Technical Report

Enterprise Vault 8.0 E-Mail Archive Efficiency on NetApp Storage

Nathan Walker, NetApp Senior Technical Marketing Engineer
April 2009 | TR-3765

DEDUPLICATION METHODOLOGY

This document discusses the methodology and architecture used to analyze the FAS deduplication performance of Symantec™ Enterprise Vault™ 8.0 on NetApp® storage. As always, please refer to the latest technical publications on the NOW™ (NetApp on the Web) site for updates on processes; Data ONTAP® command syntax; and the latest requirements, issues, and limitations.

TABLE OF CONTENTS

1   EXECUTIVE SUMMARY

Symantec Enterprise Vault 8.0 is today's leading e-mail and content archiving solution. Enterprise Vault reduces the amount of storage required for e-mail and file systems by managing content using automated, policy-controlled archiving to online stores for active retention and seamless retrieval of information. Companies using Enterprise Vault on NetApp storage can expect to reduce their e-mail storage requirements by 50% or more, while enabling satisfaction of regulatory and legal requirements. Now you can streamline your operations and minimize your risk with solutions from NetApp.

2   INTRODUCTION

IT organizations are finding ways to intelligently manage exploding e-mail repositories. These groups are transparently moving less frequently accessed e-mail to lower-cost storage using proven technologies. As mail is moved from a flat storage model to a sophisticated archive solution, more data can be managed at a lower cost per megabyte. While reducing IT operations costs, these efforts also establish a foundation to satisfy compliance, records retention, and legal hold requirements. When combined with SnapLock®, archive data permanence is assured on WORM volumes. Innovative storage and data management technology solutions from NetApp and Symantec provide an optimized stack to archive corporate data. Duplicate mails and mail attachments can be stored as a single instance within Enterprise Vault. Duplicate blocks of data can be removed to provide a massively dense and highly efficient archive storage tier for the e-mail archive. You can achieve the most efficient e-mail archiving platform using proven and reliable solutions built upon Symantec and NetApp storage solutions.

3   AUDIENCE

This paper is intended to serve as a strategic planning guide for organizations with data archiving and retention requirements. Readers of this document should have a basic understanding of Symantec Enterprise Vault, Microsoft® Exchange, Microsoft SQL Server®, and NetApp storage systems. This paper is not intended as a replacement for vendor documentation or proper product training. Consult with those vendors for product feature and operating details.

4   SCOPE

The primary intention of this paper is to analyze the storage efficiency capabilities of NetApp deduplication for FAS when used with an Enterprise Vault 8.0 optimized single-instance storage model, or OSIS, to store archived mail content. A comparison will be made between this archiving paradigm and the more familiar paradigms of personal storage table (PSTs) and Microsoft Exchange. While in the Enterprise Vault 8.0 archive, only the native Symantec compression was used; no attempt was made to evaluate third-party compression alternatives, although the modular architecture of Enterprise Vault 8.0 does permit such decisions. Content archives on WORM storage are discussed in this paper, but were not tested for dedupe performance as the algorithms are the same for WORM and non-WORM storage.

5   TECHNOLOGY INTRODUCTION

Enterprise Vault 8.0 is the latest version of the Symantec industry-leading e-mail and content archiving platform. This release was designed with a new set of capabilities and features to provide optimized storage, management, and discovery of corporate data. Among its updated features is OSIS, which is designed to keep single copies of individual e-mails or files regardless of the number of times they occur or from what content source they originate. In addition, the storage interface for Enterprise Vault has been rewritten to make sure that new data blocks are written to align with block boundaries of the storage subsystem. This key improvement facilitates deduplication of data at the storage volume tier. When data writes start at the beginning of the block, there is a higher probability of duplicate blocks for archived content. This means data archives written after the Enterprise Vault 8.0 upgrade will have a greater contribution to data deduplication than those written before.

NetApp deduplication is a fundamental component of Data ONTAP. NetApp has the first deduplication capability that can be used broadly across many applications, including primary data, backup data, and archival data. NetApp is the only vendor with a dedupe WORM-compliant storage system. Data ONTAP 7.3.1 and above feature the ability to dedupe a SnapLock FlexVol® volume in either compliance or enterprise mode. NetApp offers the only suitable platform for strict adherence to SEC 17a-4, NASD 3110, DOD 5015, Sarbanes-Oxley, and HIPAA requirements.

6   TESTING OBJECTIVES

The objective of this paper was to analyze the storage efficiency of Enterprise Vault 8.0 on NetApp storage. The paper discusses the findings and results of a production archive migration from Enterprise Vault 2007 to Enterprise Vault 8.0, with storage deduplication turned on. Additionally, comparisons will be made to the same mail in the Microsoft PST format as well as an Exchange 2003 message store.

7   TESTING ENVIRONMENT METHODOLOGY

A production e-mail archive was used to evaluate the storage efficiencies of the combined Symantec and NetApp storage efficiency solution. Due to time and resource constraints, the entire 1.3TB of the source Enterprise Vault 2007 archive was not processed. The Enterprise Vault 2007 vault store, index, and SQL Server database were copied to a lab environment using Volume SnapMirror®. The disaster recovery procedures in the Enterprise Vault administrator's guide included with the application binaries were used to bring up the archive in a lab environment with all guest OSes configured identically on three IBM xSeries servers functioning as VMware® ESX hosts. Figure 1 shows a representation of the environment.
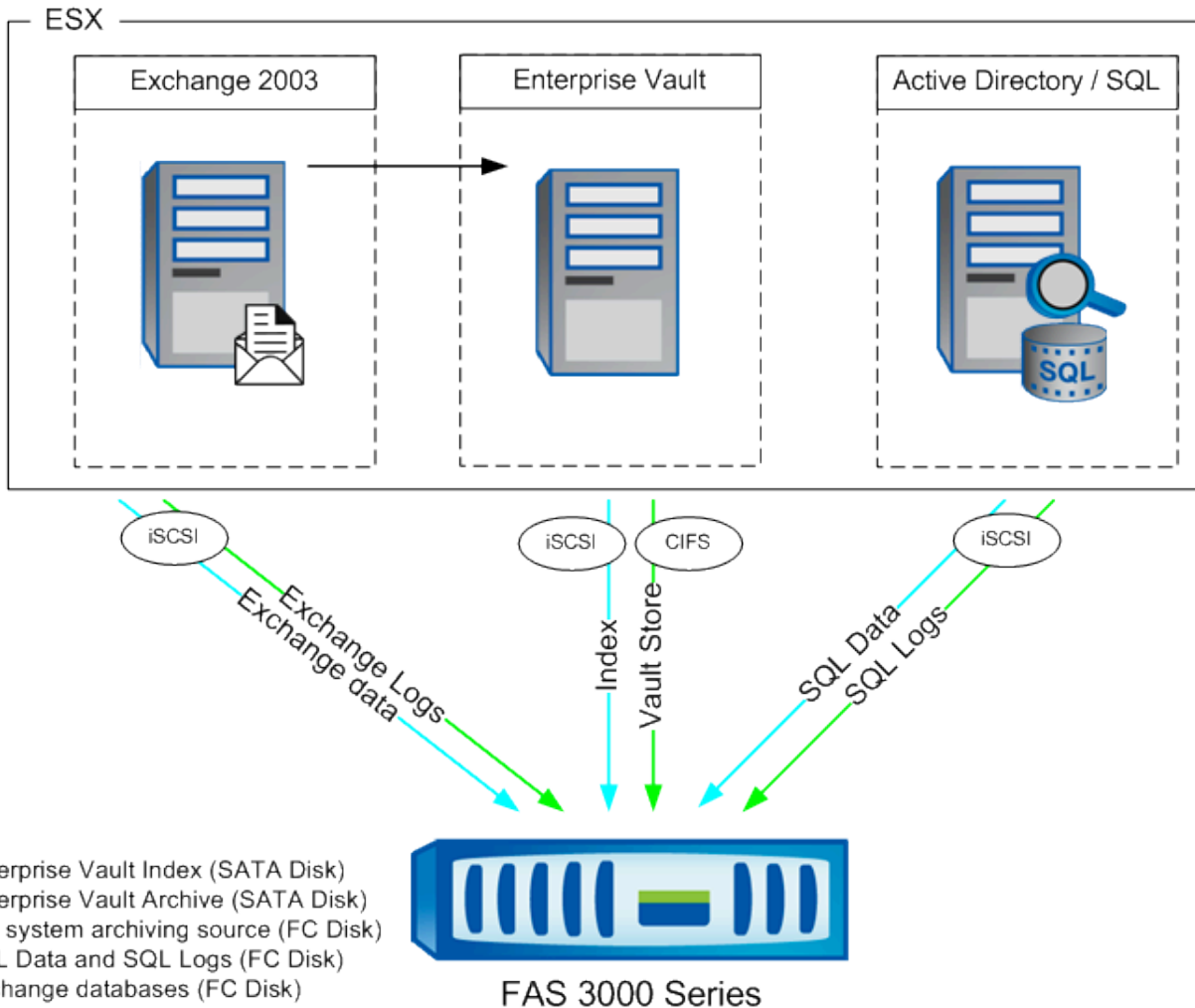
Figure 1) Enterprise Vault 8.0 lab logical architecture.

Once the copy of the production archive was recovered and brought online, a FlexVol volume was created to store the PST files extracted from the recovered archive. The export archive wizard was then used to extract mail from all archives within a 62-day window. This produced 100GB of PSTs from 859 mail archives. Not all archives in the source vault store had archived content during that period. Next, two organization units (OUs) were created in Active Directory for both application scenarios. Finally, Microsoft's csvde utility was used to quickly populate each of the OUs with 859 Active Directory mailbox-enabled users. The usernames were concatenated with ".2007" or ".8" to clearly distinguish each account. Because a disaster recovery scenario of an Enterprise Vault vault store does not require recovery of Active Directory, no attempt was made to recreate the original AD topology or user accounts. Only the relevant user accounts were created in the lab environment.

A screenshot of the Active Directory Users and Computers console is shown below, with the names of the users blocked out for privacy purposes.
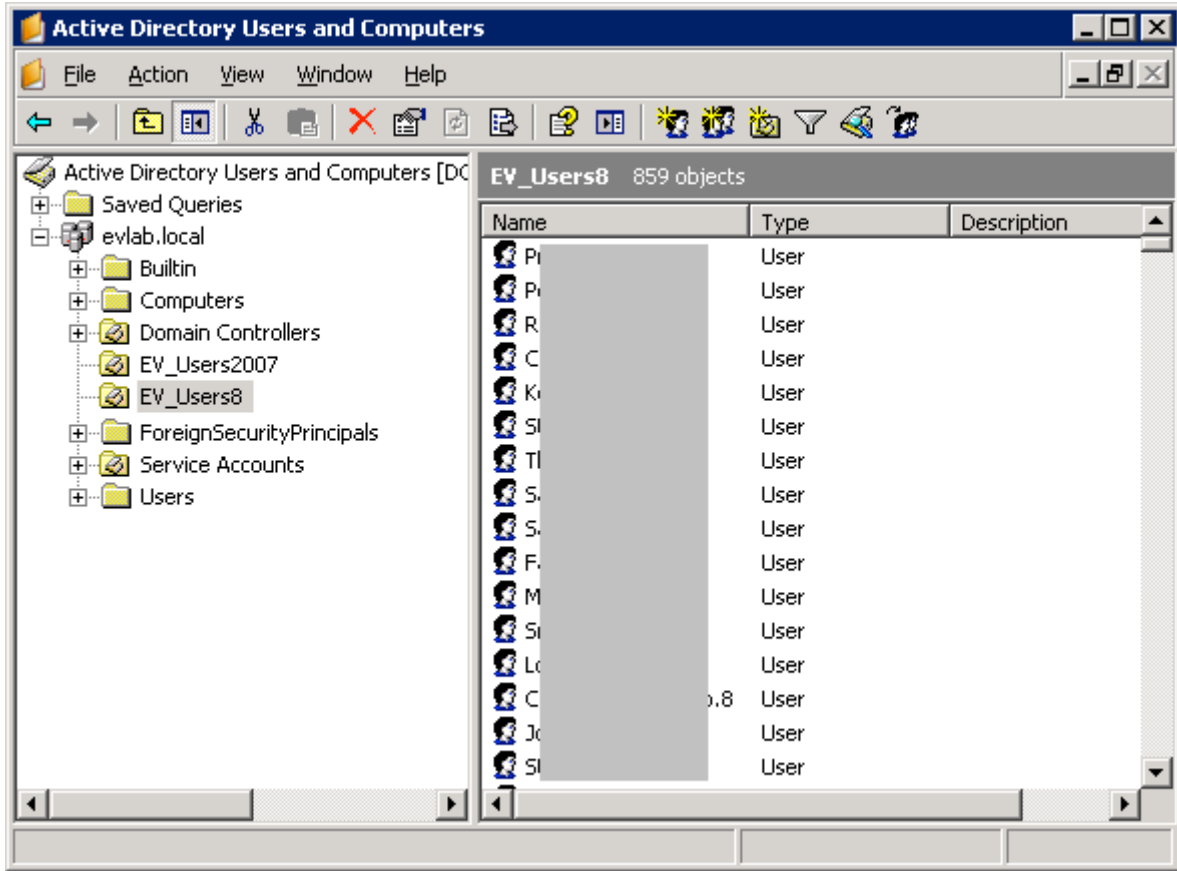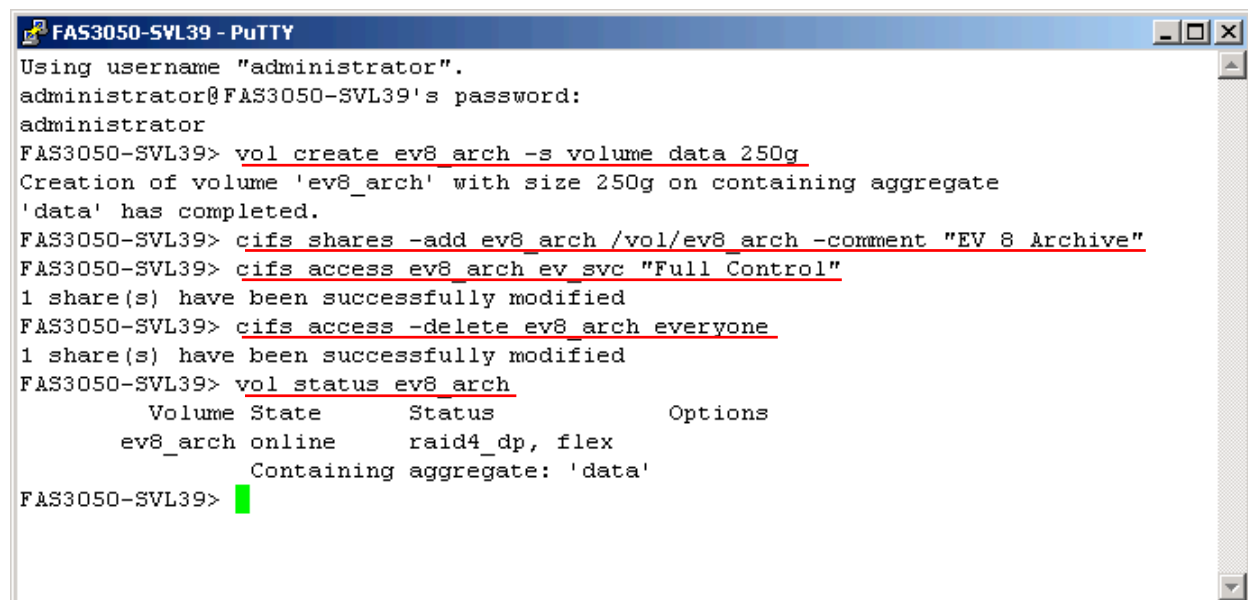

Figure 2) Enterprise Vault 8.0 Active Directory OUs.

To house the vault store partitions for proper analysis, two FlexVol volumes were created on the NetApp aggregate, named ev2007_arch and ev8_arch, accordingly using the command line interface.

```
FAS3050-SVL39 - PuTTY
Using username "administrator".
administrator@FAS3050-SVL39's password:
administrator
FAS3050-SVL39> vol create ev8_arch -s volume data 250g
Creation of volume 'ev8_arch' with size 250g on containing aggregate
'data' has completed.
FAS3050-SVL39> cifs shares -add ev8_arch /vol/ev8_arch -comment "EV 8 Archive"
FAS3050-SVL39> cifs access ev8_arch ev_svc "Full Control"
1 share(s) have been successfully modified
FAS3050-SVL39> cifs access -delete ev8_arch everyone
1 share(s) have been successfully modified
FAS3050-SVL39> vol status ev8_arch
         Volume State       Status              Options
       ev8_arch online      raid4_dp, flex
               Containing aggregate: 'data'
FAS3050-SVL39>
```

Figure 3) FlexVol copy creation.

A Windows® batch script was written to use xcacls to assign the correct NTFS permissions for each PST to its corresponding mailbox archive. The archiving policy was modified to accept all of the default Outlook object classes. Two provisioning groups were created, corresponding to each of the OUs. Then the import process was able to begin.

The Enterprise Vault import wizard was used to bring each user's 62-day window of mail from PSTs using the PST import wizard into the Enterprise Vault 2007 vault store, with the new partition on the dedicated ev2007_arch FlexVol volume. To expedite the process, shortcuts were not created for the archived content. Because the Exchange, Enterprise Vault, and Active Director and SQL Server servers were all guest OSs in ESX hosts, the import performance was considerably lower than with physical servers. Symantec fully supports ESX as a virtualization environment, but warns customers of the degradation. Refer to the Enterprise Vault 8.0 performance guide for a more detailed virtualization discussion.[1]

After verifying the state of the Enterprise Vault 2007 imports, the environment was upgraded to Enterprise Vault 8.0. A new vault store group was created for this scenario with its partition placed on the dedicated ev8_arch FlexVol volume. The storage type was set to "NetApp Device" to make sure of proper block alignment to the WAFL® file system, as shown in Figure 4.
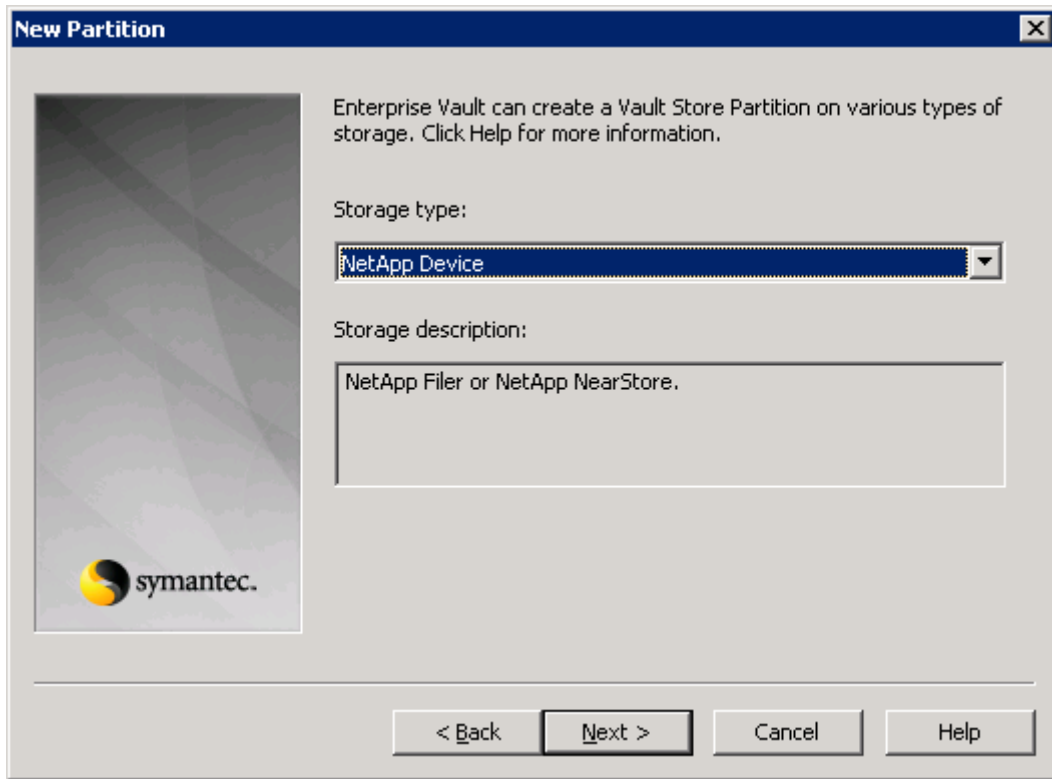

Figure 4) New partition creation on NetApp storage.

The sharing was set to be at the vault store level, as shown in Figure 5. The other partition was closed, and imports began with copies of the same PSTs used for the previous import. After the imports had finished, a thorough review of the message count per archive and event logs was completed to make sure there were duplicate objects in both vault stores.
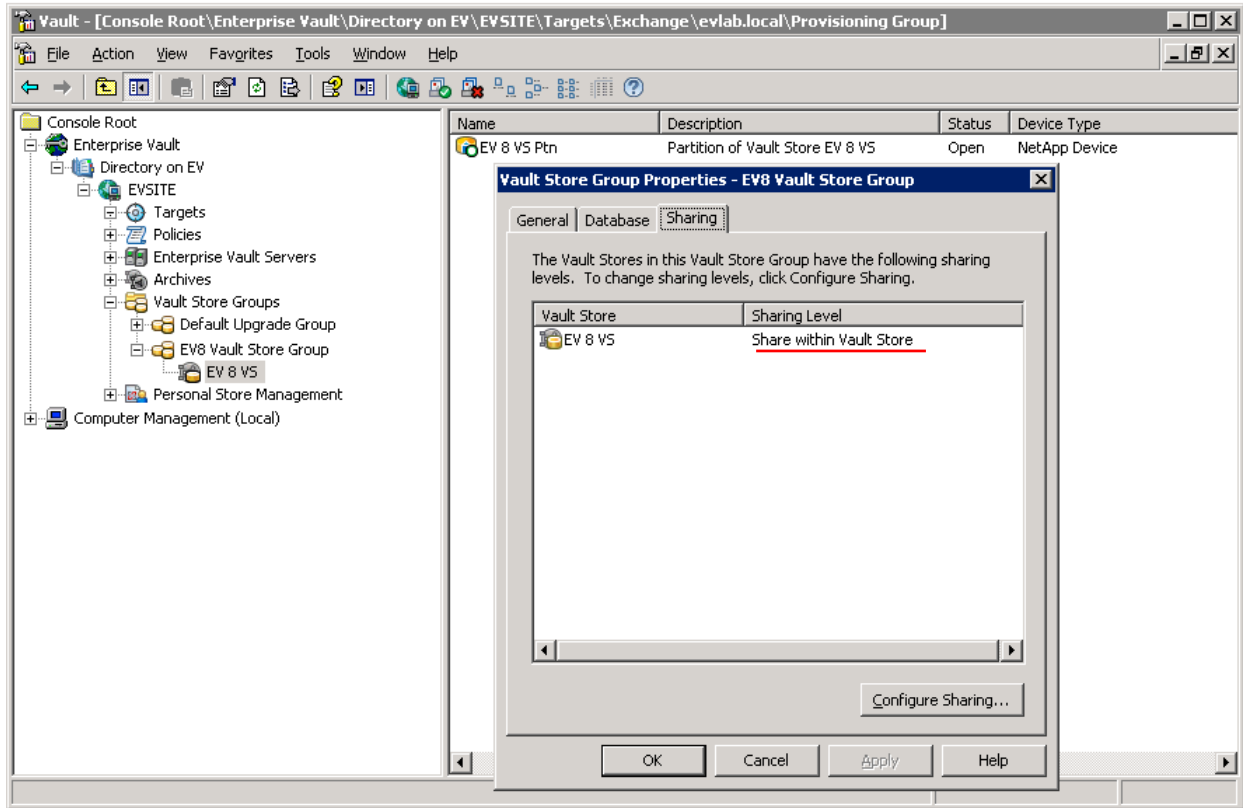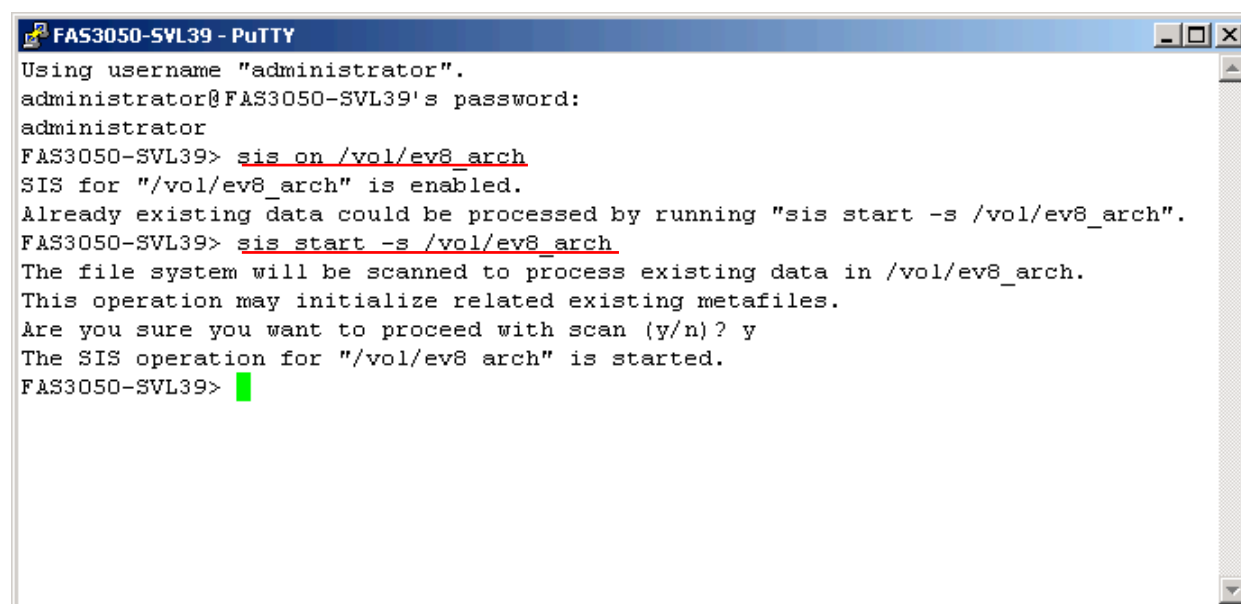

Figure 5) Vault store group properties.

The PST processing rate for the above export and import workflows averaged about 0.8GB per hour. Because some users had manually sent items to the archive of a nonstandard class, these were not recognized by the archiving policy and were therefore omitted in both the Enterprise Vault 2007 and the Enterprise Vault 8 import processes.

Finally FAS deduplication was turned on for the FlexVol volumes holding the Enterprise Vault 2007 partition and the Enterprise Vault 8 partitions. Review TR-3505 for deployment and implementation details for FAS and V-Series deduplication.[2]

```
FAS3050-SVL39 - PuTTY                                                _ | □ | ×

Using username "administrator".
administrator@FAS3050-SVL39's password:
administrator
FAS3050-SVL39> sis on /vol/ev8_arch
SIS for "/vol/ev8_arch" is enabled.
Already existing data could be processed by running "sis start -s /vol/ev8_arch".
FAS3050-SVL39> sis start -s /vol/ev8_arch
The file system will be scanned to process existing data in /vol/ev8_arch.
This operation may initialize related existing metafiles.
Are you sure you want to proceed with scan (y/n)? y
The SIS operation for "/vol/ev8 arch" is started.
FAS3050-SVL39>
```

Figure 6) Enabling FAS dedupe.

As a final comparison, all content from the Enterprise Vault 8 vault store archives was exported back to the corresponding mailboxes. Because shortcuts were not created in the mailboxes, the archives and the mailboxes had identical content. This was completed to serve as an Exchange baseline of identical content. All 859 mailboxes were in the same vault store, with the EDB file located on a dedicated FlexVol volume. The author of this document recognizes that Exchange 2003 uses single instancing within the same message store when mail is delivered by normal mechanisms. Pushing content from the archive into the Exchange mailboxes did not single-instance the mail or attachments. However, messaging architects recognize Exchange 2003 and 2007 single-instance storage as a performance feature, not as a storage optimization consideration.[3][4] Therefore, Exchange single instancing is often completely overshadowed by database white space and other fundamental Exchange characteristics.

8   LAB ENVIRONMENT DETAILS

This section lists the relevant technical details of the lab environment. As noted earlier, a single NetApp FAS3050 held the VMFS volumes, the iSCSI LUNs for the Exchange and SQL Server databases, and the CIFS shares for the Enterprise Vault archives. This particular arrangement would not be recommended for a production environment. Enterprise Vault requires either local disks or virtual local disks for the SQL Server databases and logs, Enterprise Vault, and Microsoft Exchange Server. Either SAN or IP-based SAN satisfies these requirements. This does not preclude having all these volumes on a single storage system. NetApp's multiprotocol block-based and file-based storage means iSCSI, FCP, NFS, and CIFS protocols can be used for access to SATA and FC disks, again on a single storage controller, if desired.

Table 1) Lab environment data storage details

Component	Details
NetApp storage system	FAS3050 (not clustered)
Data ONTAP version	7.3.1
Disk protection	RAID-DP®

Table 2) Lab environment VMware ESX host details.

Component	Details
Model	IBM eserver xSeries 336
Processors	1 Xeon CPU, 3.0 Ghz
Hyperthreading	Active
NIC quantity/speed	4 /1Gb (two active connections)
Internal disk	3 x 36GB
ESX version	ESX 3i, 3.5.0, 123629
RAM	6GB

Table 3) Lab environment virtual guest details.

Component	Details
Memory	4GB
CPU quantity	2 vCPU
NIC quantity/speed	2 x 1Gb

Table 4) Lab environment software details.

Component	Details
Enterprise Vault (before upgrade)	7.5.4.2534
Enterprise Vault (after upgrade)	8.0.0.1405
Enterprise Vault 8.0 hotfix	Enterprise Vault_8.0_Hotfix_Etrack_1499601_319373
E-mail	Microsoft Exchange 2003, 6.5.6944.0
Database	Microsoft SQL Server 2005, 9.00.1399.06, Standard Edition
Operating system (all)	Microsoft Windows 2003, SP2, Standard Edition
Microsoft iSCSI software initiator	2.07
NetApp iSCSI Windows host utilities	4.1.2732.1335
NetApp SnapDrive®	5.0.1

9   RESULTS AND ANALYSIS

There are a number of ways to generate synthetic content for Exchange and subsequently send it to an archive. While these methods may provide critical details on processing rates, IOPs, general throughput, and hardware sizing, they are not as effective in guiding storage architects toward effective storage efficiency using single instancing and deduplication. Synthetic content will not single instance, compress, or deduplicate in the same way that real-world data would. So while the cross-section of production Exchange mail analyzed is unique to this environment, it serves as a real-world case study into the tangible benefits of using Enterprise Vault 8.0 on NetApp storage platforms.

Factors such as attachment types, application versions used to create attachments, sharing boundaries, attachment reuse, numbers of recipients, and the underlying storage system will all introduce tremendous variability in the final Enterprise Vault storage efficiency levels.

In particular, Microsoft Office introduces a number of elements that make it challenging to calculate storage savings. When Office 2003 documents are received as attachments, the fingerprint of the document changes. Outlook modifies the document summary information when the document is inserted into a mailbox. So it is a

completely different fingerprint if originated from the file system. A scenario when production environments would be affected would be if file system archiving and mail archiving were to use the same vault store, now possible with Enterprise Vault 8.0. Office 2003 also modifies the metadata of documents when they are printed; the "printed date" is stored internally. So differences make it impossible to store what appear to be identical documents, as the fingerprints are different. This Outlook issue has been corrected with Office 2007.

Fortunately, there is a higher compression capability with Office 2007 documents. Symantec has noticed a mix of attachments that consist mainly of Office 2003 documents compresses to 60% of its original size. A mix of attachments that consists mainly of Office 2007 documents compresses to 90% of its original size. This was sampled using Enterprise Vault 8.0 for mailbox archiving, with its own compression. When there is more compression at a higher level (for example, application) the block deduplication strategies need to be more sophisticated.

The number of recipients who are copied on the messages also plays an important role in the storage reduction. A custom SQL Server script was used to calculate the SIS ratio, or the number of references to SIS parts divided by the total number of SIS parts. For the imported mail content the ratio was 2.16. A higher number means there is more shared content, and OSIS will play a more significant role in reducing Enterprise Vault 8.0 content, compared to earlier application versions.

Another custom SQL Server script was used to generate the following table, demonstrating the number of shared parts and the percentage by which they were compressed. NULL type extensions are messages or messages with small attachments that are stored as sharable parts.

Additional details on both of these SQL Server scripts are available through your NetApp sales representative.

Table 5) Top 15 SIS object types.

File Type Extension	Total Stored SIS Parts	Compressed Percentage
*.vsd	261	20.16%
*.msg	324	57.83%
*.mht	336	93.63%
*.log	368	93.48%
*.bmp	432	92.13%
*.gif	722	0.00%
*.txt	808	88.65%
*.zip	1,546	0.00%
*.gz	2,737	0.00%
*.ppt	4,509	24.79%
*.jpg	8,816	0.00%
*.pdf	9,792	18.21%
*.xls	10,224	76.87%
*.doc	13,732	51.82%
NULL	142,618	70.05%

Across all attachments, the average file was compressed by Enterprise Vault to 46% of its original size. Historically, all content in Enterprise Vault was compressed without any administrator visibility or control. New to Enterprise Vault 8 is the ability to configure compression through the checkboxes on the volume tab of the partition properties. The first box controls a little more than is indicated by the online documentation. Discussions with Symantec engineers have clarified the actions resulting from selecting the first checkbox: compressed DVS files and uncompressed DVSSP files are created. Enterprise Vault still performs its own single-instance storage, but because the DVSSP files are not compressed, other applications can dedupe with these files. If the second checkbox is selected, no files are compressed at all.
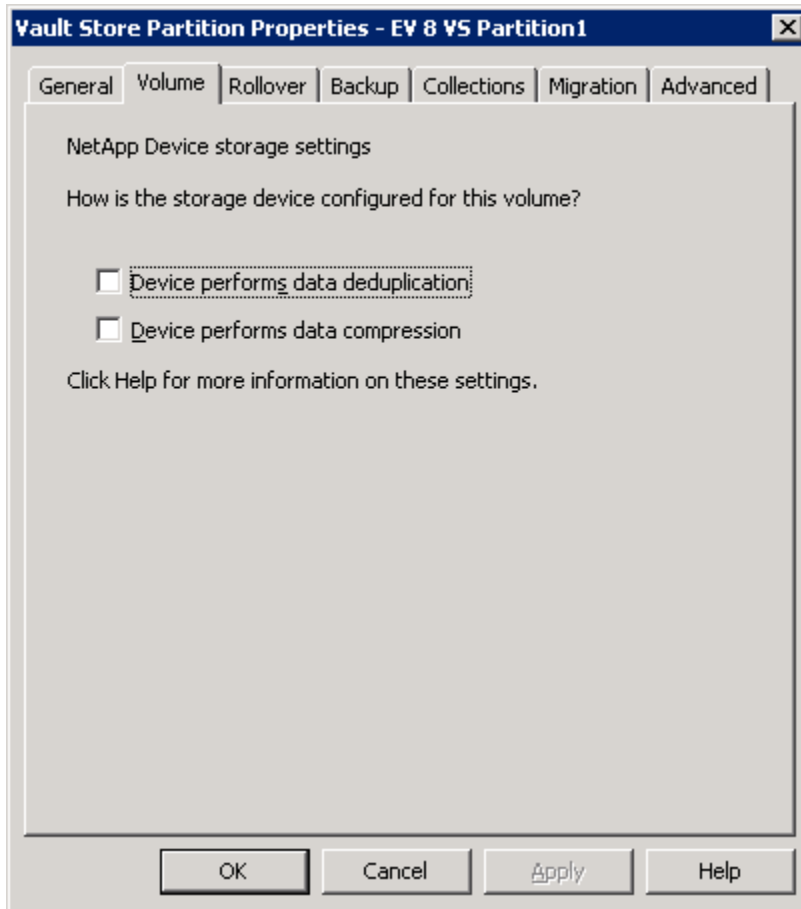

Figure 7) Volume tab of partition properties.

Early tests with the beta version of Enterprise Vault 8.0 indicated maximal storage efficiency was achieved when both checkboxes were left unselected. Even with Enterprise Vault making every attempt to remove redundancy and compress content, there are still opportunities for the underlying storage system to make significant reductions in the storage footprint by performing block deduplication.

Table 6 summarizes the same mail content in each of the four environments. To provide a reliable comparison, great efforts were taken to make sure each location had identical content. The disk space used in the Enterprise Vault 8.0 scenario includes block deduplication. Previous to Enterprise Vault 8.0, the archive deduplication savings were negligible as data was not properly aligned to the beginning of the block on the storage system. All storage sizes in the following table are given where 1kB = 1,024 bytes, 1MB = 1,024kB.

Table 6) Efficiency summary: Enterprise Vault 8.0 on NetApp storage.

Environment	Native Disk Space
PST	102,898MB
Exchange 2003 mail store	80,382MB
Enterprise Vault 2007	50,475MB
Enterprise Vault 8.0	41,347MB

The reduction between the mail data in PST format and the deduplicated, compressed, and single-instance format in Enterprise Vault 8.0 is nearly 60%. Likewise, converting the vault store data from the Enterprise Vault 2007 format to the Enterprise Vault 8.0 format realizes nearly an 18% reduction in the data footprint on disk when using NetApp FAS dedupe. Your particular results may vary according to the unique messaging and user characteristics of your environment. The storage savings in your environment may be greater or lesser than what was measured in our testing.

Following an upgrade to Enterprise Vault 8.0, all archived content is stored according to the optimized single-instance storage paradigm. Customers have two choices about the legacy archives:

- Those who do not require maximal storage savings can perform a simple upgrade and verify that open vault store partitions are located on a NetApp FlexVol volume with dedupe licensed and enabled.

- Archive administrators who require maximal storage efficiency using block-level deduplication and optimized single-instance storage will have to manually export and import all archives following the upgrade. Internal NetApp lab testing has confirmed the archive export and import processes proceed at a rate of 1GB to 2GB per hour on dedicated, not virtualized servers. Third-party solutions can automate such workflows, but the processing rates are estimated to be the same because of Enterprise Vault application internals.[5]

10 CONCLUSION

Enterprise Vault 8.0 on NetApp provides the best solution to help today's companies store and manage content archives. Whether mail is stored in PST files, Microsoft Exchange, or an earlier version of Enterprise Vault, migrating to Enterprise Vault 8.0 provides measurable storage efficiencies to help today's companies do more with less. NetApp is the only vendor with a truly dedupe-capable WORM-compliant storage system. Companies using Enterprise Vault on NetApp storage can expect to reduce their e-mail storage requirements by 50% or more, while enabling satisfaction of regulatory and legal requirements. Our innovative technology continues to meet the market's increasing demand for efficient storage products. We help you take your business further, faster. In a marketplace where customers must choose between performance and efficiency, only solutions based on NetApp can both accelerate business performance and provide outstanding cost efficiency. Our commitment to the success of our customers is found throughout our company and in everything we do.

11 REFERENCES

(1) http://seer.entsupport.symantec.com/docs/312319.htm.

(2) http://www.netapp.com/us/library/technical-reports/tr-3505.html.

(3) Boswell, William. Learning Exchange Server 2003. (Addison-Wesley, 2005), p. 310.

(4) http://support.microsoft.com/kb/198673.

NetApp

www.netapp.com

15   Enterprise Vault 8.0 E-M