White Paper

Building Practical Data Protection Strategies

Fred Moore, Horison Information Strategies

Abstract

Data protection has become the most critical IT discipline and businesses have historically chosen the simplest approach to perform this task. However, the simplest approach, such as creating a full backup to tape, is no longer economical nor does it provide the highest availability when compared to the next generation solutions.

This paper compares traditional data protection options and provides an overview of the impact data deduplication is having on the storage industry. Deduplication is quickly gaining momentum in the growing disk-to-disk backup market and offers several significant improvements over traditional data protection methods.

Deduplication can significantly change the economics of storage, dramatically reducing both backup recovery times and making Wide Area backup a operational reality.

Building Practical Data Protection Strategies

Table of Contents

The Origin of Backup and Recovery

In 1952, the world's first successful tape drive was delivered, the IBM 726 with 12,500 bytes of capacity per reel. In 1956 the world's first disk drive was delivered, the Ramac 350 with 5 megabytes of capacity. Though no one knew it at the time, two key events in the storage industry had just occurred; 1) the first storage hierarchy was created with online disk and offline tape storage, and 2) the first storage management applications were born, namely backup and recovery. Backup and recovery were to become the primary storage management applications for the next 50 years as protecting valuable data became increasingly important. Will these traditional data protection applications survive the demands for ultra-high availability and the growing requirements for nearly instantaneous recovery of the twenty-first century? The answer to this question lies with the next generation backup and recovery solutions that are now gaining significant momentum.

Data protection is quickly becoming the most critical piece of IT strategies today. There are four fundamental stages in the lifecycle of digital data:

1) data creation

2) data access

3) data archive and

4) data deletion/destruction.

The deletion/destruction phase no longer applies to all data types as considerable amounts of data is being stored indefinitely if not forever for a variety of reasons. The overarching goal of data protection is to protect information that cannot be easily replaced or replaced at all throughout its meaningful lifecycle.

Different levels of data protection exist and as expected, higher levels of data protection cost more to implement. Software errors, human errors, natural disasters, the increasing number of power failures, building damages, and destructive intrusion such as worms, viruses' and spy-ware have turned data protection into a complex data management process. Improved data protection and security levels have evolved over the years from consistently improving the MTBF of hardware devices to implementing a variety of local and remote strategies to address the numerous causes of downtime.

Traditional Data Protection Options

There are multiple data protection options that traditionally cover the spectrum of RPO and RTO time frames. They are classified into the following categories: Backup, Disk Mirroring, Snapshot Copy, CDP, and VTL.

Backups are usually performed once a day during the backup window. A significant investment has occurred over the past 20 years to ensure that applications, especially database related, can be interrupted or paused during this window so that a reusable copy of the data can be made. Without this critical host software and backup process investment, the usability of the copy for recovery purposes, may be severely compromised since there's no guarantee that all data for all transactions have been captured in a way that is understandable by the application.

Will these traditional data protection applications survive the demands for ultra-high availability and the growing requirements for nearly instantaneous recovery in the twenty-first century?

Terms Used to Classify the Impact of Different Levels of Data Protection

TERM	DEFINITION
RPO (Recovery Point Objective)	The desired amount of time between data protection events.
RTO (Recovery Time Objective)	The time required to recover from a data loss event and return to service. In other words, this requires classifying data or an application by its criticality or value to the business and determining how long the business can survive without having this data available. There is immense pressure to greatly reduce the amount of time it takes to recover a file or an application.
DPW (Data Protection or Backup Window)	The maximum amount of time available for an application to be interrupted or out of service while data is copied to another physical location for backup purposes. This "time out" component is seeing many new developments to reduce the amount of time needed for the traditional backup window.

Over the years, various approaches have evolved to shrink the time that applications are offline. This continues to be an area of investigation, however field testing over the application portfolio and diverse customer environments are critical with each new approach. Building the administrative processes, minimizing complexities in deploying the solution, ensuring coverage across all applications and gaining confidence in new approaches are some of the primary reasons why customers refuse or delay adopting new data protection technologies.

The Causes of Downtime and Data Loss

TYPE OF DISRUPTION	PERCENT OF FAILURES	SOLUTIONS
Software (bugs and corruption)*	18%	Snapshots, transaction logs
Hardware*	38%	
Disk		RAID 1–6, backups
Tape		Backups
Server		Clusters, failover architecture
Network*	23%	Redundancy
Natural disasters, power, flood, fire	7%	Off-site facilities
Theft	2%	Encryption
Intrusion/security	8%	Firewalls, authentication, anti-virus, filtering
Other	4%	

* Includes operational failures

** Unplanned downtime (unpredictable)

Source: Estimates by Horison Information Strategies

Full backup/restore applies to all data types and is the most common data protection method. This approach copies 100% of the data, usually a complete file or full volume, from primary disk to either tape or disk for backup. ***The backed-up copy is not executable and must be restored to become accessible by an application***. In most cases, traditional backup and restore from a backup copy causes the application to be impacted or even stop for the duration of the process. The larger the object being backed up or restored, the longer the application and its customers must wait. For mission critical or revenue generating applications, any amount of time spent waiting for a backup or recovery operation to complete becomes very costly.

Full file or full volume backup and restores are the most time consuming of all the data protection techniques and may be difficult to schedule. Years of research have shown that much of the data being backed up is the same unchanged data that was previously backed up. Backing up large amounts of unchanged data creates a significant amount of unnecessarily duplicated or redundant data however full backup and restore continues to be used since it provides a high level of confidence as a data protection strategy.

Because of the time-consuming nature of the backup and recovery process, traditional backup/recovery is morphing into several derivative techniques. Tradeoffs exist when choosing from these options to build an effective backup strategy and they should be carefully reviewed.

For incremental backups, only the data that has changed since the last full or most recent incremental backup is backed up. Since only changed data, files or blocks depending on the implementation, is backed up, this method minimizes the amount of data backed up and reduces the amount of time needed for the "backup window." However a full restore takes longer as each incremental backup will normally have to be restored in sequence to return all changes to their last known state. This can become a complex and time-consuming process. Often a full backup will be performed weekly while an incremental backup is performed daily. Incremental backup minimizes the amount of backup time but makes recoveries more difficult.

For differential backup, the same changed data that was backed up on the previous differential backup plus any new changes are also backed up on the next differential backup. That's why differentials typically grow in size each day between full backups. This means that daily backups gradually become larger and therefore take longer, but the restore process is simpler and usually shorter compared to full or incremental backups. A full restore only requires the last full backup and the last differential copy to complete. Differential backups take longer than incrementals but the recovery times are faster.

Incremental backup minimizes the backup time and differential backup minimizes the restore time and the specific application may favor one or the other. These tradeoffs are often confusing and time consuming for storage administrators. Most businesses want to shrink the painful amount of backup and reduce the recovery time, not just one or the other. Both incremental and differential backups are application and backup software specific. Multiple applications may require several implementations adding complexity.

Disk mirroring applies to all data types and is implemented as a block-for-block replica of a file, a logical unit, or a physical disk volume using either local or remote disk drives for all copies. Once the mirrored data element is established by copying the original data element, all subsequent write operations are stored in two (or more) places creating identical or nearly identical copies. In the event of a failure of one of the copies, disk mirroring provides nearly instant access to the secondary copy allowing the application to continue without downtime. Mirroring is costly as it doubles the amount of disk storage required adding significant acquisition and ongoing operational expense. Storage administrators also must choose to implement either asynchronous or synchronous mirroring and tradeoffs exist for each case.

Backing up large amounts of unchanged data creates a signicant amount of unnecessarily duplicated or redundant data.

Synchronous mirroring is frequently used in mainframe and other high availability environments given the critical nature of its applications. In synchronous mirroring, both the source and the target devices must acknowledge the write is completed before the next write can occur. This can degrade application performance but keeps the mirrored elements synchronized as true (exact) mirror images of each other.

For asynchronous mirroring, the source and target devices do not have to synchronize their writes thereby allowing the second and subsequent writes to occur independently. Therefore, asynchronous mirroring is faster than synchronous mirroring but the secondary copies are slightly out-of-synch with the primary copy. Asynchronous mirroring is sometimes used in the wide-area backup market to replicate data to locations hundreds of miles away. This becomes very costly as the amount of data remains large and high-speed bandwidth is expensive. In reality, the secondary data element is rarely any more than one minute behind or out-of-synch with the primary copy. This can become a significant exposure for certain mission critical or write-intensive applications.

Mirroring is used for many mission critical applications and it is the fastest way to access data from subsystem hardware or device failure since the recovery operation occurs in a few seconds or less by automatically switching to a mirrored copy. Note that mirroring does not help protect against data corruption problems such as hackers, worm, virus, intrusion, human or software errors as it generates two or more copies of corrupted data. For best practices, mirroring should always be accompanied by other data protection schemes that can permit a recovery or restore to occur from clean data that existed before the corruption occurred. Disk mirroring is defined and commonly referred to as RAID 1.

Given the many tradeoffs and limitations in these traditional data protection methods, several other techniques are gaining momentum to reduce some of the traditional tradeoffs.

Snapshot copy presents a consistent point-in-time view of changing data. Snapshot technology is available from a variety of data storage vendors in a variety of implementations, but not all snapshots are created equal. When using snapshot copy and write operations occur, the changed areas (writes) are saved in a separate area or partition on disk of disk storage specifically reserved for snapshot activity. Here the old value of the affected area or block can be saved in case the new block(s) are corrupted or to permit a fuzzy data image that can be used for a non-disruptive backup. Storage administrators must manage the number and currency of snapshots. Snapshots provide data protection from intrusion and data corruption but not from a device failure that contains the source copy of data. Again, tradeoffs exist. The challenge for snapshot copy is that it is not easy to find the exact snapshot copy just before the corruption takes place and complexity increases as the number of snapshots grows. Multiple application and software dependent implementations complicate the use of snapshot.

CDP (Continuous Data Protection) enables data recovery where every write and update operation is continuously written to a disk device that may or may not be the same as the primary device. If Snapshot Copy is a series of still images, then CDP is like a movie. Unlike mirroring however, the secondary copy is a sequential history of write events with a timestamp. All write operations are queued to the secondary disk or the journal device. Journals are typically kept as a continuous history for 2-4 days covering the period of maximum likelihood for a data recovery action to occur. Journals are especially good for protecting from intrusion and data corruption enabling restores to go back in time to a point before the corruption occurred. Sparse journaling is available to journal writes for only the more critical tasks. Tradeoffs exist. CDP has the same challenges as snapshot copy and does not replace traditional backup or provide protection in the event of data center loss. CDP is ideally suited to perform quick recoveries in case of data corruption for environments that have a small amount of changing data. Multiple application and software dependent implementations complicate the use of CDP and many applications do not support CDP.

VTL (Virtual Tape Library) solutions have proliferated and over 25 companies presently offer a variety of virtual tape implementations for the backup/recovery market. Virtual tape concepts for mainframe computers were pioneered and popularized in 1997 by IBM and StorageTek as the effective utilization of tape cartridges was historically low. The subsequent benefits of virtual tape implementations are well documented and VTLs are now becoming available for non-mainframe computer systems primarily being used as a fast disk cache.

VTLs are now becoming available for non-mainframe computer systems primarily being used as a fast disk cache.

A virtual tape library consists of a disk array that appears as several tape drives attached to a tape library but doesn't include an actual physical tape library. Two general classes of VTLs are available: integrated and standalone. An integrated virtual tape library combines disk arrays as a front-end to an automated tape library whereas a standalone does not directly control the physical tape library. The disk storage serves as a cache or buffer for more active data for the larger-capacity and lower-cost tape library. The device images presented to the operating system appear as multiple tape drives rather than physical disk drives, therefore "virtualizing" the disk by making it look like something other than it really is.

Virtual tape systems store multiple virtual tape volumes on a single physical tape cartridge, commonly enabling cartridge utilization levels to reach 80 percent or more. This is becoming increasingly important as tape cartridge capacities rapidly increase. Pre-established policies such as file size, capacity utilized, and frequency-of-usage patterns determine when the data is moved directly to and from the disk buffer and the automated library. In this case, the HSM (Hierarchical Storage Management)-like policy based functionality resides outboard of the host or application server, directly controlling bi-directional data movement between disk and tape storage thus enabling direct data transfer between both tier 2 and tier 3 storage.

Backup/Recovery Strategy Comparison

BACKUP SCHEME	TRADEOFFS	DATA TYPES SUPPORTED	DATA COMPRESSED ON BACKUP	STANDARD STORAGE DEVICE	FAILOVER OR RESTORE ARCHITECTURE
Full backup	Straightforward process, maximum storage consumption, can be used on all applications	All data supported	Yes	Tape	Restore
Incremental backup	Reduces backup window, all incremental copies used in recovery making recovery longer than differential	Application specific, not supported by all applications	Yes	Disk, tape	Restore
Differential backup	Reduces recovery time, only last differential copy used in recovery making backup time longer than incremental	Application specific, not supported by all applications	Yes	Disk, tape	Restore
Mirror	Doubles disk costs, protects from disk failures, does not protect against intrusion or corruption	All data types supported	No, one-for-one copy	Disk only	Failover
Snapshot	Very fast, does not protect against disk failures, does not protect against intrusion or corruption, can take time to determine the recovery point	Application specific, not supported by all applications	No	Disk only	Restore
Continuous (CDP)	Is often application specific, can take time to determine the recovery point, fast recovery, protects against intrusion and corruption	Application specific, not supported by all applications	No	Disk only	Restore
VTL (Virtual Tape Library)	Mature for mainframes, gaining momentum in open systems	All data types supported	Limited to tape compression capabilities, significantly reduces performance	Low-cost disk with optional automated tape library	Restore
Deduplication	Very fast backup and recovery, greatly reduces disk storage requirements and expense over all other methods, compute intensive requiring appliance	Application and backup software neutral, avoid only non-repeating data types	Yes, provides further data reduction via eliminating redundant data	Disk only	Rapid Restore

The major benefits of virtual tape include higher performance (as data can frequently be accessed from the disk cache), significantly increased tape-cartridge utilization, simpler backup policy management by defining more tape drives than physically exist increasing parallel I/O activity, and a financial savings resulting from the reduction in the physical number of tape drives and media. Again, tradeoffs exist. There are numerous implementations and choosing the optimal VTL for your business can be time consuming.

Next Generation Data Protection Solutions Arrive as Deduplication Takes Off

Backup solutions and enhancements continue to evolve since its early days. Many improvements have been made in both hardware and software to improve performance. However, these solutions barely assist in keeping pace with the growing amount of valuable data that needs to be backed up and protected.

Deduplication is quickly gaining momentum in the growing disk-to-disk backup market and offers several significant improvements over traditional data protection methods.

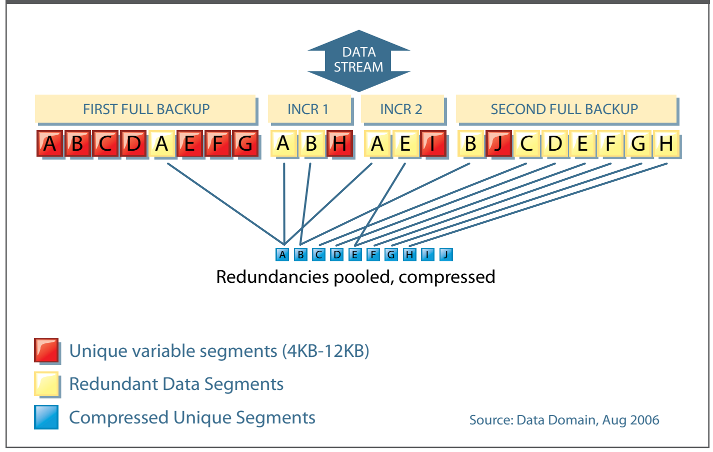
Deduplication can significantly change the economics of storage, dramatically reducing both backup recovery times and making Wide Area backup a operational reality. Deduplication goes by many different names including global compression, commonality factoring, capacity optimization, single instance store and referential integrity. Deduplication can also be used in-line as the data is backed up or can be used after the backup has been completed. Next generation data protection solutions are disk-based and have presently become the rage of the storage industry.

In-line deduplication reduces the amount of storage capacity required since it only stores unique occurrences of data. Combining deduplication with data compression makes the capacity savings even more compelling. Realistic results from leading vendor's initial deployments indicate significant reductions for backed up data. In many cases, incremental storage is less than 1% of the original size of the backup. Therefore the biggest advantage of deduplicated data besides the reduction in physical storage is the significantly reduced amount of time spent in data recovery process. Tradeoffs exist. Data deduplication is computationally intensive and can cause some performance degradation depending on the size of the storage pool. To resolve this issue, some companies have implemented a deduplication architecture by using an in-line appliance. Multiple appliances may be required as the size of the storage pool increases. Nonetheless, the increased speed of today's microprocessors has made deduplication a reality and the overall benefits including storage space savings and the improved RTO should significantly outweigh this tradeoff in most all cases. For a comparison of traditional and next generation data protection options, see the Backup/Recovery Strategy Comparison table above.

Data Reduction is Becoming the Key

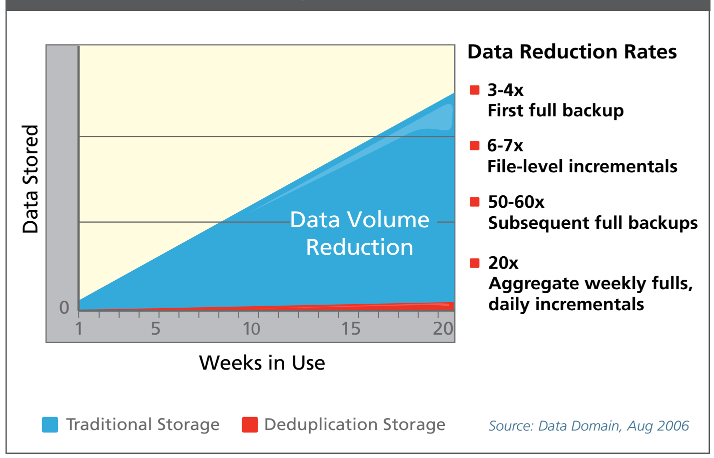
Reducing redundant copies of data can significantly shrink storage requirements. This lowers storage costs since fewer disks are needed, reduces power, cooling and floor space, shortens backup/recovery times and enables WAN-based disaster recovery since there can be far less data to transfer. Deduplication works best when there is a high degree of duplicate data to be backed up each day. Most data does not significantly change each day and therefore can benefit from deduplication.

Deduplication: Under the Hood



In-line deduplication segments the incoming data stream, uniquely identifies the data segments, and then compares them to segments previously stored. If an incoming data segment is a duplicate of what has already been stored, the segment is not stored again but a pointer is created for it. If the segment is deemed to be unique, it is then further compressed with conventional algorithms for an average 2:1 size reduction, and stored to disk designated for deduplication storage.

Compression Effect



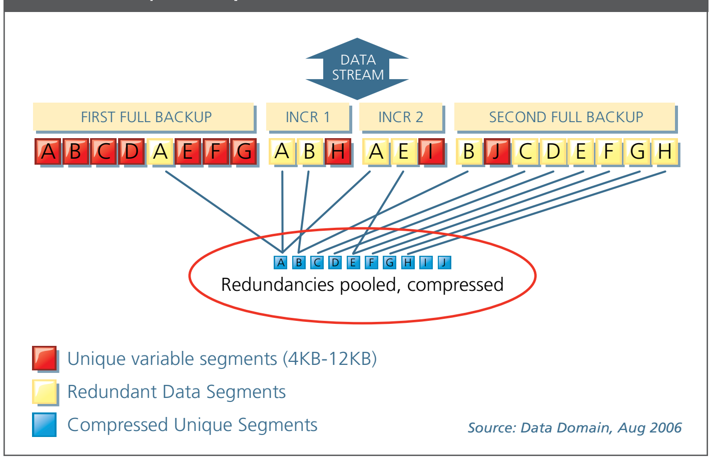
The degree of storage reduction from deduplication various by data type and by backup policy. Realistic results from leading vendor's initial deployments show the following spectrum of data reduction rates. Over time for a given backup policy of daily incrementals and weekly fulls, a 20x reduction or more is possible.

File deduplication is another space-saving technology intended to eliminate redundant (duplicate) files on a storage system. By saving only a single instance of data or a file, disk space can be significantly reduced. For example, if a 20 gigabyte file or volume with average daily data changes rates is backed up seven times in a week, a significant amount of unnecessary duplicate data is stored. This is costly in terms of storage hardware and often difficult to sort out for timely recovery purposes. Is there a valid reason that duplicate data should be stored seven times? Deduplication algorithms can be used to reduce this unnecessary over head. Deduplication algorithms analyze the data and create pointers for six of the seven redundant copies of that file thereby reducing the total storage required.

Deduplication Redefines the WAN (Wide Area Network) Market

At last, backing up and recovering data to and from remote locations can be done quickly and can become an affordable reality. By shrinking the amount of data sent by as much as 99%, deduplication significantly reduces the time required to send data to remote locations by vaulting only new and unique data segments in a compressed format (see diagram below). This enables enterprises to asynchronously backup/replicate selected business data from a primary location over a WAN to a secure off-site location. This process occurs automatically as backup data is stored, so there is no need for administrators to allocate additional time each day. WAN vaulting, enabled by the massive data reduction capability of deduplication, significantly reduces the bandwidth costs and security risks associated with human intervention and physical transportation associated with removable media. After years of struggling with expensive broadband costs for wide area data transfer, the massive data reduction capability provided by deduplication makes using networks to move backup data offsite for disaster recovery economically and operationally feasible.

Only Deduplicated Data Moves Across WAN



Preparing for deduplication makes the implementation process easier. The checklist below highlights some helpful considerations to make this task easier.

Deduplication Checklist

ISSUE	THINGS TO CONSIDER
Compatibility with your existing backup applications	What changes, if any, are required to your existing backup processes? Assess difficulty of the implementation. Are you prepared? Certify the backup vendor works with the deduplication solution.
Appliance or software based solution	Software-only solutions allow enterprises to re-use existing or older disk systems, appliances normally include disk storage and simplify installation.
Sizing performance requirements	Significant reductions in storage may require several iterations. Determine time needed to achieve the maximum deduplication effect. What is the net expected storage savings compared to full, incremental or differential backups?
Scalability for performance and capacity	Assess instrumentation is provided to measure performance and capacity thresholds. Insure additional capacity and processing power be added to keep up with data growth or to improve backup and restore speeds. Verify upgrades non-disruptive.
Pointer security management	Insure the pointers to the unique data segments are protected from hardware, software and intrusion threats. Losing these pointers means that the backups can no longer be retrieved.
Supports remote replication for disaster recovery	Data reduction not only reduces the amount of storage required locally but also reduces the amount of network traffic during replication, which can result in significant bandwidth savings for remote backup and recovery.
Pricing fit to budget	Do you pay for storage capacity, by number of installed solutions, or another scheme? What are the warranty period and maintenance pricing considerations? Remember not to confuse the initial acquisition costs (Capex) and the ongoing operational expenses (Opex) when evaluating a deduplication solution.
Vendor readiness and positioning	Check out vendor reference installations. Verify the vendor's roadmap will address your future data protection requirements.
Environmental savings	Determine power, footprint and facilities savings for use in cost-justification process

Data Domain: Deduplication Storage for the Distributed Enterprise

Data Domain has developed a unique non-disruptive approach to deduplication. Recognizing the growing magnitude of both the local and Wide Area backup and recovery problem, Data Domain has become the fastest growing provider of next generation data protection solutions. Data Domain has implemented deduplication via in-line arrays and appliances significantly reducing the storage needed for disk-based backup. This reduction makes disk-based storage economically feasible for backup and recovery.

The requirement to have data stored at geographically dispersed locations is becoming more critical as site damages and the frequency of natural disasters increase daily. Deduplication is the critical technology that finally enables businesses to economically replicate backup data over WANs for much faster and reliable remote site disaster recovery. Typical percentages of a full volume or source copy of data needed for various replication techniques are:

Typical Data Reduction Rates

TYPE OF BACKUP	DATA REDUCTION RATE
Full Backup	100%, all data backed up each time
Incremental/differential	5% for files, 100% for databases
Mirroring	5%, changes absorbed into each mirror
Snapshot, CDP	5%, only changes stored
Deduplication	< 1%, eliminates redundant data and compresses original occurrence

These typical data reduction rates demonstrate the magnitude of Data Domain's deduplication architecture.

Seamless integration into the existing backup infrastructure is especially critical for new data protection solutions since organizations are resistant to experimenting with their data protection strategy. All Data Domain Enterprise Protection Storage systems meet this requirement and work without change to existing backup software and operational procedures. By working with existing backup software. Data Domain also enables businesses to consolidate multiple backup products into one backup/recovery architecture.

Data Domain Highlights

PRODUCT FEATURES	CUSTOMER BENEFITS
Reduces storage and bandwidth costs with Data Domain 20x Global Compression™ technique	Makes WAN network vaulting affordable. Global Compression reduces data sent to remote location while protecting data over any bandwidth and distance.
Full support for all backup software packages	Data Domain requires no changes to existing backup software providing ease of installation and ongoing operations. Using a transparent approach consolidates existing backup software processes.
Ease of operations	Administrators can spend time on other important tasks.
Continuous data verification	Ensures data can always be recovered. If an error occurs during transfer, Replication software recovers automatically and restarts data transfer.
Reduces the amount of disk needed for backup data, provides concurrency	Storage expenses are reduced. Both the originator and replicated data remain fully accessible while transfers are in progress without impact to availability.
Eliminates redundant data needed for backup and recovery	Improves recovery times for local and remote operations.

Conclusion

We are beginning to see the next generation of backup and recovery solutions appear. The time consuming backup and recovery processes of the past are becoming obsolete as the amount of time spent in the traditional backup window has steadily decreased. Data protection has become the most critical IT discipline and businesses have historically chosen the simplest approach to perform this task. However the simplest approach, such as creating a full backup to tape, is no longer economical nor does it provide the highest availability when compared to the next generation solutions.

Today's IT environments are demanding a more comprehensive strategy for data protection, security and high-availability than ever before based on numerous causes of data loss. Data recovery options must align with application and business requirements in order to yield the highest availability. New data protection solutions led by Data Domain's deduplication architecture are quickly gaining market momentum improving data and application availability while increasing the probability for a business to survive most types of outages. This is increasingly critical since most businesses can no longer survive without their IT function. Ensuring that the IT infrastructure is resilient to machine failures, intrusions, natural disasters, human mistakes, accidents and the digital crime wave makes implementing the best possible data protection solution well worth it.

About the Author

In 1998, Fred Moore founded Horison Information Strategies, an information strategies consulting firm in Boulder, Colorado, that specializes in marketing strategy, industry analysis and business development for the IT industry. Fred began his 21-year career at StorageTek as the first systems engineer and concluded as corporate vice president of Strategic Planning and Marketing. Fred served as Storage Editor for West World Productions and has written numerous books, articles and reports for the storage industry. He is a 1989 recipient of the Distinguished Alumnus Award and a 2004 recipient of the Arts and Science Scholar-In-Residence Award at the University of Missouri where he received a bachelor's degree in mathematics and a master's degree in computer applications in physical geography. A sought-after IT speaker, motivator and writer worldwide, Fred completed the Berkeley Executive Program in 1997. He currently serves on a few select boards in the storage networking industry.

Data Domain | 2421 Mission College Blvd., Santa Clara, CA 95054 | 866-WE-DDUPE, 408-980-4800