ACHIEVING STORAGE EFFICIENCY WITH DATA DEDUPLICATION

Dell NX4

Dell Inc.

Visit dell.com/NX4 for more information and additional resources

TABLE OF CONTENTS

Executive Summary 3

Introduction 4

- Audience 4
- Terminology 5

Dell NX4 Data Deduplication Overview 6

- Detailed Overview 6
  - The space reduction process 7
  - Minimizing client impact 8
  - Management 9
  - Client input/output to space-reduced files 11
- Architecture 12
- Performance 13
- Interoperability with other Dell NX4 features 13
  - Network and NDMP PAX (tar and dump) 13
  - NDMP Volume Backup (NVB) 14
  - Point-in-time views of file system 14
  - Replication 14
  - Quotas 15
  - Celerra FileMover archiving 15
  - Celerra File-Level Retention 15
- Deployment considerations 16

Conclusion 17

EXECUTIVE SUMMARY

As storage usage increases exponentially, improving storage efficiency is critical for many data centers today. There are many viable solutions to achieve this objective. These include data migration, thin provisioning, content and quota management, and data deduplication. However, the solution you implement is based on what you believe is a suitable or effective approach to achieve storage efficiency in your production environment. If the intent is to reduce the cost of storing data at the file system level by achieving space savings without affecting end-user experience, and to propagate these space savings within the storage environment at the file system level, an appropriate technology to consider is data deduplication.

For a file server, the main purpose of data deduplication is to increase the file storage efficiency by eliminating redundant data from files stored on a file system hosted by the file server. Though there are many product offerings that feature data deduplication, the objective of any type of deduplication should be to decrease the need for storage space intelligently while being mindful of the user impact. Celerra® Data Deduplication is introduced in the The Dell NX4 as a deduplication offering combines file-level single instancing and file-level compression into one feature. It increases storage efficiency with average space savings of up to 30 percent to 40 percent for typical network-attached storage (NAS) file system data.

The Dell NX4 offers many features that optimize storage efficiency. For example, the virtual (thin) provisioning for NAS and Internet SCSI (iSCSI) reduces the possibility of allocating storage resources beyond the available capacity. In addition to this, compared to other competitive offerings, it significantly lowers the inherent overhead that is associated with managing data structures. Celerra Data Deduplication adds a new approach to reduce the cost of storing data on a NAS file system further by automatically targeting files that are the best candidates for compression and single instancing in terms of the level of frequency of file access and file size. Celerra Data Deduplication is designed to improve storage efficiency by reducing storage needs intelligently. It filters files that are too small, too big, and accessed very often from being processed. Further, Celerra Data Deduplication avoids compressing files if the space savings are minimal and negatively impact the file access service levels. By eliminating redundant data from user file systems without significantly affecting the end user or administrator experience, Celerra Data Deduplication allows customers to reduce the amount of storage they need, and design and deploy a tiered storage solution that achieves storage efficiency while maintaining the existing file access service levels for end users.

The purpose of deduplication is to achieve storage efficiency. Storage savings is a means to that purpose. By optimizing the use of the existing storage environment through single instancing and compression, deduplication can save storage costs and lower future storage projections for the data center. Deduplication is enabled at the file system level and is transparent to access protocols.

INTRODUCTION

This paper is intended not only to introduce Celerra Data Deduplication but also to draw attention to the feature's product design, which offers practical value to the customer in terms of ease of management, intelligent storage efficiency, and flexibility. This paper illustrates the concepts of Celerra Data Deduplication. It also shows how a user can use and manage deduplication on Dell NX4. The feature is designed to be a simple but powerful addition to an already strong portfolio of NX4 features. With this in mind, the paper discusses the interoperability between deduplication and other Dell NX4 features.

Audience

This paper is intended to be used by Dell field personnel and customers who are familiar with Dell NX4 technology but are not familiar with Celerra Data Deduplication. The "Terminology" section next covers some key terms that are used in this paper, but it does not include terms that are generic to basic management and administration of the Dell NX4.

Terminology

Common Internet File System (CIFS) — File-sharing protocol based on the Microsoft Server Message Block (SMB). It allows users to share file systems over the Internet and intranets.

deduplicated file or space-reduced file — A file that has had its associated data single-instanced, compressed, or both.

deduplication — Process used to compress redundant data, allowing space to be saved on a file system. In Celerra Data Deduplication, only one unique copy of each file is stored if the file data is represented more than once in the file system. The file data is also compressed to further improve storage efficiency.

dump — Backup format in PAX that traverses a file tree in mixed width-first and depth-first order.

Internet SCSI (iSCSI) — Protocol for sending SCSI packets over TCP/IP networks.

Network File System (NFS) — Distributed file system providing transparent access to remote file systems. NFS allows all network systems to share a single copy of a directory.

NDMP Volume Backup (NVB) — Dell NX4-specific type of NDMP backup mechanism that backs up data blocks at a volume level, rather than at a file level. NVB reads a set of disk data blocks in an efficient manner compared to the method used for traditional, file-based backups. NVB works only with qualified vendor backup software. NVB is commonly known as VBB.

production file system (PFS) — Production file system on a Dell NX4 Network Server.

reduplication — Process to undo the effect of Celerra deduplication on a file. If the file was compressed, it will be decompressed. If there are multiple instances of the file data, then a copy of the file data is made so that blocks are

not shared between instances of the file. This process consumes additional space in the file system. Therefore, there must be sufficient free space in the file system to hold an additional copy of the original file for this process to complete.

SavVol — Standard Celerra metavolume to which SnapSure™ writes checkpoint data. This volume can be smaller than the production file system upon which the checkpoint is based.

single instance — When multiple files have identical data, the file system only stores one copy of the file and shares that between multiple files. Different instances of the file can have different names, security attributes, and timestamps. None of the metadata is affected by deduplication.

tar — Backup format in PAX that traverses a file tree in depth-first order.

virtual provisioning — Configurable Dell NX4 file system feature that can only be used in conjunction with Automatic File System Extension. This option lets you allocate storage based on longer-term projections, while you dedicate only the file system resources you currently need. Users—NFS or CIFS clients and applications—see the virtual maximum size of the file system of which only a portion is physically allocated. Combined, the Automatic File System Extension and virtual provisioning options let you grow the file system gradually on an as-needed basis.

DELL NX4 DATA DEDUPLICATION OVERVIEW

The main objective of data deduplication is to increase file storage efficiency by eliminating redundant data from files located on the file system. You can increase the amount of information stored at a given cost by reducing the amount of redundant data stored in the file system. There are various data reduction technologies that are implemented in other data deduplication solutions, such as fixed-block deduplication, variable-block deduplication, file-level deduplication, and file compression. Celerra Data Deduplication combines file-level deduplication (also known as file-level single instancing) and data compression technologies to provide maximum benefit for the resources that are required to provide storage efficiency, while still minimizing the client impact on mission-critical files.

Detailed Overview

There are a number of technologies classified under data reduction or deduplication. Table 1**Error! Reference source not found.** on page 6 lists the four major data reduction technologies and the space they save when applied to a typical file server or NAS data set[1]. An estimate of the resources required to implement, deploy, and exercise each technology is also included in the table.

Table 1 Data reduction technology comparison

Technology	Typical space saving	Resource footprint
File-level deduplication	10%	Low
Fixed block deduplication	20%	High
Variable block deduplication	28%	High
Compression	40%–50%	Medium

- **File-level deduplication**, also known as file-level single instancing, provides relatively modest space savings. It does not require many CPU and memory resources to implement.
- **Fixed-block deduplication** provides better space savings, but it requires more CPU and memory resources. It requires more CPU resources because of the processing power that is required to calculate hashes for each block of data, and more memory resources to hold the indices used to determine if a given hash has been seen before.

[1] Each technology was applied to an approximate 900 GB data set taken from shared file systems on the EMC corporate network. The data set consisted of a mix of office documents, media files, source code, binaries, and other file types typically found on file systems hosted by corporate file servers. The assumption was that the EMC internal file server data was sufficiently “typical” to be meaningful.

- **Variable-block deduplication** provides slightly better space savings than fixed-block deduplication, but the difference is not significant when applied to file system data. Variable-block deduplication is most effective when applied to data sets that contain repeated but block-misaligned data, such as backup data in backup-to-disk or virtual tape library (VTL) environments. The resource footprint of variable-block deduplication is similar to fixed-block deduplication. It requires similar amounts of memory and slightly more processing power.
- **Compression** is often considered to be different from deduplication. However, compression can be described as infinitely variable, bit-level, intra-object deduplication. It is another technique that alters the way in which data is stored, mainly to improve storage efficiency. In fact, it offers, by far, the greatest space savings of all the techniques listed for typical NAS data, and is relatively modest in terms of its resource footprint. It is relatively CPU-intensive but requires very little memory.

These four data reduction technologies are not mutually exclusive. It is perfectly possible and even reasonable to apply two or more of these techniques to the same set of data.

It is interesting to note that approximately 75 percent of the savings that you achieve by applying all four techniques together can be achieved by applying compression alone. Compression's modest resource demand suggests that it is a clear winner in this comparison. However, compression has a disadvantage over other techniques, that is, a potential performance "penalty" associated with decompressing the data when it is read or modified.

This decompression "penalty" can work both ways. Reading a compressed file can often be quicker than reading a non-compressed file. The reduction in the size of data that you must retrieve from the disk more than offsets the additional processing required to decompress the data.

Celerra Data Deduplication combines the benefits of file-level deduplication and compression to provide maximum space savings for the required resources, when applied to file system data.

The space reduction process

Celerra Data Deduplication has a flexible policy engine that specifies data for exclusion from processing and decides whether to deduplicate specific files based on their age. When enabled on a file system, Celerra Data Deduplication periodically scans the file system for files that match the policy criteria and then compresses them. The compressed file data is hashed to determine if the file has been identified before. If the compressed file data has not been identified before, it is copied into a hidden portion of the file system. The space that the file data occupied in the user portion of the file system is freed and the file's internal metadata is updated to reference an existing copy of the data. If the data associated with the file has been identified before, the space it occupies is freed and the internal file metadata is updated. Note that the Dell NX4 detects non-compressible files and stores them in their original form. However, these files can still benefit from single-instancing.

Celerra Data Deduplication employs SHA-1 (Secure Hash Algorithm) for its file-level deduplication. SHA-1 can take a stream of data less than 2^{64} bits in length and produce a 160-bit hash, which is designed to be unique to the original

data stream. The likelihood of different files hashing the same value is so substantially low that a collision rate has been reported only after 2^{69} hash operations. Unlike in compression, you can disable single instancing in Celerra Data Deduplication.

Minimizing client impact

The Dell NX4 performs all deduplication processing as a background asynchronous operation that acts on file data after it is written into the file system. It does not process data while file data is being written into the file system. This is to avoid latency in the client data path, because access to production data is sensitive to latency.

In addition to doing all the processing in the background, you can configure Celerra Data Deduplication to avoid processing the “hot” data in the file system. The “hot” data is any file that is in active use by clients. Note that “hot” data is defined by how recently clients accessed or modified the files. By not processing active files, you avoid introducing any performance penalty on the data that clients and users are using to run their business. Surveys of file system data profiles show that typically only a small proportion of the data in a file system is in active use. This means that Celerra Data Deduplication processes the bulk of the data in a file system without affecting the production workload. A comprehensive data management strategy often involves archiving files that are used for some time to an alternative tier of storage. If you use this strategy, Celerra Data Deduplication maximizes storage efficiency for those files that are no longer actively used but are active enough not to qualify for archival out of the file system. If you combine both deduplication and archiving, you can potentially create a multi-tiered storage solution that provides greater storage efficiency.

There is a defined default policy that the policy engine uses to scan the files in a deduplication-enabled file system. This is based on the investigation and analysis of how typical files age from active use to disuse in various company settings of different industry and sector types. However, the default policy may not meet every company’s information lifecycle due to its unique workflow or setting. Celerra Data Deduplication provides the flexibility and granular control to allow administrators to define what “hot” data is to them, though careful planning is highly recommended before changing the policy. Deduplication policies are set at the Data Mover level.

Administrators can configure several policy parameters to determine what constitutes active and inactive files in their environment. If they change the policy, it will also change the policy that is run on all deduplication-enabled file systems on the Data Mover. Administrators can manually define the policy to adjust the values for last access and modification time thresholds.

File size is also a criterion to decide whether a file is a candidate for deduplication. Administrators can define the minimum and maximum size parameters to define the file size range that files must meet for deduplication.

Administrators can also define filters for NX4 to avoid processing files based on the file extension.

There are features to reduce the burden of managing Celerra Data Deduplication. The impact of the deduplication processing on Dell NX4 is controlled through automated scheduling of the process and self-throttling based on the CPU load. Each Data Mover scans and deduplicates only one file system at a time. While scanning and deduplicating files, if the NX4 detects that the CPU load of the Data Mover exceeds a user-defined threshold, the process throttles its activity to a minimal level until the time it detects that the CPU load has decreased to less than a low-activity threshold. This means that the deduplication and reduplication processes effectively consume CPU cycles that will otherwise be idle and avoid affecting the system's ability to satisfy client activity.

Celerra Data Deduplication targets aged files and avoids new files that are considered active. Therefore, this feature scans each file system that is marked to be processed no more than once a day. Administrators can adjust this maximum frequency, if required, and they can prompt the system to scan a specific file system immediately, if required.

Management

You can manage the Celerra Data Deduplication feature through the Celerra Manager graphical user interface (GUI) or Control Station command line interface (CLI).

Enabling

To enable Celerra Data Deduplication on a file system, select the checkbox in the **New File System** window of Celerra Manager when creating new file systems. You can also select **On** in the **File System Properties** window of Celerra Manager for existing file systems, as shown in Figure 1 on page 8.

Deduplication state

After you enable a file system for deduplication, Celerra Data Deduplication scans it periodically and looks for more files to deduplicate. You can query the state of the deduplication process for each file system through the Control Station CLI, and view it in the **File System Properties** window of Celerra Manager as shown in Figure 1 on page 8.

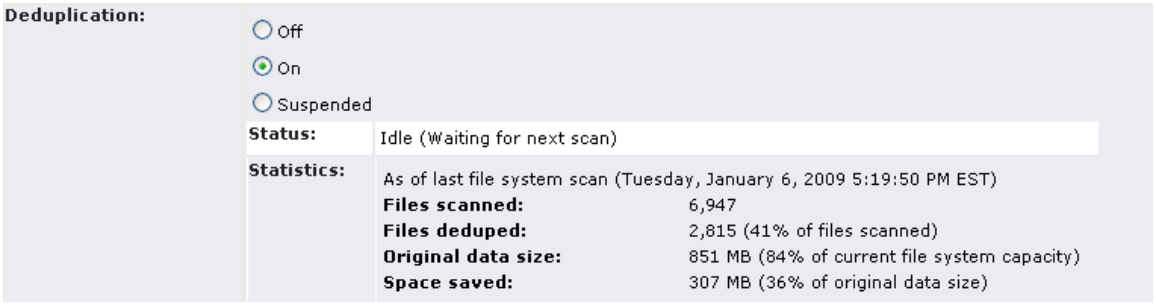

Figure 1 Celerra Data Deduplication section in Celerra Manager

The default Celerra Data Deduplication state for a file system is "Off." When a file system is in the "Off" state, it has no deduplicated files and the policy engine does not scan for files to deduplicate.

The “On” state indicates that the Celerra Data Deduplication processing is enabled for the file system. When a file system is in the “On” state, it may contain deduplicated files, and the policy engine scans the file system for more files to deduplicate on its next scheduled run.

The “Suspended” state means that the Celerra Data Deduplication processing is suspended for the file system. When a file system is in the “Suspended” state, it may contain deduplicated files. However, the policy engine does not scan to look for any more files to deduplicate.

The Dell NX4 administrators can switch between deduplication states at any time. If they switch the deduplication state for a file system from “On” or “Suspended” to “Off,” it prompts the system to reduplicate all deduplicated files in the file system. The system checks whether there is sufficient space available in the file system to complete this process (without filling the file system completely) before accepting the request to turn deduplication “Off.” If the space is not sufficient, the system informs the administrator about the amount of additional space that is required to complete the operation, and recommends the extension of the file system.

Deduplication status

As shown in Figure 1, Celerra Data Deduplication also reports whether the status of a given file system is:

- **Scanning** — Currently being processed. An estimate of its progress through the file system is provided.
- **Idle** — Waiting to be processed.
- **Reduplicating** — Currently reduplicating deduplicated files. An estimate of its progress is provided.

Deduplication statistics

As shown in Figure 1, the Dell NX4 displays the results of the deduplication process on the data in the file system with the following statistics:

- When the last successful scan of the file system was completed.
- **Files scanned** — Total number of files that the deduplication policy engine looked at when it last scanned the file system.
- **Files deduplicated** — Number of files processed by the deduplication policy engine to save space. It also shows the percentage of deduplicated files versus scanned files.
- **Original data size** — Space required to store the data in the file system if it is not de-duplicated. This number might exceed the capacity of the file system, in which case the file system is said to be overprovisioned. This is shown by the ratio of the original data size to the file system capacity, which is also displayed.
- **Space saved** — Amount and percentage of space saved by deduplication. This is calculated by subtracting the actual space used to store data after deduplication from the original data size.

When scanning a file system enabled for deduplication the first time, the statistics shown are reported in real time. After the first scan, statistics are reported as static values based on the last successful scan.

Client input/output to space-reduced files

The Celerra Data Deduplication feature does not affect the client input/output (I/O) to files that have not been deduplicated. The feature does not introduce any additional overhead for access to files that it has not processed. The default policy is designed to filter out files that have frequent I/O access and thus avoid adversely affecting the time required to access those files.

Read access to deduplicated files is satisfied by decompressing the data in memory and passing it back to the client. The Dell NX4 does not decompress or alter any data on disk in response to client-read activity. In addition, random reads require decompression of the requested portion of the file data and not of the entire file data. Reading a file that is not deduplicated can take longer than reading a deduplicated file, because of the decompression activity. However, the opposite may also be true. Reading a deduplicated file is sometimes faster than reading a file that is not deduplicated, because less data needs to be read from the disk, which more than offsets the increased CPU activity associated with decompressing the data.

A client request to write to or modify a deduplicated file causes the file to reduplicate (decompress) in the file system. A write to or a modification of a deduplicated file that is single-instanced causes a copy of the file data to be reduplicated for this particular instance while preserving the deduplicated data for the remaining references to this file. This means that the initial write I/O to a deduplicated file takes longer than subsequent I/Os to the file. The following three factors mitigate this effect:

- Most applications do not modify files. They typically make a local copy, modify it, and when finished, write the entire new file back to the file server, discarding the old copy in the process. Therefore, the file is never reduplicated on the file server; it is just replaced.
- The Dell NX4 avoids processing active files (accessed or modified recently) based on policy definitions. Hence, deduplicated files are less likely to be modified and, if they are, performance is less likely to be a critical factor.
- By default, NX4 avoids deduplicating files larger than 200 MB. This is because the Common Internet File System (CIFS) redirector on Windows clients does not allow an I/O to a file if it does not receive a response within 25 seconds. Performance testing has shown that NX4 can reduplicate a 200 MB file in less than 25 seconds even when it is under heavy load. This feature helps to avoid Windows client timeouts if the clients modify a deduplicated file in the production file system (PFS). Note that if the environment does not contain any Windows clients or applications that modify files in the PFS, the maximum file size can be increased, as required, to provide potentially greater space savings.

ARCHITECTURE

Figure 2 is a block diagram representation of the Celerra Data Deduplication software architecture within the data access in real time (DART) operating system.

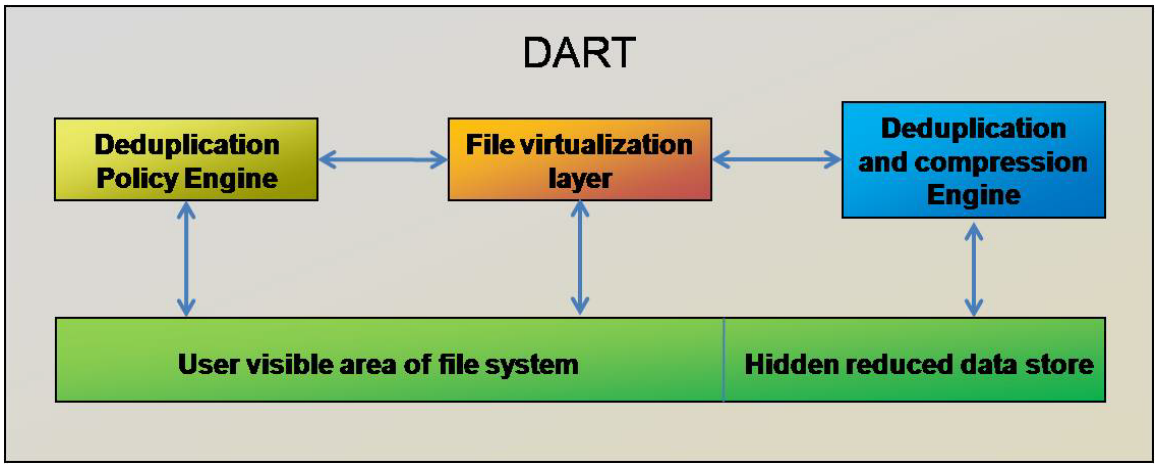

Figure 2 Celerra Data Deduplication software architecture

The deduplication policy engine searches the user-visible portions of file systems for files to be processed based on the policy parameters. When it finds a file to process, it informs the file virtualization layer within DART, which passes the data associated with the file to the deduplication and compression engine.

The deduplication and compression engine determines if the data associated with each space-reduced file has been seen and stored before and stores the associated data in the hidden and reduced data store portion of the file system. The deduplication and compression engine determines if a given body of data has been seen before by calculating a SHA-1 hash of the data and comparing that hash to the indices it maintains in the hidden and reduced data store portion of the file system.

The deduplication and compression engine also stores data in the hidden and reduced data store portion of the file system in a compressed form and retrieves it when required. Note that the compression algorithm automatically detects if the data associated with a particular file does not reduce in size when compressed, and stores that data in its original decompressed form.

The file virtualization layer is responsible for reducing the data associated with files transparent to clients. The only indication that clients have that file data has been reduced is that its “size on disk” attribute is smaller than when the file was created or last modified.

PERFORMANCE

Celerra Data Deduplication is meant to be a non-invasive feature that can achieve storage efficiency while managing to avoid high performance usage on the Data Mover when the policy engine is active. This feature is meant to run periodically, because the policy engine is designed to look for aged files and not newly modified or created files.

Deduplication processing can scan up to a million files in three minutes by using 5 percent of CPU processing power. Depending on the system load, deduplication can process 1.5 TB (on a very busy Data Mover) to 15 TB (on an idle Data Mover) of data per week.

Because read access to a deduplicated file is a pass-through operation, users should see little impact in terms of read performance. Random reads of deduplicated files appear to users in the same way as reads of normal files. Large sequential reads of deduplicated files may take 25 percent longer than reads of the files in their unprocessed form. Because reading a deduplicated file requires about 50 percent more CPU cycles than an unprocessed file, performance can be affected if many deduplicated files are read simultaneously, though these occurrences should be minimized if the policy is properly tuned to meet the work patterns of the environment.

Writing to a deduplicated file prompts a full reduplication of the compressed file data. The reduplication process pauses the I/O until the file is fully reduplicated back to the user-visible area of the file system. Users are not likely to notice a recall time for smaller files. By default, the maximum file size for deduplication is set to 200 MB to prevent CIFS clients from timing out.

INTEROPERABILITY WITH OTHER DELL NX4 FEATURES

Network and NDMP PAX (tar and dump)

When backed up over the network through CIFS or NFS, space-reduced files are decompressed to their original size for transfer to the backup application, although the data is not decompressed on disk. This means that the space saving storage efficiency benefits realized in the Celerra deduplicated production file system do not flow through to backups when using network-based or PAX-based NDMP backups. Further, the decompression of the deduplicated data requires additional CPU overhead, which slows down the backup of individual deduplicated files. The backup of a deduplicated file takes two to four times longer than normal, with smaller files (less than 1 MB) proportionally the slowest. In a typical deduplicated file system, with a mix of deduplicated and normal files of varying sizes, this type of backup is projected to take 20 percent to 35 percent longer than the backup of the same file system before deduplication. Those who are concerned about this impact on the time taken for backup can use Celerra Network Data Management Protocol (NDMP) Volume Backups.

In addition, when restoring files from a PAX-based NDMP backup, the file is restored as a normal file and is not deduplicated. Therefore, the file requires more file system space after the restore. Depending on the amount of available space within the file system, restoring previously deduplicated files from a tape can potentially consume all the free space within the file system.

NDMP Volume Backup (NVB)

NDMP Volume Backup (NVB) can back up Celerra Data Deduplication-enabled file systems and restore them in full by using the full destructive restore method. Because NVB operates at the block level (while preserving the history of the files it backs up), it does not cause any data reduplication when backing up a deduplicated file system. The data in the file system is backed up in its reduced form. This means that the benefits of the storage efficiency realized in the production file system flow through to backups.

However, the Dell NX4 does not support a single-file or file-by-file restore of deduplicated files from NVB backups. Hence, Dell recommends that NVB backups of deduplicated file systems should be used as part of a strategy where a single-file or file-by-file restore is done from locally or remotely replicated SnapSure checkpoints and not from "tape." Because the majority of file restores happen within the first few days after their deletion, a SnapSure checkpoint is an efficient and faster way to restore most files.

Point-in-time views of the file system

The deduplication process releases space in the production file system immediately. However, blocks may be copied to the SnapSure SavVol in the process. Deduplicating data associated with a file involves copying the data within the file system to the hidden and reduced data store so that it can be compressed and single-instanced. Because the SnapSure checkpoint copy changes blocks to the save volume on first write, you need to copy the blocks that get deduplicated to the SavVol to preserve a previous checkpoint point-in-time view of the file system. These blocks are freed when the corresponding checkpoint gets deleted or refreshed and are available for reuse by other checkpoints. The number of blocks that are copied to the SavVol during the deduplication process is a function of how full the file system is, the rate of change in it, and so on, and hence it is difficult to predict. By default, NX4 is configured to abort deduplication operations on a file system before it causes the SavVol to extend. This avoids the SavVol expanding due to deduplication activity. If the deduplication process is aborted in this way, an alert is generated that explains what happened. The Dell NX4 administrator can choose to extend the SavVol or simply let the deduplication process execute again on its next scheduled run.

Replication

Deduplicating the contents of a file system before you replicate it by using Celerra Replicator™ can greatly reduce the amount of data that is sent over the network as part of the initial baseline copy process. When replication and deduplication run together, the impact of deduplication on the amount of data transferred over the network depends on the relative timing of replication updates and deduplication runs. In all but the most extreme circumstances, replication updates are more frequent than deduplication scans of a file system. This means that new and changed data in the file system is usually replicated in its non-deduplicated form first, and any subsequent deduplication of that data prompts additional replication traffic due to the block changes within the file system to the hidden and reduced data store. This is true of any deduplication solution that processes replicated data and updates remote replicas of the data because the replication updates occur more frequently than deduplication scans. The space savings realized by the production file system is reflected on the destination file system.

Quotas

The Dell NX4 can track users, groups, and directory tree quotas that use either of two quota "policies." It can track quota usage based on the logical size of files or the size that the files occupy on disk (block size policy). The process of deduplicating a file does not change its logical size and, hence, has no effect on logical file size-based quota tracking. The process of deduplicating a file reduces its size on the disk. Therefore, if the quota is based on the size on the disk, user, group, and directory tree usage quotas reduce as files are deduplicated and increase as they are reduplicated.

Celerra FileMover archiving

Celerra Data Deduplication is transparent to Celerra FileMover archiving. You can use these two features together to maximize the storage efficiency of the file storage solution. Any files archived from a NX4 file system on which Celerra Data Deduplication is enabled is written to and read from the archive storage in the non-deduplicated form. Note, however, that the archive storage system may deduplicate the archived data itself. NX4 file systems used as repositories for archived data are good candidates for Celerra Data Deduplication.

Combining FileMover and Deduplication can provide a tiered file storage solution that allows production or near-production performance for "hot" and occasionally accessed data, and storage savings for seldom-used data that can be archived over to less expensive storage. As files progress through their life cycles, they can seamlessly progress from production storage to spaced-reduced storage and then to archival storage, while decreasing the space and cost footprint in the process.

Celerra File-Level Retention

You can enable Celerra Data Deduplication on both the enterprise and compliance types of Celerra File-Level Retention (FLR) file systems without compromising on the protection offered to the data that the file systems contain.

DEPLOYMENT CONSIDERATIONS

The following need to be considered when deploying Celerra Data Deduplication:

- Celerra Data Deduplication does not deduplicate data across or between file systems.
- File systems enabled for processing by Celerra Data Deduplication can be replicated by using Celerra Replicator (V2), but not V1. The destination Data Mover is required to support Celerra Data Deduplication.
- You can back up Celerra Data Deduplication–enabled file systems by using Celerra NVB and restore in full. However, the Dell NX4 does not support a single-file or file-by-file restore of deduplicated files from NVB backups. Dell recommends the use of NVB backups of deduplicated file systems as part of a strategy where a single-file or file-by-file restore is done from locally or remotely replicated SnapSure checkpoints and not from "tape."
- Celerra Data Deduplication does not apply to iSCSI logical unit numbers (LUNs) hosted by NX4.
- Celerra Data Deduplication does not process or affect alternate data streams (also known as named attributes) associated with files and directories in the file system.
- Celerra Data Deduplication does not process files less than 24 KB in size. The overhead associated with processing such files negates any space savings achieved.
- There is no limit to the size of a file system on which Celerra Data Deduplication can be enabled. However, the file system must have at least 1 MB of free space before deduplication can be enabled. If there is not enough free space, an error message is generated and the server log is updated.
- You can enable Celerra Data Deduplication on any number of file systems.
- You can enable Celerra Data Deduplication on existing file systems.

CONCLUSION

The Celerra Data Deduplication feature adds to Dell NX4's impressive storage efficiency capabilities by intelligently reducing space usage and providing further storage efficiency. You can optimize the use of the existing storage environment for file system data through single instancing and compression.

A combination of file-level deduplication and compression represents the best technique to provide maximum benefit for the resources consumed. When applied to file system data, you can expect savings in the range of 30 percent to 40 percent for typical file-share data. This feature builds on NX4's ease of use by providing a single-click option in Celerra Manager to enable deduplication for each file system. Celerra Data Deduplication also supports almost all NX4 features and works in accordance with maximum-supported NX4 limits.