SNIA

Data Management Forum

Continuous Data Protection
Solving the Problem of Restoration

June, 2008

Table of Contents

Introduction

Today's businesses are facing an ever-increasing amount of data that is threatening to undermine their existing storage management solutions. The foundation of today's business operations is data. Data is created, stored, massaged, transformed, mined, modified, deleted and utilized continuously across an organization. This data is present on many systems, on hard disks drives in laptops, in flash memory on USB key drives, and in large server farms sharing array based storage on a SAN. What does all this data have in common? The high cost of managing it, sharing it, and protecting it from damage and corruption.

Data protection is not the simple process it once was. No longer can a few simple scripts manage the data stored on stand-alone machines. Explosive technology growth has resulted in complex environments with dedicated storage area networks housing an assortment of intelligent storage devices. Stand-alone applications have grown into complex, feature and data rich environments with a thirst for storage. As the business costs associated with downtime and data loss increase, many businesses are adopting new policies and procedures around system protection that provide recovery options to ensure timely business restart from data corruption or system failure.

Traditional backup methods have struggled to meet the data recovery point objectives (RPO) and recovery time objectives (RTO) of today's businesses. The focus of IT is shifting to new processes and technologies that deliver full, fast and reliable recovery of data. Simply throwing additional disk into the mix is not enough. This document reviews traditional methods of data protection and availability, including the premise of disk-to-disk backup, and offers insights into an alternative data protection solution - Continuous Data Protection (CDP).

Often touted for its ability to provide instant restoration and granular recovery, CDP also solves the backup window challenge by eliminating the backup window itself, enabling the creation of backup copies anytime, day or night, without affecting online operations.

CDP changes the cost structure of data protection, improves operational recovery, and provides a foundation for implementing application specific Service Level Agreements (SLAs) as well as a tiered storage model for your data processing environment.

What you will learn in this white paper includes:

1. Best practices for data protection and operational recovery
2. How CDP works to address your data protection, recovery and availability requirements
3. The different choices for implementing CDP using real life examples

The SNIA DMF Data Protection Initiative: the CDP-SIG

This paper is authored by members of SNIA's Data Management Forum (DMF) CDP Special Interest Group (SIG), whose mission is to evangelize continuous data protection by both acting as the worldwide-authority and resource on CDP and by educating IT professionals to help them make informed decisions regarding the technological uses and practical applications of CDP.

Historical look at Data Protection

Over the last thirty years, data protection has evolved as business needs and technology have changed. At first, most compute systems were stand-alone, mainframe-based systems. Disk based data storage was expensive, and tended to be limited to business critical usage. Additionally, networking and interconnectivity between systems was expensive and limited, tape was the prevalent interchange media for data, and every major system had one or more tape drives. Yet even in this environment, there was a need to protect the disk based data that resided on these systems. Out of this need arose the capability to backup and recover data from tape. As these systems evolved, so did backup technology. As systems became interconnected, it became more common to share a single tape drive or tape library between multiple systems. In these systems, one server owned the tape library, and the other systems became 'clients' and sent their data across the network.

Tape became the primary vehicle used in the 70's, 80's and 90's for backup because, at the time, tapes were more durable and reliable and less expensive then the disk-based media. Additionally, tapes could be written in a format that was portable across systems. The disadvantages associated with using only tape backup for application protection and long-term archive have evolved over time.

To backup an application and all of its data requires that the application be shutdown, so that its data files are in a consistent state. This is disruptive to the users of the applications, so backups were usually performed during a quiet time, typically midnight to 6 AM, when the applications were not in use. This time came to be called the "backup window." Today, companies competing in global markets rely on their applications 24x7; this has effectively eliminated the "backup window."

Additionally, backup operations were rarely trouble free. At times, a server would be shutdown, off-line, or an application would be in an inconsistent state, resulting in missed or corrupted backups. As data volumes grew, the time needed to backup the data increased, adding additional pressure to the ever-decreasing backup window. Backup vendors responded by creating new methods for accessing and protecting the data. They created open-file backup methods, implemented incremental and differential backup approaches, and designed their production environment to support faster, streaming tape drives.

The Introduction of "Snapshots"

On the heels of backup came the concept of Snapshots. A Snapshot is a fully usable copy of a defined collection of data that contains an image of the data as it appeared at the point in time at which the copy was initiated. A Snapshot may be either *based-on* or a *copy-of* the data it represents, and usually resides on the same array as the production volume. The two basic types of Snapshots described here include:

- The first type is a Snapshot copy, which is a full copy of the production data taken at a specific point in time. It is static, and once the Snapshot copy is created, it no longer depends on the production data and can be easily moved to other locations or servers. A Snapshot copy is the same size as the production data.

- The second is a delta Snapshot, which also represents the production data at the original point in time. However, it is tied to the production volume, and as the production volume is updated, the older data is preserved so that the duplicate can reference the original data. A delta Snapshot starts out much smaller than the production, but the data it relies on grows with every change to the production, as it needs to reference the original, unmodified data.

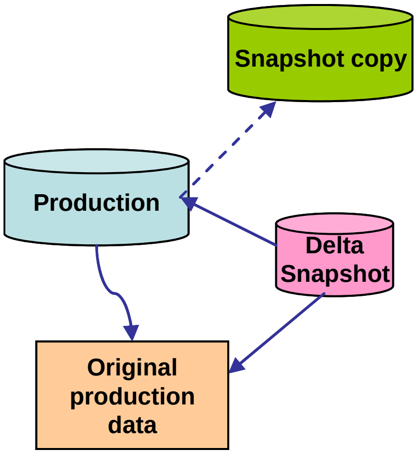

Figure 1 - Snapshot Architecture

Snapshot technology can be implemented at the host, in the storage network, or at the storage system level. Host based snapshots may be performed on the volume presented to the host (for example on disk 2), or at the file system level (for example on the E: drive). At the storage system level, snapshots are performed on the volumes or LUNs presented to the storage network. When implemented by the array, a Snapshot may be a full copy of a volume or LUN, or it may be a delta Snapshot, which just contains the changes necessary to apply to the current version of the LUN to recreate the image at a specific point in time.

With the proliferation of offerings in the storage area, many users are implementing some form of an information life cycle management strategy. By classifying the importance of different sets of data throughout their environment customers can identify the best methodology for backing up their data. For example, mission-critical data that resides on high-end arrays may have an SLA that cannot be met by solely using tape, as recovery can be slow, invasive, and costly. A better option would be to take periodic Snapshots or point-in-time copies of the data and then backup the Snapshot to near-line disk or tape. Customers looking to meet changing SLAs for data protection should look into enhancing their tape-based backup technology with support for Snapshots, near-line disk, or virtual tape libraries, or a combination of these.

Traditionally, Snapshots, which create "virtual" point-in-time copies of data, have been reserved for backup purposes, primarily of mission-critical data. Often, users use a Snapshot to create a copy and then perform a backup from the Snapshot in order to eliminate downtime for the application. However, there are many other uses for this technology. The appeal is simple: Snapshots allow users to make periodic copies of data (the frequency of the snapshots is determined by the user according to data protection requirements and space availability) which can be readily accessed in the event of a logical error (e.g., user corruption, virus, etc.).

Many applications have specific interfaces that can be invoked to place them into a Snapshot mode. For example, the Microsoft Windows Volume Shadow Copy Services (VSS) framework enables authorized backup products to invoke VSS to request that it 'quiesce' an application prior to a backup. A VSS-aware application will write in-memory data to the disk and mark the application as being in a consistent state. This helps improve the recoverability of the application's state from a Snapshot image.

Around the same time that Snapshot technology came to prominence, array vendors started promoting their replication technology for disaster recovery purposes. Operating at the array level, replication was seen as a fast, reliable and non-disruptive alternative to using backup for disaster recovery.

Array-based replication is not the solution for all data protection problems. First, array-based replication tends to be homogeneous, meaning that it is supported only on arrays from a single vendor, and sometimes not even across array families from a single vendor. For example, many companies have more than one storage array product families - a high-end or tier-one array, and a mid-range or tier-two array. For these storage vendors, array-based replication technology is commonly restricted to replication between a single array tier (tier one to tier one and tier two to tier two). Typically, vendors do not support array replication solutions across tier-one and tier-two product families. Secondly, array-based replication configuration is sometimes complex and may require a more experienced storage administrator or vendor configuration assistance.

In addition, array replication does not protect against data corruption. If data is corrupted on the production side, that change will be quickly replicated to the remote site. For example, if you accidentally drop the wrong table in a database, this change will be quickly replicated, resulting in a corrupted database at both sites. To overcome this situation, users started to combine array-based replication technology with external life cycle management support. This was commonly done by taking Snapshots of replicated volumes and then retaining those Snapshot images for pre-defined periods of time. In the example above, to recover from a dropped table, the user would extract the table from one of the earlier Snapshot copies. Of course, this is a manual process, and requires that the user keep multiple Snapshot copies, from differing points in time, available at the remote site. Keeping track of the Snapshot copies, and ensuring that they are resynchronized, if necessary, with changes to the production volumes, becomes a manual process and can tie up large amounts of an administrator's time.

Enhancing Traditional Data Protection with CDP

CDP systems offer virtually unlimited data recovery point options. They allow the user the freedom to identify a point-in-time just prior to a data corruption event for data recovery. The CDP system can also be used to create multiple copies of the data from any point in time. These time-based versions of data can easily be accessed to implement traditional, scheduled data backups to removable media, to replicate the data over distance for disaster recovery readiness, or to farm out the data for alternative uses (test systems, data mining, etc.), all without impacting ongoing application activity.

The SNIA CDP Definition

> *Continuous Data Protection (CDP) is a methodology that continuously captures or tracks data modifications and stores changes independent of the primary data, enabling recovery points from any point in the past. CDP systems may be block-, file-, or application-based and can provide fine granularities of restorable objects to infinitely variable recovery points. So, according to this definition, all CDP solutions incorporate these three fundamental attributes:*
>
> *1. Data changes are continuously captured or tracked*
> *2. All data changes are stored in a separate location from the primary storage*
> *3. Recovery point objectives are arbitrary and need not be defined in advance of the actual recovery*
>
> *A number of recognized technological approaches deliver CDP, including block-, file-, and application-based. Today, many vendors offer varying degrees of support and awareness of specific application and data environments. But regardless of the underlying technological approach utilized, CDP can offer faster data retrieval, enhanced data protection, and increased business continuity with lower overall cost and complexity.*

It is important to note that CDP systems deliver what is known as an atomic view of the data. All the data across all the disks is recorded as if it was captured at exactly the same moment in time. It is as if time stopped at that exact moment. This atomic view provides consistency and stability across databases, applications, federations and even entire datacenters. Upon recovery, the data is presented to the servers exactly as it existed at that point-in-time. Depending on the technology, in-flight transactions may have to be recreated or applications and databases may need to rollback or forward their changes to reach a stable environment. However, all of today's applications can perform this type of "crash-recovery" without the need for additional technology. CDP can be utilized to "clone" or dynamically recreate entire application environments without application involvement. In fact, using CDP gives customers the ability to stage application environments on a completely different SAN or even in a separate geographic location.

As we have established, CDP technology can, in parallel to online applications, present alternate views of data, from any point or event in time. An important use for these time-based views is as a source for creating copies using traditional, scheduled backup systems.

The Benefits of CDP over Traditional Data Protection

Efficient use of a CDP image for backup requires that users be given a process that creates a point-in-time copy, mounts the copy on the backup server, initiates a backup, dismounts and evaporates the copy – all at the touch of a button.

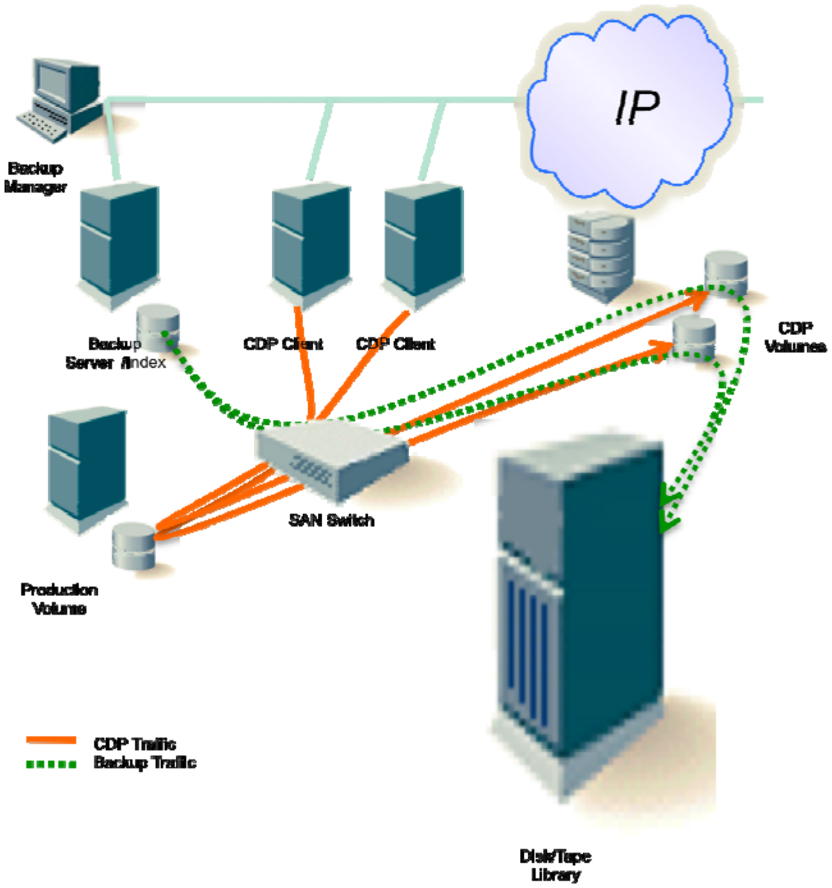

Figure 2 - CDP Architecture

Customers want CDP to be integrated with their backup product so that it can be scheduled and monitored though their standard systems, and not just a separate tool. Due to the value of CDP as a complement to backup, many of today's backup vendors have responded by integrating CDP technology into their backup solutions. This enables the users to utilize CDP technology to front end their traditional backup processes and helps them to minimize or eliminate existing backup windows. By combining CDP with traditional backup, the user gains the flexibility of any-point-in-time recovery for their latest data combined with the longer term backup or archival protection of older data in disk or tape libraries. By using a tiered protection architecture, a well integrated backup product will take the oldest CDP image, index its contents, and then store the image in the backup repository on a lower tier of storage, either on disk or in a virtual or physical tape library. When an older image needs to be restored, the backup product brings back the full CDP image or, if it is so configured or designed, it brings back the individual object.

Additional Uses for CDP Copies

There is a wide range of additional uses for point-in-time copies. These uses include:

- **Surgical Recovery of production data**
Recovery of a file or table from a database from an earlier point-in-time. The required point-in-time can be selected and the file or database table copied or extracted and moved back to the production server without effecting other data on the production volume.

- **Compliance analysis**
Using time-based views, an image can be mounted at a specific point-in-time and compliance tools (email searches, etc.) can be run against the image without impacting the production volume. This can also be useful to access data at an earlier point-in-time before it may have been changed on the production server.

- **Data Warehouse seeding**
The biggest cost when seeding or updating a data warehouse is copying the data from the production environment into the data warehouse environment. Using a point-in-time copy, data can be exported from the point-in-time copy into the data warehouse without affecting the production server.

- **End of period operations**
At the end of each period (week, month, quarter, etc.) a point-in-time image of the production data can be used to exactly represent the data as it existed at the end of the period. The image can be used for archiving a copy to disk or tape with the assurance that the archived data is an exact representation of the production data.

- **Cloning of an application environment for development and testing**
It is common for customers to have a development and test environment where updates to production software are developed and tested before they are deployed into production. A point-in-time image of the production data can be used to build the development and test environment so that the real production data can be used during the development and QA cycles, allowing the customer to expose any issues with the updates before they are pushed to production.

- **and many more…**

Data Restoration in CDP Environments

There are two general principles used to define restoration policies: Recovery Point Objective (RPO) and Recovery Time Objective (RTO). RPO is the targeted recovery point, and essentially defines the maximum amount of data loss that can be tolerated in a recovery. RTO defines how long the business can tolerate being offline or down from a failure.

An RTO objective should include the three phases of application recovery: analysis, data restoration and application recovery. Analysis is typically 5% of the overall process, data restoration 90%, and application recovery 5% . The caliber of restoration capabilities can best be judged in terms of RPO and RTO, with a note about how easily the environment can be audited to avoid restoration failures.

If a business determines that it cannot lose more than six hours of data, then a backup copy of the data must be created every six hours. Of course, because most logical data failures aren't found when they occur, but later (minutes, hours, even days), the RPO typically is implied across a time frame like two or three days. Therefore, a six-hour RPO over two days means that a backup occurs every six hours, four per day, for a total of eight.

CDP as a technology offers infinite RPO. That is what the word 'continuous' in CDP refers to: the continuous spectrum of recovery points. Any point or event in time means just that, so the RPO becomes infinitely granular. With CDP, nothing additional has to be done, the user simply determines how far back he/she wants the recovery spectrum to extend - a day, several days or longer.

Estimating Additional Storage Needs in CDP Implementations

The user must consider the disk space requirements to store CDP activity across the recovery spectrum selected. For example, if there is 120GB of production data and their assumption is that 10% of that data changes on a daily basis, then 60GB of additional storage is required to cover the 5 day period (12GB x 5 days). If 10 days of changes are needed, an additional 120GB of storage is needed (12GB x 10 days). If the CDP implementation is to cover 1 month, then an additional 360GB of storage is required (12BG x 30).

30 days is typically the maximum length of storage of CDP data. The additional storage needed by the CDP implementations may be reduced by using data compression or other data reduction technologies to reduce the overall amount of space required. This chart is helpful in estimating the additional data storage requirement to support a CDP implementation.

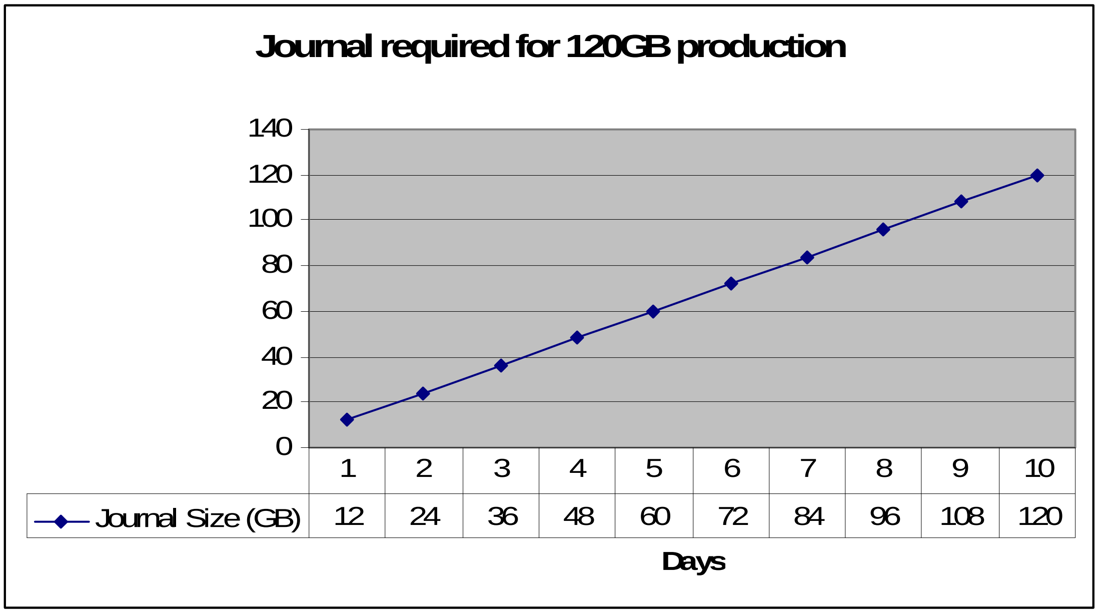

Figure 3 - Journal Space / Protection (assumes a 10% change per day)

Addressing Recovery Time Objectives with CDP

While CDP solutions offer an infinite selection of recovery points, the methods used to recover the data at the recovery point selected vary between products. The implementation method can greatly affect the actual recovery time.

Most CDP systems provide the user with a picture of the data from a restoration point of view – this picture or image is used to view the recovery spectrum and select a recovery point. Once the recovery point is determined, the data can be moved or copied into the production environment, much like restoring a Snapshot. Some CDP solutions implement virtualized recovery. This allows the CDP image, at the recovery point selected, to be mapped directly into the production environment, providing an RTO time measured in seconds.

The picture below gives an example of the RPO and RTO differences between a traditional recovery environment and a CDP recovery environment. The first example shows application recovery in a traditional backup environment. The time between the last good backup image and the failure represents the amount of data that may be lost, this may be seconds, minutes, hours or even days. At the point of the failure, the clock starts on the time required to return to normal operations. The first step is to understand why the failure occurred, and perhaps fix the problem. At that point, a restore is initiated. If the restore is from tape the process can easily take multiple hours. After the data is restored, the challenge of restarting the application server can be addressed. For some applications this may require a significant effort and may include rebuilding application server consistency and the recreation of lost data. Finally, the applications can be restarted and production can continue.

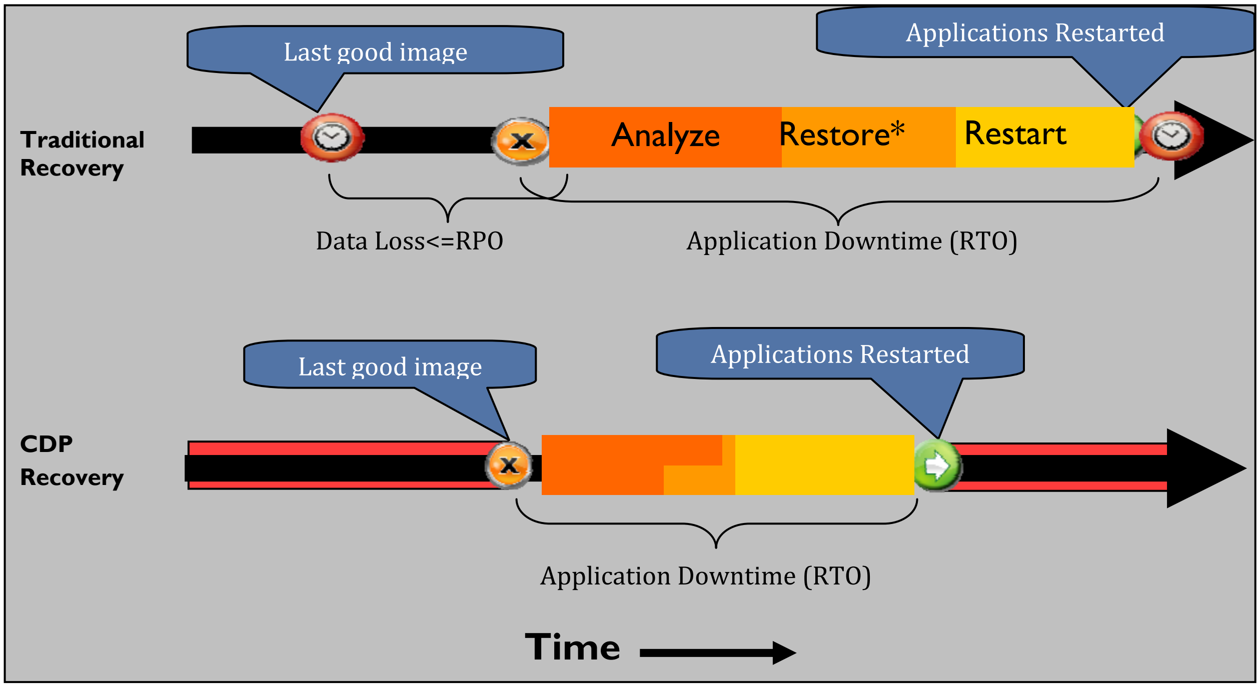

Figure 4 - RPO/RTO

The second example shows application recovery in a CDP environment. Since all data changes are protected, the user has the option to recover immediately prior to the failure event eliminating all data loss – effectively delivering an RPO of 0. The failure analysis process can overlap with the much quicker initiation of image recovery. For some implementations, the recovery time can be reduced to minutes or seconds. Finally, the application restart time can be greatly reduced as the very short window of data loss can typically be addressed with no application server rebuild time and very limited data loss.

When the two recovery processes are compared, the advantages of using CDP-based technology in environments with 24x7 application support requirements is apparent. In CDP environments, the RPO is at or near zero and, depending on implementation specifics, the RTO is significantly less than traditional recovery.

Summary

This document introduces Continuous Data Protection (CDP) concepts and benefits. It demonstrates the impact that a CDP implementation can have in reducing data loss by narrowing application recovery point and recovery time. It provides a comparison to traditional backup methods and shows how CDP images can be used to support other data center requirements, including conducting compliance searches, seeding test systems and populating data warehouses. Customers should evaluate the value of implementing a CDP solution in all 24x7 application environments and wherever they are using advanced data protection technologies to shorten the backup window and narrow recovery times. For more information on CDP visit the SNIA Data Protection Initiative website http://www.snia.org/forums/dmf/programs/data_protect_init/.

About the Data Protection Initiative

The Data Protection Initiative (DPI) was created by the Data Management Forum (DMF) to allow industry leaders and participants to come together in a community to focus on defining, implementing, qualifying, and teaching improved methods for the protection and retention of data and information. The DPI operates as an online virtual community, sharing work efforts, training programs, and outreach services such as research, whitepapers and training, and educational courses. This group's primary objective is to serve IT professionals by creating certification, education, and training programs, and by creating a world-class collaborative information portal with global influence and reach. The DPI's goal is to build the knowledge base and training capabilities to become the worldwide authority on data protection, helping all of the SNIA's constituents (vendors, IT, regulatory agencies, and channel partners) to better understand and implement data protection solutions. For more information on the DPI visit http://www.snia.org/forums/dmf/programs/data_protect_init/

About the SNIA

The Storage Networking Industry Association (SNIA) is a not-for-profit global organization, made up of some 400 member companies and 7,000 individuals spanning virtually the entire storage industry. SNIA's mission is to lead the storage industry worldwide in developing and promoting standards, technologies, and educational services to empower organizations in the management of information. To this end, the SNIA is uniquely committed to delivering standards, education, and services that will propel open storage networking solutions into the broader market. For additional information, visit the SNIA web site at www.snia.org.

About the Author

With over 25 years of storage industry experience, Gary Archer has worked for several major storage companies, including EMC, IBM, Kashya, Legato and Maranti Networks. As a founding member of the SNIA he served on their board during the early years of SNIA's existence. Gary is currently the Senior Product Marketing Manager for RecoverPoint at EMC, and is their representative on the SNIA Data Management Forum CDP Special Interest Group.

CDP Terminology

Any Point-In-Time: Any Point-in-Time (copy, replica, etc.) "APIT" refers to the ability to access or recreate the exact data state as it existed at any previous point in time. Typically unique to CDP technology.

Block-based Continuous Data Protection: Continuous Data Protection that operates at the block level of logical devices. As data blocks are written to primary storage, copies of the blocks are stored and managed by the CDP system.

Continuous Data Protection (CDP): Continuous Data Protection is a methodology that continuously captures or tracks data modifications and stores changes independent of the primary data, enabling recovery points from any non-predetermined point in the past. CDP systems may be block-, file-, or application-based, and can provide fine granularities of restorable objects to infinitely variable recovery points.

File-based Continuous Data Protection: (CDP) Continuous Data Protection that operates at the file level. All changes to files and file metadata are stored and managed by the CDP system.

Scheduled Point-In-Time: A "SPIT" is a Scheduled Point-In-Time image that is created in conjunction with external processing to create a stable recovery point for a file system, database or application.