

Special Report

Data Protection via Replication

By Steve Kenniston,
Technology Analyst, Enterprise Storage Group

Sponsored By:



This Special Advertising Section Produced By:

NetworkWorld
www.nwfusion.com

Today's organizations rely upon their data to make critical business decisions every day. The need for up-to-the minute data for critical business decision making is sometimes at risk due to the structure, or lack thereof, of some enterprise organizations' approaches to backing up this data at their remote or branch offices - the very locations where much of the most recent data sits. This Special Report takes a look at the challenges enterprises are facing in protecting data at remote locations and branch offices and identifies ways in which to meet these challenges head on.

About the Author

Steve Kenniston

Technology Analyst

Enterprise Storage Group

A Technology Analyst at Enterprise Storage Group, Steve has over 15 years of real-world enterprise-class experience in the Software and Services business with the last 10 specifically in the Storage Industry.

Table of Contents

Data Sprawl and the Inherent Problems of Protecting It 3

Solving Distributed Data Protection Problems 4

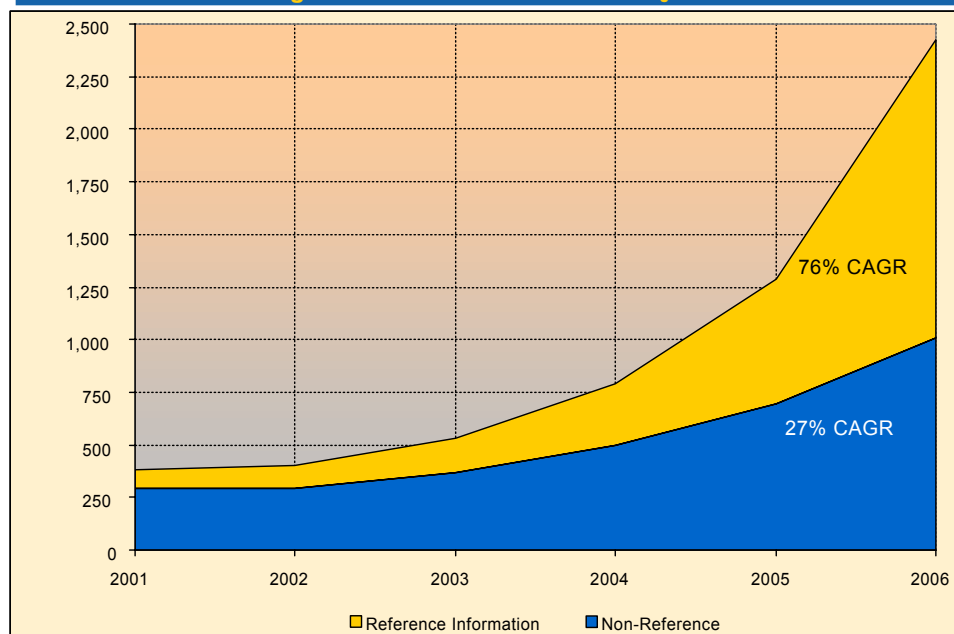
Data Protection and Consolidation 6

Data Sprawl and the Inherent Problems of Protecting It

We have entered a digital age where by users have become digital pack rats. The storage of reference information and/or unstructured data is growing at a staggering pace. In 2001, The Enterprise Storage Group (ESG), a storage analyst firm headquartered in Milford, MA, performed a study — The Data Protection Research Project — that identified the growth of unstructured data in the enterprise. Figure 1 below identifies how reference information is outpacing the growth of traditional non-reference or transactional data in the enterprise (growth of amount of data tied to where it lives).

By 2006, Reference Information will represent 58% of Corporate & Governmental information

Figure 1: Forecasted Growth in Petabytes



Source: Enterprise Storage Group 2003

Reference information is defined as a “digital asset retained for active reference and value.” It includes, but is not limited to; electronic documents such as contract, e-mail and e-mail attachments, presentations, CAD/CAM designs, source code and Web content; certain digitized information such as check images, blueprints, historical documents, medical images, geophysical, satellite and surveillance information, computer-gen-

erated imaging, genomics, proteomics, bioinformatics, video, photographs and voice data. As the volume of information continues to grow, the locations in which it lives also continues to sprawl. Information is created digitally to provide a better way to share it. These files are electronically distributed to branch offices and desktops inside and outside the corporation. At some points along the way, the information is altered and resaved, and the new information has the same or an increased value once the changes are made. The big challenge for IT today is how best to protect this information and the complexities associated with protecting it because of where it lives in the enterprise. A recent census bureau survey stated that each organization or enterprise has more than 50 remote or branch offices. Each one of these offices has an average number of eight to 10 people, all working eight hours per day on their computer, creating or altering information and resav-

ing that information. Every night this information needs to be protected. In some instances, nightly data protection is not enough. Losing a whole day's worth of information can be very costly to some businesses.

Data growth and sprawl is only part of the issue however. The other component is the human component associated with protecting this information. There are a number of issues here. First, with the down turn in the economy over the past 24 months, businesses are trying to do more with less. This means less people. Considering data protection is one of the most difficult challenges for IT, throwing less people at the problem doesn't seem to be a way to solve the challenge. Depending on the size of the remote office or distributed data center, the first technical people to be let go were folks in these offices.

Couple this with the fact that data protection and backup in general, can be a complicated task even for the more experienced IT professional. Having the ability to install and configure even the simplest of back-up products when this level of technical expertise is not your forte can be difficult. And once it is set up, are you sure it is operating properly? Who is managing the process? Who is ensuring jobs are being completed? Who is taking the tapes off-site? Who is test-

ing restores? Typically, the answer to these questions is “no one”. If anyone is doing anything, they are only changing tapes on a nightly basis and probably bringing the tapes to their house for “safe keeping”. Additionally, there is a good deal of time associated with properly protecting information. If IT is not a function manned in the remote office, then chances are, the data is not being properly protected, and something is being missed. The only reason to back data up is to ensure that in the event of a failure, it can be recovered. If log files are not checked to ensure backups complete successfully, or if test recoveries are not performed on a regular basis, then the actual act of doing the backups is almost a moot point. If there is some sort of failure, whether it is a tape failure due to bad media or for some other technical reason, the problem has to be solved. Again, there typically no expertise in the remote offices to solve the issue. Someone gets on the phone with a technical person at corporate and tries to troubleshoot the problem. This takes time away from folks doing their real job. It takes time to properly protect data and in these remote and distributed offices, there is no expertise providing this function.

Data protection at remote data centers also can be costly. Because of the lack of expertise, which leads to poorly implemented storage architecture, a good deal of direct attached storage, additional tape libraries and back-up licenses are purchased when they may not need to be purchased. For example, a file and print server at a remote data center runs out of capacity. A call would be placed to IT corporate which would order another Windows server and have it shipped to the remote site. They also would order a tape library and a backup license for that server. The license and the library can be as costly as the server itself. The reality is there are other configurations just like this in the remote office. Each server has a library. While each library may be able to accomplish the back-up every night with ease, it probably could handle backing up multiple servers. This is a costly implementation considering some of the alternatives that exist today.

Data protection outside the core data center is difficult, but it is growing even more difficult. Nightly backups for some of the applications that live outside the glass house of the main data center are just not enough. In a recent ESG data-protection survey, more than 75% of the respondents stated that e-mail was their No. 1 application. They said they couldn't tolerate more than four hours of down time with this application and felt that there was a very high level of exposure with current data-protection practices around this application. This means that just doing backups at these remote sites is not good enough.

Solving Distributed Data Protection Problems

Now that we have outlined how much data is outside the data center and that protecting it is difficult to manage, let us take a look at some approaches to solve these problems.

One of the first major trends IT is migrating to is the use of disk for better data protection. Disk is playing multiple roles in data protection from VTL (virtual tape libraries) to snapshot targets, mirroring and replication targets. There are a combination of software technologies (snapshots and replication), new hardware technologies (ATA, SATA disk) and the lowering costs of infrastructure (pipes to move data between locations) that are playing a key role at how IT Professionals are looking at new ways to protect data. Disk provides two key roles when it comes to data protection. First, disk can help to shrink backup windows providing backup administrators more time to back-up more data. Second, it provides faster recovery of data. This is key for IT professionals looking to decrease their RTO (recovery time objectives). Moving data to and from disk is much faster than moving it to and from tape. The platform or media type is one key component to better data protection, from a hardware perspective, but there is also software.

First, let us define how IT sets the definitions for data protection. There are two ways, RPO (Recovery Point Objective) and RTO (Recovery Time Objective). The RPO is the point in time

Quick Recovery is Vital

Tape backup = Days

Data Replication = Minutes

Microsoft Windows Storage Server 2003 and Double-Take deliver solutions for affordable, quick recovery.

Double-Take Microsoft Windows Storage Server 2003

NSI Software

Special Report

which data must be able to be recreated. This time is usually measured in hours or minutes old or ago. In other words, "Our mail system cannot be more than four hours old." The recovery time objective is the amount of time it would take to recover data. In other words, "We cannot be down for more than four hours." These factors are key determinants in how data should be protected and hence drive the appropriate technology for this data protection.

Snapshots

One technology getting a good deal of press today is snapshots. A snapshot is a picture of what a file system looked like at a given point in time. A snapshot contains a list of pointers to the files as they exist on a given file system. When a file in the file system changes, a copy of the original file is moved to the snapshot, thereby maintaining the same file subset that was created at the time of the snapshot. This technique is called "copy on write," and is a very common way of doing snapshots. Snapshots are usually controlled and managed by the volume management software used on your system or by the array itself. They are usually set up to happen on some periodic basis such as hourly, every four hours, twice a day, etc. This period is dependant on how long IT wants to have their data exposed. For example, if data within an environment cannot be more than four hours old, then traditional tape backups can address this issue. (Assuming that by traditional we mean nightly incremental backups and weekend full backups.) Snapshots help to alleviate this problem. By making copies of the data throughout the day, IT can recover information that is not as old. Recovery from snapshots takes very little time. The snapshot is mounted, as a volume, and a journal is played in which all files that had been unchanged are moved into the volume. The volume is then ready for use with the data that was captured at the time of the snapshot. This process is much faster than trying to recover yesterday's backup from tape. In addition, that backup would only be as current as the previous evening. Depending on the failure in the day, a full day's worth of work could be lost.

Replication

Replication technology is also getting a lot of attention these days. Following the events of Sept. 11, a number of vendors approached the end-user community with disaster-recovery solutions. The reality is vendors had replication products and were trying to pass them off as disaster recovery solutions. These replication products were rudimentary at best. Over the past two years however, vendors have really picked up the

development of replication technologies and made them more reliable, easier to deploy and take advantage of low bandwidth and have better integration with important applications such as e-mail and databases.

Replication is the process of creating a second copy of data on a remote array. There are hardware-based solutions where the array will transmit the data from one array to another, and there are software-based solutions where software that resides on a host will transfer data from one array to another. The primary difference between these two is cost. There are many types of replication as well, but two are most predominant, and of these, one is much more widely used. The two types are synchronous and asynchronous replication, and synchronous replication is used only in special circumstances when a second exact copy of the data is a must.

Synchronous Replication. This form of replication adds a degree of latency to the overall data protection process. However, it does ensure that each array has the exact same copy of the data. How does it work? When the application writes a piece of data, the replication engine intercepts the write and sends it to the primary array as well as the secondary array. Only when both systems have confirmed that the write has made it to disk does the replication engine inform the application that it is okay to accept another write and process it. As you can see, for a high transactional application, this process can slow things down. Additionally, if this type of replication is required, bandwidth for transferring data becomes the issue, and increasing the size of the pipe between the two arrays is often necessary, and expensive. There are certain applications or certain business requirements that call for this type of replication. A number of large trading firms on Wall Street use synchronous replication when procuring trades. Synchronous replication is typically block-based replication, where changed blocs of a database are replicated to a remote location. This location is usually close in proximity because of latency issues and costs associated with the infrastructure used to get the data from Point A to Point B in such a way it makes the application useful. There are also issues with synchronous replication. If a write to the secondary array fails or the link between the two arrays goes away, the application has to stop (or special policies have to be put in place to tell the application that should some piece of the infrastructure fail, it is OK to keep processing).

Asynchronous Replication. Asynchronous replication is the most common type of replication because of the nature in which it operates. As data is written to the primary array, the replication engine grabs that write and sends it to the secondary array. As this happens, and unlike synchronous replication, the application can continue to operate, not waiting for a response from the secondary array to acknowledge the write made it. Most replication tools manage the process and consistency of what data is where. In fact, when running in an asynchronous mode, most replication products can set a number of parameters to deal with the existing infrastructure limitations. Replication tools can be set up to use only a certain amount of the link between the two arrays, it can be set such

Data Protection Highlights

In a recent survey of over 250 end users, ESG found the following:

- 75% of IT professionals believe that e-mail is their most critical application
- 27% of IT professionals believe that their most critical application cannot be unavailable for more than four hours
- 32% of IT professionals believe that their data is at risk or exposed
- 40% of IT professionals will be increasing their incremental back-up frequency to more than once per day

that it can be a certain number of writes behind the primary array. It can be set such that buffer writes and sends them at certain times. Additionally, because of the nature of asynchronous replication, other techniques can be used for moving data from one loca-

tion to another. The ability to move data at a given point in time during a day, to move only a certain amount of data at a time, or the ability to utilize only a portion of the pipe between replication sites are all critical to helping IT save money when it comes to data replication.

Now that we know the different types of replication, how can they help IT professionals deal with some of the challenges they are facing today? We have already discussed protecting the edge data centers, and now will get into back-up consolidation and disaster recovery.

Data Protection and Consolidation

Protection and Consolidation. As we discussed earlier, data protection at remote data centers and remote offices is a growing concern for IT. Today, IT spends a good deal of money trying to protect remote offices and data centers. There are application servers, back-up servers, back-up licenses, tape libraries and people all responsible for ensuring that the information at the edge is protected and protected properly. The problem is IT does not know that the information is properly protected and because they do not know, they are exposed. Ultimately, IT would like to be able to take advantage of the best practices within the data center that they use on a daily basis to ensure they are meeting the data-protection needs of the lines of business and the company. Additionally, they would like to be able to do all of this in such a way that they can reduce costs, reduce data exposure, simplify the way in which data protection is rolled out to remote data centers, which would make it easier to protect information when new servers or applications are deployed. How can all of this happen? A good place to start is with replication. How can replication help? There are a number of things replication can do to help including removing back-up server costs, tape library costs, back-up license costs, actual tape costs and administrative costs.

Quick Recovery is Vital



Tape backup = Days



Data Replication = Minutes

Microsoft Windows Storage Server 2003 and Double-Take deliver solutions for affordable, quick recovery.



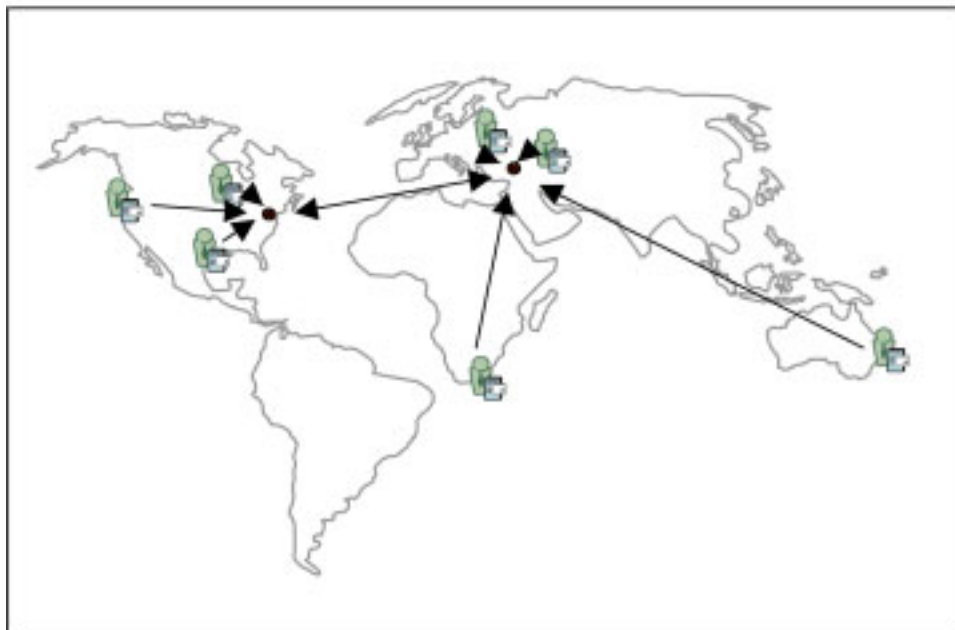

[NSI Software](#)

Disaster Recovery. Another hidden benefit of doing replication for data protection is disaster-recovery protection. Having an additional copy of the data replicated to a remote location greatly increases an organization's ability to get back up and running in the event of some type of disaster. Statistics show that a large number of small organizations do not recover (and even go out of business) from a disaster if they have not put the proper data protection components in place. As stated earlier, a number of vendors discussed products that were available to help with disaster recovery, but it is only today that IT folks actually have real solutions to help them with disaster recovery.

Data Migration / Planned Outages.

Replication also can be used to help IT centers move data centers or even consolidate data centers. Most data center consolidations take place over a weekend, whereby IT has done a full backup (typically without testing the recovery of that backup) on a Friday evening. Then they bring up servers and applications at the new site and perform a recovery of the data they backed up. This means that in addition to installing the application software, they have to install the backup software and tape libraries to perform the recovery. IT has to cross their fingers to make sure all of this goes as planned. A number of things can happen: improper drivers for the tape library and server operating system, initial backup did not work properly, as well as a host of other issues. Many of these problems are solved with replication. Installing the replication software on the host to receive the data is probably the only additional step required. As long as IT has done their homework to ensure that the version of the operating system the server is running supports the replication software, they are off and running. With replication, there is no worry about data loss or a potentially incomplete backup. The data that was replicated to the target server is there and IT knows this before they perform the migration, or shut a system down to put in a new system. Once the new site or system is up and running, all IT needs to do is change the direction of flow of the data to be

Figure 2: Replication for Data Protection and Disaster Recovery



from the target to the new primary and let the data migrate to the new location or server.

Technology Integration

We have discussed snapshots and we have discussed replication as new additional ways to protect data. We also have highlighted in the Data Protection Highlights section that 40% of the IT professionals that took part in ESG's Data Protection Research Project are going to be moving to more than one incremental backup per day. The number of folks that do weekly full backups is moving to more than weekly full backups by 10%. There is only one way folks will get here. By combining technologies such as snapshots and replication, IT will be able to meet all the data protection needs they have for instantaneous recovery at remote sites and long-term data protection needs. One benefit of integrating these two technologies that is key is one of the pitfalls for replication. The adage always said, "It's not the computer's fault; it did exactly what you told it to do." applies here. As with replication, a replication engine does not intelligently know what it is replicating, so if corrupt data is put into the replication engine or if a piece of data is deleted, the engine will replicate the corrupt data or replicate the dele-

Special Report

tion. By taking snapshots of data and replicating the snapshots, the ability to roll back to a point in time where data was known to be valid is a bonus. And as long as this window meets the RPO of the line of business, then the solution fits for local data protection and recovery, meets the challenges for traditional backup and meets any and all requirements for disaster recovery.

The Real World

Where does technology like this fit? Banks with remote office branches, software corporations with remote development offices, CAD/CAM design centers that are remote from the head design center, or any company with remote sales offices that are important to the business. Each of these scenarios are good examples where the current data-protection practices could use some assistance.

Let us look at the example of a bank with remote offices. Typically, these branches do a number of transactions per day into some database application that tracks user account numbers, deposits and withdrawals, and any other transaction that may be associated with a customer. Today, a teller may be responsible for putting the back-up tape into the back-up library and ensuring, the next morning, the backup was successful. This person is probably not qualified to make sure that the back-up was successful. Also, what happens to the tape to which the data was saved? These tapes should be stored off-site. Is the teller also responsible for testing that the backup was successful against the database? These tasks today are probably not taken care of.

In the new world of replication, IT can send an IT professional to the site once to install the replication software on the production database servers and enable replication. Once this is done, data is now replicated to a volume at the main data center where IT can perform a local backup on that volume for data protection.

Another scenario for a crucial database application, could be that IT set up snapshots for the application. At points in time, every two or four hours, a snapshot of the database can be taken. Then, the replication engine can replicate the snapshot volume. Now the local branch is protected should a problem happen during the day. IT could log onto the branch site, point the application to the latest valid snapshot with valid data, and the branch would be up and running again without losing much data or causing the tellers to have to enter too much data back into their system. Additionally, the last valid snapshot

that is replicated at the end of the evening becomes the full day's incremental backup that IT then moves off to archive tape. This combination of technologies has been implemented at many different types of transactional remote offices and data centers for better data protection and eases data recovery for the administrator.

Another key feature is that IT can perform a backup on the replicated database without having to have special software called "open file managers" to backup a running database. The target for replication does not have a database running on top of it, so all files are closed. This allows more constant use of databases at the local site.

Another scenario is for remote offices with users who generate a number of personal productivity files such as Word, Excel, Powerpoint or even CAD/CAM files. Traditionally a back-up product is installed at the remote site. A master server license is installed on the back-up server and then client licenses are installed at each of the desktops. The master backs up the clients to the master server with a local tape drive. All of the problems of someone having to ensure that the tapes are in the library and that the next morning testing is done to make sure the backup was successful, and that the tapes make it to a "safe" location still exist.

For less money, IT could set up a NAS device in the remote office or data center and have each employee store their most important files on this device. Then, a replication engine could migrate the data from this one box back to the main data cen-

Quick Recovery is Vital



Tape backup = Days



Data Replication = Minutes

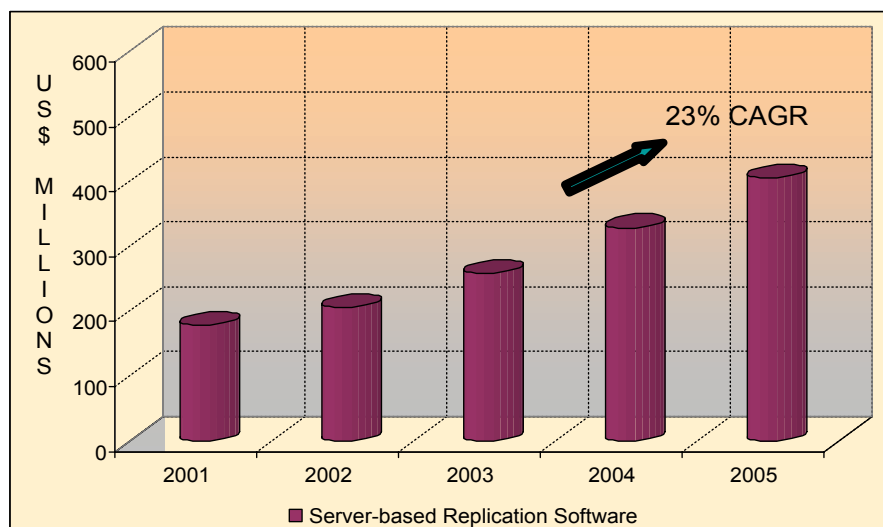
Microsoft Windows Storage Server 2003 and Double-Take deliver solutions for affordable, quick recovery.




[NSI Software](http://www.nsi-software.com)

Figure 3

The Worldwide market for Server-based Replication Software will grow from \$177M in 2002 to \$405M in 2006.



Source: Enterprise Storage Group 2003

ter where again, IT can do a backup to tape for archive purposes. Should an end user request to recover a lost or deleted file, IT just needs to go to the replication target and move the file back to its original location. These solutions save IT a great deal of time and money.

The Emergence of Snapshots and Replication

In a report delivered in May of 2003 by the Enterprise Storage Group, they predicted that the software-based replication market was going to grow from \$177M in 2002 to \$407M in 2006 with a 23% CAGR. Additionally, the overall replication market, hardware and appliance based, would grow from \$1.2 billion in 2002 to \$2.6 billion in 2006 with an annual 23% CAGR. Additionally, this growth rate is 70% of the growth rate of traditional back-up technologies in the market, which really reflects the fact that companies are serious about replication these days.

The other question that comes up is the age-old debate between disk and tape as it comes to data protection. The interesting thing here is that there is no debate. The reality is that both will have to co-exist for a long time. Tape today is the most economical way to archive data, and archiving data is very important. Replication can play a large role in protecting data more

efficiently and more economically. The combination of replication to disk (within the data center, where there are best practices) and archive to tape complement one another in a cohesive data-protection strategy.

By taking the burden of backup and recovery away from less-qualified folks at remote offices or data centers and putting it in the hands of the folks that do data protection for a living makes sense and keeps as many folks within the remote offices and data centers busy working on more productive tasks that are efficient to the business.

At the end of the day, it is all about protecting the valuable information assets in the most economical way possible, including efficient use of human capital. Information availability is critical to every business — whether it be for online product decision — and this

information can make transactional trading or the ability to do some analysis and review market trends for the next product decision, can make or break business, don't let it break yours.

© 2003 Network World, Inc. and Enterprise Storage Group.
All rights reserved.



[Request Reprint](#)

[Request a reprint of this report.](#)