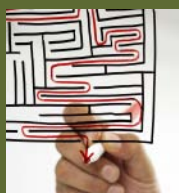


STORAGE

SEARCHSTORAGE.CO.UK

essential guide to

Data deduplication



Deduplication can save you big money in future capacity purchases, but there are lots of products, implementation methods and features to consider.



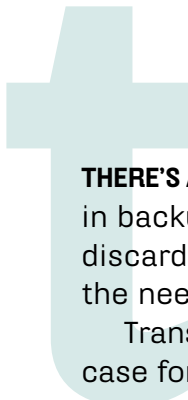
INSIDE

- Myriad variables in dedupe choice
- Backup dedupe product classes
- Selecting data for primary dedupe
- Why you need global deduplication
- Virtualisation: Changing the equation



Sorting through the data dedupe choices

Use this Essential Guide to learn what choices you'll face in a data deduplication project and to determine how to best implement deduplication technology in your data centre.



Myriad variables in dedupe choice

Backup dedupe product classes

Selecting data for primary dedupe

Why you need global deduplication

Virtualisation: Changing the equation

THERE'S ABSOLUTELY NO doubt that [data deduplication](#) can bring enormous savings in backup disk space. By assigning a unique identifier to chunks of data and discarding identical instances that come after them, data [deduping](#) can reduce the need for disk capacity by magnitudes of up to 90% or more.

Translate those kinds of capacity savings into cash, and there is a compelling case for applying the power of data deduplication to your backups and potentially even your primary data. After all, spending on disk capacity is the single largest item in a storage manager's budget (see our [2011 Purchasing Intentions survey](#)), so saving such large percentages makes dedupe a very attractive option.

But, as they say, there's no such thing as a free lunch, and the potential variables that can affect the outcome of data deduplication product selection and implementation are many, varied and often interlinked. That's because in the space of the few short years in which data deduping has gone from breakthrough technology to a plateauing of adoption levels, numerous methods of carrying out data deduplication have arisen in product form. In short, just about every data deduplication vendor does it differently.

Probably the first thing you'll need to assess is the nature of the backup regime into which you want to fit data deduplication. This is going to have a huge impact on the type of product(s) you buy.

If, for example, you aim to use dedupe as part of a remote office backup strategy or to back up virtual machine image files, this puts you in the market for [source deduplication](#). Source dedupe carries out its work at the source server(s), as the name suggests, which makes it suited to use cases where you want to reduce the data volume before transmitting it over the wire.

By contrast, [target deduplication](#) is suited to cases where you are happy to transmit all your data before it is processed by the dedupe engine.

Which of these methods is going to suit you depends on your use case,

the network bandwidth you possess and backup window time available.

Similar factors are at play when you come to decide whether to deduplicate data [inline or post-process](#), and here the choices begin to pile on top of one another. You can, for example, deduplicate inline at the source or at the target, and you can deduplicate at the target via inline or post-process methods.

These are just two of the major decision points on the road to purchasing and implementing data deduplication technology. Between these and a final decision, there are—or should be—numerous interlinked questions, including:

- Should you implement software or hardware dedupe?
- What types of data do you want to deduplicate?
- Can you [dedupe your primary data](#)?
- What complications are there of using deduplication in a virtualised server environment?
- Do you need global deduplication, and, if so, how many nodes will you need?

In this Essential Guide, we walk you through the decisions you'll need to make when starting down the road to data deduping. It's a complex process, but the rewards can be well worth it. ☺

Antony Adshead is bureau chief of [SearchStorage.co.UK](#).

Myriad
variables in
dedupe choice

Backup dedupe
product classes

Selecting data
for primary
dedupe

Why you
need global
deduplication

Virtualisation:
Changing
the equation



DEDUPE FOR BACKUP:

Product classes and decision points

Learn about source vs target deduplication, inline vs post-processing deduplication, how global dedupe fits into the equation, and how to determine which technology is right for your environment. BY TODD ERICKSON

DATA DEDUPLICATION IS a technique to reduce storage needs by eliminating redundant data in your backup environment. Only one copy of the data is retained on storage media, and redundant data is replaced with a pointer to the unique data copy. Dedupe technology typically divides data sets into smaller chunks and uses algorithms to assign each data chunk a hash identifier,

Myriad variables in dedupe choice

Backup dedupe product classes

Selecting data for primary dedupe

Why you need global deduplication

Virtualisation: Changing the equation

which it compares with previously stored identifiers to determine if the data chunk has already been stored. Some vendors use delta-differencing technology, which compares current backups with previous data at the byte level to remove redundant data.

Dedupe technology offers storage and backup administrators a number of benefits, including lower storage space requirements; more efficient disk space use; and less data sent across a WAN for remote backups, replication, and disaster recovery. Jeff Byrne, a senior analyst for the Taneja Group, said deduplication technology can have a rapid ROI. “In environments where you can achieve 70% to 90% reduction in needed capacity for your backups, you can pay back your investment in these dedupe solutions fairly quickly.”

"In environments where you can achieve 70% to 90% reduction in needed capacity for your backups, you can pay back your investment in these dedupe solutions fairly quickly."

—JEFF BYRNE, senior analyst, Taneja Group

While the overall data deduplication concept is relatively easy to understand, there are a number of different techniques used to accomplish the task of eliminating redundant backup data, and it’s possible that certain techniques may be better suited for your environment. So when you’re ready to invest in dedupe technology, consider the following technology differences and data deduplication best practices to ensure that you implement the best solution for your needs.

In the following, we’ll cover source vs target deduplication, inline vs post-processing deduplication, and the pros and cons of global deduplication.

SOURCE VS TARGET DEDUPLICATION

Deduping can be performed by software running on a server (the source) or in an appliance where backup data is stored (the target). If the data is deduped at the source, redundancies are removed before transmission to the backup target. “If you’re deduping right at the source, you get the benefit of a smaller image, a smaller set of data going across the wire to the target,” Byrne said. Source deduplication uses client software to compare new data blocks on the primary storage device with previously backed-up data blocks. Previously stored data blocks are not transmitted. Source-based deduplication uses less bandwidth for data transmission, but it increases server workload and could increase the amount of time it takes to complete backups.

Myriad variables in dedupe choice

Backup dedupe product classes

Selecting data for primary dedupe

Why you need global deduplication

Virtualisation: Changing the equation

Lauren Whitehouse, a senior analyst with the Enterprise Strategy Group, said source deduplication is well suited for backing up smaller and remote sites because increased CPU usage doesn't have as big of an impact on the backup process. Whitehouse said virtualised environments are also "excellent use cases" for source deduplication because of the immense amounts of redundant data in virtual machine disk (VMDK) files. However, if you have multiple virtual machines (VMs) sharing one physical host, running multiple hash calculations at the same time may overburden the host's I/O resources.

Most well-known data backup applications now include source dedupe, including Symantec's Backup Exec and NetBackup, EMC's Avamar, CA's ArcServe Backup, and IBM's Tivoli Storage Manager (TSM) with ProtecTier.

Target deduplication removes redundant data in the backup appliance—typically a NAS device or virtual tape library (VTL). Target dedupe reduces the storage capacity required for back-

up data but does not reduce the amount of data sent across a LAN or WAN during backup. "A target deduplication solution is a purpose-built appliance, so the hardware and software stack are tuned to deliver optimal performance," Whitehouse said. "So when you have large backup sets or a small backup window, you don't want to degrade the performance of your backup operation. For certain workloads, a target-based solution might be better suited."

Target deduplication may also fit your environment better if you use multiple backup applications and some do not have built-in dedupe capabilities. Target-based deduplication systems include Quantum's DXi series, IBM's TSM, NEC's Hydrastor series, FalconStor Software's File-interface Deduplication System (FDS) and EMC's Data Domain series.

"A target deduplication solution is a purpose-built appliance, so the hardware and software stack are tuned to deliver optimal performance."

—LAUREN WHITEHOUSE, senior analyst,
Enterprise Strategy Group

INLINE VS POST-PROCESSING DEDUPLICATION

Another option to consider is when the data is deduplicated. Inline deduplication removes redundancies in real time as the data is written to the storage target. Software-only products tend to use inline processing because the backup data doesn't land on a disk before it's deduped. Like source deduplication, inline increases CPU overhead in the production environment but limits the total amount of data ultimately transferred to backup storage. Asigra's Cloud

Myriad variables in dedupe choice

Backup dedupe product classes

Selecting data for primary dedupe

Why you need global deduplication

Virtualisation: Changing the equation

Backup and CommVault Systems' Simpana are software products that use inline deduplication.

Post-process deduplication writes the backup data into a disk cache before it starts the dedupe process. It doesn't necessarily write the full backup to disk before starting the process; once the data starts to hit the disk, the dedupe process begins. The deduping process is separate from the backup process so you can dedupe the data outside the backup window without degrading your backup performance. Post-process deduplication also allows you quicker access to your last backup. "So on a recovery that might make a difference," Whitehouse said.

However, the full backup data set is transmitted across the wire to the deduplication disk staging area or to the storage target before the redundancies are eliminated, so you have to have the bandwidth for the data transfer and the capacity to accommodate the full backup data set and deduplication process. Quantum's DXi-series backup systems use both inline and post-process technologies.

Content-aware or application-aware deduplication products that use delta-differencing technology can compare the current backup data set with previous data sets. "They understand the content of that backup stream, and they know the format that the data is in when the backup application sends it to that target device," Whitehouse said. "They can compare the workload of the current backup to the previous backup to understand what the differences are at a block or at a byte level." Whitehouse said delta-differencing-based products are efficient but they may have to reverse-engineer the backup stream to know what it looks like and how to do the delta differencing. Sepaton's DeltaStor system and ExaGrid Systems' DeltaZone architecture are examples of products that use delta-differencing technology.

"They understand the content of that backup stream, and they know the format that the data is in when the backup application sends it to that target device."

—LAUREN WHITEHOUSE, senior analyst,
Enterprise Strategy Group

GLOBAL DEDUPLICATION

Global deduplication removes backup data redundancies across multiple devices if you are using target-based appliances and multiple clients with source-based products. It allows you to add nodes that talk to one another across multiple locations to scale performance and capacity. Without global deduplication capabilities, each device dedupes just the data it receives. Some global systems can

Myriad variables in dedupe choice

Backup dedupe product classes

Selecting data for primary dedupe

Why you need global deduplication

Virtualisation: Changing the equation

be configured in two-node clusters, such as FalconStor Software's FDS High Availability Cluster. Other systems use grid architectures to scale to dozens of nodes, such as ExaGrid Systems' DeltaZone and NEC's Hydrastor.

The more backup data you have, the more global deduplication can increase your dedupe ratios and reduce your storage capacity needs. Global deduplication also introduces load balancing and high availability to your backup strategy and allows you to efficiently manage your entire backup data storage environment. Users with large amounts of backup data or multiple locations will benefit most from the technology. Most backup software providers offer products with global dedupe, including Symantec NetBackup and EMC Avamar, and data deduplication appliances, such as IBM's ProtecTier and Sepaton's DeltaStor offer global deduplication.

As with all data backup and storage products, when evaluating potential deduplication systems, the technologies used are only one factor you should consider. In fact, according to Whitehouse, the type of dedupe technologies vendors use is not the first attribute many administrators look at when investigating deduplication solutions. Price, performance, and ease of use and integration top deduplication shoppers' lists, Whitehouse said. Both Whitehouse and Byrne recommend first finding out if your current backup product has deduplication capabilities. If not, analyse your long-term needs and study the vendors' architectures to determine if they match your workload and scaling requirements. ☉

Todd Erickson is a news and features writer for TechTarget's Storage Media Group.

Myriad
variables in
dedupe choice

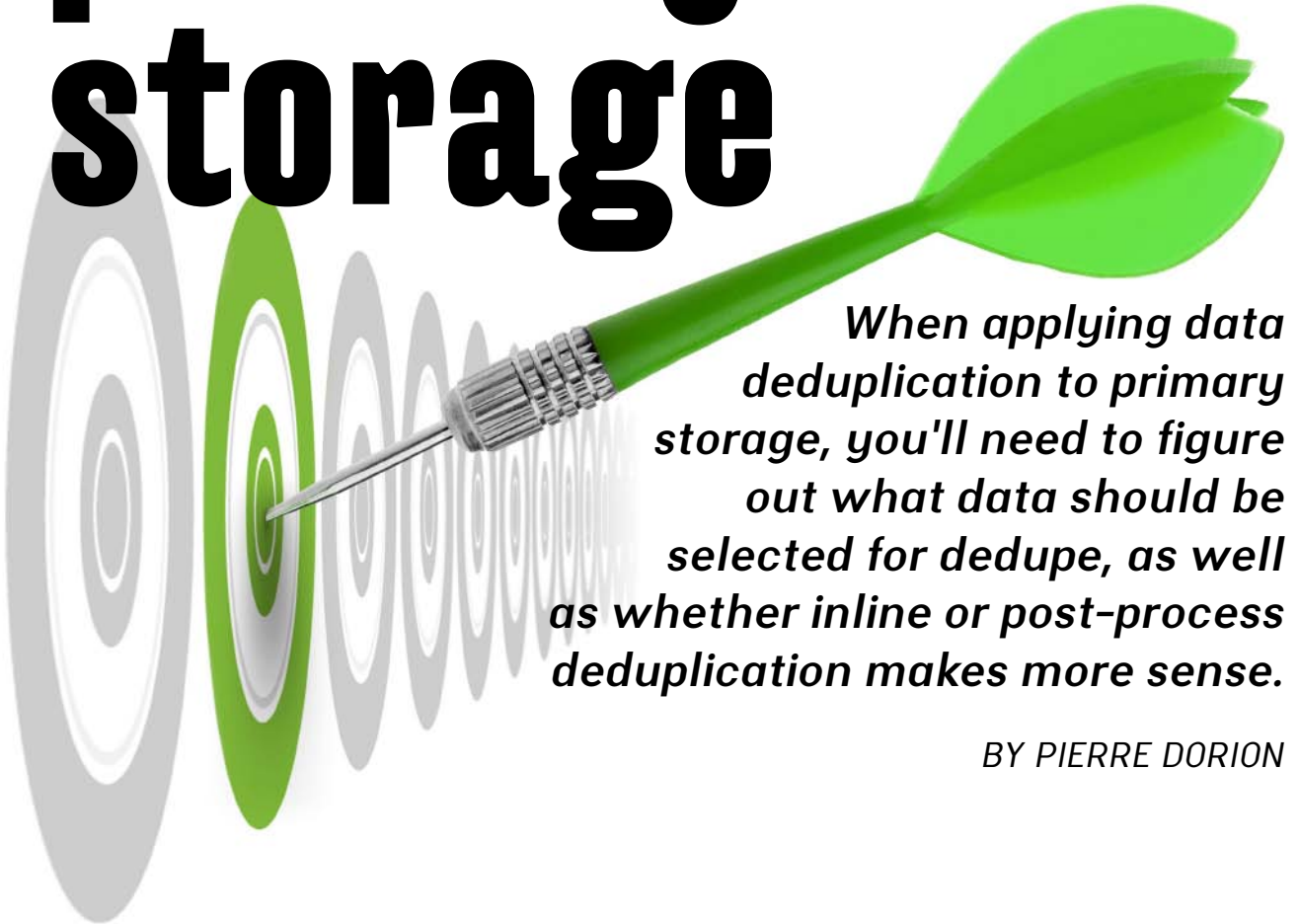
Backup dedupe
product classes

Selecting data
for primary
dedupe

Why you
need global
deduplication

Virtualisation:
Changing
the equation

Deduplication for primary storage



When applying data deduplication to primary storage, you'll need to figure out what data should be selected for dedupe, as well as whether inline or post-process deduplication makes more sense.

BY PIERRE DORION

DATA DEDUPLICATION HAS been a hot topic and a fairly common practice in disk-based backups and archives. Users' initial wariness seems to have given way to adoption, and a deeper focus on the technology has opened up more ways to leverage the benefits of deduplication. The next frontier for deduplication is in the realm of primary storage.

But how do you determine which primary data is a good fit for deduplication?

This is where the difference between structured and unstructured data comes into play. A database can be a significantly large file, subject to frequent and random reads or writes. For that reason, the majority of this data can be considered active. That means any processing overhead associated with deduplication could significantly impact I/O performance.

In comparison, if we examine data on a file server, we quickly see that only a small portion of files are written to more than once and usually only

[Myriad variables in dedupe choice](#)

[Backup dedupe product classes](#)

[Selecting data for primary dedupe](#)

[Why you need global deduplication](#)

[Virtualisation: Changing the equation](#)

for a short period of time after they were created. That means a very large portion of unstructured data is rarely accessed, making it a prime candidate for deduplication. This allows rules to be set to deduplicate data based on a “last access” time stamp. Shared storage for virtual servers or desktop environments also presents good opportunities for deduplication because many operating system files aren’t unique.

Other data selection criteria include format and data retention. Encrypted data and some imaging or streaming video files tend to yield poor deduplication results because of their random nature. In addition, data must reside in storage for some time to generate enough duplicate blocks to make deduplication worth the effort.

Encrypted data and some imaging or streaming video files tend to yield poor deduplication results because of their random nature.

Transient data that’s only staged to primary for a short period—such as message queuing systems or temporary log files—should be excluded.

Primary data that requires frequent access with optimum write performance won’t be a good fit for data deduplication. Data that’s difficult to deduplicate due to its format can be stored on a no-deduplication, lower-performance disk array to keep costs down. The remaining unstructured data that doesn’t require frequent or high-performance access (such as application or user file data) can be stored on a deduplication-enabled primary storage array.

Next up is the decision of inline vs post-process deduplication. Let’s say you’ve excluded encrypted data, streaming video and transient data, and you’ve established rules to determine “last access” and retention. You’ve identified primary data that’s a good fit for deduplication. This is when you’ll have to choose between inline or post-process deduplication.

The ability to deduplicate files once they’ve been inactive, or not accessed for some time, would favour post-process deduplication over inline because only selected data can be processed at a later time based on specific criteria and after it has been written to disk. This contrasts with inline deduplication, which would process all data as it’s written and may impact performance of certain types of data. Although inline deduplication processes all data immediately, it doesn’t always make it a poor choice for implementation on primary storage. It just means that storage tiering—determining where you need the best performance—is a crucial first step before deciding to apply deduplication technology to primary storage. ☉

Pierre Dorion is data centre practice director and a senior consultant with Long View Systems in Phoenix, Arizona, specialising in business continuity and disaster recovery planning services and corporate data protection.


Myriad variables in dedupe choice

Backup dedupe product classes

Selecting data for primary dedupe

Why you need global deduplication

Virtualisation: Changing the equation



Global deduplication: What it is and how it saves money

If you have target-based deduplication, you might be missing out on global deduplication's operational and acquisition cost benefits.

BY W CURTIS PRESTON

[Myriad variables in dedupe choice](#)

[Backup dedupe product classes](#)

[Selecting data for primary dedupe](#)

[Why you need global deduplication](#)

[Virtualisation: Changing the equation](#)

GLOBAL DATA DEDUPLICATION removes redundant data when backing up data to multiple deduplication devices. With global dedupe, when data is sent from one node to another, the second node recognises that the first node already has a copy of the data and doesn't make an additional copy.

While global dedupe isn't the only criterion in the complicated job of choosing a deduplication system, it certainly should be a high-priority one.

Unfortunately, support for global deduplication is missing (or meagre) in many target deduplication systems today. This insufficiency increases both the acquisition cost and the operational cost of such systems.

Before we drill down into why that's the case, let's explain exactly what global dedupe is and is not.

HASH-BASED VS DELTA-BASED DEDUPLICATION SYSTEMS

Consider the hash-based deduplication systems shown in “Hash-based and delta-based deduplication products” below. Hash-based deduplication slices data into chunks, creates a hash for that chunk and then performs a hash table lookup to see if it has ever seen that hash before. Delta-based deduplication systems compare an entire backup (eg, saveset, image, dump) to another backup that it is similar to. For example, they look for the delta between the most recent full backup of Elvis to the previous full backup of Elvis.

Vendors of hash-based dedupe products often tout how they compare every incoming backup to every other backup they have ever seen, saying that their dedupe is more “global.” However, delta-based vendors tout how their dedupe is more granular. Hashing is more efficient with dissimilar data; deltas are more efficient with similar data. Comparisons of the deduplication ratio of the two methods often result in a draw.

“HASH” DOES NOT EQUAL “GLOBAL”

Since hash-based vendors consider their dedupe to be more global, some representatives of those vendors refer to what their products do as global dedupli-

Hash-based and delta-based deduplication products		
Vendor	Hash/Delta	Global dedupe support
EMC Data Domain Global Deduplication Array	Hash	2 nodes, NetBackup OST only
EMC Data Domain (All other products)	Hash	None
ExaGrid Systems EX Series	Delta	10 nodes
FalconStor VTL	Hash	4 nodes
IBM ProtecTier	Delta	2 nodes
NEC Hydrastor	Hash	55 nodes
Quantum DXi	Hash	None
Sepaton	Delta	8 nodes

Myriad variables in dedupe choice

Backup dedupe product classes

Selecting data for primary dedupe

Why you need global deduplication

Virtualisation: Changing the equation

cation. The matter is further complicated by EMC Data Domain, which often uses *global compression* to describe what it does. (In all fairness, Data Domain was using *deduplication* long before it came into common usage.) However, this is not what we mean when we use *global deduplication*.

The concept of global dedupe comes into play when you purchase multiple dedupe appliances (ie, nodes). If a vendor supports global dedupe (also known as multi-node dedupe), it will have a cluster, or grid, of multiple nodes that work together as one. Data sent to one node in the grid is compared with previous data sent to that appliance and with data sent to any other node in that grid. This allows the customer to load balance backups across all nodes in the grid, while being assured that data common to more than one node will be stored only on one node. All known source deduplication systems support global dedupe; it is with target deduplication systems that this discussion is most relevant.

As shown in “Hash-based vs delta-based deduplication products,” most target deduplication vendors offer some level of global deduplication. The more nodes a vendor supports in a globally deduped cluster, the easier things will be for their customers. NEC supports 55 nodes, ExaGrid Systems supports 10 nodes, Sepaton supports eight nodes, FalconStor Software supports four nodes, IBM supports two nodes, and EMC Data Domain supports two nodes but only with its fastest system (the DD890) and only for NetBackup OST customers. Quantum is the only target dedupe vendor to offer no support for global deduplication.

If one can treat all of their global dedupe nodes as one grid, and load balance all of their backups across all nodes in that grid, configuration of the backup system should be extremely easy.

HOW GLOBAL DEDUPLICATION LOWERS OPERATIONAL AND ACQUISITION COSTS

The easiest thing to understand is how global dedupe lowers operational costs. If one can treat all of their global dedupe nodes as one grid, and load balance all of their backups across all nodes in that grid, configuration of the backup system should be extremely easy. However, if you purchase multiple nodes that don't work together, you must create and maintain multiple subsets of your backups, and point each subset at only one node. The creation and maintenance of these backup subsets makes the backup system harder and costlier to maintain, and the requirement to point each subset to only one node makes the backup system less reliable. The latter also increases management cost due to manual workarounds that must be performed in the case of a node failure.

Myriad variables in dedupe choice

Backup dedupe product classes

Selecting data for primary dedupe

Why you need global deduplication

Virtualisation: Changing the equation

The existence of global deduplication also reduces acquisition cost in three ways. First, consider the differences between the type of nodes that are used to build ExaGrid, FalconStor, NEC and Sepaton, and the nodes that are used to build Data Domain, IBM and Quantum systems. The former are able to use less expensive nodes because their performance is not capped at one or two nodes. The latter tend to use much more expensive CPUs and RAM because the power of each node is what drives their performance numbers. Buying the latest and greatest CPUs and components is more expensive than buying the previous generation.

The second way that global deduplication reduces acquisition cost is by allowing a customer to buy today what they need today, and buy tomorrow what they need tomorrow—without throwing away anything they bought yesterday. Customers buying target dedupe systems that do not have global dedupe (eg, Quantum) or from vendors that offer it only for two nodes (eg, IBM, EMC Data Domain) or only for certain customers (eg, EMC Data Domain) are faced with a very different proposition. Because they cannot grow their current system by simply adding more nodes, they are forced to do one of three things.

Their first choice is to buy today what they need tomorrow. For example, a customer that needs a Quantum DXi 7500 today will probably be advised to buy a DXi 8500 instead, even if the customer won't need its performance or capacity for a year or more. In the world of ever-decreasing cost of disk, buying today what you need tomorrow will always cost you more money. Data Domain customers have a slightly different choice. If a customer needs a DD690 today but grows into a DD890 in a year or so, it can purchase just the head of the DD890 appliance and replace the DD690 head with the DD890 head while keeping the disk it already purchased. The DD690 head will, of course, go to waste. Finally, a customer could “upgrade” any of these systems by simply purchasing another system and using both of them side by side. Besides the operational cost mentioned above of performing backups with multiple discrete nodes that don't work together, there is also the acquisition cost created by the difficulty of properly sizing each node. Since it is impossible to get it right, and under-sizing such a system would be even worse, resulting in even more waste, customers tend to oversize such systems.

The final way that global deduplication helps reduce acquisition cost is by ensuring all backups are compared with previous backups—regardless of which node they were sent to. Since backup systems are always in a state of flux and customers must constantly change their backup policies in order to meet various needs, it would be nice if at least they do not have to worry about messing up their deduplication ratio by changing their backup configuration. Global deduplication should ultimately result in the best deduplication ratio; it also reduces the amount of disk the customer must purchase. ☺

W Curtis Preston is an independent backup expert. He is the webmaster of Backup-Central.com and the author of “Backup and Recovery” and “Using SANs and NAS.”

Myriad variables in dedupe choice

Backup dedupe product classes

Selecting data for primary dedupe

Why you need global deduplication

Virtualisation: Changing the equation



Virtualisation's impact on deduplication

Learn which dedupe technique makes the most sense in a virtualised server environment and what to watch out for when implementing the technology.

MORE AND MORE businesses are showing an interest in implementing data deduplication technology in their virtualised environments. In this Q&A with Jeff Boles, senior analyst with the Taneja Group, learn whether target or source deduplication is better for a virtualised environment, what to watch out for when using dedupe for virtual servers, and what VMware's vStorage APIs have brought to the scene.

SEARCHSTORAGE.CO.UK: **Have you seen more interest in data deduplication technology among organisations that have deployed server virtualisation? And, if so, can you explain what's driving that interest and the benefits people might see from using dedupe when they're backing up virtual servers?**

BOLES: Absolutely. There's lots of interest in using deduplication for virtualised environments because there's so much redundant data in virtual server environments. Over time, we've become more disciplined as IT practitioners in how we deploy virtual servers.

We've done something we should've done a number of years ago with our general infrastructures, and that's creating a better separation of our core OS

Myriad variables in dedupe choice

Backup dedupe product classes

Selecting data for primary dedupe

Why you need global deduplication

Virtualisation: Changing the equation

data from our application data. And consequently, we see virtualised environments that are following best practices today with these core OS images that contain most operating system files and configuration stuff. They separate that data out from application and file data in their virtual environments, and there are so many virtual servers that use very similar golden image files with similar core OS image files behind a virtual machine. So you end up with lots of redundant data across all those images. If you start deduplicating across that pool you get even better deduplication ratios even with simple algorithms than you do in a lot of non-virtualised production environments. There can be lots of benefits from using deduplication in these virtual server environments just from a capacity utilisation perspective.

What kind of data deduplication is typically being used for this type of application? Do you see source dedupe or target, and does one have benefits over the other?

There are some differences in data deduplication technologies today. You can choose to apply it in two places—either the backup target or you can choose to apply it at the source through the use of technologies like Symantec’s Pure-Disk, EMC Avamar or some of the other virtualisation-specialised vendors out there today.

Source deduplication is being adopted more today than it ever has before and it’s particularly useful in a virtual environment. First you have a lot of contention for I/O in a virtualisation environment. Generally, when folks start virtualising, they try to stick with the same approach, and that’s with a backup agent that’s backing up data to an external media server to a target, following the same old backup catalogue jobs, and doing it the same way they were in physical environments. But you end up packing all that stuff in one piece of hardware that has all these virtual machines (VMs) on it, so you’re writing a whole bunch of backup jobs across one piece of hardware. You get a whole lot of I/O contention, especially across the WANs, and more so across LANs. But any time you’re going out to the network you’re getting quite a bit of I/O bottlenecking at that physical hardware layer. So the traditional backup approach ends up stretching out your backup windows and messes with your recovery time objectives [RTOs] and recovery point objectives [RPOs] because everything is a little slower going through that piece of hardware.

So source deduplication has some interesting applications because it can chunk all that data down to non-duplicate data before it comes off the VM. Almost all of these agent approaches that are doing source-side deduplication push out a very continuous stream of changes. You can back it up more often because there’s less stuff to be pushed out, and they’re continually tracking changes in the background; they know what the deltas are, and so they can minimise the data they’re pushing out.

Myriad variables in dedupe choice

Backup dedupe product classes

Selecting data for primary dedupe

Why you need global deduplication

Virtualisation: Changing the equation

Also, with source-side deduplication you get a highly optimised backup stream for the virtual environment. You're pushing very little data from your VMs, so much less data is going through your physical hardware layer, and you don't have to deal with those I/O contention points, and consequently you can get much finer-grained RTOs and RPOs and much smaller backup windows in a virtual environment.

Does data deduplication introduce any complications when you use it in a virtualised environment? What do people have to look out for?

When you're going into any environment with a guest-level backup and pushing full strings of data out, you can end up stretching out your backup windows. The other often-overlooked dimension of deduplicating behind the virtual server environment is that you are dealing with lots of primary I/O that's pushed into one piece of hardware now in a virtual environment. You may have many failures behind one server at any point in time. Consequently, you may be pulling a lot of backup streams off of the deduplicated target or out of the source-side system. And, you may be trying to push that back on the disk or into a recovery environment very rapidly.

Dedupe can have lots of benefits in capacity but it may not be the single prong that you want to attack your recovery with because you're doing lots of reads from this deduplicated repository. Also, you're pulling a batch of disks simultaneously in many different threads. There may be 20 or 40 VMs behind one piece of hardware, and you're likely not going to get the recovery window that you want—or not the same recovery window you could've gotten when pulling from multiple different targets into multiple pieces of hardware. So think about diversifying your recovery approach for those “damn, my virtual environment went away” incidents. And think about using more primary protection mechanisms. Don't rely just on backup, but think about doing things like snapshots, where you can fall back to the latest good snapshot in a much narrower time frame. You obviously don't want to try to keep 30 days of snapshots around, but have something there you can fall back to if you've lost a virtual image, blown something up, had a bad update happen or something else. Depending on the type of accident, you may not want to rely on pulling everything out of the dedupe repository, even though it has massive benefits for optimising the capacity you're using in the backup layer.

A few years ago, VMware released the vStorage APIs for Data Protection and some other APIs as a part of vSphere. Are you seeing any developments in the deduplication world taking advantage of those APIs?

The vStorage APIs are where it started getting interesting for backup technology in the virtual environment. We were dealing with a lot of crutches before then, but the vStorage APIs brought some interesting technology to the table. They

Myriad variables in dedupe choice

Backup dedupe product classes

Selecting data for primary dedupe

Why you need global deduplication

Virtualisation: Changing the equation

have implications for all types of deduplication technology, but I think they made particularly interesting implications for source-side deduplication, as well as making source-side more relevant. One of the biggest things about vStorage APIs was the use of Changed Block Tracking (CBT); with that you could tell what changed between different snapshots of a VM image. Consequently, it made this idea of using a proxy very useful inside a virtual environment, and source-side has found some application there, too. You could use a proxy with some source-side technology so you can get the benefits of deduplicating inside this virtual environment after taking a snapshot, but it only deduplicates the changed blocks that have happened since the last time you took a snapshot.

Some of these vStorage API technologies have had massive implications in speeding up the time in which data can be extracted from a virtual environment. Now you can recognise what data has changed between a given point in time and you can blend your source-side deduplication technologies with your primary virtual environment protection technologies and get the best of both worlds. The problem with proxies before was that they were kind of an all-or-nothing approach. You use the snapshot, and then you come out through a proxy in the virtual environment through this narrow bottleneck that will make you do a whole bunch of steps and cause compromises with the way you were getting data out of your virtual environment.

You could choose to go with source-side, but you have lots of different operations going on in your virtual environment. Now you can blend technologies with the vStorage APIs. You can use a snapshot plus source-side against it and get rapid extraction inside your virtual environment, and a finer application of the deduplication technology that's still using source-side to this one proxy pipe, which mounts up this snapshot image, deduplicates stuff and pushes it out of the environment. vStorage APIs have a lot of implications for deduping an environment and blending deduplication technologies with higher-performing approaches inside the virtual environment. And you should check with your vendors about what potential solutions you might acquire out there in the marketplace to see how they implemented vStorage APIs in their products to speed the execution of backups and to speed the extraction of backups from your virtual environment. ☺

Myriad
variables in
dedupe choice

Backup dedupe
product classes

Selecting data
for primary
dedupe

Why you
need global
deduplication

Virtualisation:
Changing
the equation