Managing the information that drives the enterprise

STORAGE

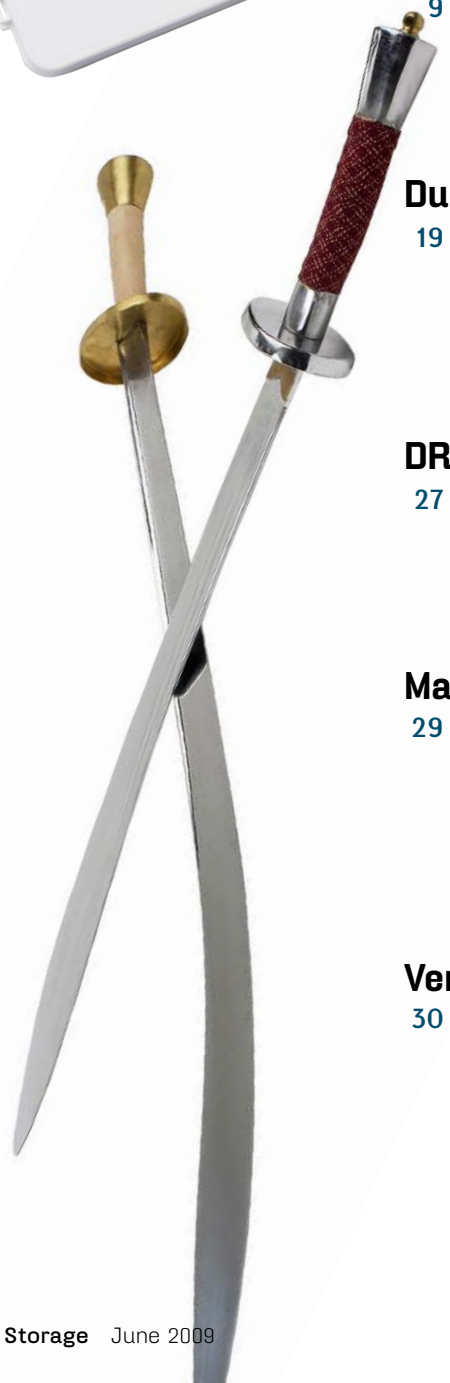
Vol. 8 No. 4 June 2009

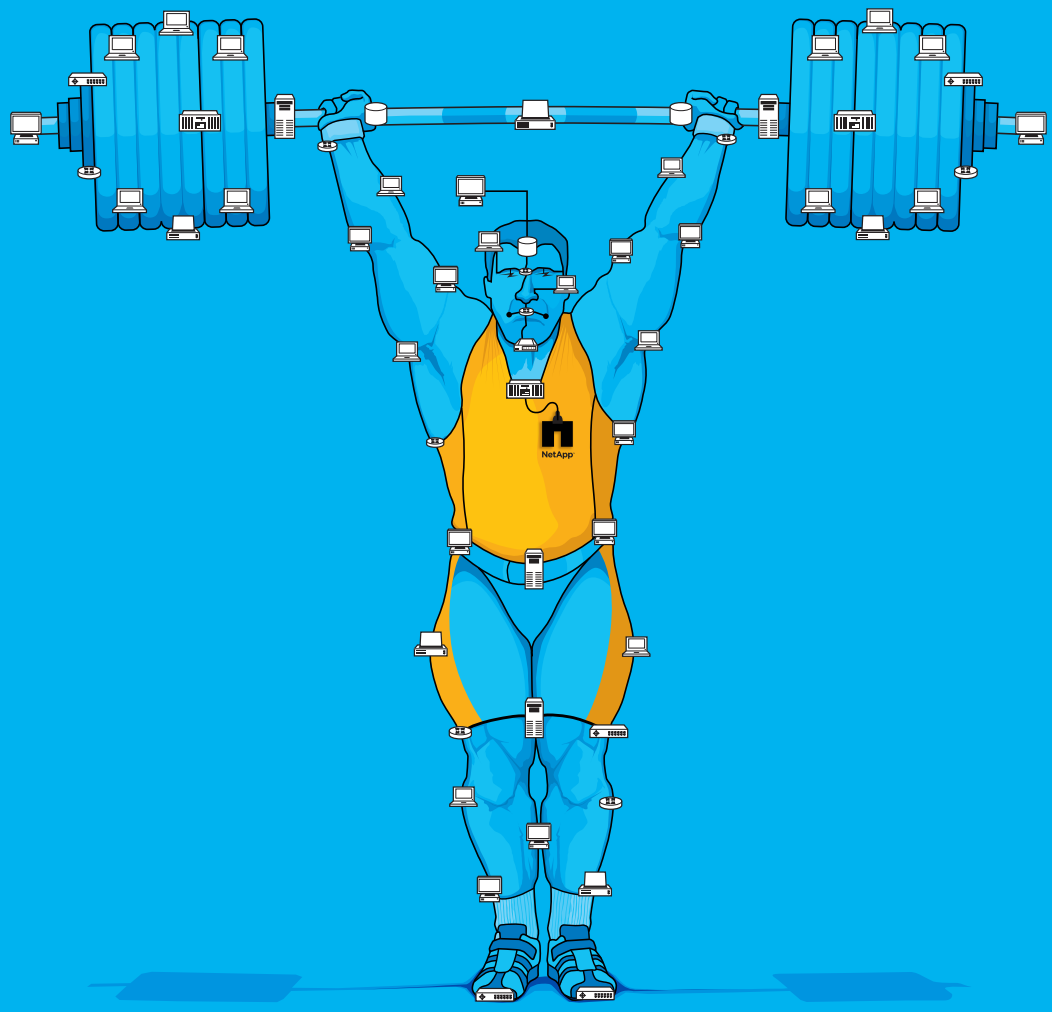
Heal Thyself

No more disk drive failures or endless RAID rebuilds—that's the promise of a new breed of self-healing storage systems. page 9

ALSO INSIDE

TechTarget
The Technology Media ROI Experts





With NetApp® at the heart of your business, you can

EFFORTLESSLY HOLD YOUR DATA

at a fraction of the cost and footprint.

Imagine storage and data management solutions smart enough to support the data you need, and not a lot of dead weight. It's possible when you partner with NetApp. Our industry-leading solutions use deduplication and other space-saving technologies to help you store data efficiently and reduce your footprint by 50% or more. So you can manage exponential growth while minimizing your storage investment—all with the support of a team that will exceed your expectations. See how we can help your business go further, faster. Find out how you can use 50% less storage, guaranteed,* at netapp.com/efficiency.

NetApp™
Go further, faster

Vendors must adapt to a new environment

Spiraling capacity growth, a deteriorating economy—are there any vendors ready to step up and deliver the storage efficiency tools we need now?

WHAT DOES IT TAKE to light a fire under storage vendors? That might sound like some kind of nerdy joke, but there's no punch line here. The whole bigger, faster (and more expensive) thing is bred into storage vendors—it's part of their DNA. But if storage managers want to survive in that world, they're the ones who will have to adapt. So what will make the vendors change? How about a recession coupled with unrelenting data growth?

That's how evolution usually works. When profound changes in the environment occur, the existing life forms have to adapt to a new habitat or perish. But storage vendors are the mammoths of this ecosystem, lumbering along on their chosen paths and seeming to take little notice of the changes going on around them.

When thin provisioning started showing up as a check-box on storage system RFPs, vendors finally got the message.

Storage managers, though, are survivors. It's part of the job description. They might not have had to deal with such intimidating economic issues before, but for the most part they've been there, done that.

That should spell opportunity, but many storage vendors seem a bit slow to recognize that a new world order—even if it's only a temporary one—could make their previously popular products seem old-fashioned and out of step with the times. They were slow to embrace some of the efficiency technologies that have become de rigueur in many storage shops today, like thin provisioning. 3PAR was the thin provisioning pioneer, but for a long time the technology was seen more as an exotic option than a key part of a storage management system. Storage managers, on the other hand, saw the value right away; thin provisioning offered a quick and easy solution to the age-old problem of overprovisioned and underused disk capacity.

A lot of storage vendors, however, translated thin provisioning into fewer disk sales, which is true to a certain extent. But thin provisioning doesn't eliminate the need for additional capacity; it forestalls those purchases by giving storage managers a tool to better use their installed

capacity. The disk sales will come, just maybe not as soon or as quickly as in the past. When thin provisioning started showing up as a checkbox on storage system RFPs, vendors finally got the message—but it was the user community that forced their hand.

Not all vendors are sluggish in their response to changing environments. Dave Raffo, our senior news director, recently blogged about how Compellent bucked this year's quarterly report trend by posting an increase in its profits; a modest increase, to be sure, but up is always better than down. I can't say for sure what accounts for Compellent's success, but I'd guess that storage management efficiency played a part. The company builds efficiency tools into its systems in the form of automated data migration that makes storage tiering easier.

Kudos, too, for NetApp, a company busy proving that deduplication does, indeed, have a place in primary storage systems. And they're doing that by essentially giving it away free with their operating system.

Perhaps the greatest opportunity for storage vendors today is power efficiency.

Perhaps the greatest opportunity for storage vendors today is power efficiency. They've been talking a good game for a couple of years by slapping "green" labels on their products, but outside of a handful of vendors tinkering with MAID or MAID-like technologies, most of the green stuff we've seen seems more like marketing mumbo-jumbo than real energy savings.

That might have been OK a year or two ago when most storage managers had higher priorities than cutting power costs. Today, however, power consumption is squarely within most storage managers' sights.

But the same problem persists: Storage vendors conveniently use their own metrics to rationalize their products' green-ness. We get kilowatts per I/O, kilowatts per terabyte, kilowatts per tile and a host of other predictably favorable ratings. But no common and comparable metric exists simply because storage vendors don't want you to do an apples-to-apples comparison.

In this age of opportunity, some enterprising storage vendor may have the nerve to step up and produce a set of meaningful metrics. All sorts of consortia and industry groups have been waving the green flag for years now, with very little to show for all the hoopla. But wouldn't it be nice, for example, if a vendor with a real power conservation story to tell came out with a broad scale of power consumption numbers that matched every other vendor's particular twist on the topic? You'd get your apples-to-apples comparisons, some truth in advertising and maybe just one more step in the evolution toward storage efficiency. ⊙

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

** Click here for a sneak peek at what's coming up in the July/August issue.*

Managing the information that drives the enterprise

STORAGE

The Essential Guide to Data Deduplication

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.

Exclusive Extra for Storage Subscribers

The Essential Guide to Data Deduplication is a brand new expert resource brought to you by the Storage magazine team. This guide outlines the fundamental aspects you need to know about data dedupe and provides a number of tips and best practices from some of the leading experts in the field. Get a handle on data deduplication with 24/7 access to this new Essential Guide in an easy-to-access, easy-to-read, easy-to-print online format.

Get Your FREE Copy Today: **www.SearchDataBackup.com/Dedupe_Guide**

COMING IN JULY/AUGUST

Backup Applications and VMware

Most backup applications have announced some form of integration with VMware to make backups easier, cheaper and more granular. This article provides best practices for virtual server backups and looks at the various products to determine how well integrated they are and what benefits users should expect from that integration.

Quality Awards IV: Tape Libraries

The fourth annual *Storage* magazine/SearchStorage.com Quality Awards focus on user satisfaction with tape libraries. We analyze the results of our exclusive survey where users describe their experiences related to their tape libraries' initial quality, features and reliability, as well as the technical support and sales competence of library vendors.

Virtual Desktops and Storage

The desktop is the last frontier of virtualization. We examine just how significantly storage systems will be affected as desktops get virtualized. We'll address issues such as whether VDI will require more capacity, if certain types of storage systems are better suited for VDI, and if thin provisioning and deduplication will help or hinder a VDI installation.

And don't miss our monthly columns and commentary, or the results of our Snapshot reader survey.

***Storage* magazine**
117 Kendrick Street,
Suite 800
Needham, MA 02494
editor@storagemagazine.com

Subscriptions:
www.SearchStorage.com

Managing the information that drives the enterprise

STORAGE

The Essential Guide for DR Planning

Developing and Refining a DR Plan

Preparing a disaster recovery (DR) plan can be a complex and resource-consuming project. Use these tips to simplify the process and put together a predictable and comprehensive DR strategy.

Exclusive Extra for Storage Subscribers

Preparing a disaster recovery (DR) plan can be a complex and resource-consuming project. Fortunately, The Essential Guide to Developing and Refining a DR Plan highlights a number of expert tips to simplify the process and put together a predictable and comprehensive DR strategy. Discover how to take the complexity out of DR planning with 24/7 access to this new Essential Guide in an easy-to-access, easy-to-read, easy-to-print online format.

Get Your FREE Copy Today: **www.SearchDisasterRecovery.com/DR_Planning**

Storage, heal thyself

Several storage system vendors claim their products can detect and repair hard disk problems automatically. Here's how they do it and the low-down on how well they work.

By Marc Staimer

FUNDAMENTAL CHANGE in the basic building blocks of storage is occurring, one that's as groundbreaking today as RAID was when it was introduced 20 years or so ago. The revolutionary development is commonly referred to as "autonomic self-healing storage," and it promises greater reliability from disk systems than ever before.

Autonomic self-healing storage might sound more like a trumped-up term than a fundamental change. After all hasn't it been around for a while in the form of RAID, redundant array of independent nodes (RAIN), snapshots, continuous data protection (CDP) and mirroring?

If you define self-healing as the ability to restore from a failure situation, you'd be right. All of those familiar technologies are designed to restore data from a failure situation. But to be a bit more precise, those technologies are actually self-healing data, not self-healing storage. They restore data when there's a storage failure and mask storage failures from the apps—they don't restore the actual storage hardware.

Self-healing storage is more accurately defined as transparently restoring both the data and storage from a failure. That might seem like splitting hairs, but it's not. It's the difference between treating the symptoms and fixing the cause.

WHAT HAPPENS WHEN A DISK FAILS

The lowest common denominator in standard storage systems today is the hard disk drive (HDD). The hard disk drive is the only electro-mechanical device in the storage system, and it has the highest probability of failure or lowest mean time between failures (MTBF) (see "MTBF: The odds of failure," below). It's well documented that the HDD component is the

MTBF: THE ODDS OF FAILURE

A DISK MANUFACTURER'S hard disk drive (HDD) mean-time between failures (MTBF) rating enables you to forecast the useful operational life of a hard disk drive. When there are a lot of HDDs in the system, the probability of HDD failures increases. The general formula for calculating average time between drive failures within a system is as follows:

$$\frac{\text{HDD MTBF}}{\text{Number of HDDs in the storage system}}$$

Using the manufacturers' MTBF numbers (approximately 1.5 million hours for enterprise-class Fibre Channel and SAS HDDs, and approximately 600,000 hours MTBF for SATA HDDs), a system with 240 enterprise drives should expect a hard disk drive failure every 260 days: 1,500,000/240 = 6,250 hours or about 260 days (roughly two HDDs per year or approximately a 0.8% replacement rate). If the HDDs are SATA, the system should expect a HDD failure every 104 days (roughly four HDDs per year or approximately a 1.67% replacement rate).

Unfortunately, manufacturer MTBF numbers don't reliably reflect real-world MTBFs. The Computer Science department at Carnegie Mellon University in Pittsburgh ran stress tests of 100,000 Fibre Channel, SAS and SATA hard disk drives. Their published testing results determined that a typical drive (Fibre Channel, SAS or SATA) has a realistic MTBF of approximately six years or 52,560 hours. Using Carnegie Mellon's MTBF numbers, a storage system with 240 HDDs can expect a drive failure approximately every nine to 10 days (approximately 40 HDDs per year or an annual replacement rate of 16.67%).

Achilles' heel of a storage system.

The unprecedented growth of data in most companies has led to an explosion of storage systems and hard disk drives. It's statistically proven that as HDDs proliferate so will the number of hard disk drives failures, which can lead to lost data. Analyzing what happens when a HDD fails illustrates the issue:

1) A hard disk dive fails
2) The drive must be physically replaced, either manually or from an online pool of drives.
3) Depending on the RAID set level, the HDD's data is rebuilt on the spare:
   a) RAID 1/3/4/5/6/10/60 all rebuild the hard disk drive's data, based on parity
   b) RAID 0 can't rebuild the HDD's data
4) The time it takes to rebuild the HDD's data depends on the hard disk drive's capacity, speed and RAID type.
   a) A 1 TB 7,200 rpm SATA HDD with RAID 5 will take approximately 24 hours to 30 hours to rebuild the data, assuming the process is given a high priority.
   b) If the rebuild process is given a low priority and made a background task to be completed in off hours, the rebuild can take as long as eight days. The RAID group is subject to a higher risk of a second disk failure or non-recoverable read error during the rebuild, which would lead to lost data. This is because the parity must read every byte on every drive in the RAID group to rebuild the data. (Exceptions are RAID 6, RAID 60 and NEC Corp. of America's D-Series RAID 3 with double parity.)
      (1) SATA drives typically have a rated non-recoverable read error rate of 10^{14}: roughly 1 out of 100,000,000,000,000 bits will have a non-recoverable read error. This means that a seven-drive RAID 5 group with 1 TB SATA drives will have approximately a 50% chance of failing during a rebuild resulting in the loss of the data in that RAID group.
      (2) Enterprise-class drives (Fibre Channel or SAS) are rated at 10^{15} for non-recoverable read errors, which translates into less than a 5% chance of the RAID 5 group having a failure during a rebuild.
      (3) RAID 6 eliminates the risk of data loss should a second HDD fail. You pay for that peace of mind with decreased write performance vs. RAID 5, and an additional parity drive in the RAID group.

Eventually, the hard disk drive is sent back to the factory. Using the

MTBF example in "MTBF: The odds of failure," p. 10, this suggests that there'll be approximately 40 HDD "service events" per year.

Most storage admins might be surprised by what happens when a HDD is sent back to the factory. After being run through the factory's failure analysis process, the results for the vast majority of failed hard disk drives (somewhere between 67% and 90%) will be "no failure found"—the HDD is fine. But the service event still took place and the RAID data rebuild still had to occur. That's a lot of operational hassle for "no trouble found."

UNDETECTED DATA CORRUPTION

Another problem with HDDs that's rarely mentioned but actually quite prevalent is "silent data corruption." Silent data corruptions are storage errors that go unreported and undetected by most storage systems, resulting in corrupt data being provided to an application with no warning, logging, error messages or notification of any kind.

Most storage systems don't detect these errors, which occur on average with 0.6% of SATA HDDs and .06% of enterprise HDDs over 17 months (from "An Analysis of Data Corruption in the Storage Stack," L.N. Bairavasundaram et al., presented at FAST '08 in San Jose, Calif.). Silent data corruption occurs when the RAID doesn't detect data corruption errors such as misdirected or lost writes. It also occurs with a torn write—data that's partially written and merges with older data, so the data ends up part original data and part new data. Because the hard disk drive doesn't recognize the errors, the storage system isn't aware of it either so there's no attempt at a fix.

Inside ANSI's T10 DIF spec

The American National Standards Institute's (ANSI) T10 DIF (Data Integrity Field) specification calls for data to be written in blocks of 520 bytes instead of the current industry standard 512 bytes. The eight additional bytes or "DIF" provide a super-checksum that's stored on disk with the data. The DIF is checked on every read and/or write of every sector. This makes it possible to detect and identify data corruption or errors, including misdirected, lost or torn writes. ANSI T10 DIF provides three types of data protection:

- Logical block guard for comparing the actual data written to disk
- Logical block application tag to ensure writing to the correct logical unit (virtual LUN)
- Logical block reference tag to ensure writing to the correct virtual block

When errors are detected, they can then be fixed by the storage system's standard correction mechanisms.

AUTONOMIC SELF-HEALING SYSTEMS

Among this new breed of systems, some tackle end-to-end error detection and correction, including silent data corruption. Other systems take the same approach, but add sophisticated algorithms that attempt to "heal-in-place" failed HDDs before requiring a RAID data rebuild. A final group of systems matches those capabilities and ups the ante with the new concept of "fail-in-place" so that in the rare circumstance when a HDD truly fails (i.e., it's no longer usable), no service event is required to replace the hard disk drive for a RAID data rebuild.

Among this new breed of systems, some tackle end-to-end error detection and correction, including silent data corruption.

END-TO-END ERROR DETECTION AND CORRECTION

Vendors and products offering end-to-end error detection and correction include DataDirect Networks Inc.'s Silicon Storage Architecture (S2A) with its QoS and SATAssure; EMC Corp.'s Symmetrix DMX-4 with its Double Checksum; NEC's D-Series support of the American National Standards Institute's new T10 DIF (Data Integrity Field) standard for enterprise Fibre Channel or SAS HDDs, and their proprietary Extended Data Integrity Feature (EDIF) for SATA hard disk drives; Panasas Inc.'s ActiveStor with Vertical Parity for SATA HDDs; Sun Microsystems Inc.'s Zettabyte File System (ZFS)-based systems when volumes are mirrored; and Xiotech Corp.'s Emprise 5000 (aka Intelligent Storage Element), which is also based on the T10 DIF standard (see "Self-healing storage products," p. 14).

T10 DIF is a relatively new standard and only applies to SCSI protocol HDDs (SAS and Fibre Channel) (see "Inside ANSI's T10 DIF spec," p. 12). The T10 DIF standard is being incorporated into quite a few storage systems scheduled for release in 2009 and 2010. However, there's no standard spec for end-to-end error detection and correction for SATA hard disk drives at this time. That's why DataDirect Networks, EMC and NEC devised their own SATA end-to-end error detection and correction methodologies.

DataDirect Networks' S2A SATAssure software does a Reed-Solomon error-correction calculation on every read operation and then compares HDD data to parity to ensure data consistency. SATAssure repairs the data if an inconsistency is detected, then passes it back to the requesting app and rewrites it to the HDD. All of this happens in real-time. EMC DMX-4 uses a double checksum that's very similar to Oracle Corp.'s industry-proven double checksum that minimizes database corruptions.

NEC's D-Series EDIF is modeled on ANSI T10 DIF. The difference is that EDIF is specifically modified for SATA's Integrated Disk Electronics (IDE) protocol.

Panasas' Vertical Parity is designed to maintain individual hard disk drive reliability. Vertical Parity isolates and repairs (using redundant information in the horizontal RAID stripe) torn, lost or misdirected writes on SATA HDDs at the disk level before they're seen by the RAID array.

Sun's ZFS is now used in several unified storage systems (Sun's 4500 and 7000 Series, and the new OnStor Inc. Pantera LS 2100). ZFS utilizes its own end-to-end error-detection algorithms to sniff out silent data corruption. It requires mirrored volumes and corrects the detected silent data corruption by copying the uncorrupted data from the good volume.

DOES END-TO-END ERROR CORRECTION WORK?

User evidence over the past 18 months suggests that HDD error-correction methods work. Interviews with IT organizations storing petabytes of storage (where silent data corruption is statistically more likely to be noticed)

SELF-HEALING STORAGE PRODUCTS

	Applies to HDD format	Atrato Inc. Velocity1000	DataDirect Networks Inc. Silicon Storage Architecture (S2A)	EMC Corp. Symmetrix DMX-4	NEC Corp. of America D-Series	Panasas Inc. ActiveStor	Sun Microsystems Inc. Zettabyte File System (ZFS); Sun 4500 and 7000 Series; OnStor Inc. Pantera LS 2100	Xiotech Corp. Emprise 5000 (ISE)
End-to-end error detection and correction	Fibre Channel (FC) 3.5"		✔	✔	✔		✔	✔
	SAS 3.5"		✔				✔	
	SATA 3.5"		✔	✔	✔	✔	✔	FATA
	SAS 2.5"							✔
	SATA 2.5"	✔						
Heal-in-place (autonomic self-healing)	FC 3.5"		✔		✔			✔
	SAS 3.5"		✔					
	SATA 3.5"		✔		✔	✔		FATA
	SAS 2.5"							✔
	SATA 2.5"	✔						
Fail-in-place	FC 3.5"							✔
	SAS 3.5"							
	SATA 3.5"							FATA
	SAS 2.5"							✔
	SATA 2.5"	✔						
Warranty		3 years	2 years	3 years	3 years	1 year	1 to 3 years depending on model	5 years

in mission-critical applications such as government labs, high-energy particle research, digital film/video production and delivery, seismic processing and so on, have revealed high levels of satisfaction. Perhaps the most telling remark came from an IT manager who wishes to remain anonymous: "I don't worry about silent data corruption anymore because it's no longer an issue for us."

HEAL-IN-PLACE

Sector errors in traditional disk subsystem designs mark the HDD as failed. A failed HDD initiates a RAID data rebuild process that degrades performance and takes a long time. It can also be expensive, as there may still be useful life in the hard disk drive.

A heal-in-place system goes through a series of automated repair sequences designed to eliminate or reduce most of the "no failure found" HDD failures, as well as the subsequent unnecessary and costly RAID data rebuilds. As of now, there are five systems that provide heal-in-place capabilities: Atrato Inc.'s Velocity1000 (V1000), DataDirect Networks' S2A series, NEC's D-Series, Panasas' ActiveStor and Xiotech's Emprise 5000. Each provides a proven, albeit completely different, heal-in-place technology.

DataDirect Networks' S2A's heal-in-place approach to disk failure attempts several levels of HDD recovery before a hard disk drive is removed from service.

Atrato's V1000 uses fault detection, isolation and recovery (FDIR) technology. FDIR continuously monitors component and system health, and couples it with self-diagnostics and autonomic self-healing. Atrato uses FDIR to correlate SATA drive performance with its extensive database of operational reliability testing (ORT) performed on more than 100,000 SATA hard disk drives. FDIR uses decision logic based on that extensive ORT history, stress testing and failure analysis to detect SATA HDD errors. It then leverages Atrato Virtualization Software (AVS) to deal with detected latent sector errors (non-recoverable sectors temporarily or permanently inaccessible). AVS' automated background drive maintenance commonly prevents many of these errors. When it doesn't, it remaps at a sector level using spare capacity on the virtual spare SATA HDDs. This enables many of those SATA HDDs with sector errors to avoid being forced into a full failure mode permanently, and allows those SATA hard disk drives to be restored to full performance.

DataDirect Networks' S2A's heal-in-place approach to disk failure attempts several levels of HDD recovery before a hard disk drive is removed from service. It begins by keeping a journal of all writes to each HDD showing behavior aberrations and then attempts recovery operations. When recovery operations succeed, only a small portion of the HDD

requires rebuilding using the journaled information. Having less data to rebuild greatly reduces overall rebuild times and eliminates a service event.

NEC's D-Series Phoenix technology detects sector errors, but allows operation to continue with the other HDDs in the RAID group. If an alternative sector can be assigned, the hard disk drive is allowed to return to operation with the RAID group avoiding a complete rebuild. Phoenix technology maintains performance throughout the detection and repair process.

Panasas' ActiveScan feature continuously monitors data objects, RAID parity, disk media and the disk drive attributes. When it detects a potential problem with HDD blocks, the data is moved to spare blocks on the same disk. Future hard disk drive failure is predicted through the use of HDD SMART attribute statistical analysis, permitting action to be taken that protects data before a failure occurs. When a hard disk drive failure is predicted, user-set policies facilitate preemptively migrating the data to other HDDs. This eliminates or mitigates the need for reconstruction.

Xiotech's Emprise 5000, or ISE, is architected to proactively and re-actively provide autonomic self-healing storage. ISE preventive and remedial component repair takes place within its sealed DataPacs (storage capacity modules). It never requires manual intervention to pull failed drives. ISE provides in-place automatic data migration (when required), power cycling, factory remanufacturing and component re-calibration; only the surfaces of affected heads with allocated space, as opposed to entire disk drives, are rebuilt in very fast parallel processes. The result is the equivalent of a factory-remanufactured HDD, and the only components ever taken out of service are those that are beyond repair. Everything else is restored to full activity and performance.

Based on interviews with users and on vendors' historical service data, autonomic self-healing works.

DOES AUTONOMIC SELF-HEALING WORK?

Based on interviews with users and on vendors' historical service data, autonomic self-healing works. The numbers show a decrease in RAID data rebuilds and service calls by as much as 30% to 50%. For Atrato and Xiotech, there are never any HDD replacement service calls because of their fail-in-place technologies.

FAIL-IN-PLACE

Fail-in-place is a fairly new concept aimed at resolving some prickly side effects of hot-plug or hot-swap HDDs in storage systems. An example of these difficult side effects include pulling the wrong drive and causing inadvertent data loss; delaying the replacement of a failed HDD, which defers rebuild starts and increases data loss risk; or using spare drives

that may not have been recently tested, which may result in a second hard disk drive failure.

The basic concept of fail-in-place is to redefine and increase the smallest field-replaceable unit (FRU) from being a HDD to being a storage pack. A storage pack is a collection of hard disk drives operating in concert with a certain percentage of capacity allocated for sparing. HDD failures are automatically rebuilt from the allocated capacity.

There are currently only two vendors supplying fail-in-place storage systems: Atrato (with its V1000) and Xiotech (with the Emprise 5000 or ISE). Both systems feature end-to-end error detection and correction, as well as autonomic self-healing.

The basic concept of fail-in-place is to redefine and increase the smallest field-replaceable unit (FRU) from being a HDD to being a storage pack.

Both vendors' product architectures are based on the concept of available user capacity being tightly coupled with enclosure lifecycle within a single FRU. An enclosure's lifecycle is the timeframe in which the enclosed raw capacity will be available to an application. The total enclosure capacity also includes an allowance for anticipated sparing requirements over the warranted capacity life of the enclosure (three years for Atrato and five years for Xiotech).

The differences between the two implementations are reflective of each vendor's product philosophies. Atrato makes 2.5-inch SATA drives enterprise-capable with their ORT, end-to-end error correction and detection, autonomic self-healing, high densities per enclosure, and with clever vibration and cooling methods. They improve performance by combining 160 drives for up to 80 TB in a 3U enclosure that provides up to 12,500 IOPS and 1.5 GBps throughput from a single enclosure.

Xiotech's focus is on providing increased reliability and performance from enterprise Fibre Channel and SAS 3.5-inch and 2.5-inch drives. The baseline FRU is a sealed DataPac of 10 3.5-inch or 20 2.5-inch Fibre Channel or SAS HDDs for up to 16 TB in 3U. Each ISE has dual removable DataPacs, power supplies with cooling, 96-hour battery backup and active-active RAID controllers. Unlike standard storage subsystems, ISE DataPacs feature innovations such as sophisticated vibration reduction and improved cooling; Xiotech exploits the internal structure of all of the components to fully leverage very advanced drive and system telemetry. DataPac drives feature special firmware that relieves the burden of device compatibility required of all other storage subsystems. The result of the tightly knit control within the DataPac is a highly reliable "super disk" that has demonstrated more than a 100-fold increase in reliability vs. a typical

storage system drive bay (based on Xiotech's test of 208 ISEs containing 5,900 drives for 15 months with no service events).

DOES FAIL-IN-PLACE WORK?

Atrato and Xiotech have proven that fail-in-place definitely works. Their product testing and customer testimonials indicate these technologies can virtually eliminate HDD replacement service calls. That translates to lower costs, less risk of lost data and fewer application disruptions.

Self-healing storage solves tangible operational problems in the data center. It reduces services events, costs, management, data loss risk and application disruptions. And most importantly, it works. Ten years from now, self-healing storage will be considered a minimum requirement just as RAID is today.

Marc Staimer is president of Dragon Slayer Consulting.

Dueling directors

Director-class switches have been used to maintain performance in large storage networks for years, but soon they'll act as the hubs that unify data center networks.

By Jacob Gsoedl

FIBRE CHANNEL (FC) as a technology has been relatively static over the past 10 years, and FC switch innovation has been incremental—from bandwidth support and additional features to increased resilience and availability—culminating in the high-end FC director platform. Aside from a few failed incursions by vendors like QLogic Corp. and others, Brocade Communications Systems Inc. and Cisco Systems Inc. now almost exclusively own the high-end Fibre Channel switch and director market. With their top-of-the-line platforms (the Brocade DCX Backbone and Cisco's MDS 9500 Multilayer Director Series), they duel for customers who require a combination of high FC performance and high availability. Features and suitability for existing infrastructure have typically been the primary director selection criteria. Enhancements related to Fibre Channel over Ethernet (FCoE), as well as the convergence of FC and Ethernet into a unified data center protocol, have added roadmaps and vendor strategies as relevant purchasing considerations. Protecting the new investment and ensuring its future are of paramount importance considering how profoundly storage-area networks (SANs) and data centers will be transformed.

The Brocade DCX Backbone and Cisco MDS 9500 Series have much in common. They're both chassis based and can be scaled by adding or changing hot-swappable line cards. With all components redundant and hot-swappable (blades, fans, power supplies), they present no single point of failure. From 1 Gbps/2 Gbps /4 Gbps /8 Gbps Fibre Channel to FICON, FC over Internet Protocol (FCIP) and Internet Protocol over FC (IPFC), and connectivity options for iSCSI (DCX via an iSCSI gateway and the MDS natively), they support all relevant storage networking protocols. With Brocade's Fabric Application Platform option and the Cisco MDS 9000 Storage Services Module (SSM), both vendors can turn their switches into intelligent app platforms for third-party services, such as EMC Corp.'s Invista for fabric-based storage virtualization and EMC RecoverPoint for fabric-based data protection. By way of the Brocade Data Center Fabric Manager (DCFM) and Cisco Data Center Network Manager (DCNM), both provide commensurate management applications to provision and proactively manage devices across multiple fabrics. Both vendors also support N_Port ID Virtualization (NPIV) to simplify connectivity, management and monitoring of proliferating virtual server environments. In addition, Cisco's ability to create completely isolated fabric topologies via its virtual SAN (VSAN) technology has been matched by Brocade's new Virtual Fabrics feature, which enables organizations to create logical groups of separately managed devices, ports and switches within a physical SAN.

> "With the support of Virtual Fabrics, Brocade eliminated one of the competitive advantages Cisco had with VSANs."
>
> —Bob Passmore, research vice president, Gartner Inc.

"With the support of Virtual Fabrics, Brocade eliminated one of the competitive advantages Cisco had with VSANs," said Bob Passmore, research vice president at Stamford, Conn.-based Gartner Inc. Notwithstanding a long list of commonalities, Brocade and Cisco differ in some key areas and features, as well as in product strategy.

PRODUCT OVERVIEW

In an attempt to establish a product category that resides above traditional directors, Brocade doesn't categorize its DCX Backbone as a director-class product. However, with four times the chassis bandwidth, significant energy efficiency and several feature enhancements, the DCX Backbone can be viewed as a "green" Brocade 48000 Director on steroids. Among the DCX Backbone enhancements are inter-chassis links (ICL) that allow cascading two DCX Backbone switches via a 512 Gbps pipe, integrated Fibre Channel routing, fabric-based encryption, and the separation of control processor and core switching blades. Unlike the 48000 Director and the Cisco MDS 9500 Series, the DCX Backbone doesn't support

iSCSI natively and depends on external iSCSI gateways to interface with iSCSI SANs.

The DCX Backbone is available in two modular form factors. Built for large enterprise networks, the 14U Brocade DCX Backbone has eight vertical blade slots to provide up to 384 FC ports using Brocade-branded 4 Gbps or 8 Gbps small form-factor pluggables (SFPs). Built for midsized networks, the 8U Brocade DCX-4S Backbone has four horizontal blade slots to provide up to 192 Fibre Channel ports. Additionally, each chassis has two slots for redundant control processor blades and two more slots for redundant core switching blades that move traffic between blades.

With the Cisco MDS 9100 Series rack-mountable fabric switches at the low end and the midrange MDS 9200 Series—which provides MDS 9500-equivalent features at a smaller form factor and lower cost by sacrificing redundancy and scalability—the MDS 9500 director family tops Cisco's coherent FC switch product line. Cisco offers three MDS 9500 models: The MDS 9513, supporting a record 512 8 Gbps FC ports, with a total of nine slots for line cards and two slots for redundant supervisor modules, is targeted at enterprise networks with high port count requirements. For smaller networks and as edge devices, Cisco offers the MDS 9509 featuring nine slots, and the MDS 9506 with six slots.

Introduced by Cisco in 2002, the MDS 9500 is on track to follow the Catalyst 6500 Ethernet switch family in prolonging its life toward a decade, displaying longevity that's rare in the fast moving high-tech arena where assets usually depreciate in three to five years. "MDS 9506 and 9509 switches that shipped in 2002 still support 8 Gbps FC and will support FCoE," confirmed Paolo Perazzo, senior product line manager, Cisco's Data Center Switching Technology Group. In contrast, during the same period, Brocade has brought five different director platforms (12000, SilkWorm

FEATURE COMPARISON

	BROCADE DCX BACKBONE	CISCO SYSTEMS INC. MDS 9500
Models	*DCX Backbone (eight port blades) *DCX-4S Backbone (four port blades)	*MDS 9513 (11 port blades) *MDS 9509 (seven port blades) *MDS 9506 (four port blades)
Maximum port count	384	512
Bandwidth per slot	256 Gbps	96 Gbps
Biggest benefit	Fastest Fibre Channel (FC) director on the market. Significant lead in the total number of FC ports sold.	High degree of investment protection, long product lifecycle and coherent product line that uses the same NX-OS operating system
Fibre Channel over Ethernet (FCoE) support	Will be available once the standard is ratified	Available now via the Nexus 5000 switch; for the MDS 9500, Nexus 2000 and Nexus 7000, it will be released on standard ratification

24000, 48000, DCX Backbone and McData i10K) to market.

"Brocade is a smaller but very innovative company with shorter product lifecycles, partnering with best-of-breed vendors. Cisco, on the other hand, has a record of designing architectures that last for many years," noted Bob Laliberte, an analyst at Enterprise Strategy Group (ESG) in Milford, Mass.

A case in point is the DCX Backbone. "The main difference between the 48000 and DCX is the separation and rearchitecture of the core switching and control processor blades into separate blades, which required a new chassis design," said Bill Dunmire, senior product marketing manager at Brocade.

THE HIGH COST OF HIGH-END SWITCHES

PLATFORM	PRICE
Brocade DCX Backbone with eight 48-port 8 Gbps blades and full redundant configuration	$628,000
Cisco MDS 9513 with eight 48-port 8 Gbps blades and full redundant configuration	$593,000
Brocade DCX-4S with four 48-port 8 Gbps blades and full redundant configuration	$328,000
Cisco MDS 9506 with four 48-port 8 Gbps blades and full redundant configuration	$320,000

(List prices provided by Brocade/Cisco reseller)

COMPARING ARCHITECTURES

The Brocade DCX Backbone is based on a shared memory architecture where data moves from switching ASIC to switching ASIC along multiple internal ISLs that make up the path from an ingress port to an egress port. To load balance between these inter-ASIC links within the switch, the DCX Backbone relies on either exchange- or port-based routing. "Besides fewer components on blades, which reduces the likelihood of failure, in a shared memory architecture ASICs on the core switching blades talk to ASICs on port blades using the same protocol, minimizing protocol overhead," Dunmire explained.

In comparison, the Cisco MDS 9500 leverages a crossbar architecture where frame forwarding is directly performed in ASICs on the line cards. The crossbar manages forwarding of packages, and a central arbiter ensures fairness and prioritization. While the MDS 9506 and MDS 9509 provide the fabric switching module and central arbiter on the supervisor blade, the MDS 9513 uses a separate pair of switching modules located in the back of the MDS 9513 chassis. "Unlike a shared memory architecture where traffic moves across internal switching ASICs along varying paths, resulting in varying latencies, in a crossbar architecture the latency between ports is consistent across all ports within the switch," said Omar Sultan, solution manager, data center switching, data center solutions at Cisco.

Even though each vendor claims its architecture is superior, they each have their pros and cons. With the exception of a few vendor specific peculiarities, both platforms can be used to power the most mission-critical and largest SANs with comparable results and user experience; this is substantiated by Brocade and Cisco splitting the director market almost evenly. "The two products work very well and

by having competition, they're pushing new features and, in the long term, bring down the price," ESG's Laliberte said.

Nevertheless, there are noticeable differences between the two platforms. The DCX Backbone supports local switching, which allows traffic between ports on the same blade to be directly switched instead of having to go through the core switching module; this means lower latency for devices connected to the same blade and improved scalability by reducing the amount of traffic that has to pass through the core switching blades. Although Cisco rebuffs the local switching benefit, emphasizing bigger latency variances as a result of local switching, support for local switching in its latest Nexus platform suggests that the lack of local switching support in the MDS 9500 is a disadvantage.

In addition to reliability, performance and throughput are the most relevant attributes of a director platform. The Brocade DCX Backbone currently wins the raw throughput comparison with 256 Gbps throughput per slot vs. 96 Gbps for the Cisco MDS 9500. When combined with local switching, it can concurrently operate more ports at full 8 Gbps utilization than the MDS 9500, as verified by a February 2009 Miercom lab test (Report 090115B). As a result, the MDS 9500 depends to a greater degree on oversubscription than the DCX Backbone. In practical SAN reality, however, not all ports will operate at full 8 Gbps rate, and the use of oversubscription combined with traffic prioritization and QoS makes the throughput difference less significant. In the past, increases in port and chassis throughput benefited mostly ISLs and, to a lesser degree, servers; but now the proliferation of virtual server environments definitely makes bandwidth capacity more relevant. "Server virtualization is a game changer, making oversubscription more problematic because physical servers running many virtual machines are more likely to fully utilize a SAN link," Gartner's Passmore said. Cisco confirmed that it's working on a next-generation switch fabric module that will match the DCX's 256 Gbps slot throughput; existing customers will be able to upgrade by simply replacing the existing switch fabric module. "Replacing the switch fabric module costs an order of magnitude less than a forklift upgrade," noted Bill Marozas, business development manager, Cisco Data Center Solutions.

In addition to reliability, performance and throughput are the most relevant attributes of a director platform.

Despite each vendor's claim that its platform requires less SAN architecting, each director platform has idiosyncrasies a SAN designer needs to take into consideration to ensure optimal performance. In the case of the MDS 9500, the SAN design effort will likely be related to managing oversubscription and traffic prioritization. Correspondingly, the DCX

Backbone requires SAN architects to take latency variances between different ports within the same chassis into account, as well as its use of port- and exchange-based routing to load balance inter-ASIC links. While both Brocade and Cisco support port- and exchange-based routing over external ISL links, Brocade's use of these protocols inside the switch has been somewhat controversial. Customers need to make a choice between one of two routing modes; despite repudiation by Brocade, benchmarks like the December 2008 Miercom report (Report 081215B) have shown slower performance if the switch is used with port-based routing instead of the default exchange-based routing; and some array vendors advise their customers to stay away from the DCX's default exchange-based routing for some of their arrays.

"HP does not typically make specific recommendations regarding switch routing, but we recommend using port-based routing with the StorageWorks Continuous Access EVA solution since exchange-based routing doesn't guarantee in-order frame delivery all the time across exchanges," said Kyle Fitze, marketing director for the StorageWorks Storage Platforms Division at Hewlett-Packard (HP) Co. Conversely, EMC and NetApp confirmed that all of their arrays work flawlessly using the DCX default exchange-based routing mode.

FCoE AND CEE/DCE

Compelled by the prevalence of Ethernet and its enhancements, and the success and simplicity of iSCSI, Brocade and Cisco have embarked on bringing Ethernet into the well-guarded FC domain via Fibre Channel over Ethernet. FCoE is a T11 standard expected to be ratified later in the year. It uses Converged Enhanced Ethernet (CEE)—Cisco calls it Data Center Ethernet (DCE)—as the physical network transport to deliver Fibre Channel payloads. However, unlike its Ethernet brethren, it's lossless and appears as native Fibre Channel to the operating system and apps. Unlike iSCSI, it's not routable and is designed as a low-latency, high-performance Layer 2 data center protocol.

The value proposition of FCoE and converged Ethernet is lower infrastructure cost realized by simplifying cabling and reducing the number of adapters from two host bus adapters (HBAs) and two network interface cards (NICs) to two redundant Converged Network Adapters (CNAs) through which the converged local-area network (LAN) and FC traffic traverses. Instead of connecting two NICs to an Ethernet switch and two HBAs to a Fibre Channel switch, the two CNAs terminate into a CEE/DCE-capable switch that delivers Ethernet traffic to the LAN and FC traffic to the SAN. Although FCoE and CEE/DCE are expected to eventually be used from core to edge, its initial use is primarily at the access layer to connect servers to CEE/DCE-capable switches.

Both Brocade and Cisco are committed to FCoE, but with different strategies. Brocade won't ship Converged Enhanced Ethernet products until the standard is ratified; at that point, Brocade will support FCoE and

CEE in its DCX Backbone via new blades. Older Brocade Fibre Channel products, such as the 48000 Director, will connect through the DCX Backbone or a new top-of-rack switch to interface with CEE components.

With the Nexus 5000 Series top-of-rack switch, Cisco is the first vendor to offer a pre-standard FCoE product. For the MDS 9500 director family, as well as the Nexus 2000 Series Fabric Extenders and Nexus 7000 Series switches, DCE and FCoE support won't be available until standard ratification, similar to Brocade's plans.

Overall, Cisco has a more coherent product strategy to support a unified data center protocol and has been working on it for the past five years. With the Nexus 7000 core switch and its formidable 15 Tbps planned throughput designed to support 100 Gbps Ethernet, its unified NX-OS operating system that's used by both the MDS 9000 family and the new Nexus platform, and its Data Center 3.0 initiative to unify computing systems, Cisco is amazingly ready for the battle for the unified data center. On the other hand, with the recent acquisition of Foundry Networks Inc., Brocade isn't standing still and, despite integration challenges, a legion of loyal Brocade and Foundry customers are likely to side with Brocade. Even though the battle has begun, broad adoption of CEE isn't expected until late 2010. "With the exception of some early adopters, broader adoption of CEE won't happen until 18 to 24 months from now," said Greg Schulz, founder and senior analyst at StorageIO Group in Stillwater, Minn.

MAKING A CHOICE

With much in common, including pricing (see "The high cost of high-end switches," p. 22), the most important director selection criteria are which platform fits best into your existing infrastructure and which vendor's strategy aligns more closely with your roadmap.

The most important director selection criteria: which platform fits best into your existing infrastructure and which vendor's strategy aligns more closely with your roadmap.

"We chose Brocade directors mainly because we already had McData and Brocade departmental switches and our staff was familiar with managing them. The DCX performance was icing on the cake," said Michael Kilmer, primary storage administrator at Daktronics Inc., a leading manufacturer of large format display systems in Brookings, S.D.

"The fact that we are an all-Cisco shop as well as the VSAN feature were the main reasons we went with the MDS 9500 platform," explained Colleen Rhode, systems analyst at East Carolina University in Greenville, N.C.

For Keith E. Moore, director of technology services at Bellevue, Wash.-based Applied Discovery Inc., a division of LexisNexis, scalability and features were the decisive factors for selecting Brocade. "One of the

reasons Applied Discovery chose Brocade is that Brocade is known to handle heavy traffic while sustaining wire speed. Our Cisco FC network consistently had performance issues under heavy load," Moore said.

David Turner, vice president of technical operations at MobiTV Inc., a leading provider of video to mobile devices in Emeryville, Calif., is evaluating the Nexus 5000 to complement the firm's MDS 9509 directors to lower costs and take advantage of copper cabling. "I decided to move off Brocade to Cisco for the sole purpose of taking advantage of the modularity of the MDS switches. It's much more cost-effective to get to the next version with the Cisco platform. Unlike Brocade, it doesn't require expensive forklift upgrades," Turner said.

Fernando Mejia, senior manager of IT infrastructure at the Independent Purchasing Cooperative (IPCoop) Inc., the purchasing arm of the Subway franchise in Miami, acquired a Cisco Nexus 7000 instead of a Catalyst 6500 because of its high performance, scalability and the ability to replace his stackable Brocade FC switches once FCoE becomes available.

MATURING DIRECTORS

Regardless of whose product you choose, both platforms will reliably power your SAN, which is confirmed by the myriad storage-area networks currently powered by Brocade and Cisco. Both vendors are embracing the converged Ethernet paradigm in their product roadmaps, but unless you're willing to debug the initial CEE/DCE flaws as an early adopter, you're well advised to wait for at least another year until the standard and products have matured. ⊙

Jacob Gsoedl is a freelance writer and a corporate director for business systems. He can be reached at jgsoedl@yahoo.com.

DR readiness in a deduplicated world

One of the big benefits of employing data deduplication for secondary disk targets is its ability to reduce the size of data sets, enabling replication with lower bandwidth requirements.

IF YOU'RE A STORAGE PRO, you should be familiar with the phrase "time to protection." This is the time required to complete all of the activities that must occur between the initiation of a backup and the arrival of the backup copies at an offsite location for disaster recovery (DR) purposes. For tape-based DR schemes, this includes the time it takes to execute the backup, prepare offsite tape copies and transport them to a remote location.

For disk-based DR strategies, this would be the time it takes to back up to disk and move the data offsite via replication, which can vary depending on the amount of data to transfer and the available bandwidth. One of the big benefits of employing data deduplication for secondary disk targets is its ability to reduce the size of data sets and enable replication with lower bandwidth requirements. This makes automated data electronic vaulting less time-consuming and less costly.

Once the data has been duplicated at a second site, how much time is required to restore data from the deduplicated DR copy?

So we know dedupe helps, but does it also hinder? The added process of identifying and eliminating redundant data could affect performance between initiation of a backup and initiation of replication. Deduplicating during the backup process (inline, before data is written to disk) could impact backup performance, while deduplicating after the backup process is complete (post-process) could delay replication.

THE PATH TO DR READINESS

When it comes to recovery, there are two points in the data path to focus on: the point of local protection, which is when a copy of production data is onsite for operational recovery; and time to protection, which is the point at which you have a copy offsite for DR.

Systems with inline dedupe capabilities—such as those from Data Domain Inc., Hewlett-Packard (HP) Co. (with its StorageWorks D2D Backup

Systems), IBM Corp. (Diligent) and NEC Corp.—promote the efficiency of enabling replication initiation as soon as data "hits" the disk, allowing for fast time to protection. Post-process approaches take a different point of view. Vendors, including ExaGrid Systems Inc., FalconStor Software Inc., HP (with its Virtual Library System) and Sepaton Inc., would maintain it's more important for backup to disk to complete at wire speed and that initiating dedupe outside the backup window guarantees better backup service-level agreements (SLAs). Replication initiation varies here—some vendors begin within a few minutes, while others have a longer lag time.

EMC Corp. and Quantum Corp. fall into both the inline and post-process camps because their products let an admin decide when dedupe occurs. By offering choice, policies can be set for specific backup workloads. And flexibility is good because there's a place for each approach. For example, if you have workloads where you expect a lot of redundant data, then in-line dedupe may be preferred. If the workload has a lot of new data or if the backup window is small, then a post-process approach may be better.

Another dimension to consider is time to recovery. Once the data has been duplicated at a second site, how much time is required to restore data from the deduplicated DR copy? How quickly can data be read and reconstituted to an application-usable state? Some vendors keep a non-deduplicated full backup image just for this scenario. This approach will aid in providing more rapid recovery, but will use additional storage capacity.

CAN THE PROCESS BE ACCELERATED?

For Symantec Corp. Veritas NetBackup 6.5 customers, the Symantec OpenStorage (OST) option can help. Veritas NetBackup OST, when used in conjunction with an OST-enabled dedupe storage system (Data Domain, FalconStor and Quantum are currently certified), eliminates many of the challenges associated with the creation and management of duplicate backup images, the transportation of backup copies to an alternate site and the centralized creation of tape-based copies for long-term retention. In this case, Veritas NetBackup maintains knowledge and control of backups written to the OST-interface disk storage units of vendors' devices. Its "optimized duplication" technology improves performance for creation of replicas stored at the secondary site. For example, Data Domain, the first vendor with a certified OST interface, has been able to demonstrate replication performance improvements of 75% or more in OST environments.

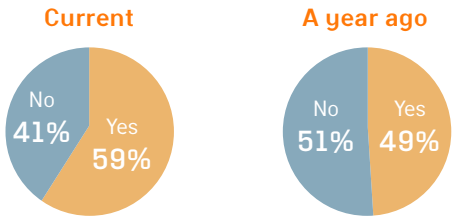
The business benefits of storage capacity optimization via data dedupe are well-regarded. But dedupe can also enable significant efficiency when it comes to disaster recovery. When making an investment decision in dedupe, in addition to evaluating products based on local dedupe processing and operational recovery on the premises, it makes sense to investigate the product's viability to provide DR readiness offsite. ⊙

Lauren Whitehouse is an analyst focusing on backup and recovery software and replication solutions at Enterprise Strategy Group, Milford, Mass.

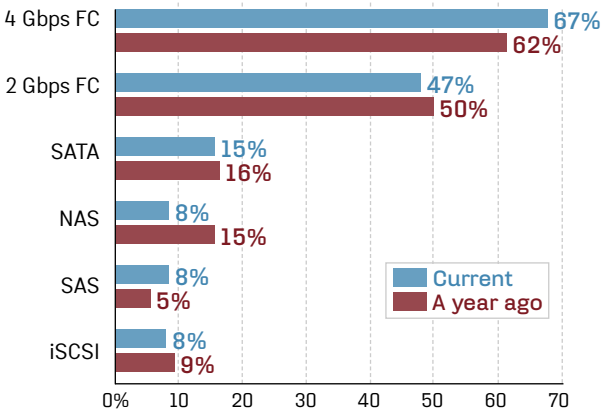
Majority of users opt for tiered storage

STORAGE TIERING—putting data in its proper place—can be a huge money saver for a storage shop by getting less-frequently accessed data off expensive primary systems and onto cheaper disks. A year ago, less than half of the *Storage* readers surveyed had a tiering process in place, but nearly 60% of those who weren't tiering said they planned to do so. Apparently, a lot of them carried through on their plans, as our most recent survey shows 59% now use tiered storage. In addition, half of those not yet sold on tiering said they'll take the plunge (approximately 61% will do so over the next 12 months). Those implementing storage tiers tend to keep their schemes simple: 75% limit the number of tiers they support to two or three, while 76% keep half of their data or less on primary storage. The biggest pain point related to tiering remains classifying data so it can be properly placed. Forty-two percent of respondents said it's the biggest hitch in the process vs. 41% last year. *—Rich Castagna*

Have you implemented a tiered storage architecture (excluding tape libraries)?



What does your tier 1 storage consist of?*



** Choose all that apply*

What's your biggest pain point related to your tiered storage system?

42%	Classifying data so that it's sent to the right tier
24%	Moving data between tiers
18%	Keeping track of where data currently resides
8%	Poor performance on lower tiers

36%

Respondents who move data manually from tier to tier

> "Successful tiered storage will be predicated on automation and granularity. Manually moving LUNs around the farm is a no-win situation."
>
> *—Survey respondent*

Check out the following resources from our sponsors:

NetApp, page 3

Use 50% Less Storage With NetApp. Guaranteed.

See How Much Space You Can Save With Our Calculator