

Backup *Deep Dive*



Modernize Your Backup Infrastructure

Copyright © 2009 InfoWorld Media Group. All rights reserved.



Modernize Your Backup Infrastructure

The technologies have arrived to vastly improve backup and recovery performance and reliability. Here's how to put them to good use.

By W. Curtis Preston

EVER GET THE FEELING that your backup system is behind the times? Do you read trade magazines and wonder if you're the only one still using an antiquated backup system? The first thing you should know is that you're not the only one. But your backup system could probably use some modernization.

New technologies have changed the nature of the backup game in a fundamental way, with disk playing an increasingly important role and tape moving further into the background. Many of the liabilities and performance issues that have dogged datacenter backups forever now have plausible technology solutions, provided those solutions are applied carefully and dovetail with primary storage strategy. It is truly a new day.

Before you contemplate a modernization plan, you need a working understanding of new high-speed disk-based solutions; schemes that reduce the volume of data being replicated; and how real-time data protection techniques actually work. With that under your belt, you can start to apply those advancements to the real-world data protection problems every datacenter faces.

THE DISK IN THE MIDDLE

D2D2T (disk-to-disk-to-tape) strategies have gained popularity in recent years due to the great disparity between the devices being backed up (disks), the network carrying the backup, and the devices receiving the backup (tape).

The average throughput of a disk drive 15 years ago was approximately 4MBps to 5MBps, and the most popular tape drive was 256KBps, so the bottleneck was the tape drive. Fast-forward to today, and we have 70MBps disk drives, but tape drives that want

120MBps. Disks got 15 to 20 times faster, but tape drives got almost 500 times faster! Tape is no longer the bottleneck; it's starving to death. This is especially true when you realize that most backups are incremental and hold on to a tape drive for hours on end – all the while moving only a few gigabytes of data.

D2D2T strategies solve this problem by placing a high-speed buffer between the fragmented, disk-based file systems and databases being backed up and the hungry tape drive. This buffer is a disk-based storage system designed to receive slow backups and supply them very quickly to a high-speed tape drive.

The challenge faced by some customers (especially large ones) was that many backup systems didn't know how to share a large disk system and use it for backups. Sure, they could back up to a disk drive, but what if you needed to share that disk drive among multiple backup servers? Many backup products still can't do that, especially Fibre-Channel-connected disk drives.

Enter the virtual tape library, or VTL. It solved this sharing problem by presenting the disk drives as tape libraries, which the backup software products have already learned how to share. Now you could share a large disk system among multiple servers. In addition, customers more familiar with a tape interface were presented with a very easy transition to backing up to disk.

Another approach to creating a shareable disk target is the intelligent disk target, or IDT. Vendors of IDT systems felt the best approach was to use the NFS or CIFS protocol to present the disk system to the backup system. These protocols also allowed for easy sharing among multiple backup servers.

But both VTL and IDT vendors had a fundamental problem: The cost of disk made their systems cost effective as staging devices only. Customers stored a single



night's backups on disk and then quickly streamed them off to tape. They wanted to store more backups on disk, but they couldn't afford it. Enter deduplication.

THE MAGIC OF DATA DEDUPLICATION

Typical backups create duplicate data in two ways: repeated full backups and repeated incrementals of the same file when it changes multiple times. A deduplication system identifies both situations and eliminates redundant files, reducing the amount of disk necessary to store your backups anywhere from 10:1 to 50:1 and beyond, depending on the level of redundancy in your data.

Deduplication systems also work their magic at the subfile level. To do so, they identify segments of data (a segment is typically smaller than a file but bigger than one byte) that are redundant with other segments and eliminate them. The most obvious use for this technology is to allow users to switch from disk staging strategies (where they're storing only one night's worth of backups) to disk backup strategies (where they're storing all onsite backups on disk).

There are two main types of deduplication. Target dedupe systems allow customers to send traditional backups to a storage system that will then dedupe them; they are typically used in medium to large datacenters and perform at high speed. Source dedupe systems use different backup software to eliminate the redundant data from the very beginning of the process and serve to back up remote offices and mobile users.

BACKING UP AS YOU GO

CDP (continuous data protection) is another increasingly popular disk-based backup technology. Think of it as replication with an Undo button. Every time a block of data changes on the system being backed up, it is transferred to the CDP system. However, unlike replication, CDP stores changes in a log, so you can undo those changes at a very granular level. In fact, you can recover the system to literally any point in time at which data was stored within the CDP system.

A near-CDP system works in similar fashion except that it has discrete points in time to which it can recover. To put it another way, near-CDP combines snapshots with replication. Typically, a snapshot is taken on the system being backed up, whereupon that snapshot is

replicated to another system that holds the backup. Why take the snapshot on the source before replication? Because only at the source can you typically quiesce the application writing to the storage so that the snapshot will be a meaningful one.

PROTECTING TRANSACTION SYSTEMS

Disk-to-disk backup systems, deduplication, and CDP were all developed to solve specific problems. So let's have a look at the challenges of today's datacenters to see how these technologies can help.

The first challenge: high-volume transaction systems that are intolerant of data loss. Most industries experience double-digit increases in the volume of transactions every year. That's just the nature of computing. And along the way, organizations have grown increasingly worried about data loss, thanks to high-profile customer data debacles that have created one public relations nightmare after another.

Depending on the volume of transactions and tolerance of downtime, companies that want to minimize risk turn to internal disk-based systems rather than tape (which has an unfortunate tendency to escape the datacenter) as their primary backup target. The question is whether or not they use traditional backup software to get there.

Switching from tape to disk as the primary target – while still using traditional backup software – makes it easier to create backups of transaction logs that can be used to rebuild those transactions easily in case of data loss. In addition, the use of disk allows those transaction log backups to be replicated offsite so they can be used even in the case of disaster. Using disk as the primary target for backups can also help in full recovery of large databases, as the aggregate performance of the disk system can be easily matched to the recovery time objective (RTO) of the restores you are likely to perform.

But the true power of disk in a recovery system can be realized only by switching from a traditional backup system to CDP or near-CDP. Traditional backup still suffers from the laws of physics. If you've got a 20TB database to restore and a five-hour RTO, you need to be able to restore more than 4TB per hour, leaving a little time to replay the appropriate amount of transaction logs. A CDP or near-CDP system solves this by presenting an already-recovered image during an outage.



Both CDP and near-CDP systems can present to a recovery server a read-write image of the most recent backup of the system to be recovered. This includes presenting a read-write image of the latest version of the operating system and application to the server or virtual server that will be used in a recovery scenario. In fact, some CDP software systems even use incremental restore capabilities to continually keep a VMware image up to date to be used in a recovery.

The number of lost transactions a business can tolerate in a recovery scenario will determine the recovery point objective (RPO), and the amount of downtime it can afford will determine the RTO. The more aggressive a business's RTO, the more it will be led to choose CDP or near-CDP. The more aggressive its RPO, the more it will need to choose CDP over near-CDP. Many near-CDP systems cannot do any better than a one-hour RPO, because that's how often they can take a snapshot; customers looking for a one-minute RPO are usually forced to choose a true CDP solution that does not rely on snapshots.

PROTECTING E-MAIL SYSTEMS

The next challenge we'll take a look at is backing up and recovering e-mail systems. Most modern e-mail systems are at their heart database systems, so the backup systems work in similar ways. But the typical recovery request of an e-mail system differs markedly from that of a database. Databases are rarely recovered, but when they are, they are fully recovered up to the point of failure. Rarely are databases restored in part; that is, rarely do you restore a single table in a database (database recovery mechanisms do not even have that ability). Usually, your only choice is to restore the entire database to an alternate location and then export the table you need to restore.

E-mail systems, on the other hand, often receive recovery requests for a single table or even a single row in that table. In other words, they are often asked to restore an individual user's mailbox, folder, or even a single e-mail message. Oddly enough, in the first half of the current decade, the only way to restore at this granular level mirrored what you did with databases: Restore the entire e-mail application to an alternate location, and then drag and drop the mailbox, folder, or e-mail that you needed to the appropriate server.

The advent of recovery storage groups in Exchange

changed all that. With recovery groups, admins can restore only the storage group containing the affected user or e-mail, then drag and drop what they need. To take advantage of this feature, however, you must ensure that Exchange admins split their servers into multiple storage groups from the start.

Other advancements in the e-mail department include the ability of some backup software to extract user-level information from Exchange Information Store backups. This is another benefit of using disk-based backups. Placing all e-mail backups on disk allows the backup software to do its own querying and extraction of parts of the backup, facilitating quick restores of users and folders.

Unfortunately, these advancements in e-mail backup and recovery are unlikely to help you with electronic discovery. As you know, companies now commonly receive electronic discovery requests as part of a lawsuit or government investigation. Most backup software is so ill-equipped to handle such requests, the best way to modernize your backup system for electronic discovery is to not use it for discovery purposes.

Install an e-mail archive system instead. These systems make it much easier to extract all kinds of information from your e-mail system. This is especially true when you are asked for e-mails with various search criteria over long periods of time. Take, for example, a request to create a PST file of all e-mails containing the words "square," "circle," "rectangle," or "triangle" that were sent by Joe Smith to Fred Barney between February 2005 and March 2009. Satisfying that request with backup software would be nearly impossible; doing it with e-mail archive software is a piece of cake.

REDUCING RISK WITH ENCRYPTION

While encryption is not a new technology, high-speed encryption for large-volume backup systems is new. Today's backup encryption systems can encrypt as fast as the backup target can run, and include a variety of key management systems to meet a number of different environments' needs.

Everyone has heard tales of woe about unencrypted backup tapes being lost or stolen. While this has always been a problem, it's a bigger issue now, because government regulations won't let you sweep such incidents under the rug. Instead, you are required by law to notify



customers when their data goes AWOL – except if that data has been encrypted.

The strategy is simple: Either encrypt tapes or don't send them anywhere. There are a number of encryption options, including backup software encryption, SAN appliance encryption, and tape drive encryption. Pick one of these methods, and if a tape is stolen or lost, it won't be readable – and you won't have to notify anyone (with most laws on the books, anyway).

Alternatively, if deduplication and replication are in place, you can forgo sending tapes anywhere at all. Just back up to a dedupe system – and replicate over a high-speed connection to another dedupe system off-site. Now you have on-site and off-site backups and you haven't touched a tape. If you want to make tapes, you can do so at the off-site facility, so the tape never needs to be shipped. It can be locked in a tape library, a locked cage, a locked datacenter, or a locked building. With the right rules in place, those tapes don't even need to be encrypted.

PROTECTING VIRTUALIZED ENVIRONMENTS

Virtual servers can be a big help in recovery scenarios. They make it much easier to create a set of recovery servers at the recovery site that match the computing capabilities in the datacenter. The recovery servers may not be as fast, but they will have the same operating system, and they will at least think they have the same hardware, which solves an important part of the recovery problem.

Backing up is another story, because you run smack into the laws of physics. When you put 20 physical servers into one physical server as VMs (virtual machines), everything runs fine until you need to back them up – at which point the fact they're sharing the same physical storage becomes painfully apparent.

VCB (VMware Consolidated Backup) was the first "solution" to this problem, but it never really solved much. VCB may have made backups slightly faster, but it cost a lot and required another physical server plus staging storage to create image-level backups.

To modernize backup infrastructure to support virtual environments, organizations running VMware should

upgrade to vSphere and look for a backup product that supports its vStorage API. This removes the need for a physical proxy server, a staging area, two-step backups, two-step restores, and full backups in order to get incremental backups. You can use a VM as your proxy server, you don't need a staging disk, and you get change block tracking, which allows the backup app to ask a VM what blocks have changed since the last backup.

BACKING UP IS HARD TO DO

One of the biggest challenges of managing a backup infrastructure is that no one wants the job. In large companies, the backup admin position is an ever-revolving door staffed time and time again with junior people. In smaller companies, backing up the infrastructure is a peripheral duty that is often ignored. The result is the same in both cases: bad backups.

One solution to this problem is cloud backup services – or managed backup services, depending on your preferred terminology. The idea is simple: Outsource this undesirable part of IT to a company whose staff specializes in it and you'll never look back.

Cloud backup services take advantage of many of the technologies mentioned here, but allow customers to use the service without having to manage the process. Instead, customers simply install a piece of software on the systems being backed up, and the cloud backup service does the rest. But as with any backup system, make sure you have a way to verify that backups are working the way they're supposed to be working.

The unglamorous world of backups is like the rest of IT, only more so: You never hear from anyone until something goes wrong. Modernizing your infrastructure, when planned and executed carefully, can reduce your liability dramatically. But as you make those improvements, remember the backup mantra: Test everything and believe nothing.

W. Curtis Preston has specialized in designing data protection systems since 1993. A prolific writer and frequent lecturer, he is best known by the nickname "Mr. Backup." He has written three books for O'Reilly, the latest of which is entitled Backup & Recovery.