DELL EMC DATA DOMAIN EXTENDED RETENTION SOFTWARE

A Detailed Review

ABSTRACT
This white paper introduces Dell EMC® Data Domain® Extended Retention software that increases the storage scalability of a Data Domain system to enable cost-effective long term retention of backup data on deduplicated disk. In addition to the high-speed, inline deduplication needed to satisfy the data protection needs of the enterprises, Data Domain systems now offer significant optimization for long term cost efficiency and granular fault-containment and recovery. Using DD Extended Retention software, customers can leverage Data Domain systems for long term backup retention and minimize reliance on tape infrastructure in the data center.

August, 2017
# Table of Contents

EXECUTIVE SUMMARY .................................................................................................................. 3
  INTRODUCTION .......................................................................................................................... 3
  AUDIENCE .................................................................................................................................. 3
DATA DOMAIN EXTENDED RETENTION SOFTWARE OVERVIEW ........................................... 3
  STORAGE TIERS OF THE DATA DOMAIN SYSTEM ENABLED BY DD EXTENDED RETENTION SOFTWARE .................................................................................................................. 4
  DATA MOVEMENT POLICIES ..................................................................................................... 5
  SINGLE SCALABLE FILESYSTEM ................................................................................................. 5
  TECHNICAL IMPLEMENTATION ................................................................................................. 6
  SPACE RECLAMATION ON RETENTION TIER ............................................................................ 7
  FLEXIBLE DISASTER RECOVERY CONFIGURATIONS ............................................................... 7
  ENCRYPTION OF DATA AT-REST ............................................................................................... 8
  DATA DOMAIN RETENTION LOCK ............................................................................................ 8
  COST-EFFECTIVE SCALABILITY ................................................................................................. 8
TYPICAL DEPLOYMENT SCENARIOS ......................................................................................... 9
  LONG TERM RETENTION OF BACKUP DATA .......................................................................... 9
CONCLUSION ............................................................................................................................... 10
EXECUTIVE SUMMARY
Many companies have minimized the use of tape automation in their IT infrastructure by deploying deduplication storage for backup and operational recovery – Dell EMC Data Domain deduplication storage systems have been the market leaders in this category. In general, operational recovery includes retention periods from a few weeks to a few months. For longer backup retention requirements to meet legal mandates, many users simply keep backups on tape longer, for example, users may keep weekly or monthly backups on tape for 7 years. Due to the perceived cost advantages, the most widely adopted approach for long term backup retention has historically been keeping tape backups longer.

There are several drawbacks to using tape for long term backup retention. Although tape cartridges have a relatively low acquisition cost, the cartridges make up just a small fraction of the total cost of ownership. Tape automation, transport and storage space are expensive and ongoing. This is particularly true when upgrading and replacing large tape libraries, which are expensive to purchase and take up big chunks of precious data center real estate. Also, companies expend significant resources to manage the tape infrastructure that could otherwise be creating business value. In addition, there are operational challenges with a tape infrastructure. Accessing a retained file stored on tape takes considerable time, especially if it is offsite. The measured failure rate for restoring data from tapes is also high when compared to disk recoveries. Also, offline tapes do not provide online information access. Finally, the risk of theft from data on tapes during transportation also poses serious security challenges.

Dell EMC Data Domain Extended Retention software enables Data Domain systems to be the industry's first deduplication storage system for long term retention of backup data. DD Extended Retention increases the retention capacity of a Data Domain system and transparently incorporates a large-capacity retention tier dedicated to static data. This software provides additional architectural enhancements that allow the Data Domain system to incorporate very large capacities that can expand over time, reduce system cost, and ensure long term availability and integrity of data. DD Extended Retention software enables Data Domain systems to provide a high throughput, cost-viable and reliable alternative to tape for long term backup retention.

INTRODUCTION
This white paper introduces the Dell EMC Data Domain Extended Retention software on Dell EMC Data Domain deduplication storage systems and explains how it extends the Data Domain architecture with an internal tiering approach designed to enable cost effective long term retention of backup data on deduplicated disk. Read this white paper to find out how DD Extended Retention software enables customers to leverage Data Domain systems for long term retention of backups and minimize reliance on tape infrastructure in the data center.

In the following sections, this paper will describe the unique features of the DD Extended Retention software, including enabling petabyte scalability, data movement, space reclamation, and fault isolation. It will also cover deployment scenarios as well as comparisons to alternative long term retention options.

AUDIENCE
This white paper is intended for Dell EMC customers, system engineers, partners, and members of the Dell EMC and partner professional services community who are interested in learning more about the Data Domain Extended Retention software option.

DATA DOMAIN EXTENDED RETENTION SOFTWARE OVERVIEW
Data Domain Extended Retention software addresses the long term backup retention requirements of scale and cost by extending the proven Data Domain architecture with an internal tiering approach. Data Domain systems with DD Extended Retention software have two tiers of storage, each sharing a common controller, management and namespace:

- An active storage tier, which is a group of storage shelves used for operational backup and recovery
- A retention storage tier is a group of storage shelves, separate from the active storage tier, where backup data is internally migrated to when a policy threshold (time of most recent modification) is met.
These tiers are logical divisions of the storage attached to a single controller. Backup data is stored first to the active tier, and once it has been on the active tier without modification for a user-defined period of time, it is automatically moved to the retention tier. The retention tier incorporates improved manageability and enhanced compression to provide the most cost effective approach in storing long term data of backups.

Since backup data is separated and stored in different tier, the system can scale to a larger capacity for long term retention of backup data. With the additional storage capacity provided by the retention tier, the average cost of the system per gigabyte decreases as the system scales making Data Domain systems a cost-effective solution for long term retention of backups.

DD Extended Retention is a software option that can be enabled on the following Dell EMC Data Domain storage systems: DD4200, DD4500, DD6800, DD7200, DD9300, DD9500, and DD9800.

**STORAGE TIERS OF THE DATA DOMAIN SYSTEM ENABLED BY DD EXTENDED RETENTION SOFTWARE**

The **active tier** is a grouping of storage shelves within the Data Domain system meant for short-term operational backup and recovery and is sized based upon the same guidelines. For example, one could size the active tier to hold weekly fulls and daily incrementals for up to 90 days.

The **retention tier** is a logical grouping of storage shelves, with its own deduplication context just like the active tier. A user-defined, policy-based process runs periodically to move aged backup data from the active tier and into the retention unit. It continues to do this until that retention unit is determined to be full. See Figure 1, which shows backup data being moved from the active tier to retention unit in the retention tier.

Note that tiers are transparent to end-users and applications, which just see a larger Data Domain system.

---

**Figure 1:** A Data Domain system enabled with DD Extended Retention software, with its controller, active tier, and one retention unit in its retention tier. The internal data movement process is shown moving aged backup data from the active tier to the retention unit currently being filled.
DATA MOVEMENT POLICIES

The system administrator can configure the policy that guides data movement from the active tier to the retention tier on a Data Domain system. Data movement policies are managed on an MTree basis and only those files that meet the policy threshold in that MTree are moved to the retention tier. An MTree is a logical partition of the namespace in a Data Domain filesystem that can be used to group a set of files for management purposes (for more information on MTrees, please refer to the Data Domain User’s Guide). If different backup data sets require different data movement policies on the same Data Domain system, they should be placed in different MTrees. However, a single MTree spans both the active tier and retention tier, so when the data is migrated from active tier to the retention tier, it is still under the namespace of a single MTree.

For example, a dataset that needs to be retained only for 30 days can be placed on an MTree (M1) and another dataset that needs to be retained for 7 years can be placed on a different MTree (M2). M1 would not have any data movement policy defined so that all the data in M1 stays in the active tier. The data movement policy on M2 can be set to 14 days so that all the data in this MTree would be moved from active tier to the retention tier after 14 days (and retained for 7 years), but all data in M2 would remain in the same namespace.

The data movement policies are driven by two user-defined elements:

- A selected last-modified-time of a file (e.g. after 90 days without modification)
- A periodic schedule for moving the data (e.g. every week before the cleaning process)

During the data movement process, files are moved out of the active tier and into the retention unit currently being filled but all data remains in the same MTree. Note that each backup file is only moved once. Since the data movement process is a background process that can run when the system is relatively idle, there is opportunity to apply further compression on this data. Specifically, during data migration into retention tier, the system can implement more processing-intensive compression algorithms, so that when data is moved out of the active tier, it can be recompressed and packed more tightly into the retention tier as it is moved. This enables even greater storage efficiency for data stored in the retention tier.

The data movement process runs on a periodic basis and moves individual backup files from the active tier to the target retention unit based on a user-defined file age policy. However only segments that are unique to that retention unit are physically copied enabling faster data movement. The periodic data movement can be scheduled to run at a specified time and can be stopped, restarted or throttled.

For example, a user may keep weekly fulls and daily incrementals for the duration of the backup cycle (e.g. 90 days) and after that period, the user may want to keep only monthly full backups for a specified long term retention period (e.g. 7 years). To configure this data movement policy, an administrator sets the data movement for this MTree to run every week and to move all files that were last-modified 90 days prior to that date. The backup retention policy can be set from the backup application to ensure only monthly fulls are kept beyond that date. During each weekly data movement sweep, the remaining files aged beyond 90 days will be moved to the retention tier.

SINGLE SCALABLE FILESYSTEM

All Data Domain systems present a single, filesystem across all storage to enable ease-of-use and simple management. A Data Domain system enabled with DD Extended Retention software does not expose the different units and tiers to the application or the user and simply looks like a much larger Data Domain system.

This single filesystem can be completely or partially exposed as CIFS shares (for Windows), NFS mount points (for Unix/Linux), VTL for open systems and IBMi and/or through advanced integration with Data Domain Boost (for Dell EMC Avamar, Dell EMC Networker, Dell EMC Greenplum, Veritas NetBackup, Veritas Backup Exec, Oracle RMAN and Quest vRanger).

For environments that have requirements for long term retention of backups, the frequency of access to aging data is expected to decline significantly over time. DD Extended Retention software enables Data Domain systems to scale to store petabytes of backup data. When the system reaches full scale, reads of long term backup data on the retention tier may experience a slight delay.
Since DD Extended Retention software is aimed at minimizing tape usage in customer environments, most users considering the system compare this low-probability delay that only affects the oldest, least accessed data - to the delay of accessing data from offline tapes. Offline tapes are typically sitting on a shelf or shipped offsite to an external facility and accessing the data on these tapes could take anywhere from hours to days.

**TECHNICAL IMPLEMENTATION**

Within a Data Domain system, there are several levels of logical data abstraction above the physical disk storage, as illustrated in Figure 2.

![Logical and Physical System Layers](image)

**Figure 2:** Data Domain Operating System filesystem – protocol-specific namespaces are presented to clients/applications for accessing the logical filesystem layer. The files and directories in each MTree, as well as MTree snapshots, all reference the same pool of unique segments, called a collection, which is made up of log-structured containers that organize the segments on disk to optimize throughput and deduplication effectiveness.

This figure illustrates several components and layers of the Data Domain Operating System filesystem:

1. **Protocol-specific namespaces:** As an external interface to applications, there are protocol namespaces, such as CIFS/NFS file shares (over Ethernet), virtual tape libraries (over Fibre Channel) and DD Boost storage units (SUs). A Data Domain deployment may use any combination of these simultaneously to store and access data.

2. **Filesystem MTrees, directories, files and snapshots:** Files and directories for each namespace are stored in an MTree in the DD OS filesystem. MTree snapshots in DD OS are logical and very space-efficient because they share the same underlying data segments.
3. **Unique segment collection, stored in locality containers:** A ‘collection’ is the set of files (or virtual tapes) and logical MTrees, stories. The system identifies and eliminates duplicate segments within each container and then writes compressed deduplicated segments to physical disk. Segments are unique within the collection (not including specific duplicates maintained in DD OS to enable self-healing or fast recovery). Each Data Domain system has a single collection that is stored in a log of segment locality containers. For more about segment localities, see the white paper, *Data Domain SISL™ Scaling Architecture*.

4. **RAID-enabled disk storage:** These collection containers layer over RAID enabled disk drive blocks to ensure a high level of integrity. Data Domain deduplication storage systems use Data Domain RAID-6 internal disk and storage expansion shelves to protect against dual disk failures.

Data Domain enabled with DD Extended Retention software introduces an architectural extensions to this architecture. In order to implement its internal tiering, the Data Domain system divides the storage representation of a collection into collection partitions that are used as the active tier and retention tier. Physically, each collection partition, regardless the active tier or the retention tier, is stored in a corresponding logical set of storage shelves with RAID-6 protection. The retention tier has its own deduplication index. New writes to the active tier do not check against retention tier indexes. With this design, data movement from the active tier to the retention tier can happen transparently without any visible change in the namespace of the moved files. Externally, all the manageable elements of the namespace look like a single Data Domain system.

**SPACE RECLAIMATION ON RETENTION TIER**

Once the retention period of data on the retention tier expires, the system will clean the space and customers can reuse the freed space for newer data. When backup data is deleted from the retention tier, segments referring to those files (if not referring to other files on the system) are “unused” and thus available for cleaning.

**FLEXIBLE DISASTER RECOVERY CONFIGURATIONS**

An increased frequency of natural and man-made disasters have raised awareness of the importance of comprehensive disaster recovery (DR). One of the most crucial requirements for DR is to keep a replica of the data in a separate system in a remote site. For Data Domain systems, Dell EMC Data Domain Replicator software provides simple, fast, robust WAN-based replication for the enterprise. It offers numerous replication types and policies and also supports a wide variety of topologies to meet the needs of various deployments. Refer to the *Dell EMC Data Domain Replicator whitepaper* for more details.

A Data Domain system with DD Extended Retention software supports the following replication types:

- **Collection replication** performs whole-system mirroring in a one-to-one topology, continuously transferring all the changes in the source Data Domain system to the destination Data Domain system. In this configuration, the destination system is in a read-only state and data movement policies are configured only on the source system. As data is written to the active tier on the source system, it is replicated to the active tier on the destination system. When data gets migrated to the retention tier at the source, the data gets replicated to the retention tier on the destination.

- **MTree replication** is used to replicate MTrees between Data Domain systems and can be used replicating data written via CIFS, NFS, or VTL. Unlike collection replication, MTtree replication offers flexible replication topologies like one-to-one, bidirectional, one-to-many, many-to-one and cascaded. It also offers the flexibility to selectively replicate only a portion of the data written to the source system. Files replicated from the source will land on the destination active tier first regardless if the data is on the active tier or retention tier. This provides the flexibility to the user to set different data movement policies between the source and destination.

- **Managed file replication** is used when integrating with applications via Dell EMC Data Domain Boost (e.g. Dell EMC NetWorker, Dell EMC Avamar, Veritas NetBackup, Oracle RMAN, etc.). With managed file replication, backup applications can selectively replicate specific backup images and keep track of all the copies in the application’s catalog for easy recovery. Further, the backup application can configure separate retention periods for each copy of the backup image. Managed file replication supports all the topologies supported by MTtree replication.
Like MTree replication, managed file replication replicates data from the source MTree to the destination MTree and is unaware of whether the data is in the active tier or retention tier as illustrated in figure 3. Data movement policies for migrating data from active tier to retention tier are configured independently on the source and destination MTrees. Upon replication, data always lands in the active tier on the destination, and is moved to the retention tier based on the policies configured on the destination MTree.

![Figure 3: Two Data Domain systems with DD Extended Retention software in their disaster recovery (DR) configuration, leveraging Managed File replication for DD Boost data and MTree replication for NFS, CIFS, and VTL data.](image)

**ENCRYPTION OF DATA AT-REST**

Dell EMC Data Domain Encryption software allows the user to encrypt data at rest by using RSA BSAFE FIPS 140-2 compliant libraries with standard 128-bit or 256-bit Advanced Encryption Standard (AES) algorithms. Depending on IT security policies, the block cipher modes for the AES algorithm can be selected either as Cipher Block Chaining (CBC) or Galois Counter Mode (GCM). DD Extended Retention is compatible with DD Encryption and is supported on both the active tier and retention tier.

**DATA DOMAIN RETENTION LOCK**

Dell EMC Data Domain Retention Lock software provides immutable file locking and secure data retention capabilities for customers to meet both corporate governance and compliance standards (such as SEC 17a-4(f)). DD Retention Lock comes in two editions – Dell EMC Data Domain Retention Lock Governance edition and Dell EMC Data Domain Retention Lock Compliance edition. Both editions provide the capability for IT administrators to configure minimum and maximum retention periods at the MTree level and apply retention policies at an individual file level. Files locked on the active tier will remain locked when migrated over to the retention tier.

**COST-EFFECTIVE SCALABILITY**

Another benefit of DD Extended Retention software is it enables another tier of storage with a single controller. The DD9500 with DD Extended Retention software supports up to 864 TB of capacity in the active tier, and the entire system can scale up to a total of 1.7 PB of usable capacity. Assuming backup deduplication ratios that range from 10x to 50x, DD
Extended Retention software could enable scalability up to ~86 PB for long term retention of backups. Amortized across so many storage shelves, the cost of the controller at scale becomes minimal.

While tape cartridges in a cardboard box can be cheaper than this on the surface, the operating overhead and risks of tape are suboptimal and can add unforeseen costs.

**TYPICAL DEPLOYMENT SCENARIOS**

DD Extended Retention software can be deployed on a Data Domain system that is meant as a) a target for consolidation of backups for operational recovery (30 to 90- day retention periods) and long term (years) retention of backups, and b) a consolidated and replicated target for backup data that needs to be retained from long term from other Data Domain systems.

**LONG TERM RETENTION OF BACKUP DATA**

For users seeking a solution for short and long term retention of backups, DD Extended Retention software can be deployed on a Data Domain system such that the system serves as a target for both short-term operational recovery and long term retention of backup data. This would eliminate the need for weekly, monthly or quarterly copies on tape. This deployment model, as illustrated in figure 4 below, leverages a standard backup application like Dell EMC NetWorker or Veritas NetBackup to send backups directly to the Data Domain system.

![Figure 4: Illustrates a configuration where backups are sent directly to the Data Domain system enabled with DD Extended Retention software. The active tier is sized to store the data of the short-term backup cycle, while the retention tier grows with long term retention data over time. Optionally, the system at the data center could replicate to a second Data Domain system in the remote DR site.](image)

In this configuration, the active tier would be sized for short-term operational recovery needs and the retention tier would be sized based on the required retention policy of backup data that needs to be retained for longer retention periods.

DD Extended Retention software can also be enabled on a Data Domain system to consolidate backups for long term retention across a large enterprise. As shown in figure 5 below, other Data Domain systems are replicating into one centralized Data Domain system with DD Extended Retention software. In this configuration, backup applications at each remote site would send backup data to a local Data Domain system and then leverage DD Replicator to send only unique data over the WAN to the remote Data Domain system with DD Extended Retention software for long term retention of backup data.
Unlike the first use case in figure 9, since the active tier is not being used for operational recovery in this deployment, it may require a smaller footprint given that this space is mostly to be used as landing space. In addition, with DD Boost managed file replication, replicating between systems and implementing different retention periods on the replica is simple because all management can be done from the backup application. This ease of replication management is only available for backup applications that support DD Boost managed file replication.

CONCLUSION

DD Extended Retention software enables cost-effective long term retention of backup data on Data Domain deduplication storage systems. With DD Extended Retention software, a Data Domain system benefits from the cost-effective scalability, simple management and extensible architecture that enable customers to service their long term backup retention needs.

For over a decade, Data Domain systems have been an ideal solution for eliminating tape for operational recovery, and now with DD Extended Retention software, Data Domain systems can take it to the next level to further reduce tape, including long term retention of backup data. With its high backup throughput, optimized design and extensible architecture, Data Domain system is the ultimate data preservation platform, moving to tape’s major hideout in the data protection market landscape.