run them through a cryptographic hashing algorithm, such as SHA-1, Tiger, or SHA-256, each of which create a numeric value (160 bits to 256 bits depending on the algorithm) that can be compared against the numeric values of every other segment that the dedupe system has ever seen. Two segments that have the same hash are considered to be redundant. A modified hash-based approach typically uses a much smaller hash (e.g., cyclic redundancy check of only 16 bits) to see if two segments might be the same; they are referred to as redundancy candidates. If two segments look like they might be the same, a binary-level comparison verifies that they are indeed the same before one of them is deleted and replaced with a pointer.

Delta differential systems attempt to associate larger segments to each other (e.g., two full backups of the same database) and do a block-level comparison of them against each other. The delta differential approach is only useful in backup systems, as it only works when comparing multiple versions of the same data to each other. This does not happen in primary storage; therefore, all primary storage deduplication systems use either the hash-based or the modified hash-based approach to identify duplicate data.

TARGET DEDUPLICATION VS. SOURCE DEDUPLICATION AND HYBRID APPROACHES

Where should deduplicate data be identified? This question applies only to backup systems, and there are three possible answers: target, source, and hybrid. A target deduplication system is used as a target for regular (non-deduplicated) backups, and is typically presented to the backup server as a NAS share or virtual tape library (VTL). Once the backups arrive at the target, they are deduplicated (either inline or as a post-process) and written to disk. This is referred to as target dedupe, and its main advantage is that it allows you to keep your existing backup software.

If you're willing to change backup software, you can switch to source deduplication, where duplicate data is identified at the server being backed up, and before it is sent across the network. If a given segment or file has already been backed up, it is not sent across the LAN or WAN again—it has been deduped at the source. The biggest advantage to this approach is the savings in bandwidth, making source

A target deduplication system is used as a target for regular (non-deduplicated) backups, and is typically presented to the backup server as a NAS share or virtual tape library.

dedupe the perfect solution for remote and mobile data.

The hybrid approach is essentially a target deduplication system, as redundant data is not eliminated until it reaches the target; however, it is not as simple as that. Remember that to deduplicate data, the files must be first broken down into segments. In a hash-based approach, a numeric value—or hash—is then calculated on the segment, and then that value is looked up in the hash table to see if it has been seen before. Typically, all three of these steps are performed in the same place—either at the source or the target. In a hybrid system, the first one or two steps can be done on the client being backed up, and the final step can be done on the backup server. The advantage of this approach (over typical target approaches) is that data may be compressed or encrypted at the client. Compressing or encrypting data before it reaches a typical target deduplication system would significantly impact your dedupe ratio, possibly eliminating it altogether. But this approach allows for both compression and encryption before data is sent across the network.

> In a hash-based approach, a numeric value—or hash—is then calculated on the segment, and then that value is looked up in the hash table to see if it has been seen before.

PRIMARY STORAGE DEDUPLICATION

Deduplication is also used in primary data storage, where duplicate data is not as common—but it does exist. Just as in backups, the same exact file may reside in multiple places, or end-users may save multiple versions of the same file as a way of protecting themselves against “fat finger” incidents. One type of data that has a lot of commonality between different files is system images for virtualization systems. The C: (or root) drive for one system is almost exactly the same as the C: (root) drive for another system. A good deduplication system will identify all of those common files and segments and replace them with a single copy.

Whether we’re talking backups or primary storage, the amount of disk saved is highly dependent on the type of data being stored and the amount of duplicate segments found within that data. Typical savings in backup range from 5:1 to 20:1 and average around 10:1, with users who do frequent full backups tending towards the higher ratios. Savings in primary storage are usually expressed in reduction percentages, such as “there was a 50% reduction,” which sounds a lot better than a 2:1 deduplication ratio. Typical savings in primary storage range from 50% to 60% or more for typical data, and as much as 90% or more for things like virtual desktop images.

YOUR MILEAGE WILL VARY

There is no “may” about it—your mileage will definitely vary when deduping data. Your dedupe performance and data deduplication ratio will be significantly different than that of your neighbors. A data deduplication system that is appropriate for your neighbor may be entirely inappropriate for you, since different approaches work better for different data types and behavior patterns. Therefore, the most important thing to remember when selecting a data deduplication system is to perform a full proof of concept test before signing the check. ⊙

W. Curtis Preston is an executive editor at TechTarget and an independent data backup expert.

GLOBAL DEDUPLICATION AND BACKUP: a primer

Global deduplication can reduce backup data across multiple devices, improve dedupe ratios and ease management. *By Dave Raffo*

NOW THAT DATA DEDUPLICATION is rapidly becoming a mainstream feature in data backup and recovery, a big question for any data dedupe product is, "Does it do global deduplication?"

Global deduplication reduces backup data across multiple devices and sites, so the devices act as one large system. The alternative is local dedupe, which reduces data on each device individually. Global deduplication becomes more important to your data storage environment when you have more data to back up and use more devices because it can often improve deduplication ratios. However, one of the biggest benefits of global dedupe is the ability to efficiently manage multiple devices. For instance, global dedupe allows load balancing and high availability.

The major data backup software applications with dedupe today offer global deduplication. These include Asigra Inc.'s Cloud Backup, EMC Corp.'s Avamar, CommVault Systems Inc.'s Simpana and Symantec Corp.'s NetBackup PureDisk.

Exagrid Systems Inc.'s EX Series, FalconStor Software Inc.'s File-interface Deduplication System (FDS), Hewlett-Packard (HP) Co.'s Virtual Library System (VLS), IBM Corp.'s ProtecTier, NEC Corp.'s HydraStor and Sepaton Inc.'s DeltaStor support global dedupe on virtual tape libraries and disk targets.

Historically, some of the largest deduplication vendors have supported only local deduplication, including market leader EMC's Data Domain and Quantum Corp.'s DXi-Series. Local dedupe devices each use their own repository, which limits the deduplication ratio when shifting a backup job to a different appliance. This meant a 16-controller Data Domain array, for instance, acted as 16 separate systems. Data Domain execs argued that their systems are large enough and fast enough that global dedupe isn't necessary. But, Data Domain recently announced a Global Deduplication Array that will include global dedupe for two controllers in the initial release. The product is due to ship by the end of the second quarter.

Independent backup expert W. Curtis Preston offers this global deduplication best practice. He suggests that organizations backing up more than 50 TB of data a night should use global deduplication. Also, those with smaller but rapidly growing backups should also consider global deduplication. ⊙

Dave Raffo is the Senior News Director for the Storage Media Group.

Source deduplication decreases remote-office backup data and bandwidth

Source deduplication, also called client-side deduplication, can provide efficiencies in reducing remote-office backup data and bandwidth needs.

By Christine Cignoli

THE BEST WAY to tell that data deduplication has come into its own is the variety of flavors now available. Source deduplication—also called client-side dedupe—dedupes at the backup client level instead of deduping data at the media server or appliance level. This approach works particularly well with remote offices or branch offices and laptops.

"If it wasn't for dedupe at the client, before anything else happens, the solution wouldn't work," said Gregory Fait, associate principal and director of IT infrastructure at architecture firm Perkins & Will. The firm's decentralized management setup and remote offices all over North America led them to source deduplication with EMC Corp. Avamar as part of a tape replacement project. "When you're talking about two to three terabytes at a remote site and looking at our bandwidth and the pipes we had," said Fait, "there's no choice but to do it locally before it went over the wire."

Source deduplication "is a natural evolution," said Lauren Whitehouse, senior analyst at Milford, Mass.-based Enterprise Strategy Group. "As people get comfortable with the technology and the technology improves, it's not having as much impact on production environments as everyone thinks," she said. Source deduplication is all about efficiency. "The closer you are to the source of the data," added Whitehouse, "the more efficient you're going to be in moving data around."

EASING REMOTE BACKUP WITH SOURCE DEDUPLICATION

Network engineer Andrew Harkin was also looking to ease remote backups with Avamar at Avera Health, a healthcare organization with small clinics and hospitals across the upper Midwest, comprising 56 remote servers at 21 sites. Harkin needed to cut down his data backup windows. "Traditional tape backups weren't working," said Harkin. Deduping data at the source has had "a huge impact on us," he said. "Overnight it went from 25 or 26 hours to two to three hours to get them done."

"Traditional tape backups weren't working. Deduping data at the source has had "a huge impact on us." Overnight it went from 25 or 26 hours to two to three hours to get them done."

—ANDREW HARKIN, network engineer, Avera Health

Software-based data backup and recovery vendors are in the driver's seat in the source dedupe market, according to Whitehouse. "The target-side [dedupe] vendors don't have the same opportunity as the software vendors to capture data at the source," she said. Vendors such as Asigra Inc. and Robobak incorporate source dedupe into their backup software. Storage-centered vendors are entering the source dedupe market; Whitehouse said IBM Corp. and EMC are in the best position to follow Symantec Corp. in offering dedupe in many places throughout the backup process. Symantec recently announced source deduplication with PureDisk in its newest Backup Exec 2010 and NetBackup 7 releases, adding to previously available media server and appliance options, using the OpenStorage (OST) plug-in technology.

Source dedupe for laptops and desktops is an emerging area. "We see the demand for dedupe in the data center and remote offices because there is so much duplicated information," said Mathew Lodge, senior director of product marketing at Symantec. Symantec isn't supporting laptop dedupe yet, he said. "It's more a question of timing. Laptops and desktops are interesting, but it's not the most urgent problem."

There's a high level of duplication among desktop and laptop data, said Rob Emsley, senior director, product marketing, EMC Backup Recovery Systems division. "Even more than remote offices, the amount of duplicate data that exists within those computers in the same firm is very high," said Emsley.

Though source dedupe can eliminate bandwidth headaches with regular backups, there's still the matter of that first full backup to consider. "The first one is the first one, no way around it," said Ron Roberts, president and CEO of Robobak. Most vendors recommend "seeding" to establish the first backup and avoid a huge bandwidth hit. Seeding might entail backing up files common to each user, such as the Windows operating system, which everyone then dedupes against. "Before the rollout, do a backup of the typical applications your company has," said Eran Farajun, executive vice president at Asigra. "Back up files in the data center, and now users won't have to re-backup files."

Emsley said that EMC users might seed the system by moving the Avamar backup server to the remote location, doing the initial backup of all machines at that location, then moving the server to the central data center where it will reside. Or "simply deploy the backup agent in that remote office and back up over your network connection," he said. "There are various ways of deploying it."

Perkins & Will's Fait decided to use large external USB drives to seed their remote offices with a copy of the data. "It cut roughly two to three weeks "off the setup process," Fait said. Avera's Harkin used a virtual machine to seed his system. "I built a virtual machine off a template, a very basic system with nothing on it, and I used that for the first one," he said. "After seeing it, I could have used a production server with no real problem."

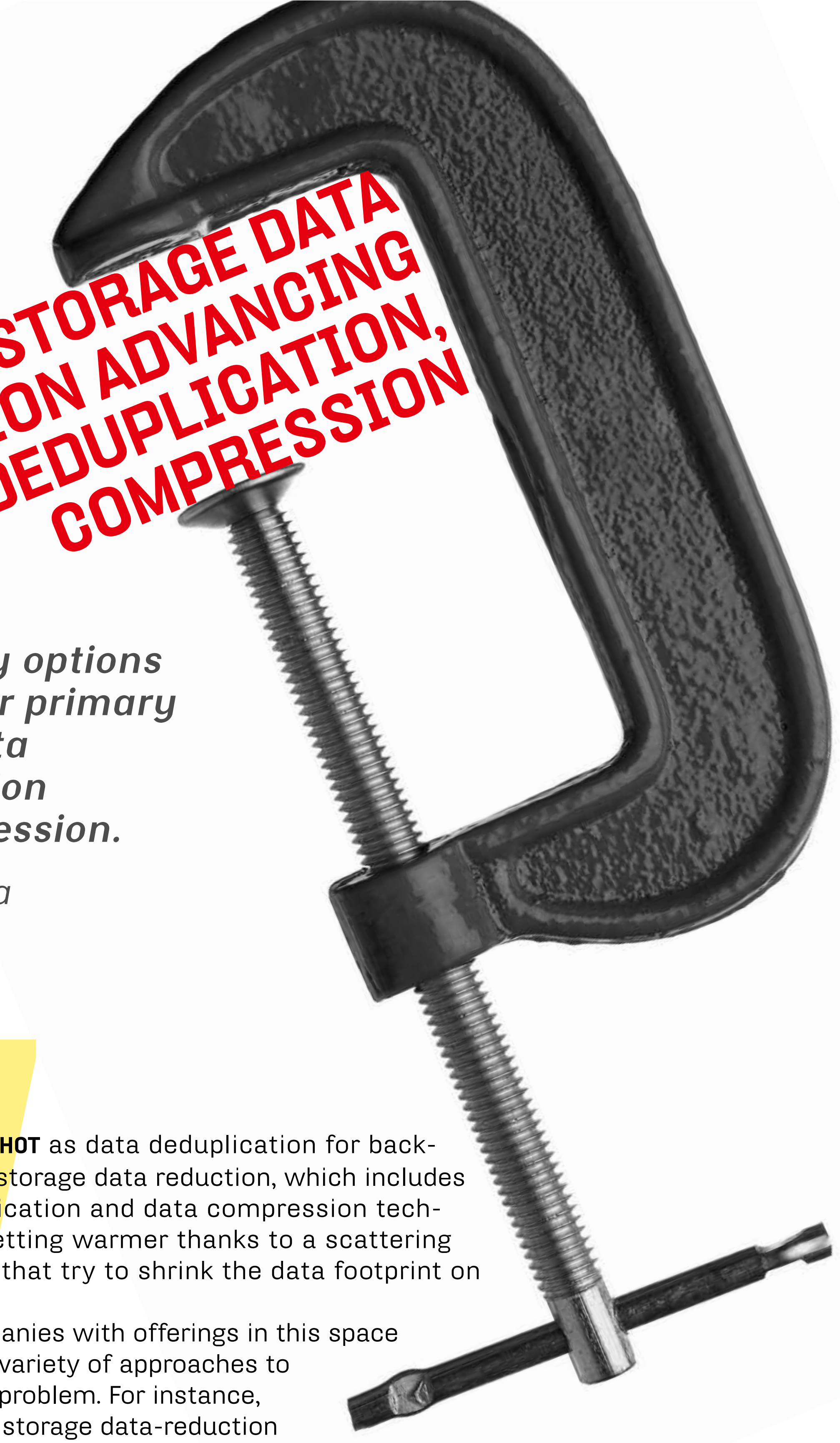
But not everyone wants to use the processing power of a production server and backup application agent for dedupe. "We deduplicate at the server level for our systems," said Al Schipani, manager of server engineering at Westchester Medical Center in Valhalla, NY. He and his team have been beta-testing Symantec's NetBackup 7, and they still prefer media server dedupe for their data, especially for virtual machines. "Since we are 24/7 we do not want to add the additional workload of deduplication on the source," he said.

Fait said next on the dedupe wish list is laptop backups, as well as some cloud-based options—"a kind of end user, self-service type of backup and restore," he said, "whether inside a corporate network or public Internet."

Harkin wants to see more integration from EMC across its backup offerings (which have come together through acquisitions). He'd also like user interfaces made a priority. "I think as an industry, people are used to ugly interfaces and commands," said Harkin. "In a real-life scenario, you don't have enough time to get things done, and having a nice, well-designed interface is paramount to getting things done." ⊙

Christine Cignoli is a Boston-based technology writer.

Vendor resources
Primary storage dedupe
Source dedupe
Global deduplication
Dedupe explained



PRIMARY STORAGE DATA REDUCTION ADVANCING VIA DATA DEDUPLICATION, COMPRESSION

Technology options advance for primary storage data deduplication and compression.

By Carol Sliwa

WHILE NOT AS HOT as data deduplication for backup, primary storage data reduction, which includes data deduplication and data compression techniques, is getting warmer thanks to a scattering of products that try to shrink the data footprint on tier 1 disk.

The companies with offerings in this space are taking a variety of approaches to address the problem. For instance, one primary storage data-reduction approach searches for duplicates at the file level, while others are more granular, comparing data blocks or byte streams, of fixed or variable sizes. Some work post-process, storing the data writes before starting the data deduplication process. And one compression specialist operates inline, in the data path.

NetApp Inc. deduplication, which operates at the block level, is the most prominent of the offerings taking aim at primary storage. The company claims that more than 8,000 customers have licensed its free deduplication technology since its 2007 release.

Rival EMC Corp. followed NetApp into primary storage deduplication in early 2009 with the release of its Celerra Data Deduplication, which actually performs compression before tackling deduplication on file-based data.

Ocarina Networks Inc. also does both compression and deduplication but takes a different path than EMC. Ocarina's ECOsystem first extracts and decompresses file-based data, then deduplicates on a variable- or sliding-block basis before compressing it.

The current list of entrants in the primary storage data reduction space also includes Storwize Inc., which has a compression-only offering. Storwize CEO Ed Walsh contends that primary storage is not the appropriate place for data deduplication.

"You dedupe what's repetitive, and you don't find in primary data the same repetition that you see in a backup data flow," said Walsh, who was formerly the CEO of Avamar, a deduplication backup software vendor acquired by EMC in 2006.

> "You dedupe what's repetitive, and you don't find in primary data the same repetition that you see in a backup data flow."
>
> —ED WALSH, CEO, Storwize

MORE PRIMARY STORAGE DEDUPLICATION PRODUCTS ON THE HORIZON

Some industry analysts expect more vendors to continue turning their attention to primary storage deduplication. Permabit Technologies Corp., for instance, offers inline, sub-file level deduplication. Permabit targets its dedupe at archiving but claims some customers use it for primary storage. Sun Microsystems Inc., now owned by Oracle Corp., late last year added built-in deduplication to its ZFS file system. Other vendors that employ the open-source ZFS technology, such as Nexenta Systems Inc., are exploiting it.

"Vendors that have solutions today in the market may not be the ones you'll see in five years," said Lauren Whitehouse, a senior analyst at Enterprise Strategy Group. "It's not that they're going to go away, but I don't think they'll even be the top ones. It might be the application vendor or the operating system vendor, someone closer to the creation of data, the storage of that data, policies around that data."

Valdis Filks, a research director for storage technologies and strategies at Gartner Inc., said he expects two or three more vendors to offer deduplication for primary storage in 2010, with more to follow in 2011. By 2012, primary storage dedupe will be ubiquitous, he predicted.

"Sometimes we say a technology turns the industry upside down or on its head. Allegorically and technically, dedupe on primary storage does that," Filks said. "We are so used to writing the data to a backup dedupe device and deduping it there. If everything is deduped at source, I expect the back-end dedupe vendors to start to have lots of trouble, and they will obviously have a marketing offensive saying primary deduplication is the wrong place."

Filks said software-intensive, modern-design storage devices with a file system and intelligent block-based architectures, which have the ability to store metadata pertaining to each data block, will be best suited to primary storage data deduplication. Performance issues can be overcome through a combination of multi-core high-speed processors, low-cost DRAM for cache and solid-state drive technology, he added.

"Designers have more performance-accelerating components in storage than they have ever had before, at an affordable price," Filks said.

In the meantime, the majority of end users have been content to hold off on primary storage data reduction.

"People are OK just buying more disk drives," said Greg Schulz, founder and senior analyst at StorageIO Group. "People understand and realize that they can go in and archive, pull the data out of databases, out of email, out of file systems, and then back it up onto a deduped disk or onto a compressed tape."

"Designers have more performance-accelerating components in storage than they have ever had before, at an affordable price."

—VALDIS FILKS, research director for storage technologies and strategies, Gartner Inc.

Understanding how the current crop of primary storage data-reduction products works and where each of their sweet spots lies can help an IT organization to decide if the technology might be a good fit to help curb the explosive growth of storage.

"Vendors doing sub-file reduction have a much higher hurdle to get over because they have to demonstrate that they can do that with very little performance impact in primary storage use cases," said Jeff Boles, a senior analyst and director, validation services at Taneja Group. ⊙

Carol Sliwa is the features writer for SearchStorage.com.

PRIMARY STORAGE DATA DEDUPLICATION AND DATA REDUCTION: Motivators and market landscape

Primary storage data deduplication products mature.

By George Crump

W

WHEN IT COMES TO the amount of attention the storage industry pays to various technologies, only cloud storage might score higher than deduplication. But hype and overhype can cause a problem for the channel, which has to separate fact from fiction and decide what customers really need and what they're actually going to spend money on. The simple truth is that most customers, especially when budgets are tight, spend money only in the areas that are causing the most pain. Is primary storage deduplication—or the broader class of products that address primary storage data reduction—one of those areas?

Before we answer that question, first let's discuss what primary storage deduplication is and what some of the motivators are for a customer selecting a primary storage deduplication product. Primary storage deduplication is largely based on the same technology as backup or archive deduplication. Redundant blocks of data are identified and stored only once. This requires some overhead to build the metadata database that manages the reference points to the data. For the overhead to be worthwhile, there should be a significant return on the investment in the form of increased capacity.

The problem is that primary storage is unlike backup storage in key ways. In a backup scenario, the same data is sent to the backup store over and over again. As a result, backup storage deduplication can deliver data reduction rates in the range of 20:1. But primary storage does not typically have a high level of redundancy and so primary storage deduplication can't deliver similar reduction rates or a similar ROI.

Beyond that, primary storage has less headroom than backup storage in which to perform the deduplication. Primary storage needs more headroom to sustain performance rates; if the headroom isn't there, application performance on that primary storage will suffer.

Finally, while primary storage is more expensive than backup storage, its capacity and cost have come down, making the cost of buying more primary storage less expensive than in the past. It's relatively easy to keep adding more shelves of storage with more capacity to primary storage. Customers may see this as the path of least resistance.

With all these factors, why would customers consider primary storage deduplication, and why should VARs pay attention to this market? One motivator relates to power. The challenge with just throwing disks at the data growth problem on primary storage is finding room on the electrical grid to power all those drives. When calculated by itself, even though the space savings from dedupe on primary storage are much lower than in a backup scenario, the ROI of squeezing more data on to the same storage may be compelling; when combined with the potential power savings, it may be irresistible.

The challenge with just throwing disks at the data growth problem on primary storage is finding room on the electrical grid to power all those drives.

The second key motivator in primary storage deduplication is the ever-growing deployment of virtual machines. In most virtual server environments, the entire server image is loaded on to a shared storage platform. This includes the operating system and other key files, all of which tend to be very similar across servers. In an environment with 100 virtual servers, there is a lot of duplicate data that wasn't there before server virtualization. This data tends to be read-heavy, so the performance impact is not as severe as write-heavy data when read from a deduplicated area. In large virtual server environments—say, more than 50 virtual machines—the amount of redundant data makes an investment in primary storage deduplication worthwhile, even without taking into consideration the power savings discussed above.

MARKET LANDSCAPE

There are a number of products that handle primary storage deduplication and data reduction. For instance, there are content-aware deduplication/

compression tools (from the likes of Ocarina Networks Inc.) available that allow for content-specific examination. For example, say you have two photos of the same image stored on a system that are identical except that one has had the "red eye" removed. To most deduplication systems, these are totally different files. A content-aware dedupe product stores almost the entire image just once, retaining separate data for the area in the photo where the images differ.

A content-aware dedupe product stores almost the entire image just once, retaining separate data for the area in the photo where the images differ.

Beyond content-aware deduplication/compression products, there are products available (from the likes of Ocarina Networks and Storwize Inc.) that can keep the data in its optimized state across storage tiers. For example, data could be examined for redundancy, compressed and then moved to a disk archive. This not only frees up the primary storage pool but it also more deeply optimizes the secondary storage tier. In some cases this data can even be sent to the backup target in its optimized format.

Another approach, from Storwize, focuses on compression rather than deduplication. Storwize's compression appliances sit inline in front of NAS heads. Although they don't deduplicate at all, they compress data universally (as opposed to deduplication, which obviously acts only on duplicate data). Interestingly, in almost every test case, the Storwize appliance has not impacted storage performance, primarily because with compression, while there's processing required to compress the data, there's less data to transport, cache and compute.

Another inline primary storage data reduction method that uses deduplication, along with compression, is WhipTail Technologies Inc.'s Racerunner SSD appliance. The product's use of deduplication and compression means it won't be as fast as more traditional SSDs, but it may be a happy medium for many customers. Those that need more performance than what mechanical drives can offer but not the extreme performance of traditional SSDs are good candidates. Racerunner SSD is the only product that does block-based in-line primary storage deduplication.

While those two examples address inline primary storage data reduction, most primary storage data reduction tools are post-process and work on near-active data—data that is idle but is not quite ready to be archived to a secondary or archive tier. Many customers decide that they don't want to or can't migrate to that secondary tier at all and so this form of optimization may be ideal for them.

NAS systems are the most common deployment of this technique. Companies like EMC Corp., NetApp Inc. and Nexenta Systems Inc. all have a deduplication component in their offerings now (NetApp as part of its Data OnTap operating system, EMC within its Celerra, and Nexenta with a product based on Sun's ZFS). In most cases, files are examined when NAS utilization is low. The deduplication component will identify files or a block of files that have not been accessed in a period of time, compare them at the appropriate level with other data segments on the NAS for similarities and then eliminate redundant segments. This process could produce savings ranging from 3:1 to 5:1 depending on the dataset.

The downside to most of these NAS implementations is that efficiencies are gained only on a platform-by-platform basis. If your customer has a mixed NAS environment, you may want to look at an external solution that can work across different systems. Today, Ocarina Networks and Cofio Software Inc. have products in this space. Both companies' products can identify redundant segments of data and store only one copy of that segment. Both also have the capability to move data from primary storage to secondary storage and maintain the deduplication efficiency.

So, are we "there" with primary storage deduplication? To a large extent, yes. The solutions are beginning to mature and customer need—thanks to constricting access to power, rampant growth in virtualization and unstructured data—is increasing quickly. Now is an ideal time for resellers to develop a strategy around primary storage deduplication and data reduction. ⊙

George Crump is president and founder of Storage Switzerland, an IT analyst firm focused on the storage and virtualization segments.

Check out the following resources from our sponsors:

FalconStor Software

Video: GHG leverages FalconStor VTL with Dedupe to improve backup and recovery performance

WhitePaper: Demystifying Data Deduplication: Choosing the Best Solution

SAN for Dummies: Changing the Rules of Deduplication

hp intel

ROBO and regional data centre data protection solution scenarios using HP Data Protector software, HP VTL systems and Low Bandwidth replication

Exploring the ROI of VTL Deduplication Solutions: How HP StorageWorks Stands Up Against the Competition

VMware backup decisions - A guide to understanding what VMware backup method is best for your environment

Quantum

Quantum DXi6500 Family Addresses Mid-Market Deduplication Needs; Cost, Existing Backup Integration, Implementation and Scalability

Quantum goes beyond deduplication

D2D2T Backup Architectures and the Impact of Data De-duplication