Technical Report

Storage Savings with Domino and NetApp Deduplication

John B. Spinks, NetApp
April 2010 | TR-3843

ABSTRACT

NetApp® deduplication technology provides block-level intelligent data compression by eliminating redundant data. Only one unique instance of the data is retained on storage. Redundant data is replaced with a pointer to the unique data copy. Combining NetApp deduplication and Lotus Domino offers customers not only lower storage space requirements and lower capital expenditures but also huge savings in terms of lower energy bills and storage footprint. Deduplicated or compressed data also means longer disk retention periods and less data transfer on the corporate WAN for backup, replication, and disaster recovery. This paper describes the use of NetApp deduplication in Domino environments.

TABLE OF CONTENTS

1 EXECUTIVE SUMMARY

With IT budgets contracting, IT managers are charged with delivering an ever-increasing level of service to users and showing tangible financial savings as part of cost-cutting measures. NetApp deduplication is a storage technology that enables IT managers to realize significant savings and deliver on the "store more with less" promise. This technology delivers intelligent data compression by eliminating redundant data at the storage-block level. Only one unique instance of the data is retained on the storage. Redundant data is replaced with a pointer to the unique data copy. More efficient use of disk space with deduplication allows longer disk retention periods, better recovery time objectives, and less data transfer across the WAN for disaster recovery. This technology helps companies to achieve their green initiative goals by lowering power requirements for cooling and operations. Other benefits are a smaller data center footprint and less manpower required to manage the storage infrastructure. NetApp deduplication is an integral part of Data ONTAP® and is suitable for all storage tiers, including primary, backup, and archival storage.

This document walks you through the steps required to configure NetApp deduplication on a NetApp storage system in an IBM Lotus Domino environment. For this paper, we assume that NetApp deduplication is implemented in an existing NetApp and Domino environment; that your NetApp storage system and IBM Lotus Domino environments are already configured and are in working condition; and that you are familiar with the basic operations and terminology of a NetApp storage system, Domino server, and your host operating system.

2 DEDUPLICATION OVERVIEW

Data deduplication is a method of improving storage utilization by eliminating redundant data. In the data deduplication process, one unique copy of the data is retained while redundant data is replaced with a pointer to the unique copy.

NetApp deduplication, part of NetApp's storage efficiency offerings, provides block-level deduplication within an entire flexible volume on NetApp storage systems. Deduplication is supported on FAS and V-Series systems that are using Data ONTAP 7.2.5.1 or later for customers running Data ONTAP 7.2, or Data ONTAP 7.3P1 or later for customers running Data ONTAP 7.3.

For a complete understanding of deduplication, read TR-3505: "NetApp Deduplication for FAS and V-Series Deployment and Implementation Guide[1]." This paper contains portions of TR-3505.

Note: To access the NetApp Data ONTAP manuals and reference guides, you must have a NOW™ (NetApp on the Web) user name and password. Register for free at http://now.netapp.com.

2.1 HOW DEDUPLICATION WORKS

The core enabling technology of deduplication is fingerprints—unique digital "signatures" for every used 4KB block in the flexible volume. NetApp deduplication stores only unique blocks in the flexible volume. When deduplication runs for the first time on a flexible volume with existing data, it scans the blocks in the volume and creates a fingerprint database, in the form of a fingerprint file. This file contains a sorted list of all fingerprints for used blocks in the flexible volume. After the fingerprint file is created, fingerprints are checked for duplicates. If a duplicate is found, a byte-for-byte comparison is done of all bytes in the block. If an exact match is found, the block's pointer is updated to the existing data block and the new (duplicate) data block is released.

In real time, as additional data is stored on the deduplication-enabled flexible volume, a fingerprint is created for each new block and is written to a change log file. When deduplication is run subsequently, the change log is sorted and its sorted fingerprints are merged with those in the fingerprint file. Then the deduplication processing occurs. The fingerprint and change log files both exist in the flexible volume metadata information. During the deduplication process, some temporary files are also created. These temporary metadata files are deleted after the deduplication process completes.

Figure 1 is a high-level overview of how NetApp deduplication works.

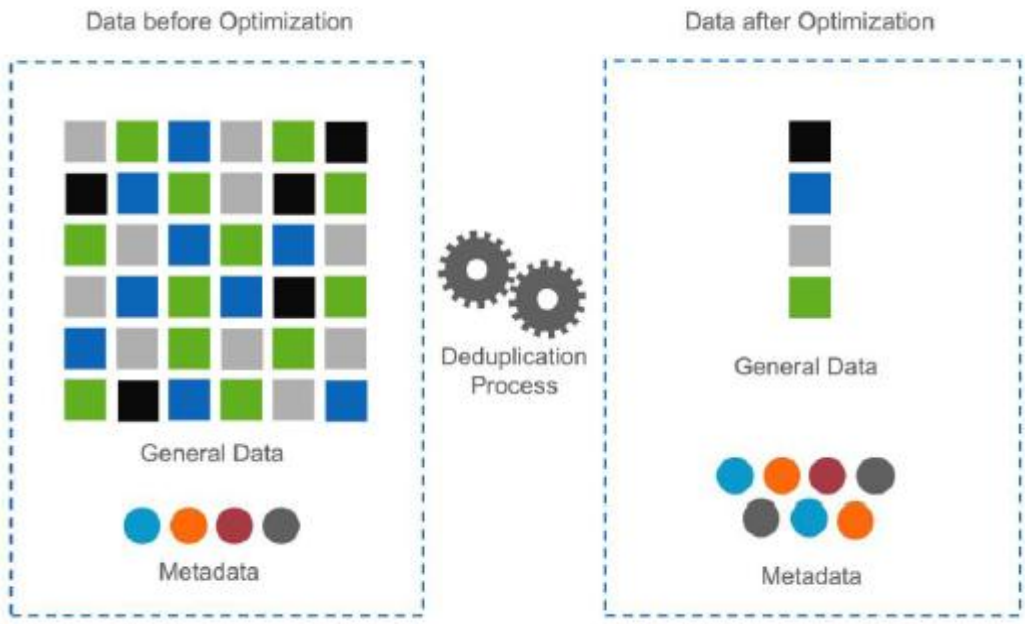

Figure 1) How NetApp deduplication works.

The deduplication feature is enabled on a per flexible volume basis and can be enabled on any number of flexible volumes in a storage system.

Deduplication can be run in four ways:

- Scheduled on specific days and at specific times
- Manually via the command line
- Automatically when 20% new data has been written to the volume
- Automatically on a destination volume when used with SnapVault®

2.2 SPACE PLANNING

Before you implement deduplication, you need to plan the storage space required for metadata. The deduplication process creates fingerprint, change log, and temporary metadata files. The total storage used by the deduplication metadata files is approximately 1% to 6% of the total storage space of the data in the volume. The breakdown of space requirements associated with the deduplication metadata is as follows:

- **Fingerprint metadata file:** Space requirement for fingerprint database file is less than 2%.
- **Change log file:** There are two deduplication change log files, their size fluctuates depending on how frequently deduplication is run and the rate of change of the data, but they generally account for a less than 2% storage space.
- **Temporary files:** Deduplication creates some temporary metadata files while running. These files are deleted when the deduplication process completes. You should allocate approximately 2% of storage space for temporary files.

In Data ONTAP 7.2.x, all deduplication metadata files reside in the volume. If you take a Snapshot™ copy, this metadata is captured and locked in the Snapshot copies of the volume.

Starting with Data ONTAP 7.3, a portion of the metadata still resides in the volume, and part of it resides in the aggregate outside the volume. The fingerprint database and the change log files are located outside the volume in the aggregate and therefore are not captured in Snapshot copies. However, the temporary metadata files created during the deduplication operation are still kept inside the volume. These temporary files are deleted when the deduplication operation completes, but if a Snapshot copy is taken while deduplication is running; the temporary metadata files can be locked into the Snapshot copies.

Therefore you need to plan a little differently for Data ONTAP 7.2.x than for Data ONTAP 7.3:

If you're running Data ONTAP 7.2.X, leave about 6% extra space inside the volume on which you plan to run deduplication.

If you're running Data ONTAP 7.3, leave about 2% extra space inside the volume on which you plan to run deduplication, and about 4% extra space outside the volume in the aggregate, for each volume running deduplication.

2.3 SPACE SAVINGS ESTIMATION TOOL

In order to estimate space savings, NetApp offers the Space Savings Estimation Tool (SSET). You can use this tool to analyze the actual data set and determine the effectiveness of deduplication. Although the space savings in SSET are only estimates, use and testing thus far indicate that the actual results are within +/- 5% of the estimated results.

SSET is available to NetApp system engineers and partners to perform nonintrusive testing of data. The tool is available for Linux® and Windows® systems and is limited to evaluating a maximum of 2TB of data. See the SSET readme file for complete usage information. SSET can be downloaded from the NetApp Field Portal under the Tools tab in the Estimator category. The Field Portal is limited to NetApp partners and employees.

3 CONFIGURATION OVERVIEW

Deduplication is an integral part of Data ONTAP. To activate deduplication you must have a deduplication license. Additionally, if you are running on a FAS system, you must have a NearStore® license. There is no additional charge for either license; just contact your NetApp sales representative.

No changes to your Domino environment are required to enable deduplication.

3.1 INSTALLING AND LICENSING DEDUPLICATION

Deduplication is included with Data ONTAP; it just needs to be licensed. To add the deduplication license, use the following command:

license add [a_sis license key]

To run deduplication on any of the FAS platforms, you also need to add the NearStore option license:

license add [nearstore_option license key]

Note: Parameters shown in angle brackets (< >) are optional. Parameters and options shown in square brackets ([ ]) are required and must be provided. A comma followed by an ellipsis (…) indicates that the preceding parameter can be repeated multiple times.

3.2 ACTIVATING DEDUPLICATION

Once the storage environment is properly licensed for deduplication, you need to enable deduplication on your flexible volume:

```
sis on <vol>
```

Next scan and deduplicate your existing data:

```
sis start -s <vol>
```

When deduplication is first enabled on a flexible volume, a default schedule is configured, running nightly at midnight. To modify this schedule, use the following command:

```
sis config [-s sched] <vol>
```

Deduplication can be manually run at any time with this command:

```
sis start <vol>
```

Monitor deduplication status with this command:

```
sis status <vol>
```

Monitor space savings with this command:

```
df -s <vol>
```

4 DOMINO AND DEDUPLICATION

The storage savings that you can expect vary widely with the type (e-mail, applications, and so on) and redundancy of data in your environment. In the NetApp test lab, running SSET on a fresh installation of Domino, a space saving of approximately 22% was obtained, before any actual data was managed. NetApp customers using Domino have reported anywhere from 25% to 60% savings in their Domino environment.

Figure 2 shows some examples of NetApp deduplication savings.

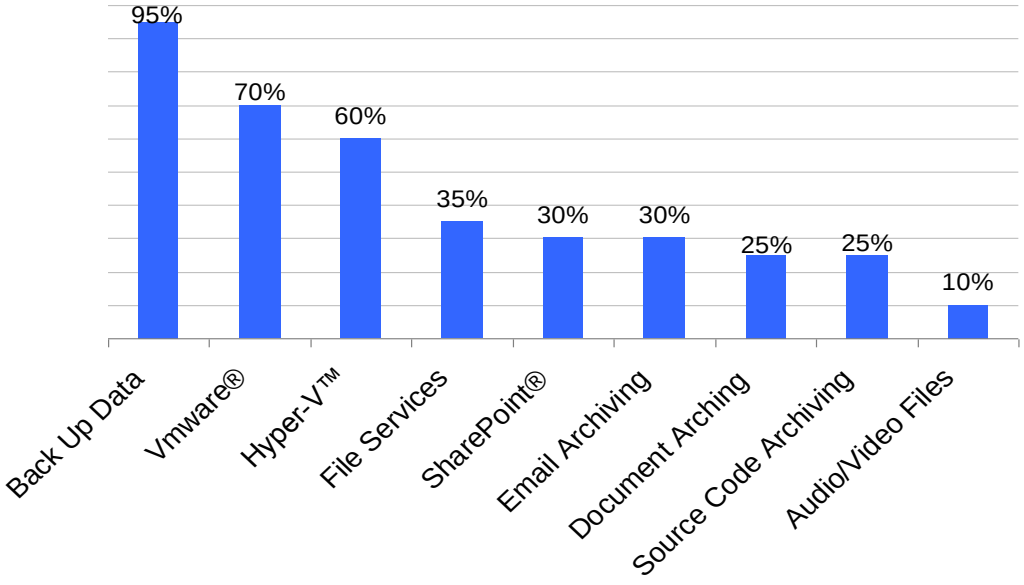

Figure 2) Examples of NetApp deduplication savings.

4.1 DOMINO ATTACHMENT AND OBJECT SERVICE

Domino 8.5 introduces a new feature, Domino Attachment and Object Service (DAOS). DAOS shares data identified as identical among multiple databases. Typically, in a Domino environment a separate and complete copy of each attached document is kept in each database that receives it. With DAOS, the server saves a reference to each attached file in an internal repository and refers back to the same file from multiple documents and/or databases.

IBM estimates that savings with DAOS can average 25% of space consumed. Many companies have seen 60% or even greater storage savings by enabling DAOS in their environments.

NetApp deduplication is still effective when DAOS is enabled, but it is expected that the reported space savings from NetApp will be lower because DAOS has already performed much of the work. The overall space savings should be greater because of the use of two complementary technologies to deduplicate the data set.

4.2 DOMINO ENCRYPTION

Domino data encryption adds extra information for each block and therefore each block has a unique fingerprint. If Domino database encryption is enabled for all or the majority of databases, the space savings from deduplication are expected to be very small.

4.3 QUOTAS

Domino quotas are not affected by deduplication. A mailbox with a limit of 1GB cannot store more than 1GB of data in a Deduplicated volume, even if the data consumes less than 1GB of physical space on the storage system.

4.4 PERFORMANCE

The performance for a deduplication-enabled environment varies widely based on the type of data, amount of redundant data, and average file size as well as the storage platform, types of disks, and number of disk spindles in the aggregate.

Because of these factors, NetApp recommends that you size your environment appropriately, keeping your performance and deduplication goals in mind. Your NetApp representative can use NetApp sizing tools to estimate the optimal number of disks to use with your platform.

5 GENERAL DEDUPLICATION CONSIDERATIONS

5.1 PROCESSING

Only one deduplication process can be run on a flexible volume at a time, and a maximum of eight deduplication processes can run concurrently on the same NetApp storage system

5.2 PERFORMANCE

- NetApp recommends that before deploying deduplication in your environment, you carefully asses the performance impact due to deduplication, measure it in a test setup, and take sizing into consideration.
- Deduplication runs as a low-priority background process on the system; however, it can still affect the performance of user I/O and other applications running on the system

5.3 SNAPSHOT TIMING

If possible, run deduplication, and make sure that it is complete, before creating a Snapshot copy. During deduplication temporary files are created in the volume, and if a Snapshot copy is taken while deduplication is running you may be taking a copy of this temporary data, thus reducing your space savings.

6 SUMMARY

NetApp deduplication is based on a proven technology with a wide customer adoption. NetApp deduplication improves data center efficiency by allowing longer data retention on disk and improved space, power, and cooling requirements. Deduplication is suitable for a variety of applications, including Lotus Domino. Lotus customers running Domino 8.5 or later with Domino Attachment and Object Service (DAOS) enabled will see greater overall storage savings because they are combining two deduplication technologies, but the savings reported by NetApp deduplication will be lower than if DAOS is not in use. NetApp deduplication is an excellent choice for use with Lotus customers running any version of Domino.

7 ACKNOWLEDGEMENTS

Any paper such as this is never solely the work of the author; it is the work of the collective team that helped such a paper come together. I would like to acknowledge the efforts of Carlos Alvarez; author of TR-3505, which is the source of most of the information on NetApp deduplication. I would also like to thank Jawahar Lal and Michelle Nguyen for taking the time to review and refine this technical report. Additionally, I would like to thank the editing, legal, and creative services teams that work so hard on polishing and making this information available.

APPENDIX A: DEDUPLICATION COMMAND SUMMARY

This command summary is copied from TR-3505.

`sis on <vol>`	Enables deduplication on the specified flexible volume.
`sis start -s <vol>`	Begins the deduplication process on the flexible volume specified and performs a scan of the flexible volume to process existing data. This option is typically used upon initial configuration and deduplication on an existing flexible volume that contains undeduplicated data.
`sis start <vol>`	Begins the deduplication process on the flexible volume specified.
`sis status [-l] <vol>`	Returns the current status of deduplication for the specified flexible volume. The `-l` option displays a long list.
`df –s <vol>`	Returns the value of deduplication space savings in the active file system for the specified flexible volume. Use this command to see how much space has been saved.
`sis config [-s sched]\ <vol>`	Creates an automated deduplication schedule. When deduplication is first enabled on a flexible volume, a default schedule is configured, running it each day of the week at midnight.
`sis stop <vol>`	Suspends an active deduplication process on the flexible volume specified.
`sis off <vol>`	Deactivates deduplication on the flexible volume specified. This means that there will be no more change logging or deduplication operations, but the flexible volume remains a deduplicated volume and the storage savings are kept.
`sis check <vol>` (This command is available only in Diag mode.)	Verifies and updates the fingerprint database for the specified flexible volume; includes purging stale fingerprints.
`sis stat <vol>` (This command is available only in Diag mode.)	Displays the statistics of flexible volumes that have deduplication enabled.
`sis undo <vol>` (This command is available in Advanced and Diag modes.)	Reverts a deduplicated volume to a normal flexible volume.

NetApp provides no representations or warranties regarding the accuracy, reliability or serviceability of any information or recommendations provided in this publication, or with respect to any results that may be obtained by the use of the information or observance of any recommendations provided herein. The information in this document is distributed AS IS, and the use of this information or the implementation of any recommendations or techniques herein is a customer's responsibility and depends on the customer's ability to evaluate and integrate them into the customer's operational environment. This document and the information contained herein must be used solely in connection with the NetApp products discussed in this document.

www.netapp.com