

Symantec Enterprise Vault™

Performance Guide

9.0

Last updated: November 23, 2011



Symantec Enterprise Vault™: Performance Guide

The software described in this book is furnished under a license agreement and may be used only in accordance with the terms of the agreement.

Last updated: November 23, 2011.

Legal Notice

Copyright © 2011 Symantec Corporation. All rights reserved.

Symantec, the Symantec Logo, VERITAS, Enterprise Vault, Compliance Accelerator, and Discovery Accelerator are trademarks or registered trademarks of Symantec Corporation or its affiliates in the U.S. and other countries. Other names may be trademarks of their respective owners.

This Symantec product may contain third party software for which Symantec is required to provide attribution to the third party ("Third Party Programs"). Some of the Third Party Programs are available under open source or free software licenses. The License Agreement accompanying the Software does not alter any rights or obligations you may have under those open source or free software licenses. Please see the *Third Party Software* file accompanying this Symantec product for more information on the Third Party Programs.

The product described in this document is distributed under licenses restricting its use, copying, distribution, and decompilation/reverse engineering. No part of this document may be reproduced in any form by any means without prior written authorization of Symantec Corporation and its licensors, if any.

THE DOCUMENTATION IS PROVIDED "AS IS" AND ALL EXPRESS OR IMPLIED CONDITIONS, REPRESENTATIONS AND WARRANTIES, INCLUDING ANY IMPLIED WARRANTY OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE OR NON-INFRINGEMENT, ARE DISCLAIMED, EXCEPT TO THE EXTENT THAT SUCH DISCLAIMERS ARE HELD TO BE LEGALLY INVALID. SYMANTEC CORPORATION SHALL NOT BE LIABLE FOR INCIDENTAL OR CONSEQUENTIAL DAMAGES IN CONNECTION WITH THE FURNISHING PERFORMANCE, OR USE OF THIS DOCUMENTATION. THE INFORMATION CONTAINED IN THIS DOCUMENTATION IS SUBJECT TO CHANGE WITHOUT NOTICE.

The Licensed Software and Documentation are deemed to be commercial computer software as defined in FAR 12.212 and subject to restricted rights as defined in FAR Section 52.227-19 "Commercial Computer Software - Restricted Rights" and DFARS 227.7202, "Rights in Commercial Computer Software or Commercial Computer Software Documentation", as applicable, and any successor regulations. Any use, modification, reproduction release, performance, display or disclosure of the Licensed Software and Documentation by the U.S. Government shall be solely in accordance with the terms of this Agreement.

Symantec Corporation
350 Ellis Street, Mountain View, CA 94043
www.symantec.com

Contents

Chapter 1	Introduction.....	9
Chapter 2	New storage model	11
	Vault Store Groups.....	11
	Single instancing rules.....	11
	Advantages of the new storage model	12
Chapter 3	Hardware	13
	Servers and processors	13
	Recommended hardware	13
	Recommended memory.....	13
	Hyperthreading	14
	64-bit Windows.....	14
	Windows 2008	14
	Storage.....	14
	Enterprise Vault Store.....	14
	Indexes	15
	Local disks	15
	Network	16
Chapter 4	SQL Server	17
	Hyperthreading	18
	Fingerprint database	18
	Accelerator database	18
Chapter 5	Exchange mailbox archiving.....	19
	Choice of CPU	19
	Calculating disk space	20
	Disk space used by vault stores.....	20
	Disk space used by indexes.....	21
	Disk space used by databases.....	22
	Network usage	23
	Communicating with and copying data from the Exchange Servers.....	23
	Communicating with SQL	23

	Writing to the storage medium	23
	Reading and writing indexes.....	24
	Summary.....	24
	Effect on the Exchange Server.....	25
	Sparsely populated mailboxes	25
	Tuning parameters for Exchange mailbox and journaling	25
Chapter 6	Exchange journaling.....	31
	Introduction	31
	Choice of CPU.....	31
	Calculating disk space.....	32
	Disk space used by vault stores	32
	Disk space used by indexes.....	33
	Disk space used by databases.....	34
	Network usage.....	35
	Communicating with and copying data from the Exchange Servers.....	35
	Communicating with SQL.....	35
	Writing to the storage medium	35
	Reading and writing indexes.....	36
	Summary.....	36
	The impact of journal report decryption on journaling	36
	Tuning parameters for Exchange mailbox and journaling	37
Chapter 7	PST migration	39
	Introduction	39
	Choice of CPU.....	39
	Location and collection.....	40
	Increasing the number of concurrent migrations	41
	Calculating disk space.....	41
	Disk space used by vault stores	42
	Disk space used by indexes.....	43
	Disk space used by databases.....	43
Chapter 8	Domino mailbox archiving.....	45
	Choice of CPU.....	45
	Adjusting the number of threads	46
	Calculating disk space.....	47
	Disk space used by vault stores	47
	Disk space used by indexes.....	49
	Disk space used by databases.....	49
	Network traffic.....	50

	Retrieving archived items.....	51
Chapter 9	Domino journaling	53
	Choice of CPU	53
	Number of concurrent connections to the Domino server.....	53
	Calculating disk space	54
	Disk space used by vault stores.....	55
	Disk space used by indexes.....	55
	Disk space used by databases.....	56
	Network traffic	57
Chapter 10	NSF migration	59
	Introduction.....	59
	Choice of CPU	60
	Calculating disk space	60
	Disk space used by vault stores.....	61
	Disk space used by indexes.....	61
	Disk space used by databases.....	62
	Network traffic	63
Chapter 11	File System Archiving.....	65
	Choice of CPU	65
	Calculating disk space	67
	Disk space used by vault stores.....	67
	Disk space used by indexes.....	68
	Disk space used by databases.....	68
	Network usage	69
	Communicating with the file server.....	69
	Communicating with the SQL database.....	70
	Transferring data to the storage medium and retrieving for indexing.....	70
	Reading and writing indexes.....	70
	Summary	70
	File types	71
	File System Archiving and SMTP Archiving	71
	Upgrading FSA metadata after the upgrade to Enterprise Vault 9.0.....	71
Chapter 12	SharePoint.....	73
	Introduction.....	73
	Choice of CPU	73
	Calculating disk space	74
	Disk space used by vault stores.....	74

	Disk space used by indexes.....	75
	Disk space used by databases.....	75
	Network traffic.....	76
	Retrieving items.....	77
Chapter 13	Enterprise Vault Discovery Collector	79
	Choice of CPU.....	80
	Calculating disk space.....	80
	Disk space used by vault stores	81
	Disk space used by indexes.....	81
	Disk space used by databases.....	81
	Network usage.....	82
	Communicating with the collector server.....	83
	Communicating with the SQL database	83
	Transferring data to the storage medium and retrieving for indexing.....	83
	Reading and writing indexes.....	83
	Summary.....	84
Chapter 14	Search and download.....	85
	Types of response time	85
	Viewing items online.....	85
	FSA - opening placeholders.....	86
	Searching indexes.....	86
	Archive Explorer.....	87
Chapter 15	Virtual Vault	89
	Overview	89
	Initial synchronization	90
	Incremental synchronization	91
Chapter 16	Move Archive	93
	Overview	93
	Setting Move Archive parameters	93
	Moving small number of users	94
	Moving large number of users.....	94
	General notes	96
Chapter 17	Archiving to Centera	97
	Archiving with and without Centera collections	97
	Centera sharing model.....	98

Choice of Enterprise Vault server	98
Additional disk space used by databases.....	99
Centera settings	99
Centera limits	99
Self-healing.....	100
NTFS to Centera migration.....	100
Chapter 18 Archiving to a storage device through the Enterprise Vault Storage Streamer API.....	101
Dell DX Object Storage Platform	101
Choice of Enterprise Vault server	102
Additional disk space used by databases.....	102
Chapter 19 Accelerators	103
Accelerator performance	103
Customer's tasks performance	104
Factors influencing search performance	105
Factors influencing export performance.....	111
Customers' service performance	113
Network sockets	114
Discovery Analytics' service performance.....	114
Accelerator database server	116
Chapter 20 Document conversion	121
Converting to HTML or text	121
Excluding files from conversion.....	122
Conversion timeout	123
Chapter 21 Monitoring your system	125
Using Performance Monitor.....	125
Using SQL Performance Monitor counters to measure archiving rate	128
Useful SQL.....	128
Hourly archiving rate	128
Archiving rate/minute	129
Archived file sizes	130
Using Log Parser	131
Log Parser and IIS logs.....	131
Getting index statistics	134
Chapter 22 Miscellaneous.....	137

VMware ESX Server 137

NTFS collections 137

Export to PSTs..... 138

Storage expiry 138

Provisioning 138

Index rebuild 138

Chapter 23 32-bit and 64-bit platforms..... 141

32-bit Windows 141

64-bit Windows: Windows x64 142

64-bit Windows: Windows Itanium 143

Introduction

This document provides guidelines on expected performance when running Symantec Enterprise Vault.

Every customer has different needs and different environments. Many factors influence the performance and needs of an archiving system. These include the type and size of data that is archived, the file types, the distribution lists, and so on. In addition, most systems expect growth in both volume and size of data, and indeed the very existence of an archiving service may lead users to allow their mail or files to be archived rather than delete them. All this leads to the need to be very cautious when sizing a system.

The new storage model introduced in Enterprise Vault 8.0 offers considerable scope for savings in storage.

This guide has a separate section for each of the archiving agents.

New storage model

Enterprise Vault 8.0 introduced a new storage model that allowed more space saving. This section briefly describes the new model. The methods to calculate storage sizes in the agent sections are based on the new model but are simplified to make the calculations easier.

Vault Store Groups

Previous versions of Enterprise Vault shared items within a partition. Enterprise Vault 8.0 onwards may share items within a Vault Store Group, depending on the sharing level specified. The heart of the Vault Store Group is the fingerprint database, which keeps a fingerprint of every item that is shared. Every archived item is fingerprinted for resilience and recovery purposes, but the only fingerprints stored in the database are for item parts that are shareable.

A Vault Store Group may include many partitions and the partitions may be located on different storage devices and even different types of storage device. For performance reasons it is advisable to confine the Vault Store Group to one physical location. A connectivity test is supplied when setting up a partition that checks whether the connection is suitable.

Single instancing rules

Any attachments over 20 KB are detached from the message. These are eligible for sharing. They are compressed and stored as DVSSP files. A mix of attachments that consist mainly of Office 2003 documents compresses to 60% of their original size. A mix of attachments that consists mainly of Office 2007 documents compresses to 90% of their original size.

If the attachments have content that can be indexed, a file that contains the converted content in text or HTML form is created separately as a DVSCC file. Typically, the total size of the converted content files is around 5% of the total size of the DVSSP files. This is because many files contain minimal converted content, such as images, or the text content is only a fraction of the size of the total file.

Messages that are over 20 KB in size after the eligible attachments have been removed also have shareable parts separated as DVSSP files. Any other messages are stored as a single DVS file.

The average size of DVS files is 14 KB and takes up on average 16 KB on disk.

The same rules apply to File System Archiving, but in this case the whole file is generally stored as a shareable item in a DVSSP file and the information specific to that instance of the file is stored as a DVS file. The DVS file is typically 3 KB in size and takes up to 4 KB on disk.

The limit at which items are considered eligible for sharing is termed the “SIS threshold”, and it is currently set at 20 KB. If the value was higher, the level of sharing would be reduced. If the value was lower then there would be more overhead in processing shared parts with little actual gain in space because of the overhead in creating two separate files. The size of the item is calculated by Enterprise Vault and may differ from the size seen in Outlook or other viewers. Messages whose apparent size is smaller than 20 KB may have shareable parts when calculated by Enterprise Vault.

Advantages of the new storage model

Some of the advantages of the new storage model are as follows:

- Shareable parts are single-instanced across partitions.
- Shareable parts are single-instanced between mailbox and journal archives.
- Items archived by different agents are single-instanced. For example, this is the case with files archived by FSA and attachments to Exchange messages.
- Stored items are not updated once they have been created. Items on WORM devices will be shared.
- There is less network traffic to the storage devices because items are not updated and because there are separate converted content files.
- New partitions or Vault Stores may be created without loss of sharing.

Hardware

The most critical factor in the performance of Enterprise Vault is the specification of the system that is used—the servers, the disk systems, the memory, and the network.

Servers and processors

Recommended hardware

In general, the more powerful the processor, the higher the ingest and retrieval rates. The other components of the system—the disks, memory, network, and archiving source—need to match this power.

Enterprise Vault makes good use of multi-core processors, and quad-core processors are recommended.

Recommended memory

4 GB of memory is recommended. Future versions of Enterprise Vault may require more memory to allow improvements in indexing and search. The Enterprise Vault servers should be capable of easy upgrades to memory.

The operating system boot flag /3GB must not be used as this does not provide any benefit and can result in running out of system page table entries. See “32-bit and 64-bit platforms” on page 141 for a fuller explanation.

Note: The use of Windows x64 enables more than 4 GB of physical memory to be provided. This allows the system file cache to grow and improves the I/O to index files. Increased memory should be used when many concurrent searches are expected.

Hyperthreading

Turning hyperthreading on or off makes no noticeable difference to the overall performance of Enterprise Vault, so we recommend that the manufacturer's default setting is not changed.

64-bit Windows

There is no performance difference when running Enterprise Vault on 64-bit Windows (WOW 64) when compared with 32-bit Windows. See "32-bit and 64-bit platforms" on page 141 for a fuller explanation.

Windows 2008

There is no performance difference when running Enterprise Vault on a Windows 2008 operating system when compared with Windows 2003 or when archiving from an Exchange Server on a Windows 2008 system.

Windows 2008 introduces version 2 of the Server Message Block (SMB) protocol. The practical effect is that files are copied faster between Windows 2008 servers. The transfer speed is not generally an issue when ingesting data, but retrievals may be faster. This will be especially apparent during bulk retrievals where transfer rates may be up to 50% faster between Windows 2008 servers.

Storage

Enterprise Vault Store

One of the purposes of Enterprise Vault is to allow cheaper storage to be used to archive data. The primary requisite is that the archived data is secure and retrievable.

In terms of storage cost savings, there is most benefit in keeping archived data on cheaper network-attached storage (NAS). However, you can also make some savings when keeping archived data on more expensive storage, such as a Storage Area Network (SAN), due to the additional compression and single-instance storage that Enterprise Vault provides.

Most NAS devices or Centera offer quick archiving and retrieval while providing space, reliability, and security for archived data. Storage systems from most of the major vendors have been tested for performance and found to be suitable for fast bulk storage and retrieval of data.

Some vendors offer storage solutions based on optical disk or tape using a disk cache for recently stored and retrieved items. Where certified, these have a

minimum ingest rate of 20,000 items an hour and acceptable retrieval rates for recently stored items. However, they are not suitable for bulk retrieval of archived items (for example, when exporting an archive to PST or rebuilding an index).

Some storage devices have a limit on the number of files that may be held in folder before performance degrades. Enterprise Vault 8.0 has addressed this by holding fewer files in each folder, and this is no longer an issue.

Some storage vendors offer devices with block-level deduplication. Many of these vendors have tested their devices with Enterprise Vault and have recommendations on the best way to save storage.

Indexes

The storage required for AltaVista indexes depends on how they are used. If fast concurrent searches are required because Enterprise Vault Discovery Accelerator or Compliance Accelerator products are used, fast storage for the indexes is needed—for example, a SAN or direct attached storage (DAS). On the other hand, if users are prepared to wait for searches then you can use slower systems or NAS.

Note the following about indexes:

- On NAS devices, you must turn off opportunistic locking.
- Indexes become fragmented whatever the type of device, and this slows down both searching and indexing. You must regularly defragment indexes, ideally while the indexing service is stopped so that defragmentation does not conflict with updates. This is very important if you are using the Accelerator products. See “Accelerators” on page 103 for guidelines on how to tune the Accelerator products.

Local disks

Archiving generates I/Os on local disks. The primary causes of these are temporary files used when archiving and conversion, and I/Os generated by MSMQ. To isolate the I/Os that MSMQ causes, place the MSMQ files on a local disk separate from the system disk. The disk does not have to be large. MSMQ is used during Exchange Archiving and Journaling but not for File System Archiving, Domino Journaling and Domino Mailbox archiving, PST Migration, or SMTP Archiving.

Blade servers generally have slow local disks that are not suitable for high I/O loads. If possible, place the MSMQ files on a fast external disk and do not use the internal disks for vault stores or indexes.

Network

It is rare that the network is the limiting factor on performance except when some component of the system is on a wide-area network. For example, there may be a remote Exchange Server or directory database. 100-Base-T is normally sufficient, but see also the sections on network usage for the various archiving agents to calculate what bandwidth is required.

One result of the new storage model in Enterprise Vault 8.0 is that the network traffic between the Enterprise Vault server and storage devices holding the partitions is reduced by approximately 50%. Items are no longer retrieved and rewritten when new sharers are added. Indexing only retrieves the converted content for attachments that have been separated out and not the item itself.

SQL Server

The SQL Server is the heart of the Enterprise Vault system and needs to be properly specified.

A new fingerprint database is used in Enterprise Vault 8.0 to support the new storage model. The new database has tables with entries for every SIS part. This does not require a separate server.

For best performance, a standard SQL Server with four cores and a minimum of 4 GB of RAM is the recommended specification. For medium or large environments, 8 GB of RAM is recommended. You should also tune the performance of the SQL Server using standard methods, such as the following:

- Using an x64-based 64-bit platform provides more efficient memory utilization and brings performance benefits. The appropriate edition of Windows Server 2003 and SQL Server 2005 must be installed to support the capacity of memory installed, but no other tuning options need be set. SQL Server 2008 offers equivalent performance to SQL Server 2003.
- If a 32-bit database server will be used then the server should be carefully tuned to make best use of available memory. These tuning options depend upon using the appropriate edition of Windows and SQL Server for the installed capacity of memory.

If the database server has more than 4GB of physical RAM:

- Enable the operating system Physical Address Extensions boot flag (/PAE).
- Enable Address Windowing Extensions (AWE) memory in SQL Server using the following script:

```
sp_configure 'show advanced options', 1
RECONFIGURE
GO
sp_configure 'awe enabled', 1
RECONFIGURE
GO
```

SQL databases need maintenance. As part of the maintenance plan, we suggest that you take the following actions weekly:

- Rebuild indexes.
- Update statistics.

Shrink databases only when it is necessary to reclaim space on the disk, and always allow 10% free space to allow for future growth.

Many factors influence the number of Enterprise Vault servers that one SQL Server can support. The following table shows the recommended number of Enterprise Vault servers that one SQL Server supports.

Enterprise Vault servers per SQL Server	Recommended in this case
4	There is no maintenance plan, and the SQL Server has a similar specification to the Enterprise Vault servers.
8	The databases are regularly maintained, or the SQL Server has 8 GB of memory.

Hyperthreading

Hyperthreading may not be beneficial to SQL Server environments and should be carefully tested before enabling.

Fingerprint database

The fingerprint database is used by all the Vault Stores within a Vault Store Group that are participating in sharing. The database is divided into separate files allowing the I/O rate to be spread by placing them on different disks.

Accelerator database

The Accelerator database server must be well specified and should have a minimum of four cores and 4 GB of RAM. (At least 8 GB is recommended, and using x64 architecture can improve the efficiency of the memory utilization.) Split the database log and data files over separate very high speed striped arrays (RAID 10 rather than RAID 5) containing multiple high speed disks and a large array controller cache (512 MB controller cache is recommended for the SQL Server). You can potentially place the log and data files on different SAN partitions. However, the performance has not been measured.

Exchange mailbox archiving

In most cases, when you are choosing servers for email archiving, the most critical factor is the ingest rate. For email archiving, there is normally a limited window during the night to archive, and everything must be archived during this period. This window fits in between daily Exchange usage, backups, defragmentation, and all the other background activities that take place on the server.

The archiving window can be extended by archiving during the day and weekends. It is suggested that this is done during quiet user times such as lunch times or early evening.

The following figures apply to Exchange 2003, Exchange 2007, and Exchange 2010.

It is a prerequisite of Enterprise Vault that a version of either Outlook 2003 or Outlook 2007 is installed on the Enterprise Vault servers. The archiving task can retrieve items from an Exchange Server more quickly when Outlook 2003 is installed. If Outlook 2003 is already installed on the Enterprise Vault servers, we recommend that you do not upgrade it. However, if you have installed Outlook 2007, you can achieve the ingest rates described below when ingest is from more than one Exchange Server at a time.

Enterprise Vault 9.0 allows archiving from Exchange 2010 servers. Throughput is best when the recommended ratio of Mailbox to Client Access servers is maintained. The current recommendation is 4:3 (Mailbox:Client Access). For more information, see the Microsoft Exchange TechCenter at <http://technet.microsoft.com/en-gb/library/bb124558.aspx>. If the number of Client Access Servers is less than the recommended number, the ingest rate is affected.

Choice of CPU

The following rates are conservative and should easily be achieved with a properly specified and tuned system.

The following table shows the expected ingest rate for numbers of physical cores where the average message size including attachments is 70 KB.

Number of cores	Hourly ingest rate (70 KB)
2	25,000
4	40,000
8	60,000

The average size of mail messages has an effect on the throughput. The observed effect is that when the average message size is doubled, throughput is reduced by one third.

Calculating disk space

This section deals with the space used and how to size for it. When archiving, Enterprise Vault uses three areas of permanent space:

- The Vault Store partition, which is used to hold the DVS (saveset), DVSSP (saveset shared part) and DVSCC (saveset converted content) files. If collections are enabled, they are stored as CAB files. If Centera is the storage medium, it stores the files in its own format.
- The index area. The current index technology uses AltaVista indexes. Each index is stored in a separate folder and consists of up to 285 files.
- The SQL database, which is used to hold the Directory and Vault Store and fingerprint databases.

Disk space used by vault stores

The Single Instance model has changed in Enterprise Vault 8.0. The principal changes are as follows:

- Items are shared within a Vault Store group. A Vault Store group may contain many Vault Stores and partitions. The partitions may be on different device types, but note that items on Centera are not shared with other devices.
- Shareable parts of a message that exceed the SIS threshold of 20 KB are shared. This includes attachments and message bodies. User information and shareable parts below the SIS threshold are not shared.

The following gives some general rules for estimating the amount of space used. This is a simple calculation that does not take into account some of the complexities.

1. Multiply the number of items to be archived by 16 KB to get the total size of the DVS files. Count all messages, including those with attachments. These are the files that are not shared.
2. Take 60% of the size of attachments (or 90% where the mix of attachments is mostly Office 2007 documents). This is the size of attachments after compression. A rule of thumb is that 20% of files have attachments, and the average attachment size is 250 KB.
3. Divide by the number of sharers of each attachment across the Vault Store Group. This is the size of the DVSSP files after sharing.
4. Take 5% of the size of DVSSP files. This is the size of the DVSCC files.

The total space used is the sum of the DVS, DVSSP and DVSCC files.

If items in the mailboxes have already been journaled, and the journal and mailbox partitions participate in sharing within a Vault Store Group, the shared parts have already been stored and will not be stored again. The only additional space is that used to store the DVS files. There are some exceptions to this rule, and some extra DVSSP files may be created.

If items are archived to more than one partition, more shared parts will be stored on the partition where the archiving task runs first. Some partitions may grow faster than others.

Note: These recommendations do not apply to Centera, which uses a completely different sharing model. See “Archiving to Centera” on page 97 for more details.

Disk space used by indexes

Calculate the expected index size as follows:

1. Take the size of the original data.
2. Take a percentage of this according to the indexing type.

Indexing type	Percentage
Brief	3%
Medium	8%
Full	12%

The percentages for Medium and Full will be less if there is little indexable content. This is often the case where there are large attachments such as MP3 or JPEG files.

Disk space used by databases

Metadata is added to the database for every item archived. Temporary space is used to hold information on items that have not been backed up or indexed. Certain records are held for longer for users who have been enabled for Vault Cache or Virtual Vault. Permanent space is also used to hold data in the Directory database. We suggest that an extra 4 GB is allowed for this, or 8 GB where millions of items are archived between backups or Vault Cache is enabled for thousands of users. The extra space is added once only.

Future versions of Enterprise Vault may include more information in the database, and the following calculations allow room for expansion.

To calculate the space for Exchange Mailbox and Journal archiving:

1. Take the number of items archived.
2. Multiply by 500 bytes.
3. Add 4 GB.

A new database is introduced into Enterprise Vault 8.0 to hold fingerprints of the shareable parts of archived items. There is one entry for every shareable part that creates a separate DVSSP file. The number of shareable parts depends on the number of attachments or messages that exceed the SIS threshold and the number of times each is shared. A rule of thumb is that 20% of messages participate in sharing.

To calculate the space used in the fingerprint database:

1. Take the number of items archived.
2. Divide by 5.
3. Divide by the average number of sharers of each shareable item (2 if unknown).
4. Multiply by 500 bytes.

There are no additional entries in the database if mailbox archiving takes place after the items have been journaled within the same Vault Store Group, if the mailbox and journal partitions participate in sharing.

There are no specific limits on the number of items that you can store in one Vault Store database, but we recommend that a Vault Store does not contain more than 100 million items. This makes it easier to perform database maintenance activities such as rebuilding indexes and backing up databases.

There are no specific limits on the number of items that you can store in one fingerprint database. The fingerprint database is designed to allow growth and scalability by scaling out across disks and servers. Future versions of this guide will provide more information on this.

These calculations do not take into account the space required by the SQL log files or space used if regular housekeeping is not run on the databases.

Network usage

The network is used for the following purposes while ingesting items from Exchange user or journal mailboxes:

- Communicating with and copying data from the Exchange Servers.
- Accessing the SQL database.
- Transferring archived data to the storage medium (for example, NAS or Centera).
- Retrieving archived data from the storage medium for indexing.
- Reading and writing data to and from the index storage medium.
- Background activity, such as communication with the domain controller, user monitoring, and so on.

Communicating with and copying data from the Exchange Servers

Assume that the network traffic between the Exchange Server and the Enterprise Vault server is double the total size of the documents transferred.

Communicating with SQL

A rule of thumb is that 160 kilobits of total data is transferred between the SQL Server and the Enterprise Vault server for every item archived. If the Directory database is on a different server, 40 kilobits of this is transferred to the Directory database. More data is transferred to and from the Directory database when empty or sparsely populated mailboxes are archived or when mailboxes have many folders.

Writing to the storage medium

There is a reduction in the network traffic between the Enterprise Vault server and the storage media when compared with versions before Enterprise Vault 8.0. Data is written in compressed form as DVS, DVSSP and DVSCC files. When a new sharer is added to a DVSSP file, the DVSSP file and its corresponding DVSCC file are not

retrieved or rewritten. Items are read back for indexing, but where a DVSSP file has a DVSCC file, only the smaller DVSCC file is retrieved.

When Centera is the storage medium, items are not read back for single instancing. If Centera collections are enabled, indexable items may be read back from local disk rather than Centera.

Reading and writing indexes

When an index is opened, some of the index files are transferred to memory in the Enterprise Vault server. On completion of indexing, the files are written back. Sometimes the files are written back during indexing. The amount of data transferred depends on the number of indexes opened and the size of those indexes. For example, if only one or two items are archived from each mailbox, indexes are constantly opened and closed and a lot of data is transferred, especially if the indexes are large. It is therefore difficult to predict the traffic to the Index server. A rule of thumb is that the network traffic between the Index location and the Enterprise Vault server is equal to the size of the original item for every item indexed.

Summary

The following table shows the expected kilobits per second (kbps) when archiving messages of 70 KB.

In accordance with normal usage, network traffic is expressed in bits per second rather than bytes per second.

	Hourly ingest rate		
	25,000	40,000	60,000
Enterprise Vault server ↔ Exchange Server	8,000	12,500	20,000
Enterprise Vault server ↔ SQL Server (Vault Store)	1,250	1,800	2,700
Enterprise Vault server ↔ SQL Server (Directory)	560	900	1350
Enterprise Vault server ↔ SQL Server (Fingerprint)	70	110	160
Enterprise Vault server ↔ Storage medium	3,000	5,000	7,000
Enterprise Vault server ↔ Index location	4,000	6,250	10,000

Effect on the Exchange Server

Most of the time, mailbox archiving is done while mailbox users are not active. There may be occasions when mailbox archiving is needed at the same time as normal Exchange usage. This could be planned to deal with a backlog of archiving or because there is a need to archive during the day.

Enterprise Vault does not take precedence over the active mailbox users. It is not possible to extract items from an Exchange Server at the same rate when other mailbox users are active. This is generally good because archiving has less of an impact on active users. If you want to increase the archiving rate at the expense of users' response times or decrease the archiving rate, adjust the number of concurrent connections to the Exchange Server used by the archiving process. This is a property of the archiving task.

The effect on the Exchange Server can be seen in increased CPU usage and IO per second, and in longer response times. This is discussed in the Symantec Yellow Book on Enterprise Messaging Management for Microsoft Exchange (<http://www.symantec.com/enterprise/yellowbooks/index.jsp>).

The principal effect on Exchange Server is on the storage system. Any effect on active users while archiving is closely related to how well specified that system is.

There are fewer I/Os on Exchange 2007 running on 64-bit, and consequently less of an impact on users of that system. Items can be extracted for archiving faster from an Exchange 2007 system.

Sparsely populated mailboxes

There is an effect on the ingest rate when mailboxes are eligible for archiving but contain few archivable items. It takes about one hour to pass through 10,000 mailboxes without archiving anything.

Tuning parameters for Exchange mailbox and journaling

The rate at which items are archived depends mainly on the specification of the system; that is, the processing power, the IO capacity, and the network. There are some parameters that can be changed for particular problems. It is not suggested that any of the following are used as a matter of course.

Message queues

The main processes when ingesting items are as follows:

- Extracting mail from Exchange.
- Inserting items into the archives.

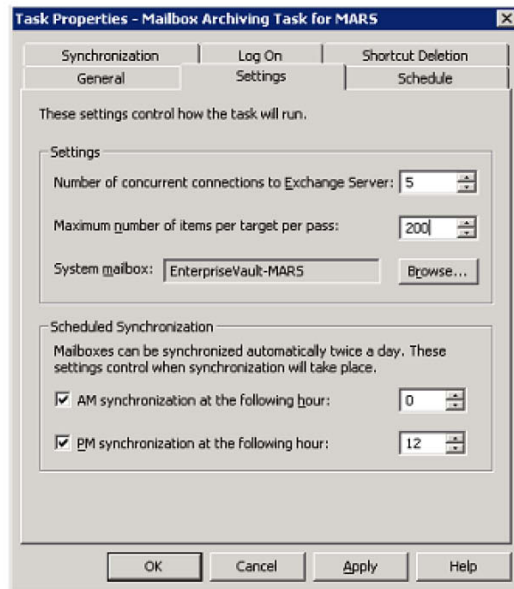
Items get passed between these processes by message queues. The queues used by mailbox archiving are:

- A1 - Changing pending item to shortcut and errors.
- A2 - Manual archive requests and error handling/retry from A3 queue.
- A3 - “Run Now” requests made from the administration console – one entry for each mailbox to be archived.
- A4 - Error handling/retry from A5 queue.
- A5 - Scheduled archive – one entry for each mailbox to be archived.
- A6 - Requests to update folders with items that have moved within a mailbox.
- A7 - Mailbox synchronization request.
- Storage Archive – Store item in an archive.

The queues used by Journal Archiving are:

- J1 - Delete message (after successful archive) or change pending item back (on error).
- J2 - Items to process.
- J3 - Mailbox to process.
- J4 - Synchronize mailboxes
- Storage Archive – Store item in an archive

Some of these queues can be used to see whether to adjust the number of concurrent connections to the Exchange Server or the number of storage archive processes.

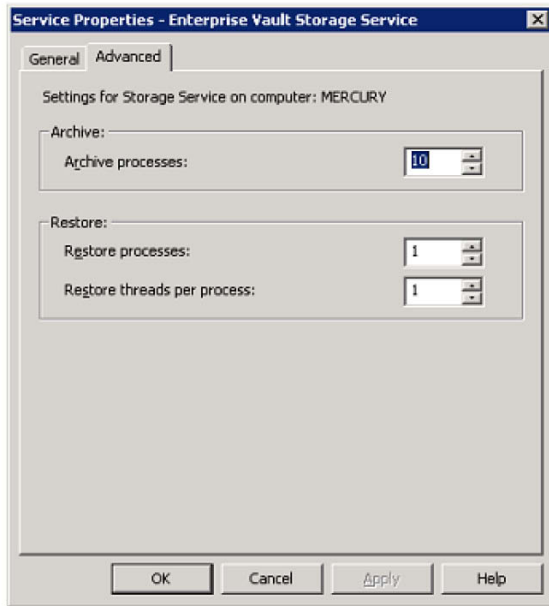


Setting the number of storage archive processes

The default number of archive processes is five. On more powerful servers it can be advantageous to increase this number. For a server with four or more cores, it is usually beneficial to increase this to 10. The indications that the archiving rate could benefit from increasing this value are as follows:

- The number of items on the MSMQ queue "enterprise vault storage archive" constantly exceeds the threshold value of 1,500.
- While archiving, the total CPU used by the system averages less than 80%.

The effect of increasing this value is that more processes are available for processing the items on the queue. This leads to more context switching and memory use and may not always be beneficial, especially on less powerful servers.



Storage queue threshold limits

When archiving, the Archiving Service takes items from the Exchange Server and puts them onto the MSMQ queue "enterprise vault storage archive". To prevent an excessive number of items from being put on the queue, a limit is set to the number of items. When this limit is exceeded, the archiving tasks pause to give the Storage Archive processes time to clear the queue. By default, this limit is set at 1,500. The number of items may exceed this because the queue is not checked constantly, so a second higher limit of 4,500 is also set.

By default, when the lower limit is exceeded, the archiving tasks pause for one minute. When the upper limit is exceeded, the archiving tasks pause for five minutes. On fast systems, the queues may be cleared in less time with the result that the system remains idle for a few seconds until the minute elapses. The storage processes run faster during this time because they are not competing with the idle archiving tasks.

The number of items on the Storage Archive Queue can be monitored using Performance Monitor.

"MSMQ Queue(evserver\private\$\enterprise vault storage archive)\Messages in Queue"

If you see this behavior, raise the following registry values.

Value	Key	Content
ThrottleLowerThreshold	HKEY_LOCAL_MACHINE \SOFTWARE \KVS \Enterprise Vault \Storage	DWORD value set to an integer value. Default is 1500. Raise this to 5000.
ThrottleUpperThreshold	HKEY_LOCAL_MACHINE \SOFTWARE \KVS \Enterprise Vault \Storage	DWORD value set to an integer value. Default is 4500. Raise this to 10000.

Setting the number of connections to the Exchange Server

If you want items to be extracted at a faster rate, you can increase the number of concurrent connections to the Exchange Servers. The indications to do this are as follows:

- You are achieving less than the required archiving rate.
- The Storage Archive queue frequently dips to zero.
- For Journal archiving, the J2 queue frequently dips to zero.

If this is the case, increase the number of concurrent connections to 10 from the default of 5. This is most beneficial when archiving from Exchange Server 2007 on 64-bit.

You can also reduce the number to minimize the impact of archiving on Exchange.

Changing the distribution list cache size

Enterprise Vault caches distribution lists and holds up to 50 distribution lists in cache. Large organizations are likely to have more than 50 lists. To keep distribution lists in cache longer, the following registry value can be changed. This has an impact on the process’s memory use.

Value	Key	Content
DLCacheSize	HKEY_LOCAL_MACHINE \SOFTWARE \KVS \Enterprise Vault \Agents	DWORD value set to an integer value. Default is 50 Raise this to 400.

Exchange journaling

Introduction

Exchange Journal archiving and Exchange mailbox archiving act in the same way, and for the most part the same factors that influence the performance of mailbox archiving also influence the performance of journal archiving. There are some differences that make journaling more efficient than Exchange Mail Archiving. These stem from the fact that only a small number of mailboxes are archived to only a small number of archives.

The main differences are as follows:

- There are fewer connections created and deleted to the Exchange Server, and Enterprise Vault archives from one folder only. This leads to more efficient use of Exchange.
- There are fewer calls to the Directory database as permissions are checked less frequently for mailboxes and folders.
- There are fewer indexes opened.

Enterprise Vault 9.0 allows archiving from Exchange 2010 servers. Throughput is best when the recommended ratio of Mailbox to Client Access servers is maintained. The current recommendation is 4:3 (Mailbox:Client Access). For more information, see the Microsoft Exchange TechCenter at <http://technet.microsoft.com/en-gb/library/bb124558.aspx>. If the number of Client Access Servers is less than the recommended number, the ingest rate is affected.

Choice of CPU

The following rates are conservative and should easily be achieved with a properly specified and tuned system.

The following table shows the expected ingest rate for numbers of physical cores where the average message size including attachments is 70 KB.

Number of cores	Hourly ingest rate (70 KB)
2	25,000
4	40,000
8	60,000

The average size of mail messages has an effect on the throughput. The observed effect is that when the average message size is doubled, throughput is reduced by one third.

The following figures apply to Exchange 2003 and Exchange 2007. The archiving task can retrieve messages more quickly from Exchange 2007 servers. This is primarily of benefit where ingest is from a single Exchange Server whose specification is less than the Enterprise Vault server.

Calculating disk space

This section deals with the space used and how to size for it. When archiving, Enterprise Vault uses three areas of permanent space:

- The Vault Store partition, which is used to hold the DVS (saveset), DVSSP (saveset shared part) and DVSCC (saveset converted content) files. If collections are enabled, they are stored as CAB files. If Centera is the storage medium, it stores the files in its own format.
- The index area. The current index technology uses AltaVista indexes. Each index is stored in a separate folder and consists of up to 285 files.
- The SQL database, which is used to hold the Directory, Vault Store and Fingerprint databases.

Disk space used by vault stores

The Single Instance model has changed in Enterprise Vault 8.0. The principal changes are as follows:

- Items are shared within a Vault Store group. A Vault Store group may contain many Vault Stores and partitions. The partitions may be on different device types, but note that items on Centera are not shared with other devices.
- Shareable parts of a message that exceed the SIS threshold of 20 KB are shared. This includes attachments and message bodies. User information and shareable parts below the SIS threshold are not shared.

The following gives some general rules for estimating the amount of space used. This is a simple calculation that does not take into account some of the complexities.

1. Multiply the number of items to be archived by 16 KB to get the total size of the DVS files. You need to count all messages, including those with attachments. These are the files that are not shared.
 2. Take 60% of the size of attachments (or 90% where the mix of attachments is mostly Office 2007 documents). This is the size of attachments after compression. A rule of thumb is that 20% of files have attachments and the average attachment size is 250 KB.
 3. Divide by the number of sharers of each attachment across the Vault Store Group. As a rule of thumb, each attachment is shared between 1.5 and 3 times. If there is more than one journal mailbox, there is a fan-out effect.
 - For one journal mailbox, fan-out = 1.00
 - For two journal mailboxes, fan-out = 1.75
 - For three journal mailboxes, fan-out = 2.11
 - For four journal mailboxes, fan-out = 2.3

You need to divide by this factor to get the total number of sharers across all the journal mailboxes that participate in sharing within the Vault Store Group.

This is the size of the DVSSP files after sharing.
 4. Take 5% of the size of DVSSP files. This is the size of the DVSCC files.
- The total space used is the sum of the DVS, DVSSP and DVSCC files.

Note: These recommendations do not apply to Centera, which uses a completely different sharing model. See “Archiving to Centera” on page 97 for more details.

Disk space used by indexes

Calculate the expected index size as follows.

1. Take the size of the original data.
2. Take a percentage of this according to the indexing type.

Indexing type	Percentage
Brief	3%
Medium	8%
Full	12%

The percentages for Medium and Full will be less if there is little indexable content. This is often the case where there are large attachments such as MP3 or JPEG files.

Disk space used by databases

Metadata is added to the database for every item archived. Temporary space is used to hold information on items that have not been backed up or indexed. Permanent space is also used to hold data in the Directory database. We suggest that an extra 4 GB is allowed for this, or 8 GB where millions of items are archived between Vault Store backups. The extra space is added once only.

Future versions of Enterprise Vault may include more information in the database and the following calculations allow room for expansion.

To calculate the space for Exchange Mailbox and Journal archiving, proceed as follows:

1. Take the number of items archived.
2. Multiply by 500 bytes.
3. Add 4 GB.

A new database is introduced into Enterprise Vault 8.0 to hold fingerprints of the shareable parts of archived items. There is one entry for every shareable part that creates a separate DVSSP file. The number of shareable parts depends on the number of attachments or messages that exceed the SIS threshold and the number of times each is shared. A rule of thumb is that 20% of messages participate in sharing.

To calculate the space used in the fingerprint database:

1. Take the number of items archived.
2. Divide by 5.
3. Divide by the average number of sharers of each shareable item (2 if unknown).
4. Multiply by 500 bytes.

There are no specific limits on the number of items that you can store in one Vault Store database, but we recommend that a Vault Store does not contain more than 100 million items. This makes it easier to perform database maintenance activities such as rebuilding indexes and backing up databases. The new storage model allows vault store databases to be rolled over regularly without loss of sharing.

There are no specific limits on the number of items that you can store in one fingerprint database. The fingerprint database is designed to allow growth and scalability by scaling out across disks and servers. Future versions of this guide will provide more information on this.

These calculations do not take into account the space required by the SQL log files or space used if regular housekeeping is not run on the databases.

Network usage

The network is used for the following purposes while ingesting items from Exchange user or journal mailboxes:

- Communicating with and copying data from the Exchange Servers.
- Accessing the SQL database.
- Transferring archived data to the storage medium.
- Retrieving archived data from the storage medium for indexing.
- Reading and writing data to and from the index storage medium.
- Background activity, such as communication with the domain controller, user monitoring, and so on.

Communicating with and copying data from the Exchange Servers

Assume that the network traffic between the Exchange Server and the Enterprise Vault server is equal to two times the total size of the documents transferred.

Communicating with SQL

A rule of thumb is that 140 kilobits of total data is transferred between the SQL Server and the Enterprise Vault server for every item archived. If the Directory database is on a different server, 20 kilobits of this is transferred to the Directory database.

Writing to the storage medium

There is a reduction in the network traffic between the Enterprise Vault server and the storage media when compared with previous versions. Data is written in compressed form as DVS, DVSSP and DVSCC files. When a new sharer is added to a DVSSP file, the DVSSP file and its corresponding DVSCC file are not retrieved or rewritten. Items are read back for indexing, but where a DVSSP file has a DVSCC file, only the smaller DVSCC file is retrieved.

When Centera is the storage medium, items are not read back for single instancing, and if Centera collections are enabled, indexable items may be read back from local disk rather than Centera.

Reading and writing indexes

When an index is opened, some of the index files are transferred to the Enterprise Vault server. On completion of indexing, the files are written back. Sometimes the files are written back during indexing. The amount of data transferred depends on the number of indexes opened and the size of those indexes. For example, if only one or two items are archived from each mailbox, indexes are constantly opened and closed and a lot of data is transferred, especially if the indexes are large. It is therefore difficult to predict the traffic to the Index server. A rule of thumb is that the network traffic between the Index location and the Enterprise Vault server is equal to the size of the original item for every item indexed.

Summary

The following table shows the expected kbps (kilobits per second) when archiving messages of 70 KB. Figures are rounded up and provide a rough guide only.

	Hourly ingest rate		
	25,000	40,000	60,000
Enterprise Vault server ↔ Exchange Server	8,000	12,500	20,000
Enterprise Vault server ↔ SQL Server (Vault Store)	1,250	1,800	2,700
Enterprise Vault server ↔ SQL Server (Directory)	280	450	700
Enterprise Vault server ↔ SQL Server (Fingerprint)	70	110	160
Enterprise Vault server ↔ Storage medium	3,000	5,000	7,000
Enterprise Vault server ↔ Index location	4,000	6,250	10,000

The impact of journal report decryption on journaling

If journal report decryption is configured on Exchange Server 2010, two messages are attached to the journal report: the original RMS-protected message, and a clear text version. A policy setting controls whether Enterprise Vault uses the clear text message or the RMS-protected message as the primary message during archiving. Enterprise Vault stores both versions of the message in the message saveset, whatever the policy setting.

Journal report decryption has an effect on the size of data archived and the rate at which items are ingested. There are two factors at work:

- The size of the journal report held in the Exchange database is doubled because two copies of the message are held.
- There is some loss of sharing either of the clear text or RMS-protected version of the message within Enterprise Vault, depending on the policy.

The following table gives guidelines on the increase in space used within the vault store partitions when clear text is used as the primary or secondary message during archiving.

Clear text policy setting	Increase in partition storage space
Clear text primary	+190%
Clear text secondary	+160%

When journal report decryption is enabled, the ingest rate falls by an average of 15%.

Tuning parameters for Exchange mailbox and journaling

See the relevant section under Exchange mailbox archiving.

PST migration

Introduction

In general, the rate at which items are migrated into Enterprise Vault from PSTs is the same or faster than the rate at which items are archived from Exchange. Large-scale migrations need careful planning because they can take several months or longer and may compete for resources with daily archiving and journaling. It is the planning that is all-important, and it is imperative not to be too ambitious for large-scale PST migrations. You must allow for downtime and for resources to be used by other processes. It is important to ensure that PSTs are fed into the migration process fast enough.

Choice of CPU

There are several methods to migrate PSTs. They all have the same performance characteristics when migrating, except for client-driven migration.

- Wizard-assisted migration.

PSTs can be migrated into user archives using the Enterprise Vault PST Migration Wizard. This is a quick way of importing a few PSTs, but it is not multi-processed. This means that it is not a suitable method for importing a large number of PSTs in a short time.

It is possible to run several wizards simultaneously, but this is awkward to set up.

- Scripted migration.

PSTs can be migrated using EVPM scripts, allowing flexibility over how each PST is processed. This automatically creates five processes when migrating.

- Locate and Migrate.

In this model, PSTs are located and collected together before migration. It does the work of discovering PSTs on the network before collecting them together into a central area for migration.

- Client-driven migration.
This is migration that the Enterprise Vault client add-in initiates. This is useful for PSTs on notebooks where the notebook is only intermittently connected to the network. This is the slowest method, migrating about 2000 items an hour, but it has little impact on the client or server. Because of the low load on the server, there can be many clients migrating in parallel with a total throughput equal to the Locate and Migrate method.

The following table shows the archiving rates per hour of items of an average size of 70 KB for the different migration methods. The throughput figures are when shortcuts are inserted into Exchange mailboxes or PSTs.

Number of cores	Wizard-assisted (single process)	Locate and Migrate/Scripted
2	15,000	25,000
4	15,000	40,000
8	15,000	60,000

Location and collection

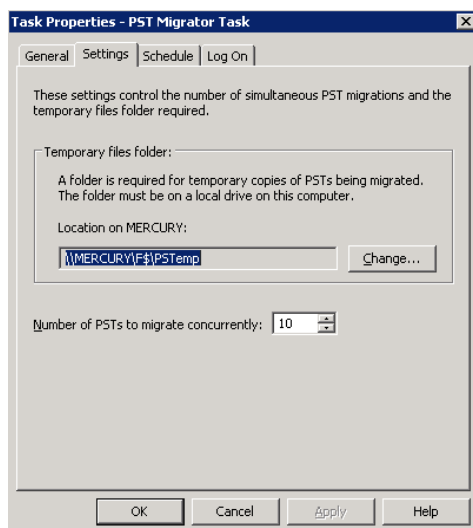
In the Locate and Migrate method, the PSTs are located before migration. The time to locate the PSTs is not predictable and may require many domains and servers to be searched. The following table shows one example.

Number of servers	Number of PSTs located	Elapsed time
40	9000	30 minutes

By default, the PST Collection task copies 100 PSTs to a holding area. As PSTs are migrated, more PSTs are collected to keep the number of PSTs in the holding area constant. This is a continual background process during migration and ensures that there are always files ready for migration.

Increasing the number of concurrent migrations

The default number of concurrent migration tasks is 10. You may find that the throughput rate is improved by increasing this to 15. This leads to higher resource usage.



You can also set the number of concurrent tasks in an EVPM script, as follows:

```
ConcurrentMigrations = 15
```

Calculating disk space

This section deals with the space used and how to size for it. When archiving, Enterprise Vault uses three areas of permanent space:

- The Vault Store partition, which is used to hold the DVS (saveset), DVSSP (saveset shared part) and DVSCC (saveset converted content) files. If collections are enabled, they are stored as CAB files. If Centera is the storage medium, it stores the files in its own format.
- The index area. The current index technology uses AltaVista indexes. Each index is stored in a separate folder and consists of up to 285 files.
- The SQL database, which is used to hold the Directory and Vault Store and fingerprint databases.

Disk space used by vault stores

The Single Instance model has changed in Enterprise Vault 8.0. The principal changes are as follows:

- Items are shared within a Vault Store group. A Vault Store group may contain many Vault Stores and partitions. The partitions may be on different device types, but note that items on Centera are not shared with other devices.
- Shareable parts of a message that exceed the SIS threshold of 20 KB are shared. This includes attachments and message bodies. User information and shareable parts below the SIS threshold are not shared.

The following gives some general rules for estimating the amount of space used. This is a simple calculation that does not take into account some of the complexities.

You should also note that PST files may come from diverse sources with few shared parts. This has to be allowed for when calculating the space used by adjusting the number of sharers. Users may have chosen to have stored more messages with attachments in PSTs, and you should allow for this by adjusting the percentage of messages with attachments.

1. Multiply the number of items to be archived by 16 KB to get the total size of the DVS files. You need to count all messages, including those with attachments. These are the files that are not shared.
2. Take 60% of the size of attachments (or 90% where the mix of attachments is mostly Office 2007 documents). This is the size of attachments after compression. A rule of thumb is that 20% of files have attachments and the average attachment size is 250 KB.
3. Divide by the number of sharers of each attachment across the Vault Store Group. This is the size of the DVSSP files after sharing.
4. Take 5% of the size of DVSSP files. This is the size of the DVSCC files.

Note: These recommendations do not apply to Centera, which uses a completely different sharing model. See “Archiving to Centera” on page 97 for more details.

Disk space used by indexes

Calculate the expected index size as follows:

1. Take the size of the original data.
2. Take a percentage of this according to the indexing type.

Indexing type	Percentage
Brief	3%
Medium	8%
Full	12%

The percentages for Medium and Full will be less if there is little indexable content. This is often the case where there are large attachments such as MP3 or JPEG files.

Disk space used by databases

Metadata is added to the database for every item archived. Temporary space is used to hold information on items that have not been backed up or indexed. Permanent space is also used to hold data in the Directory database. We suggest that an extra 4 GB is allowed for this, or 8 GB where millions of items are archived between Vault Store backups. The extra space is added once only.

Future versions of Enterprise Vault may include more information in the database and the following calculations allow room for expansion.

To calculate the space for Exchange Mailbox and Journal archiving, proceed as follows:

1. Take the number of items archived.
2. Multiply by 500 bytes.
3. Add 4 GB.

A new database is introduced into Enterprise Vault 8.0 to hold fingerprints of the shareable parts of archived items. There is one entry for every shareable part that creates a separate DVSSP file. The number of shareable parts depends on the number of attachments or messages that exceed the SIS threshold and the number of times each is shared. A rule of thumb is that 20% of messages participate in sharing.

To calculate the space used in the fingerprint database:

1. Take the number of items archived.
2. Divide by 5.
3. Divide by the average number of sharers of each shareable item (2 if unknown).
4. Multiply by 500 bytes.

There are no specific limits on the number of items that you can store in one Vault Store database, but we recommend that a Vault Store does not contain more than 100 million items. This makes it easier to perform database maintenance activities such as rebuilding indexes and backing up databases. The new storage model allows vault store databases to be rolled over regularly without loss of sharing.

There are no specific limits on the number of items that you can store in one fingerprint database. The fingerprint database is designed to allow growth and scalability by scaling out across disks and servers. Future versions of this guide will provide more information on this.

These calculations do not take into account the space required by the SQL log files or space used if regular housekeeping is not run on the databases.

Domino mailbox archiving

There is never more than one Domino Mailbox Task on an Enterprise Vault server. The task does not correspond to a Domino server but to one or more provisioning groups that contain users on one or more Domino servers. The consequence is that the task on an Enterprise Vault server may archive from one or more Lotus Domino servers, and the tasks on multiple Enterprise Vault servers may archive mailboxes on a single Domino server.

Lotus Domino 8.5 introduces single instancing. There is no noticeable effect on the ingest rate when this is enabled on the Lotus Domino servers.

The ingest rate for mail-in databases is the same as that for user mailboxes.

Choice of CPU

The default number of threads for the Domino Mailbox Task is 5 and the maximum number of threads is 15. The ingest rate is roughly in proportion to the number of threads, and the archiving rate may be improved by setting this at 15.

The following table shows the expected ingest rate for numbers of physical cores when the number of threads is set to 15 and when the average message size including attachments is 70 KB.

Number of cores	Hourly ingest rate (70 KB)
2	25,000
4 single or 2 dual processors	40,000
2 quad-core processors	60,000

The average size of mail messages has an effect on the throughput. The observed effect is that when the average message size is doubled, throughput is reduced by one third.

These throughput rates will be affected if the Lotus mailboxes are not subjected to the usual maintenance tasks such as regular compaction.

More than one Enterprise Vault server can target the same Domino Server. There is generally no performance penalty, and each Enterprise Vault Server will achieve its target throughput for up to three Domino servers.

Domino Servers are found on many different architectures. The ingest rates from ISeries, Linux and Solaris are approximately 5% faster than Windows and AIX. (The ingest rate for Windows has increased with Lotus Domino 8.5.) Check the *Compatibility Charts* at the following address to see which architectures are supported.

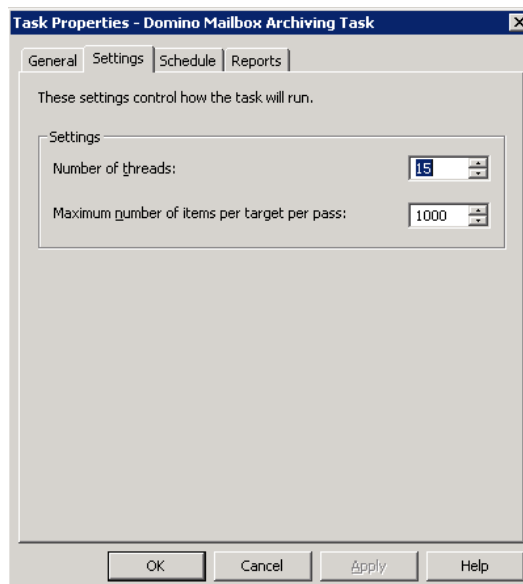
<http://entsupport.symantec.com/docs/276547>

Adjusting the number of threads

It is important to set up the optimum number of threads for Domino Archiving. The factors to take into account are as follows:

- What is the desired ingest rate for each Domino server?
- What is the impact of archiving on the Domino server?

The number of concurrent connections to the Domino server is set in the Administration Console, where it is a property of the Domino Mailbox archiving task. The default is 5 and the maximum is 15. It is recommended that the number of threads is set to a value of 15 to get the maximum throughput.



Calculating disk space

This section deals with the space used and how to size for it. When archiving, Enterprise Vault uses three areas of permanent space:

- The Vault Store partition, which is used to hold the DVS (saveset), DVSSP (saveset shared part) and DVSCC (saveset converted content) files. If collections are enabled, they are stored as CAB files. If Centera is the storage medium, it stores the files in its own format.
- The index area. The current index technology uses AltaVista indexes. Each index is stored in a separate folder and consists of up to 285 files.
- The SQL database, which is used to hold the Directory and Vault Store and fingerprint databases.

Disk space used by vault stores

Domino archiving offers considerable scope for overall space-saving because identical messages that are held in separate mail files are single-instanced by Enterprise Vault.

The Single Instance model has changed in Enterprise Vault 8.0. The principal changes are as follows:

- Items are shared within a Vault Store group. A Vault Store group may contain many Vault Stores and partitions. The partitions may be on different device types, but note that items on Centera are not shared with other devices.
- Shareable parts of a message that exceed the SIS threshold of 20 KB are shared. This includes attachments and message bodies. User information and shareable parts below the SIS threshold are not shared.

The following gives some general rules for estimating the amount of space used. This is a simple calculation that does not take into account some of the complexities.

1. Multiply the number of items to be archived by 16 KB to get the total size of the DVS files. Count all messages, including those with attachments. These are the files that are not shared.
2. Take 60% of the size of attachments (or 90% where the mix of attachments is mostly Office 2007 documents). This is the size of attachments after compression. A rule of thumb is that 20% of files have attachments, and the average attachment size is 250 KB.
3. Divide by the number of sharers of each attachment across the Vault Store Group. This is the size of the DVSSP files after sharing.
4. Take 5% of the size of DVSSP files. This is the size of the DVSCC files.

The total space used is the sum of the DVS, DVSSP and DVSCC files.

If items in the mailboxes have already been journaled, and the journal and mailbox partitions participate in sharing within a Vault Store Group, the shared parts have already been stored and will not be stored again. The only additional space is that used to store the DVS files. There are some exceptions to this rule, and some extra DVSSP files may be created.

If items are archived to more than one partition, more shared parts will be stored on the partition where the archiving task runs first. Some partitions may grow faster than others.

Note: These recommendations do not apply to Centera, which uses a completely different sharing model. See “Archiving to Centera” on page 97 for more details.

Disk space used by indexes

Calculate the expected index size as follows:

1. Take the size of the original data.
2. Take a percentage of this according to the indexing type.

Indexing type	Percentage
Brief	3%
Medium	8%
Full	12%

The percentages for Medium and Full will be less if there is little indexable content. This is often the case where there are large attachments such as MP3 or JPEG files.

Disk space used by databases

Metadata is added to the database for every item archived. Temporary space is used to hold information on items that have not been backed up or indexed. Permanent space is also used to hold data in the Directory database. We suggest that an extra 4 GB is allowed for this, or 8 GB where millions of items are archived between Vault Store backups. The extra space is added once only.

Future versions of Enterprise Vault may include more information in the database and the following calculations allow room for expansion.

To calculate the space for Lotus Domino Mailbox archiving, proceed as follows:

1. Take the number of items archived.
2. Multiply by 750 bytes.
3. Add 4 GB.

A new database is introduced into Enterprise Vault 8.0 to hold fingerprints of the shareable parts of archived items. There is one entry for every shareable part that creates a separate DVSSP file. The number of shareable parts depends on the number of attachments or messages that exceed the SIS threshold and the number of times each is shared. A rule of thumb is that 20% of messages participate in sharing.

To calculate the space used in the fingerprint database:

1. Take the number of items archived.
2. Divide by 5.

3. Divide by the average number of sharers of each shareable item (2 if unknown).
4. Multiply by 500 bytes.

There are no additional entries in the database if mailbox archiving takes place after the items have been journaled within the same Vault Store Group, if the mailbox and journal partitions participate in sharing.

There are no specific limits on the number of items that you can store in one Vault Store database, but we recommend that a Vault Store does not contain more than 100 million items. This makes it easier to perform database maintenance activities such as rebuilding indexes and backing up databases. The new storage model allows vault store databases to be rolled over regularly without loss of sharing.

There are no specific limits on the number of items that you can store in one fingerprint database. The fingerprint database is designed to allow growth and scalability by scaling out across disks and servers. More detailed advice on this will be included in future versions of the performance guide.

These calculations do not take into account the space required by the SQL log files or space used if regular housekeeping is not run on the databases.

Network traffic

The total network traffic generated by archiving an item of an average size of 70 KB is as follows. The figures show the kilobits per second (kbps) for different archiving rates.

	Hourly ingest rate		
	15,000	25,000	40,000
Enterprise Vault server ↔ Domino Server	5,000	8,000	12,500
Enterprise Vault server ↔ SQL Server (Vault Store)	850	1,400	2,000
Enterprise Vault server ↔ SQL Server (Directory)	350	560	900
Enterprise Vault server ↔ SQL Server (Fingerprint)	70	110	160
Enterprise Vault server ↔ Storage medium	6,000	10,000	14,000
Enterprise Vault server ↔ Index location	2,500	4,000	6,250

Retrieving archived items

When an archived item is read, for example by clicking on a shortcut, the request is diverted to the Enterprise Vault Domino Gateway where a temporary mail file is used to hold the retrieved item. The mail file is held in a subdirectory of the Domino mail folder called "EV".

Requests to retrieve items may come from several Domino servers and, if there is a single Enterprise Vault Domino Gateway, all retrieval requests are funneled through a single server.

To allow many concurrent users to be able to retrieve items, do the following:

- Ensure that the Domino Data folder is on a fast disk and not on the system disk. The disk should have at least 50 GB of space available for temporary Enterprise Vault mail files.
- Specify more than one Enterprise Vault Domino Gateway.

The time taken to retrieve an item onto the user's workstation is on average from 0.5 to 1.0 seconds when there are 300 concurrent users, each reading an archived item every 30 seconds. This excludes the time to display the item and assumes that the Domino Data disk can sustain 500 IO per second.

Domino journaling

In most cases, when you are choosing servers for Domino journaling, the most critical factor is the ingest rate. You need to make sure that the servers are able to journal at the required rate during the day.

Choice of CPU

The choice of CPU depends on two main factors: the ingest rate, and the file size.

For general sizing, the following ingest rates should be assumed where the average message size including attachments is 70 KB. These are rates when there is more than one Domino Journaling task.

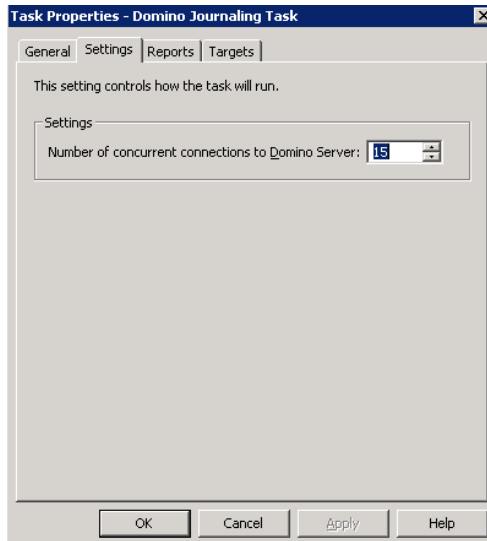
Number of cores	Hourly ingest rate (70 KB)
2	15,000
4 (including dual processors)	40,000
2 quad-core processors	60,000

Number of concurrent connections to the Domino server

It is important to set up the optimum number of connections for Domino Journaling. The factors to take into account are as follows:

- What is the desired ingest rate for each Domino server?
- What is the ratio of Domino servers to Enterprise Vault servers?
- What is the impact of archiving on the Domino server?
- Is the Domino server running on Windows?

The number of concurrent connections to the Domino server is set in the Administration Console, where it is a property of the Domino Journaling task. It is recommended that the number of threads is set to 15 for maximum throughput.



Calculating disk space

This section deals with the space used and how to size for it. When archiving, Enterprise Vault uses three areas of permanent space:

- The Vault Store partition, which is used to hold the DVS (saveset), DVSSP (saveset shared part) and DVSCC (saveset converted content) files. If collections are enabled, they are stored as CAB files. If Centera is the storage medium, it stores the files in its own format.
- The index area. The current index technology uses AltaVista indexes. Each index is stored in a separate folder and consists of up to 285 files.
- The SQL database, which is used to hold the Directory and Vault Store and fingerprint databases.

Disk space used by vault stores

The Single Instance model has changed in Enterprise Vault 8.0. The principal changes are as follows:

- Items are shared within a Vault Store group. A Vault Store group may contain many Vault Stores and partitions. The partitions may be on different device types, but note that items on Centera are not shared with other devices.
- Shareable parts of a message that exceed the SIS threshold of 20 KB are shared. This includes attachments and message bodies. User information and shareable parts below the SIS threshold are not shared.

The following gives some general rules for estimating the amount of space used. This is a simple calculation that does not take into account some of the complexities.

1. Multiply the number of items to be archived by 16 KB to get the total size of the DVS files. Count all messages, including those with attachments. These are the files that are not shared.
2. Take 60% of the size of attachments (or 90% where the mix of attachments is mostly Office 2007 documents). This is the size of attachments after compression. A rule of thumb is that 20% of files have attachments, and the average attachment size is 250 KB.
3. Divide by the number of sharers of each attachment across the Vault Store Group. This is the size of the DVSSP files after sharing.
4. Take 5% of the size of DVSSP files. This is the size of the DVSCC files.

The total space used is the sum of the DVS, DVSSP and DVSCC files.

Note: These recommendations do not apply to Centera, which uses a completely different sharing model. See “Archiving to Centera” on page 97 for more details.

Disk space used by indexes

Calculate the expected index size as follows.

1. Take the size of the original data.
2. Take a percentage of this according to the indexing type.

Indexing type	Percentage
Brief	3%
Medium	8%
Full	12%

The percentages for Medium and Full will be less if there is little indexable content. This is often the case where there are large attachments such as MP3 or JPEG files.

Disk space used by databases

Metadata is added to the database for every item archived. Temporary space is used to hold information on items that have not been backed up or indexed. Permanent space is also used to hold data in the Directory database. We suggest that an extra 4 GB is allowed for this, or 8 GB where millions of items are archived between Vault Store backups. The extra space is added once only.

Future versions of Enterprise Vault may include more information in the database and the following calculations allow room for expansion.

To calculate the space required for Domino Journal archiving:

1. Take the number of items archived.
2. Multiply by 750 bytes.
3. Add 4 GB.

A new database is introduced into Enterprise Vault 8.0 to hold fingerprints of the shareable parts of archived items. There is one entry for every shareable part that creates a separate DVSSP file. The number of shareable parts depends on the number of attachments or messages that exceed the SIS threshold and the number of times each is shared. A rule of thumb is that 20% of messages participate in sharing.

To calculate the space used in the fingerprint database:

1. Take the number of items archived.
2. Divide by 5.
3. Divide by the average number of sharers of each shareable item (2 if unknown).
4. Multiply by 500 bytes.

There are no specific limits on the number of items that you can store in one Vault Store database, but we recommend that a Vault Store does not contain more than 100 million items. This makes it easier to perform database maintenance activities such as rebuilding indexes and backing up databases. The new storage model allows vault store databases to be rolled over regularly without loss of sharing.

There are no specific limits on the number of items that you can store in one fingerprint database. The fingerprint database is designed to allow growth and scalability by scaling out across disks and servers. Future versions of this guide will provide more information on this.

These calculations do not take into account the space required by the SQL log files or space used if regular housekeeping is not run on the databases.

Network traffic

The total network traffic generated by archiving an item of an average size of 70 KB is as follows. The figures show the kilobits per second (kbps) for different archiving rates.

	Hourly ingest rate (70KB)		
	15,000	25,000	40,000
Enterprise Vault server ↔ Domino Server	5,000	8,000	12,500
Enterprise Vault server ↔ SQL Server (Vault Store)	850	1,400	2,000
Enterprise Vault server ↔ SQL Server (Directory)	350	560	900
Enterprise Vault server ↔ SQL Server (Fingerprint)	70	110	160
Enterprise Vault server ↔ Storage medium	6,000	10,000	14,000
Enterprise Vault server ↔ Index location	2,500	4,000	6,250

NSF migration

Introduction

In general, the rate at which items are migrated into Enterprise Vault from NSF files is the same or better than the rate at which items are archived from Lotus Domino. Large-scale migrations need careful planning because they can take several months or longer and may compete for resources with daily archiving and journaling. It is the planning that is all-important, and it is imperative not to be too ambitious for large-scale migrations. You must allow for downtime and for resources to be used by other processes.

One NSF file is migrated at a time. The migrator process has five threads by default but the migration process will be faster with more threads. This is controlled by the following registry value:

Value	Key	Content
MaxNSFNoteMigrationThreads	HKEY_LOCAL_MACHINE \SOFTWARE \KVS \Enterprise Vault \Agents	DWORD value set to an integer value. Default is 5. Raise this to 15.

Choice of CPU

The following table shows the archiving rates per hour assuming 15 threads and where the average message size including attachments is 70 KB.

Number of cores	Rate/Hour (70 KB)
2	25,000
4	40,000
8	60,000

Calculating disk space

This section deals with the space used and how to size for it. When archiving, Enterprise Vault uses three areas of permanent space:

- The Vault Store partition, which is used to hold the DVS (saveset), DVSSP (saveset shared part) and DVSCC (saveset converted content) files. If collections are enabled, they are stored as CAB files. If Centera is the storage medium, it stores the files in its own format.
- The index area. The current index technology uses AltaVista indexes. Each index is stored in a separate folder and consists of up to 285 files.
- The SQL database, which is used to hold the Directory and Vault Store and fingerprint databases.

Disk space used by vault stores

NSF Migration offers considerable scope for overall space-saving because identical messages that are held in separate mail files are single-instanced by Enterprise Vault.

The Single Instance model has changed in Enterprise Vault 8.0. The principal changes are as follows:

- Items are shared within a Vault Store group. A Vault Store group may contain many Vault Stores and partitions. The partitions may be on different device types, but note that items on Centera are not shared with other devices.
- Shareable parts of a message that exceed the SIS threshold of 20 KB are shared. This includes attachments and message bodies. User information and shareable parts below the SIS threshold are not shared.

The following gives some general rules for estimating the amount of space used. This is a simple calculation that does not take into account some of the complexities.

1. Multiply the number of items to be archived by 16 KB to get the total size of the DVS files. These are the files that are not shared.
2. Take 60% of the size of attachments. This is the size of attachments after compression.
3. Divide by the number of sharers of each attachment across the Vault Store Group. This is the size of the DVSSP files after sharing.
4. Take 5% of the size of DVSSP files. This is the size of the DVSCC files.

The total space used is the sum of the DVS, DVSSP and DVSCC files.

Note: These recommendations do not apply to Centera, which uses a completely different sharing model. See “Archiving to Centera” on page 97 for more details.

Disk space used by indexes

Calculate the expected index size as follows.

1. Take the size of the original data.
2. Take a percentage of this according to the indexing type.

Indexing type	Percentage
Brief	3%
Medium	8%
Full	12%

The percentages for Medium and Full will be less if there is little indexable content. This is often the case where there are large attachments such as MP3 or JPEG files.

Disk space used by databases

Metadata is added to the database for every item archived. Temporary space is used to hold information on items that have not been backed up or indexed. Permanent space is also used to hold data in the Directory database. We suggest that an extra 4 GB is allowed for this, or 8 GB where millions of items are archived between Vault Store backups. The extra space is added once only.

Future versions of Enterprise Vault may include more information in the database and the following calculations allow room for expansion.

To calculate the space required for Domino Journal archiving:

1. Take the number of items archived.
2. Multiply by 750 bytes.
3. Add 4 GB.

A new database is introduced into Enterprise Vault 8.0 to hold fingerprints of the shareable parts of archived items. There is one entry for every shareable part that creates a separate DVSSP file. The number of shareable parts depends on the number of attachments or messages that exceed the SIS threshold and the number of times each is shared. A rule of thumb is that 20% of messages participate in sharing.

To calculate the space used in the fingerprint database:

1. Take the number of items archived.
2. Divide by 5.
3. Divide by the average number of sharers of each shareable item (2 if unknown).
4. Multiply by 500 bytes.

There are no specific limits on the number of items that you can store in one Vault Store database, but we recommend that a Vault Store does not contain more than 100 million items. This makes it easier to perform database maintenance activities such as rebuilding indexes and backing up databases. The new storage model allows vault store databases to be rolled over regularly without loss of sharing.

There are no specific limits on the number of items that you can store in one fingerprint database. The fingerprint database is designed to allow growth and scalability by scaling out across disks and servers. Future versions of this guide will provide more information on this.

These calculations do not take into account the space required by the SQL log files or space used if regular housekeeping is not run on the databases.

Network traffic

The total network traffic generated by migrating an item of an average size of 70 KB is as follows. The figures show the kilobits per second (kbps) for different migration rates. The network traffic will be greater between the system holding the NSF files for migration and the Enterprise Vault server when the NSF files are not compacted.

	Hourly ingest rate (70 KB)		
	15,000	25,000	40,000
Enterprise Vault server ↔ NSF File location	10,000	16,000	25,000
Enterprise Vault server ↔ SQL Server (Vault Store)	850	1,400	2,000
Enterprise Vault server ↔ SQL Server (Directory)	350	560	900
Enterprise Vault server ↔ SQL Server (Fingerprint)	70	110	160
Enterprise Vault server ↔ Storage medium	6,000	10,000	14,000
Enterprise Vault server ↔ Index location	2,500	4,000	6,250

File System Archiving

In most cases, when you are choosing servers for File System Archiving, the most critical factor is the ingest rate. There is normally a limited window during the night to archive, and everything must be archived during this period. This window fits in between normal usage, backups, defragmentation, and all the other background activities that take place on the file server.

File System Archiving does not impose a heavy load on the file server, but there may be some I/O to the disks containing the files to be archived.

Enterprise Vault 9.0 keeps a checkpoint of items that have been archived. Previous items that have already been archived are not re-trawled. For customers with a large number of archived items, this provides a considerable performance benefit.

Choice of CPU

The choice of CPU depends on three factors:

- The ingest rate
- The file size
- The file type

For general sizing, the following ingest rates should be assumed.

Number of cores	Hourly ingest rate (100 KB)
2	25,000
4	40,000
8	60,000

The following table shows the ingest rate for some possible scenarios. These figures apply to both NTFS and NAS volumes.

Document size	Document type	Hourly ingest rate (files)	Hourly ingest rate (MB)
70 KB 2 CPU	Text	45000	3200
100 KB 2 CPU	Mixed Office	46000	4500
200 KB 2 CPU	Mixed Office	28000	5500
540 KB 2 CPU	Big PDF	3000	1500
3000 KB 2 CPU	JPEG only	7500	22500
70 KB 4 CPU	Text	55000	4000
100 KB 4 CPU	Mixed Office	60000	6000
200 KB 4 CPU	Mixed Office	46000	9000
540 KB 4 CPU	Big PDF	5400	2851
3000 KB 4 CPU	JPEG only	13000	37000
100 KB 2 CPU (deferred indexing)	Mixed Office	96000	9300

Note the following:

- Text files require no conversion, but large text files do contain more indexable content than other file types.
- Mixed Office (Word, Excel, and PDF) requires some indexing and conversion.
- PDF files are expensive to convert. You can remove some of the expense by converting PDF files to text rather than HTML.

- You can archive a large volume of data when the files are of type JPEG or are similarly unconvertible. Conversion is omitted, and indexing is limited.
- When indexing is deferred, files are not converted to text. This option is available if items are only to be accessed via placeholders and not by any search mechanism. You can create indexes later, but you must first convert the files. This can take as long as the time to archive the items in the first place.

Calculating disk space

This section deals with the space used and how to size for it. When archiving, Enterprise Vault uses three areas of permanent space:

- The Vault Store partition, which is used to hold the DVS (saveset), DVSSP (saveset shared part) and DVSCC (saveset converted content) files. If collections are enabled, they are stored as CAB files. If Centera is the storage medium, it stores the files in its own format.
- The index area. The current index technology uses AltaVista indexes. Each index is stored in a separate folder and consists of up to 285 files.
- The SQL database, which is used to hold the Directory and Vault Store and fingerprint databases.

Disk space used by vault stores

When an item is archived, it is first compressed and then metadata is added to it. The compression ratio depends on the file types that are archived.

The following gives some general rules for estimating the amount of storage needed:

1. Multiply the number of items to be archived by 4 KB to get the total size of the DVS files. These are the files that are not shared.
2. Take 50% of the size of files. This is the size of the files after compression.
3. Divide by the number of sharers of each file across the Vault Store Group. This is the size of the DVSSP files after sharing. If the number of sharers is not known, assume 1.2 per file.
4. Take 5% of the size of DVSSP files. This is the size of the DVSCC files.

The total space used is the sum of the DVS, DVSSP and DVSCC files.

50% compression of the original size applies to a mix of files containing mostly Office 2003 documents. Office 2007 documents do not compress but, with non-Office files among the files, compression averages at 80% of the original size. There is no compression for purely image files.

Note: These recommendations do not apply to Centera, which uses a completely different sharing model. See “Archiving to Centera” on page 97 for more details.

Disk space used by indexes

Files ingested through FSA usually use less indexing space than mail messages, which have far greater word content in proportion to their size than even Office-type documents. Files are usually larger than mail messages, so even brief indexing uses proportionately less space.

To calculate the expected index size as follows for Office-type documents:

1. Take the size of the original data.
2. Take a percentage of this according to the indexing type.

Indexing type	Percentage
Brief	1%
Medium	2%
Full	5%

The percentages for Medium and Full will be less if there is little indexable content. For example, if the files are all compressed image files, even full indexing will be 1%.

If files are mainly small text messages then the space used by indexing will be comparable to that used by Exchange mailbox items.

Disk space used by databases

Metadata is added to the database for every item archived. Temporary space is used to hold information on items that have not been backed up or indexed. Permanent space is also used to hold data in the Directory database. We suggest that an extra 4 GB is allowed for this, or 8 GB where millions of items are archived between Vault Store backups. The extra space is added once only.

Future versions of Enterprise Vault may include more information in the database, and the following calculations allow room for expansion.

To calculate the size of databases when using FSA:

1. Take the number of items.
2. Multiply by 3000 bytes.
3. Add 4 GB.

A new database is introduced into Enterprise Vault 8.0 to hold fingerprints of the shareable parts of archived items. There is one entry for every shareable part that creates a separate DVSSP file. For File System Archiving it is expected that all files will exceed the SIS Threshold.

To calculate the space used in the fingerprint database:

1. Take the number of items archived.
2. Divide by the number of sharers of each shareable item (1.2 if unknown).
3. Multiply by 500 bytes.

There are no specific limits on the number of items that you can store in one Vault Store database, but we recommend that a Vault Store does not contain more than 100 million items. This makes it easier to perform database maintenance activities such as rebuilding indexes and backing up databases. The new storage model allows vault store databases to be rolled over regularly without loss of sharing.

There are no specific limits on the number of items that you can store in one fingerprint database. The fingerprint database is designed to allow growth and scalability by scaling out across disks and servers. Future versions of this guide will provide more information on this.

These calculations do not take into account the space required by the SQL log files or space used if regular housekeeping is not run on the databases.

Network usage

The network may be used for the following purposes while ingesting items:

- Communicating with and copying data from the file servers.
- Accessing the SQL database.
- Transferring archived data to the storage medium (for example, NAS or Centera).
- Retrieving archived data from the storage medium for indexing.
- Reading and writing data to and from the Index Storage medium.
- Background activity, such as communication with the domain controller, user monitoring, and so on.

Communicating with the file server

A rule of thumb is that the amount of network traffic between the Enterprise Vault server and the file server is the size of the data plus 30%.

Communicating with the SQL database

A rule of thumb is to allow 20 KB for every item archived to the Vault Store database and 5 KB to the Directory database.

Transferring data to the storage medium and retrieving for indexing

The amount of data being sent and received from the storage medium depends on the single instance and compression ratios. In general, the network traffic between Enterprise Vault server and storage medium is double that of the original data.

Reading and writing indexes

When an index is opened, some of the index files are transferred to the Enterprise Vault server. On completion of indexing, the files are written back. Sometimes the files are written back during indexing. The amount of data transferred depends on the number and size of indexes. During file system archiving, there is a separate index for each archive point. For example, if only one or two items are archived from each archive point, indexes are constantly opened and closed, and a lot of data is transferred. It is therefore difficult to predict the traffic to the index location but, in general, the network traffic between the index location and the Enterprise Vault server is equal to the size of the original item for every indexed item.

Summary

The total network traffic generated by archiving an item of an average size of 100 KB is as follows. The figures show the kilobits per second (kbps) for different archiving rates.

	Hourly ingest rate (100 KB)		
	15,000	25,000	40,000
Enterprise Vault server ↔ File Server	4,500	7,250	12,000
Enterprise Vault server ↔ SQL Server (Vault Store)	700	1,200	1,800
Enterprise Vault server ↔ SQL Server (Directory)	200	300	450
Enterprise Vault server ↔ SQL Server (Fingerprint)	70	110	160
Enterprise Vault server ↔ Storage medium	9,000	15,000	23,000
Enterprise Vault server ↔ Index location	3,500	5,750	9,000

File types

There are many file types that should be excluded from archiving or included in archiving. For example, it may be only Office files that need to be archived. Again, large files such as .log files should usually be excluded from archiving to prevent the indexes from being cluttered with information that is not useful.

File System Archiving and SMTP Archiving

SMTP Archiving allows files to be diverted into a directory location and then archived by File System Archiving. SMTP mail can be written to this location almost as fast as mail can be delivered and be ready for archiving.

The folder structure of this location is Year\Month\Date\Hour. When File System Archiving runs, it creates a thread for each folder that it processes up to a maximum of 10 threads. If the SMTP location is being populated while FSA is running, there may not be enough “hour” folders populated with data to occupy all the threads, and FSA may run slower.

Upgrading FSA metadata after the upgrade to Enterprise Vault 9.0

After you upgrade to Enterprise Vault 9.0, use the FSA upgrade utility to upgrade the FSA metadata. This is described in the *Utilities* guide, which is part of the Symantec Enterprise Vault Documentation Library.

Allow one hour for every 10 to 15 million items in the FSA Vault Store database.

You can upgrade more than Vault Store at a time. This is faster than upgrading each Vault Store individually.

SharePoint

Introduction

In most cases, when you are choosing servers for SharePoint, the most critical factor is the ingest rate.

There is generally a higher ingest rate from a single SharePoint 2007 server compared with a SharePoint 2003 server.

Choice of CPU

The choice of CPU depends on three factors:

- The ingest rate
- The file size
- The file type

For general sizing, the following ingest rates should be assumed for average document sizes of 200 KB.

Number of cores	Hourly ingest rate (200 KB)
2	25,000
4	40,000
8	60,000

Calculating disk space

This section deals with the space used and how to size for it. When archiving, Enterprise Vault uses three areas of permanent space:

- The Vault Store partition, which is used to hold the DVS (saveset), DVSSP (saveset shared part) and DVSCC (saveset converted content) files. If collections are enabled, they are stored as CAB files. If Centera is the storage medium, it stores the files in its own format.
- The index area. The current index technology uses AltaVista indexes. Each index is stored in a separate folder and consists of up to 285 files.
- The SQL database, which is used to hold the Directory and Vault Store and fingerprint databases.

Disk space used by vault stores

When an item is archived, it is first compressed and then metadata is added to it. The compression ratio depends on the file types that are archived.

The following gives some general rules for estimating the amount of storage needed:

1. Multiply the number of items to be archived by 4 KB to get the total size of the DVS files. These are the files that are not shared.
2. Take 50% of the size of files. This is the size of files after compression.
3. Divide by the number of sharers of each file across the Vault Store Group. This is the size of the DVSSP files after sharing. If the number of sharers is not known, assume 1.2 per message.
4. Take 5% of the size of DVSSP files. This is the size of the DVSCC files.

50% compression applies to a mix of files containing mostly Office 2003 documents. Office 2007 documents do not compress but, with non-Office files among the files, compression will average at 80% of the original size. There is no compression for purely image files.

The total space used is the sum of the DVS, DVSSP and DVSCC files.

Note: These recommendations do not apply to Centera, which uses a completely different sharing model. See “Archiving to Centera” on page 97 for more details.

Disk space used by indexes

Files ingested through SharePoint usually use less indexing space than mail messages, which have far greater word content in proportion to their size than even Office-type documents. Files are usually larger than mail messages, so even brief indexing uses proportionately less space.

To calculate the expected index size as follows for Office type documents:

1. Take the size of the original data.
2. Take a percentage of this according to the indexing type.

Indexing type	Percentage
Brief	1%
Medium	2%
Full	5%

The percentages for Medium and Full will be less if there is little indexable content. For example, if the files are all compressed image files, even full indexing will be 1%.

If files are mainly small text messages then the space used by indexing will be comparable to that used by Exchange mailbox items.

Disk space used by databases

Metadata is added to the database for every item archived. Temporary space is used to hold information on items that have not been backed up or indexed. Permanent space is also used to hold data in the Directory database. We suggest that an extra 4 GB is allowed for this, or 8 GB where millions of items are archived between Vault Store backups. The extra space is added once only.

Future versions of Enterprise Vault may include more information in the database and the following calculations allow room for expansion.

To calculate the size of databases when using SharePoint:

1. Take the number of items.
2. Multiply by 3000 bytes.
3. Add 4 GB.

A new database is introduced into Enterprise Vault 8.0 to hold fingerprints of the shareable parts of archived items. There is one entry for every shareable part that creates a separate DVSSP file. The number of shareable parts depends on the number of attachments or messages that exceed the SIS threshold and the number of times each is shared. A rule of thumb is that 20% of messages participate in sharing.

To calculate the space used in the fingerprint database:

1. Take the number of items archived.
2. Divide by the number of sharers of each shareable item (1.2 if unknown).
3. Multiply by 500 bytes.

There are no specific limits on the number of items that you can store in one Vault Store database, but we recommend that a Vault Store does not contain more than 100 million items. This makes it easier to perform database maintenance activities such as rebuilding indexes and backing up databases. The new storage model allows vault store databases to be rolled over regularly without loss of sharing.

There are no specific limits on the number of items that you can store in one fingerprint database. The fingerprint database is designed to allow growth and scalability by scaling out across disks and servers. Future versions of this guide will provide more information on this.

These calculations do not take into account the space required by the SQL log files or space used if regular housekeeping is not run on the databases.

Network traffic

The total network traffic generated by archiving an item of an average size of 100 KB is as follows. The figures show the kilobits per second (kbps) for different archiving rates.

	Hourly ingest rate (100 KB)		
	15,000	25,000	40,000
Enterprise Vault server ↔ SharePoint Server	5,000	7,500	12,500
Enterprise Vault server ↔ SQL Server (Vault Store)	700	1,200	1,800
Enterprise Vault server ↔ SQL Server (Directory)	200	300	450
Enterprise Vault server ↔ SQL Server (Fingerprint)	70	110	160
Enterprise Vault server ↔ Storage medium	9,000	15,000	23,000
Enterprise Vault server ↔ Index location	3,500	5,750	9,000

Retrieving items

Enterprise Vault 8.0 SP3 introduced seamless shortcuts. In previous releases, items retrieved by opening shortcuts were fetched directly from the Enterprise Vault server. Now they are retrieved through the SharePoint server. Under normal conditions, retrievals take less than one second on average, depending on the size of the item. However, sites that upgrade to 8.0 SP3 or later will see an increase in network traffic to and from the SharePoint server.

Enterprise Vault Discovery Collector

Enterprise Vault 9.0 introduces the Enterprise Vault Discovery Collector.

Enterprise Vault Discovery Collector is primarily used to identify and collect data from data sources within your corporate IT environment. Collected data is then stored within Enterprise Vault for analysis and review by Discovery Accelerator.

There are four key steps in using Discovery Collector: "Index", "Copy and Collect", "Discovery Accelerator Analysis", and "Export".

This chapter provides guidelines on sizing and configuring the Enterprise Vault servers for the "Copy and Collect" step only. For the "Index" step, see the following documentation supplied with the Discovery Collector. A full list of documentation can be found with the release notes (<http://entsupport.symantec.com/docs/340192>).

Document	Comments
Symantec Enterprise Vault Discovery Collector Installation and Planning Guide	Guides you through the process of planning, sizing, and installing Discovery Collector for use on virtual machines.
Installation and Planning Template	Helps you to complete planning, configuration, and operational readiness tasks before you use Discovery Collector. The template is a Microsoft Excel macro-enabled worksheet (.xslm). You require Excel 2007 to open it.
Symantec Enterprise Vault Discovery Collector User Guide	Provides detailed information on how to configure and use Discovery Collector.
Operational Readiness Template	A Microsoft Excel template designed to help you complete operational readiness tasks prior to using Discovery Collector.

For sizing and configuration for the "Discovery Accelerator Analysis" and "Export" steps, see the *Symantec Discovery Accelerator Best Practices Guide* <Add ref>

In most cases, when you are choosing servers for Enterprise Vault Discovery Collector, the most critical factor is the ingest rate. There is normally a limited window during the night to archive, and everything must be archived during this period. This window fits in between normal usage, backups, defragmentation, and all the other background activities.

Choice of CPU

- The choice of CPU depends on three factors:
- The ingest rate
 - The file size
 - The file type
- For general sizing, the following ingest rates should be assumed.

Number of cores	Hourly ingest rate (100 KB)
2	25,000
4	40,000
8	60,000

For every doubling in the average size of items, reduce the throughput rate by 33%.

Calculating disk space

- This section deals with the space used and how to size for it. When archiving, Enterprise Vault uses three areas of permanent space:
- The Vault Store partition, which is used to hold the DVS (saveset), DVSSP (saveset shared part) and DVSCC (saveset converted content) files. If collections are enabled, they are stored as CAB files. If Centera is the storage medium, it stores the files in its own format.
 - The index area. The current index technology uses AltaVista indexes. Each index is stored in a separate folder and consists of up to 285 files.
 - The SQL database, which is used to hold the Directory and Vault Store and fingerprint databases.

Disk space used by vault stores

When an item is archived, it is first compressed and then metadata is added to it. The compression ratio depends on the file types that are archived.

The following gives some general rules for estimating the amount of storage needed:

1. Multiply the number of items to be archived by 4 KB to get the total size of the DVS files. These are the files that are not shared.
2. Take 50% of the size of files. This is the size of the files after compression.
3. Divide by the number of sharers of each file across the Vault Store Group. This is the size of the DVSSP files after sharing. If the number of sharers is not known, assume 1.2 per file.
4. Take 5% of the size of DVSSP files. This is the size of the DVSCC files.

The total space used is the sum of the DVS, DVSSP, and DVSCC files.

50% compression of the original size applies to a mix of files containing mostly Office 2003 documents. Office 2007 documents do not compress, but, with non-Office files among the files, compression averages at 80% of the original size. There is no compression for purely image files.

Note: These recommendations do not apply to Centera, which uses a completely different sharing model. See “Archiving to Centera” on page 97 for more details.

Disk space used by indexes

Calculate the expected index size as follows.

1. Take the size of the original data.
2. Take a percentage of this according to the indexing type.

Indexing type	Percentage
Brief	3%
Medium	8%
Full	12%

Disk space used by databases

Metadata is added to the database for every item archived. Temporary space is used to hold information on items that have not been backed up or indexed. Permanent space is also used to hold data in the Directory database. We suggest that an extra

4 GB is allowed for this, or 8 GB where millions of items are archived between Vault Store backups. The extra space is added once only.

Future versions of Enterprise Vault may include more information in the database, and the following calculations allow room for expansion.

To calculate the space required for Domino Journal archiving:

1. Take the number of items archived.
2. Multiply by 750 bytes.
3. Add 4 GB.

A new database is introduced into Enterprise Vault 8.0 to hold fingerprints of the shareable parts of archived items. There is one entry for every shareable part that creates a separate DVSSP file. The number of shareable parts depends on the number of attachments or messages that exceed the SIS threshold and the number of times each is shared. A rule of thumb is that 20% of messages participate in sharing.

To calculate the space used in the fingerprint database:

1. Take the number of items archived.
2. Divide by 5.
3. Divide by the average number of sharers of each shareable item (2 if unknown).
4. Multiply by 500 bytes.

There are no specific limits on the number of items that you can store in one Vault Store database, but we recommend that a Vault Store does not contain more than 100 million items. This makes it easier to perform database maintenance activities such as rebuilding indexes and backing up databases. The new storage model allows vault store databases to be rolled over regularly without loss of sharing.

There are no specific limits on the number of items that you can store in one fingerprint database. The fingerprint database is designed to allow growth and scalability by scaling out across disks and servers. These calculations do not take into account the space required by the SQL log files or space used if regular housekeeping is not run on the databases.

Network usage

The network may be used for the following purposes while ingesting items:

- Communicating with and copying data from the file servers.
- Accessing the SQL database.
- Transferring archived data to the storage medium (for example, NAS or Centera).
- Retrieving archived data from the storage medium for indexing.

- Reading and writing data to and from the Index Storage medium.
- Background activity, such as communication with the domain controller, user monitoring, and so on.

Communicating with the collector server

A rule of thumb is that the amount of network traffic between the Enterprise Vault server and the file server is the size of the data plus 30%.

Communicating with the SQL database

A rule of thumb is to allow 20 KB for every item archived to the Vault Store database and 5 KB to the Directory database.

Transferring data to the storage medium and retrieving for indexing

The amount of data being sent and received from the storage medium depends on the single instance and compression ratios. In general, the network traffic between the Enterprise Vault server and storage medium is double that of the original data.

Reading and writing indexes

When an index is opened, some of the index files are transferred to the Enterprise Vault server. On completion of indexing, the files are written back. Sometimes the files are written back during indexing. The amount of data transferred depends on the number and size of indexes. During file system archiving, there is a separate index for each archive point. For example, if only one or two items are archived from each archive point, indexes are constantly opened and closed, and a lot of data is transferred. It is therefore difficult to predict the traffic to the index location. However, in general, the network traffic between the index location and the Enterprise Vault server is equal to the size of the original item for every indexed item.

Summary

The total network traffic generated by archiving an item of an average size of 100 KB is as follows. The figures show the kilobits per second (kbps) for different archiving rates.

	Hourly ingest rate (100 KB)		
	15,000	25,000	40,000
Enterprise Vault server ↔ Collector Server	4,500	7,250	12,000
Enterprise Vault server ↔ SQL Server (Vault Store)	700	1,200	1,800
Enterprise Vault server ↔ SQL Server (Directory)	200	300	450
Enterprise Vault server ↔ SQL Server (Fingerprint)	70	110	160
Enterprise Vault server ↔ Storage medium	9,000	15,000	23,000
Enterprise Vault server ↔ Index location	3,500	5,750	9,000

Search and download

An important part of Enterprise Vault is the user experience when interacting with archived items—for example, searching for items, downloading items, or using Archive Explorer. The user cannot expect the same response times as with the native application, such as Exchange. One goal in Enterprise Vault is to allow the use of slower, cheaper media to store archived items, and retrieval from slower media will always be slower. On the other hand, the user is entitled to expect a reasonable response time, and it is right to set expectations as to when response times are slower.

Types of response time

There are two types of response time:

- The time between the server receiving a request and returning it. This response time is easy to measure and characterize.
- The user-perceived response time between the user action and the requested data being displayed. This is harder to measure and may depend on the user's workstation rather than the server.

Some user interactions require more resources on the server and some on the client server. For example, when retrieving an item, the saveset (DVS) file may be unpacked on the server or client depending on the type of retrieval.

Viewing items online

When an item is viewed online from a shortcut in the mailbox or from search results or from Archive Explorer, the DVS file and any DVSSP files are retrieved from storage and recombined on the Enterprise Vault server and returned to the client. The following are guidelines on the number of retrievals per hour and the number of concurrent users before the download time exceeds 0.5 seconds on average. A user is

defined as someone who retrieves one item per minute. The times will vary depending on the storage devices where the component parts are held.

Number of cores	Concurrent users	Hourly retrieval rate (70 KB)
2	400	25,000
4	650	40,000
8	1000	60,000

As more users are added, average retrieval rates rise and become unstable at three times the number of concurrent users.

FSA - opening placeholders

When a placeholder is opened, the File Placeholder service on the file server makes an HTTP call to the Enterprise Vault server. The number of files that can be opened by one user is purposely limited in FSA to 20 recalls within 10 seconds. If the limit is exceeded, it is not possible to open placeholder files for a few seconds. The recall process adds approximately 0.5 seconds to the time to open a file. When a placeholder is opened and the file retrieved, the file remains in its original location until the next archiving run, and further opens make no calls to the Enterprise Vault Server.

Searching indexes

There are many factors that determine the time taken to return a search result. When a user searches an index, the index is opened and some of the files are brought into memory. The first search in a series of searches takes longer. For example, typical first search response times for a user with 100,000 archived items are as follows:

First search	Subsequent searches
3 to 10 seconds	0.5 to 3 seconds

The longer times are more likely from NAS devices or from larger indexes. In most cases, subsequent search times are less than one second when up to 20 searches are in a search session. See also “Accelerators” on page 103 for more detailed information on search times.

Archive Explorer

When a folder is expanded in Archive Explorer to display a list of messages or subfolders, an index search is triggered. The retrieved information is cached on the client for 24 hours so that, if the user expands the folder again, there is no interaction with the server unless the user specifically refreshes the folder view.

When a user views an item, the message is downloaded (see “Viewing items online” on page 85).

The result is that Archive Explorer is very responsive and can support a large number of concurrent users.

Virtual Vault

Overview

The Virtual Vault functionality was introduced in Enterprise Vault 8.0 SP3. Virtual Vault integrates a view of the Vault Cache into the Outlook Navigation Pane. To users, a Virtual Vault looks and behaves in the same way as a mailbox or a personal folder. For example, users can open archived items and drag and drop items to and from the Virtual Vault.

The content strategies from which you can choose are as follows:

Content strategy	Description
Do not store any items in cache	Item headers are synchronized to Vault Cache, but the content of archived items is not stored in Vault Cache. If a user who is online opens an item in Virtual Vault, or selects an item when the Reading Pane is open, Enterprise Vault immediately retrieves the content from the online archive.
Store all items	This is the default option. Item headers are synchronized to Vault Cache, and the content of archived items is stored in Vault Cache.
Store only items that user opens	Item headers are synchronized to Vault Cache. If a user who is online opens an item in Virtual Vault, or selects an item when the Reading Pane is open, Enterprise Vault immediately retrieves the content from the online archive. The content of each item that a user opens in Virtual Vault is stored in Vault Cache.

This chapter discusses the following two content strategies:

- With-content Cache. This covers “Store All Items”.
- Contentless Cache. This covers “Do not store any items in cache” and, for all practical purposes, “Store only items that user opens”.

The chapter also deals with these two phases:

- Initial synchronization – the first time users’ Vault Caches are enabled.
- Incremental synchronization – the day-to-day synchronization of users’ Vault Cache with their archives.

Initial synchronization

When an Enterprise Vault system is upgraded to a version that supports Virtual Vault, users’ Vault Caches are synchronized with their archives.

The following table shows the expected times to complete synchronization for users with an average of 12,000 archived items of average size 70 KB on a four-core server.

	100 users	250 users	1000 users	3000 users
VC Initial Sync (contentless)	10 min	30 min	2 hrs	6 hrs
VC Initial Sync (with-content)	120 min	300 min	20 hrs	60 hrs

The following table shows the expected times to complete synchronization for users with an average of 200,000 archived items.

	1000 users	3000 users
VC Initial Sync (contentless)	31 hrs	92 hrs
VC Initial Sync (with-content)	33 days	100 days

When users are enabled for with-content Cache, the metadata that allows them to see and work in their Virtual Vault is downloaded first. Then the contents are downloaded with the most recent items first. These users can use their Virtual Vaults before synchronization has completed.

These times are elapsed times and do not take into account other activity on the servers or network delays. In most cases, it is not practical or necessary to enable all users for with-content Cache. You can take several different approaches, as follows:

- Initially enable all users for contentless Cache. This allows all users to use Virtual Vault as soon as possible.
- Enable with-content Cache for only those users who do not always have access to their online archives. Typically, these would be remote users or traveling users who work away from the office. However, for better performance, the initial synchronization should be done when the users have access to their archive over a fast network.

- Limit the size or time range of items in the with-content Cache. Users typically require access to the most recent items rather than their entire archives.
- Prioritize users and enable them in batches.

Incremental synchronization

Once the initial synchronization has completed, Virtual Vaults are updated every day while users are logged into their mail clients. Users may also start synchronization manually.

The following table shows the time taken to perform an incremental synchronization for the following daily actions:

- Download an average of 50 items.
- Upload an average of 10 items into the archive (items copied manually into Virtual Vault).
- Manually delete an average of 10 archived items using Virtual Vault.
- Create an average of two folders in the archive manually created in Virtual Vault.

	100 users	250 users	1000 users	3000 users
VC Incremental Sync - Virtual Vault client updates and item download (contentless)	4 min	10 min	40 min	2 hr
VC Incremental Sync - Virtual Vault client updates and item download (with-content)	4 min	16 min	1 hr	3 hr

The time for the incremental synchronization with-content Cache is close to the contentless Cache synchronization when the items have been preemptively cached. This is the normal case.

Move Archive

Overview

The Move Archive feature was introduced in Enterprise Vault 8.0 SP4. It allows the movement of one or more archives from one vault store to another.

Setting Move Archive parameters

By default, the Move Archive task runs at a low priority to prevent interference with other Enterprise Vault activity. If you want to increase the rate at which items are moved, you can adjust the settings on the Settings tab of the Task Properties of the Move Archive task. You can increase the following:

- Priority of the Move Archive operations in relation to other processes
- Number of concurrent move operations
- Number of threads per move operation

The total number of threads (that is, the number of concurrent move operations multiplied by the number of threads per move operation) should not exceed 20. By increasing the number of concurrent move operations, you allow more archives to be moved in parallel. By increasing the number of threads per move operation, you allow each archive to be moved more quickly. Normally, you would want a balance between the two, such as five concurrent move operations and four threads per move operation. This gives more archives a chance to complete their moves in a reasonable time without being blocked by one or two larger archives.

The effect of raising the priority and increasing the total number of threads is that the CPU and other resource usage on the Enterprise Vault Servers may reach a high level. This will have an effect on other Enterprise Vault activity, such as scheduled archiving or daily journaling.

Moving small number of users

The most common use of Move Archive is to move one user, or a few users, between vault stores, possibly across servers or sites. In this situation, the Enterprise Vault servers typically absorb the resources that are used.

Moving large number of users

Moving a large number of users requires time and planning. The process of moving an archive is equivalent to the original ingest, and it necessitates all the steps of ingest, shortcut update, and backup. In addition, the new archive must be verified and the original archive deleted. Because of the extra steps taken when moving an archive (most notably the verification phase), the total time to move an archive is likely to be longer than the original ingest.

These are the suggested steps to take to prepare for Move Archive:

1. Select a schedule for Move Archive that is different from the Mailbox Archiving Task schedule. It is suggested that the schedule is set during the day. The Move Archive process does not affect the users' use of Exchange or Lotus Notes, but it may affect interactions with Enterprise Vault when searching for or retrieving items.
2. Calculate the total time required to move the archives. To do this, consider the number of items that you want to move, and not the number of users.

The following table shows how much faster or slower a Move Archive task is than the original ingest process. It is assumed that you have raised the priority of the Move Archive process and increased the number of processes/threads.

Move type	Comparison with original ingest
Move to partition in same vault store group on different server	45% faster
Move to partition in different vault store group on different server	30% faster
Move to partition in same vault store group on same server	40% faster
Move to partition on different site	30% faster
Move from NTFS collection	25% slower
Move to/from Centera partition	As original archive

In most cases, the Move Archive process is faster than the original archive process. This is because the processing resources that the agents for Exchange or Domino archiving use are released. If the speed of the original archive is limited by a resource other than CPU, the Move Archive rate converges on the original archive rate.

3. Divide the users into blocks and prioritize those that you want to move first. The users in each block should contain the number of items that can be moved in one Move Archive session, as calculated above. When you have moved the first set of archives, you may want to adjust the number of users to be moved in a single block.
4. Add the users and allow Move Archive to take place during the scheduled period.
5. Calculate the Move Archive rate.
6. Allow the daily scheduled Archiving task to complete and update the shortcuts. The time to update shortcuts is trivial and should be absorbed into the Archiving task. Only moves within a site to new users will allow shortcuts to be updated.
7. Make a backup copy of the newly moved items. If you have moved them to a partition that is regularly backed up then this will happen automatically, but you need to allow extra time for the process to complete. Some device types have almost immediate backup or replication, and this stage will be completed quickly.
8. After items have been identified as backed up, database entries are removed from the relevant tables. To some extent, this extends the time for the StorageFileWatch process to run, but normally by a few minutes only.
9. After files have been secured, all moved items are verified to check that they have not been corrupted or altered during the move. The time taken to verify items is normally 50% of the ingest rate.
10. The archives that have been moved may be deleted from the source destination. This step is accomplished by deleting the entries marked "Completed" from the Move Archive Status list. The source archives are deleted during the next Storage Expiry run. This is normally a scheduled task. The existence of the old archives does not interfere with the use of the new archives, but if it is required to delete them quickly then you may have to extend the Expiry schedules.

General notes

- **Network.** If you copy between sites or to a different vault store sharing group on a different server, the data sent across the network is the same size as the originally archived data. This may be a factor when moving archives between remote sites over a slow link.
- **Location of moved data.** If you copy archives within a vault store sharing group, the sharable parts of the moved items are not moved. If their original storage location was the source partition, they remain there. Every item has a part that is not shared, and this is recreated on the destination partition. If items are moved to a different sharing group, new sharable parts are created on the destination partition unless a copy already exists. When the archives are deleted from the source partition, shared parts that have no references are deleted. The result is that, once all stages of the move have completed, there may not be significant space reduction on the source partition or an increase in space on the destination server.
- **Archiving source.** The archiving source (journal, archive, Exchange, Domino) makes no direct difference to the Move Archive rate.
- **Resources used.** When you move archives between servers, you use resources on both the source and destination servers. The Move Archive process disrupts other activity on these servers, such as regular ingest. The limiting resource on Move Archive is CPU, especially on both servers. There may be other factors such as network speed or disk speed that may also limit the transfer rate. These factors vary from site to site and cannot be predicted.
- **Move rates.** If you move archives between sites or differently specified systems, use the lowest specified system when you calculate the move rate.

Archiving to Centera

EMC Centera devices offer a reliable means of archiving data with the added advantage that where replica devices are involved, no backups of archived data are necessary. Replication is a continuous process that secures data on a separate Centera performing a function equivalent to backup. This allows the archiving window to be extended. (Indexes and SQL databases are not held on Centera and still require backups. In some cases, data held on Centera is both replicated and backed up.) The performance of Centera has improved with each generation. This section is based on a 16-node Gen-4 Centera with four access nodes.

The following white paper is a good point of reference when sizing and configuring Enterprise Vault with Centera:

<http://www.emc.com/collateral/hardware/white-papers/h6790-symantec-enterprise-vault-centera-wp.pdf>

Archiving with and without Centera collections

Enterprise Vault offers two methods of storing items in Centera: with collections and without collections. Centera collections are completely different from NTFS collections that can be used when storing to NTFS storage. When items are stored in Centera collections, they are first stored in a temporary area and then collected into a single object and stored on Centera. A collection is up to 100 items or 10 MB of data. Collections are recommended because they result in fewer objects on the Centera. This has several advantages:

- No fall-off in performance as the Centera gets fuller.
- Faster replication.
- Faster deletion of expired items.
- Faster self-healing in the event of a failed disk.

Items for collection are stored on a local disk before collection. This needs to be a fast disk but not large.

Retrieval of items in collections is very fast because only the item is retrieved from Centera and not the whole collection.

As the performance of Centera improves, many of these factors will have less impact, and archiving without collections is a viable solution. Customers should consult with Symantec or EMC before archiving on a Centera without collections.

Centera sharing model

The way that items are shared or single-instanced with Centera differs from other devices. On Centera, attachments are detached from the message and stored in Centera, where Centera identifies them as candidates for sharing. The exact rules are as follows:

- A saveset with an uncompressed size of 100 KB is stored unshared.
- A saveset with a compressed size of over 100 KB is examined for “streams”—indexable items or XML streams such as recipient lists—and attachments.
- If there are no streams or attachments, the saveset is stored unshared.
- If there are no streams or attachments with an uncompressed size of over 50 KB, the saveset is stored unshared.
- Any stream or attachment with an uncompressed size of over 50 KB is stored separately and is eligible for sharing.

This model had the advantage that attachments are shared even if they are attached to different messages or archived separately by File System Archiving. It also means that there is sharing across vault stores. Small messages are not shared. However, even though small messages make up the bulk of messages, messages with large shareable attachments usually make up the bulk of the size. For example, a large report might be sent or forwarded to all members of a company. Just one copy of this report is held on Centera, although there will be many copies held on the Exchange Stores or Lotus mail files in the company.

Choice of Enterprise Vault server

There is no substantial difference in performance when archiving to a Centera when compared with other storage media. Using collections does add a small CPU overhead as the collection is an extra process. Refer to the tables for each archiving type for the throughput rate. Likewise, there is little difference in retrieval times when individuals view items or perform bulk-retrieval operations.

Enterprise Vault checks for replication every 60 minutes. Therefore, shortly after an archiving run finishes, the system is fully replicated onto a local replica Centera,

and items are turned into shortcuts in users' mailboxes. Replication to a remote Centera will depend on the speed of the network link.

Additional disk space used by databases

Additional space in the database is required to store an item when the Vault Store is located on a Centera partition but is not in a Centera collection. In addition to any other calculations:

- 1. Take the number of items.
- 2. Multiply by 500 bytes.

Centera settings

Writes to the Centera are slightly slower than to other devices, but many I/Os can happen in parallel. When archiving with collections, this is not relevant because it is only collections that are written to Centera and not individual items. However, when archiving to Centera without collections, optimum performance is reached when the number of processes is increased. For example:

Number of storage archive processes	Number of PST migrators
10	20

Centera limits

Depending on the business needs, a single Centera may act as storage for many Enterprise Vault systems. The measured maximums are on a 16-node Gen-4 Centera with four access nodes are as follows:

Hourly ingest rate (inc. replication) (100 KB messages)	Hourly retrieval rate (with collections)
350,000 (from seven Enterprise Vault servers)	1,000,000 (from eight Enterprise Vault servers)

The ingest rate was limited by the number of Enterprise Vault servers available for testing. The absolute maximum is higher than this, but it is not possible to speculate what this may be. When retrieving 1,000,000 items an hour, the Centera access nodes were fully loaded.

Storage nodes may act as access nodes, and access nodes as storage nodes. There is no need to waste space by assigning nodes exclusively as access nodes, and the maximum ingest rate and retrieval rate can be increased by converting storage nodes to access nodes. There is no loss of storage capacity in doing this, but there is a cost in creating the extra connections.

Self-healing

If a disk or node fails on a Centera, the Centera goes into a self-healing state and recovers the data. The self-healing process is intensive on resources on Centera, but it does not take precedence over other activity. An example is that an index is normally rebuilt at a rate of 100,000 items an hour. While self-healing is in progress, this rate reduces to 60,000 items an hour.

NTFS to Centera migration

Items can be migrated to Centera at a high rate. The following table shows a typical example.

Metric	Hourly rate
Saveset files migrated per hour	130,000
GB (original size) migrated per hour	9

Archiving to a storage device through the Enterprise Vault Storage Streamer API

Enterprise Vault supports archiving to a range of different storage devices. Except for the EMC Centera, all of these devices have been accessed through a CIFS/SMB interface.

Enterprise Vault 9.0 introduces a new feature that allows it to use third-party storage devices that are not compatible with CIFS. For example, this is the case with content-addressable storage devices. This is achieved by adding support for a third interface; in addition to CIFS/SMB and Centera, Enterprise Vault also supports devices that implement the Enterprise Vault Storage Streamer API.

The first system to implement the Enterprise Vault Storage Streamer API is the Dell DX Object Storage Platform. This chapter is specifically about the performance of this system. It does not apply to devices that may be added in the future.

Dell DX Object Storage Platform

The Dell DX Object Storage Platform offers a reliable means of archiving data with the added advantage that, because multiple intra-cluster replicas can be configured, backups of archived data may not be considered necessary. For extra security, data may also be replicated to a Disaster Recovery DX Storage Cluster. In this case, it is suggested that a custom sizing is performed to ensure that the infrastructure is adequately sized for replication to keep pace with newly ingested data

Because items are secured without the need of backup, more time is available for other activity such as extending the archive window. Indexes and SQL databases are not held on the device and still require backups.

The device does not perform its own sharing. Enterprise Vault does single-instancing.

Choice of Enterprise Vault server

There is no substantial difference in performance when archiving to the Dell DX Object Storage Platform when compared with other storage media. Refer to the tables for each archiving type for the throughput rate.

Additional disk space used by databases

Additional space in the Vault Store database is required to store an item when the Vault Store partition is located on a Dell DX. In addition to any other calculations:

1. Take the number of items.
1. Multiply by 50 bytes.

Additional space in the Fingerprint database is required to store an item when the Vault Store partition is located on a Dell DX. In addition to any other calculations:

1. Take the number of items.
2. Multiply by 50 bytes.

Accelerators

Tuning for the Accelerator products differs in many ways from the rest of Enterprise Vault. It is the area that benefits most from active tuning, and you must consider the following recommendations for every installation.

The *Best Practices for Implementation* manual for Discovery Accelerator discusses the different aspects that you need to consider during sizing, and recommends best practices for implementation. The manual is available from the following page on the Symantec Enterprise Support site:

<http://entsupport.symantec.com/docs/326055>

Most of the advice in the manual applies to Compliance Accelerator as well.

Accelerator performance

The Accelerator service creates many threads of execution to manage the following concurrent activities:

- Customers' service
- Customer's tasks
- Discovery Accelerator Analytics' service

The customers' service might be handling potentially hundreds of requests from end-users using the client application.

The customers' tasks manage the background tasks and activity which includes potentially hundreds of search threads, possibly tens or hundreds of export threads, synchronization threads for Enterprise Vault and Active Directory, and a selection of threads to deal with other background activities.

In the case of Discovery Accelerator, the Analytics service may potentially be executing multiple instances of analytics case tasks to handle dynamic database table creation, case item transfer, original content retrieval from Enterprise Vault, message conversation structure analysis and automatic categorization.

A multiprocessor server is essential to ensure that all these concurrent activities can function with reasonable performance. The majority of installations should have a dedicated Accelerator server with four processors, and in larger customers this may need to be scaled out to multiple servers. Multi-core and multi CPU configurations can be used with equivalent performance expectations, but server sizing should not be based upon hyper-threading.

The Accelerator service can require a high memory capacity, particularly during searching when result sets of 50,000 results may be returned simultaneously from multiple Enterprise Vault index services.

A standard 32-bit Accelerator server should have 4 GB of RAM to handle the high memory requirements of the Accelerator while providing the operating system with sufficient resources.

A 64-bit Accelerator server should have a minimum of 4GB RAM, but additional RAM can be added to scale up the server to better handle the concurrent tasks. When using Discovery Accelerator Analytics it is beneficial to install at least 8GB RAM. However scaling up the server can only be done to a limited extent and scaling out to additional servers may become necessary.

It is recommended that the Accelerator database server and application servers are connected through gigabit network technology, particularly when implementing Discovery Accelerator Analytics. It is also highly beneficial to make the Enterprise Vault infrastructure available through gigabit technology.

Customer's tasks performance

The customer's tasks handle background activity such as Searching, Export and other maintenance tasks. Only a single customer tasks service can be configured for each customer, and the server hosting this service requires appropriate consideration for searching and export.

Factors influencing search performance

Many factors can influence the search throughput, including the following:

- Index server hardware:
 - Index location I/O performance (direct attached, network based, SAN).
 - Index file location isolation (dedicated or shared location).
 - Index file fragmentation.
 - Index service Windows system cache or external file cache (for example, SAN cache).
 - Index server memory availability.
 - Index server CPU availability.
 - Available Enterprise Vault index servers.
- Index structure:
 - Index compaction status (depends upon update frequency).
 - Index structure: attachment or item granularity.
 - Index size: number and size of files, index locations (roughly related to number of unique words), and number of documents.
- Search query:
 - Search term complexity.
 - Volume of results per index.
 - Distribution of indexes over index services.
 - Number of indexes searched.
 - Single index search frequency (multiple searches of single index are queued). This is most likely to occur during Compliance Accelerator searches across multiple departments. Each archive is searched for each department, resulting in a queue of potentially hundreds of searches for each archive.
- Accelerator server:
 - Server CPUs and memory.
 - Total number of Accelerator search threads (more is not necessarily better).
 - Concurrent user activity such as searching or accepting searches.

- Database server:
 - Database server I/O subsystem performance (critical to achieve best performance).
 - Database server memory (8 GB+ recommended).
 - Database server CPU capabilities (4 CPU recommended).
 - Table sizes and index fragmentation.

The overall search performance is a combination of the related index service performance and the Accelerator result processing that occurs at the database server.

The performance of searching an index is described in “Searching indexes” on page 86.

If several Enterprise Vault servers share the same storage location for indexes, the performance is likely to be significantly affected.

File fragmentation can significantly affect the Enterprise Vault indexes, unless you defragment them regularly. For example, depending on the fragmentation, a typical journal index may perform 60% faster after defragmentation. Index file fragmentation occurs in all situations, and this problem must be managed with a defragmentation maintenance plan on all types of storage device.

A single search can consume 100% of a CPU at the index server for the duration of the search, and therefore benefits from multiple processors. Concurrent searches can potentially consume all CPU resources, but this depends on the size of the indexes and the I/O capabilities of the index location storage device, which can quickly become a bottleneck. At the same time, the index server memory can be consumed by complex queries and concurrent searches, and it can start to use page file, adding to the I/O overhead.

On 32-bit Windows, the indexing service should have 4 GB of RAM installed, but more than 4 GB cannot be utilized by the Enterprise Vault services. If possible, install Enterprise Vault on Windows x64 with 12 GB of RAM (which can be used between the processes and Windows system file cache).

Search phase

The Accelerators service maintains by default 10 Accelerator search threads per Enterprise Vault indexing service. Therefore, 10 indexes per indexing service can be searched concurrently.

Compliance Accelerator application searches can result in hundreds or thousands of searches of a small number of indexes. All the searches for each unique index have to be queued, as a single index cannot be concurrently searched. This queuing reduces the overall throughput; thousands of searches of a single index (single threaded) will have a much lower throughput than a single search of thousands of

indexes (multi-threaded). Therefore, application searches must be carefully designed to ensure they are not creating unnecessary searches (searching archives or departments unnecessarily).

The structure of the index can influence the search time. Item granularity index schema is known to be potentially much faster to search and less resource-intensive. The index structure can affect how search terms are interpreted. By default, mailbox indexes are created using the attachment granularity schema. In Enterprise Vault 8.0 SP3 and later, all new journal archives are created using the item granularity schema. Changing the schema of existing archives requires reindexing.

The complexity of the query can also affect the search times, but specifying date ranges in particular helps to reduce the scope. The date range can prevent every index being opened and narrows the search within an index.

Searching for messages which are different sizes within an index does not influence the search.

The search phase requires memory, file I/O, and CPU resource. The complexity of the search affects the memory footprint and the duration. The I/O load depends on the memory available to the index server process, the Windows system file cache, and other competing processes that demand memory and system file cache resource.

Note: In 32-bit Windows, the Windows system cache policy affects how much memory is available in the system file cache (by enabling the Large System Cache). The default system cache provides up to 512 MB, depending on system requirements, and up to 960 MB if the large system cache is enabled. (The /3GB boot switch cannot be used with this setting, but /3GB should not be used on an Enterprise Vault server anyway.)

A simple search of a journal index volume on a 32-bit environment may result in 60 IOPS between index volume search iterations. With 10 Accelerator search threads, the overall I/O load in this scenario is likely be approximately 600 IOPS to the index location. This is sustained while searching through all the selected index volumes (the IO load varies when a search identifies hits that require retrieving).

The system memory and I/O bandwidth available during the search phase, together with the number of Discovery Accelerator search threads, can determine the rate at which indexes can be searched. This varies depending on the number of hits per index due to time spent retrieving the result metadata.

A simple search using a typical 4 CPU, 32-bit Enterprise Vault server with 4 GB of RAM and disks providing 1,200 IOPS with 10 Discovery threads should search around 5,000 indexes per hour (each index returning a few results). This rate is per Enterprise Vault server, so across three Enterprise Vault servers with independent

index storage the rate should be around 15,000 indexes per hour (depending on hits per index volume).

Note: The use of Windows x64 enables more than 4 GB of physical memory to be provided. This allows more concurrent index server processes to co-exist without competing for memory; each individual process still has its own 32-bit address space limits. Additional memory also enables the Windows system file cache to grow to a theoretical 1 TB. This helps to reduce the I/O overheads and contention between processes, and potentially enables the number of Discovery threads to be increased to improve the index iteration rate.

Result retrieval phase

The search result metadata retrieval phase involves iterating through the hits to obtain the full metadata detail from the index files. Essentially, this can result in one random I/O operation per result. This can determine the result retrieval rates for a single search. If the storage subsystem provides 2,000 random IOPS then the results are likely to be provided at a rate of around 2,000 results per second (but this is shared between all searches of index volumes using that storage device). However, the performance and impact depends on the memory available to the Windows system file cache, competitive demands on the system file cache, and available I/O bandwidth.

It is likely that the available IOPS is shared between 10 concurrent search threads and, on a typical 32-bit server, this does not provide the full potential throughput due to contention between the threads at the storage increasing latency.

The result metadata is passed back to Discovery Accelerator, which in turn adds the metadata to the Discovery Accelerator customer database. The database server then needs to further process the data (from all concurrent searches). Hence, it is possible that the collective Enterprise Vault servers may produce the results faster than they can be processed at the database. Therefore, no individual Enterprise Vault server needs to provide results at a rate greater than 4,000 per second.

Note: The use of Windows x64 enables more than 4 GB of physical memory to be provided. This allows the system file cache to grow and improves the I/O to index files. This reduces the I/O load to irregular peaks of possibly several hundred IOPS, with little physical I/O for the duration of the result metadata retrieval. Installing each Enterprise Vault server on a Windows x64 system with at least 12 GB to 16 GB of RAM should significantly improve performance and reduce the I/O load.

You can also distribute the load by spreading the indexes over many indexing services (with independent storage) and having several index locations at each

indexing service. However, if all the indexes are stored on a shared storage device, the storage device may become a significant bottleneck.

Result processing and search acceptance phases

When the results are returned to the Accelerator server, the processing can consume the majority of CPU and up to 2 GB of RAM. The impact depends upon how quickly the index servers are able to produce results. The greater the number of results concurrently returned, the greater the impact. There is potential to hit memory errors if many search threads return very large volumes of results at the same time. However, in these situations, the Accelerator service recovers from the error.

During searching the database log file tends to see sustained periods of 200 I/Os per second and the data files see peaks of activity between 500 and 800 I/Os per second.

When a search is accepted, the necessary processing is performed through stored procedures at the database server. With a correctly sized database server, a typical search can be accepted at a rate between 300–600 items per second. This tends to be very I/O intensive, and special attention needs to be paid to the database files (particularly the log) and their location. During acceptance, the database log file tends to see sustained periods of 900 I/Os per second, and the data files see peaks of activity between 500 and 3000 I/Os per second.

Accelerator tuning

The recommended Accelerator search thread tuning should help prevent CPU and memory overload, and will also reduce the I/O load at the index servers, improving the throughput.

The Accelerators will search up to 10 indexes per index server concurrently. However, if large volumes of results are returned from each index this can overload those index services, consuming CPU, memory, and very high levels of I/O, which will significantly degrade the performance as the search concurrency increases. Therefore, it may be necessary to reduce the number of search threads to balance the load evenly. The optimal value for this setting depends on the I/O performance of the index locations and the normal expected volume of results from each index.

To determine the best value, make several benchmark searches with typical search terms across the typical set of archives, and repeat with different numbers of threads. The number of threads can be set in the System Configuration search setting “Number of Vault Search Threads” (and requires a restart of the customer tasks). Start with a value of two threads and monitor if the overall throughput changes as the value is changed.

Enterprise Vault Index Service tuning

An index search requires a sustained high level of I/O during the search and potentially repeatedly if the search results in more than 50,000 items. So, the indexes must be placed on a high performance device to avoid any bottlenecks between the index process and storage medium. Bottlenecks in the I/O subsystem, such as a slow controller and disks, network-based storage, or third-party software that scans file accesses (such as anti-virus) can significantly impact performance.

The indexing service machines will benefit from multiple processors (minimum of two) and up to 4 GB of RAM. The recommended Accelerator search thread tuning should help prevent CPU, memory, and I/O overload.

If anti-virus software has been installed to scan file and network accesses, this feature should be disabled on the index servers. Support advises that any anti-virus software should exclude the index locations from its file scan due to known issues with anti-virus software corrupting indexes.

The index files quickly become fragmented on disk, even if there is a large amount of free storage capacity. This file fragmentation can cause severe performance problems, which must be managed on any index storage device. Use an automated background file defragmentation product or scheduled device defragmentation.

The item granularity index schema performs better than the default index structure of attachment granularity. Current Support advice is to change the index structure by adding the registry key below. However, you must rebuild any existing indexes, which can be time consuming and should be approached in a phased manner.

Value	Key	Content
SchemaType	HKEY_LOCAL_MACHINE \SOFTWARE \KVS \Enterprise Vault \Indexing	DWORD value. Set to 1.

The opportunistic file locking mechanism is known to cause problems when storing index files on NAS storage devices. Therefore, the current Support advice is to disable opportunistic locking at the NAS head.

The operating system boot flag /3GB must not be used on the Enterprise Vault indexing servers as this does not provide any benefit and can result in running out of system page table entries.

Note: These switches may be applicable to other non-Enterprise Vault servers in the environment, such as the SQL Server.

Factors influencing export performance

Many factors can affect the export throughput:

- The I/O performance of the Accelerator export location storage
- The I/O performance of the network connecting Enterprise Vault and the Accelerators
- The size of the exported data
- The export format
- The performance of the Enterprise Vault storage services and their storage devices (and use of EMC Centera)
- Pre-fetch cache setting
- Age and size of data (whether within the pre-fetch cache)
- Pre-fetch cache (native format) location I/O performance

The export is I/O bound, so any bottleneck in the network, Accelerator service export location, Enterprise Vault Storage service, and SQL Server impacts the throughput.

When exporting to the original format, the export can perform in a consistent linear manner without degrading as the volume of messages increases. The size of each exported message impacts the throughput, and larger messages reduce the overall message throughput.

For example, on a well-specified server, an export with an average message size of 70 KB from NAS-based Vault Store might export messages at an overall rate of 120,000 messages per hour. However, a larger average message size will reduce this throughput.

Exporting data that is stored in Vault Stores based on EMC Centera results in a high CPU utilization at the Enterprise Vault storage server (potentially impacting other activity such as journaling) and has a lower throughput. An example export of messages with average size of 70KB from a Centera Vault Store might export messages at a rate of 90,000 messages per hour. Again, the average size of the messages will impact the throughput.

Each export that is started is allocated 25 retrieval threads to obtain the messages from the Storage service and place them at the export location. Concurrent exports can result in degraded performance due to the Enterprise Vault Storage service, Accelerator service CPU, or export location I/O capabilities.

When the export is running at its best throughput, it consumes the majority of available CPU at the Accelerator server, which may reduce availability for other users such as interactive Web users and searches. The Accelerator export can also increase the CPU utilization at the Enterprise Vault storage services depending upon storage facilities. Retrieving from NTFS Vault Stores typically requires 30%

overall CPU, while using EMC Centera Vault Stores typically requires 80% overall CPU from a multi-processor server.

The export throughput can be improved with an adjustment in production threads but is limited by the I/O performance and CPU. The thread adjustment enables the impact of an export on the Accelerator server to be reduced and to provide a more consistent throughput. An increase in threads beyond 10 threads does not typically appear to increase throughput due to contention at the export location. The default of 25 threads may be too high for a typical environment, and perhaps should be lowered to 15 threads and tuned according to Accelerator export location disk performance, storage service performance, and export priority.

The overall throughput of concurrent exports is equivalent to a single export with the same total number of threads. Therefore, two exports with five threads provide the same overall throughput as one export with 10 threads. The overall throughput is shared between the exports, so each individual export takes longer to complete than if it were run on its own.

If the pre-fetch cache is enabled with native format, the export can obtain data from the cache if possible rather than the storage service, but this will depend if the items are still within the cache. This may improve performance if the storage service is unable to provide items at a good rate. However this will increase the I/O load at the accelerator server and could form a bottleneck if the pre-fetch cache is located on the same partition as the export location.

Exporting to PSTs passes through two distinct phases: the messages are first exported to original format at the rate described above, and then collected into PST files. The PST collection throughput will be based upon the size of the exported data. This particular phase is very I/O intensive at the Accelerator export location due to reading and writing large volumes of data to the same logical partition. The PST phase is generally faster than the export to native format phase, and exporting messages with an average size of 70 KB from an NAS Vault Store provides an overall throughput around 80,000 messages per hour and from Centera around 65,000 per hour.

When exports are concurrently generating PSTs, the overall throughput becomes more limited by the I/O capabilities of the export location.

An export to HTML with an average message size of 70 KB from a NAS-based Vault Store might export messages at a rate of 90,000 messages per hour. This will be impacted by the size of the messages and the location of the vault store (Centera).

Accelerator tuning

The recommended Accelerator thread tuning should help prevent CPU and memory overload.

Tune the number of production threads according to the disk performance, background priority, and expected export concurrency. Change the number of threads for the following circumstances:

- To ensure that the disks can keep up with the throughput. Otherwise, the unnecessary extra thread overhead and IO load will unnecessarily consume CPU and slow the overall throughput.
- To ensure better coexistence with interactive Web users by reducing export CPU utilization.
- To assist when more than one export is usually run concurrently (threads identified in point 1 divided by the number of normally concurrent exports).

You can tune the number of threads by changing the value of the system configuration setting "Number of Production Items Thread". The default is 25.

Storage service tuning

The Enterprise Vault Storage service machines will benefit from multiple processors (minimum of two) and up to 4 GB of RAM. The recommended Accelerator thread tuning should help prevent CPU overload.

Customers' service performance

The Accelerator customers' service is the main access point for the end-user client application, which provides access to all product features.

The Accelerator customers' service requires memory for preview, reviewing, and administrative tasks, and also demands high CPU requirements for concurrent multi-user reviewing tasks. If the online preview and reviewing facilities are to be used on anything other than a small customer environment, the customers' service should be hosted on a separate server.

The Accelerator customers' service benefits from multiple processors. Using the /3GB boot flag is not recommended.

A four processor server (multi-core and/or multi CPU configuration) with 4 GB of RAM should typically support up to 600 concurrent reviewers, but this does depend on the working practices of those users. More than 600 concurrent reviewers will require multiple servers to be deployed in a network load balancing (NLB) arrangement.

The customers' service does not require NLB to be configured with IP affinity.

Network sockets

You must adjust the customers' service server TCP/IP socket parameters to provide the environment with adequate network sockets at a sufficient reusable rate. To do this:

1. Locate the following key in the Windows registry:
HKEY_LOCAL_MACHINE\SYSTEM\CurrentControlSet\Services\Tcpip\Parameters
2. Update the following values, or create them if they do not already exist (assuming a dedicated customers' service server with a minimum of 2 GB of RAM):

Name	Type	Default	Recommended (decimal)
MaxUserPort	DWORD	5,000	64,512
TCPTimedWaitDelay	DWORD	240	120
MaxFreeTcbs	DWORD	2,000	65,536
MaxHashTableSize	DWORD	512	16,384

Discovery Analytics' service performance

The Discovery Accelerator analytics' service demands high CPU and memory requirements for its processing tasks at both the database server and the relevant Accelerator server during potentially long-running processing.

When a Discovery case is enabled for analysis, a set of tasks is started which run through several different phases to transfer all existing case items and their original content into a set of new per-case analytics tables.

The tasks started for each case enabled for analysis are as follows:

- Dynamic filegroup, table and FTS index creation.
- Archive property synchronization with Enterprise Vault (directory sync task).
- Initial analytics table population from source Discovery Accelerator tables (add items task).
- Original archived item content retrieval and insertion into analytics tables (ingester task).
- Message conversation structure analysis processing (conversation analyser task).
- Automated categorization rule engine processing (rule engine task).

Once the filegroup and tables have been created, the other phases start as soon as there is sufficient data to perform work. For example, as the add items task starts to populate the tables, the ingestion task is then able to start retrieving content data from Enterprise Vault. When original item data starts to be inserted, the conversation analysis can start working on conversation structure.

The new filegroup is created in one of the Analytics file locations specified during customer configuration. It is recommended that multiple locations are created to distribute the workload when multiple cases are enabled for analysis. When a case is enabled for analysis, the filegroup location is selected in a round-robin method.

Each filegroup location specified should be an independent array of high speed disks (as detailed in the Accelerator database server section). The IO activity is particularly high during the processing when a case is enabled for analysis, and therefore concurrent (parallel) case escalations can result in very high database loads. Once a case has been enabled for analysis, the IO activity for that case tends to be very low. Therefore, the main concern is balancing the load during the escalation processing. It would be worth specifying at least two independent partitions with up to as many partitions as the expected typical concurrent case escalations (parallel escalations, not the total number of cases in analysis).

The ingester and conversation analysis tasks are resource-intensive at the Accelerator server running the analytics' service. These tasks typically require 30% to 50% CPU on a four-processor system and up to 2 GB additional RAM on top of existing Accelerator requirements during escalation of a single case. Using a 64-bit server with 8 GB of RAM is recommended to handle the concurrent activity.

These tasks also place a high load upon the database server, and a single case escalation can require at least 8 GB of RAM and 80% of four processors at the database server.

The ingester task creates in memory retrieval work queues for each source Enterprise Vault server and allocates a pool of 10 threads by default to process those queues. By default, no more than two threads operate on any one queue, and the pool of 10 threads should be evenly distributed between the queues. Therefore, if there are 10 or more source EV servers then no queue has more than one thread working at any time.

The typical ingestion throughput for data of 70 KB average size and equipment as described might be up to a maximum of 200,000 top level messages (and their attachments) per hour (40,000 top level messages and their attachments per hour per retrieval thread). The throughput levels off around 200,000 per hour due to the rate at which content data can be inserted and processed at the database server.

In the situation where there are fewer than 10 source Enterprise Vault servers, a number of retrieval threads will effectively be idle due to the limit of two threads per source Enterprise Vault server queue, and the impact to the throughput becomes noticeable if there are fewer than three source Enterprise Vault servers (fewer than

six active threads). The limit of two threads per source Enterprise Vault server has been determined to ensure appropriate resource sharing and reduce the impact to the source Enterprise Vault servers.

The impact to a source Enterprise Vault server for two retrieval threads is typically around 30% CPU, 200 MB of RAM and raised level of I/O activity to the vault store partitions. The ECM API is used to retrieve the content data, and therefore the storageonlineopns.exe process will be seen to use the additional resources.

Concurrent (parallel) case escalations demand higher resources at all tiers (Accelerators, Enterprise Vault and database server).

If Discovery Accelerator Analytics is to be used in a large customer environment, the analytics' service should be hosted on a separate server or potentially scaled out to multiple servers if many concurrent (parallel) case escalations are typically expected.

The database server may need to be scaled up to handle the expected throughput of multiple case escalations alongside the normal Discovery Accelerator core operations.

The conversation analysis task runs through several phases. The first phase of collating the conversation data should complete at a similar time to the ingester task. The next phase, which constructs the conversation structure, can then run for sometime afterwards, but this depends on the total number of distinct conversations in the dataset. The worst case is that 50% of the dataset is conversation starters, and in this situation the conversation analysis can run for another 12% of the overall ingestion time.

Accelerator database server

We recommend that an Accelerator database server should have at least four processors and large customers deploying Discovery Accelerator analytics may benefit from more processors— multi-core and multi CPU configurations can be used with equivalent performance expectations, but sizing must not be based on hyper-threading.

The Accelerator database can be hosted on either 32-bit (x86) or 64-bit (x64 only) platforms. The 64-bit platform can provide performance benefits due to memory enhancements.

The Accelerator database stored procedures employ methods that are very memory-intensive. In addition, the flow of data between many tables and their indexes causes a large number of pages to be required in memory at any time, consuming large volumes of memory.

The database server requires enough memory to ensure that the data manipulation does not cause excessive paging to disk both at the Accelerator database and tempdb,

which will quickly degrade the performance. A small environment should have a minimum of 4 GB of RAM and a medium or large environment should have at least 8 GB of RAM. A customer deploying Discovery Accelerator analytics should have at least 8 GB of RAM and a large customer at least 16 GB of RAM.

The storage capacity of the Discovery Accelerator database server needs to be carefully considered using the appropriate database sizing guide. However, in initial deliberations, the following very high-level rule of thumb could be used:

$$\text{Capacity per year (GB)} = ((S * 2.7) + (E * 2.2)) / 1,000,000$$

Where:

S - Total number of items captured in all searches performed in all cases/departments per year.

E - Total number of items exported/produced in all cases/departments per year.

Note: This calculation provides a high-level calculation, but do not rely on it alone for final sizing. It does not take into account the operating system, paging and log file devices, or any transient storage required for unaccepted searches. It also does not include any additional capacity that may be required during product version upgrades.

The Discovery Accelerator analytics tables grow to a similar size as the total size of all included original items and their attachments. Analytics operates on the original item and attached item converted content provided by Enterprise Vault, so binary data such as image files will not consume space, which can reduce the overall average item size.

The converted content is encapsulated within XML and contains other metadata, and many of the analytics table columns are full text indexed which adds further overhead, which slightly raises the average item size.

Estimating the size of the analytics tables for each individual enabled case could become complicated to take into account all the different characteristics of the source data such as volume, size distribution, recipient volume and distribution, attachment volume, types and distribution, and number of unique conversations. The Discovery Accelerator source table tblIntDiscoveredItems contains some relevant details but these are not sufficiently accurate to provide estimated sizes.

However the filegroup partitions need to be sized to encapsulate multiple enabled cases of varied size with potentially varied characteristics. Therefore, a high level estimate needs to be performed based upon a high level estimated maximum number of items in analysis at any time, combined with a high level estimated average size (using 25% of total original size as the converted size) and overheads of an additional 50% of the converted size.

For example, if no more than five million items will be in analysis across all cases at any time, the total capacity of all filegroup partitions should need to be around $300 \text{ GB} + 20\% = 360 \text{ GB}$. This is based upon an original item average size of 160 KB including attachments (with 20% of messages containing 1 or 2 attachments therefore making up an extra 30% of items –6.5 million total items). Once this has been converted, and overheads added, it is equivalent to an average item size in the database of 60 KB, including attachments.

The type of storage and interfaces used must ensure that the storage does not become a bottleneck. LAN-based storage should never be used for the database files. The best devices are local storage, direct attached storage or partitions on an appropriately sized storage area network (SAN).

Each database requires the disks to be arranged for two different purposes: the database data and log files. The data files require good random access or a high number of IOs per second, and therefore a striped array of many disks should be used (using hardware-based RAID and not software). The log files require good sequential write performance so each log file should be placed on its own high speed disks with good transfer rates.

To achieve redundancy on the striped arrays (data) while maintaining performance, the RAID scheme should be carefully considered. RAID level 5 is a popular cost-effective method of achieving redundancy while maintaining striped disk read performance. However, writing incurs a cost of four physical writes per operation. Therefore, a poorly-sized RAID-5 implementation could significantly reduce the performance in the Accelerator database write-intensive environment. Correctly sizing a RAID-5 implementation may become more costly than RAID-10, and therefore a RAID-10 scheme should be considered to maintain the write performance.

To achieve redundancy on the sequential write-intensive disks (log), use a RAID-10 scheme with high speed, 15,000 rpm disks.

In the case of local or direct attached storage, multiple controllers supporting multiple channels should be used to distribute the load and provide sufficient throughput. The controllers should provide a large capacity battery-backed read and write cache. 512MB controller cache is recommended for local or direct attached storage.

Before using partitions on a SAN, the I/O load needs to be considered along with any other applications already using the SAN to ensure that the performance can be maintained. Ideally, the SQL Server implementation should be discussed with the SAN hardware vendor to ensure that optimal performance is achieved. Typically, LUNs should be created across as many suitable disks as possible, using entire disks rather than partial disks to prevent multiple I/O intensive applications from using the same disks.

The database server storage should be arranged to accommodate the different types of data. Typically, database servers should have the following partitions:

- System drive (RAID-1 array).
- Tempdb log partition (Single disk or RAID-1 array).
- Tempdb data partition (RAID-0 or 10 striped array of several drives).
- Each customer database log file partition (RAID-10 array).
- Each customer database data file partition (RAID-10 striped array of many drives).
- Discovery Accelerator only: Multiple partitions to accommodate analytics database data and FTS data (several RAID-10 striped arrays of many drives).

Ensure that these partitions only contain the related database files so that those files do not become fragmented on disk.

When the Discovery Accelerator customer database is created, alter the default data file size to a large value, perhaps representing the next year's activity. This prevents file fragmentation and wasted I/O growing the files.

Implementing an Accelerator database maintenance plan is absolutely essential to preserve the database performance. As the database is used, crucial tables and their indices fill and become fragmented. This reduces overall performance and leads to additional I/O when updating (due to page splitting). In addition, certain tables can significantly change profile each day, which results in out-of-date statistics, potentially causing inefficient query plans. Particular attention should be paid to `tblIntDiscoveredItems`, for which statistics should be frequently updated and indexes regularly maintained.

See the relevant Accelerator documentation for maintenance plan advice.

Document conversion

A proportion of the CPU power is used to convert documents to HTML for indexing. This section explains how processor power can be saved and throughput improved by changing the values of registry keys that control conversion. Registry edits should be made with care, and the registry should be backed up before making any changes.

It should be noted that all ingest rates in the document are based on a system with the default registry settings.

Converting to HTML or text

By default, items are converted to HTML. This provides text suitable for indexing and allows a formatted display when items are read in HTML from Archive Explorer or the WebApp search. The original item is also stored, and this is what is displayed when downloading an item—for example, by opening a shortcut or viewing an item from the integrated browser.

It is more CPU-intensive to convert items to HTML than to text, so you can minimize CPU usage by converting some or all items to text. For Word documents and Excel documents there are special registry keys that you can change to force this, and for all file types there is a registry key. The general registry key can also be used for Word and Excel and must be used for Office 2007 files.

Because of the expense of converting, Excel files are converted to text by default. The Office 2007 file types for Excel, Word and PowerPoint files have a different extensions—xlsx, docx, and pptx (or xlsx, docm, and pptm when macro-enabled)—and these must be added to the general registry key to force conversion to text.

Value	Key	Content
ConvertWordToText	HKEY_LOCAL_MACHINE \SOFTWARE \KVS \Enterprise Vault	String value. 0 (default) - Convert to HTML, 1 - Convert to text.
ConvertExcelToText	HKEY_LOCAL_MACHINE \SOFTWARE \KVS \Enterprise Vault	String value. 0 - Convert to HTML, 1 (default) - Convert to text.
TextConversionFileTypes	HKEY_LOCAL_MACHINE \SOFTWARE \KVS \Enterprise Vault	String value containing list of file types. The list format is: .filetype[.filetype]. Each file type must be prefixed by a period, and the list must end with a period. For example: .DOC.XLS.XLSX.XSLM. All file types can be converted to text by using the * wildcard character. For example, a value of *. converts all file types to text.

Excluding files from conversion

In order to be indexed, items that are not already text must be converted to text or HTML. Some files are excluded from conversion because they contain no textual content, such as JPEG files.

Unknown file types are opened and the first few characters are checked for textual content. Some files may look like text files because they contain valid characters, but they should not be treated as such and should be specifically excluded. One consequence of not excluding them is that the index may become full of meaningless words.

Value	Key	Content
ExcludedFileTypes	HKEY_LOCAL_MACHINE	String value containing list of file types.
FromConversion	\SOFTWARE	The list format is:
	\KVS	. filetype[. filetype].
	\Enterprise	Prefix each file type with a period and
	Vault	end the list with a period. For example:
		. GIF. JPG.

Conversion timeout

Large and complex items can take a long time to convert and slow down the whole system during conversion. To prevent such conversions from running forever and preventing other work, there is a conversion timeout mechanism. All conversions are abandoned after 10 minutes. Items are normally converted in a fraction of a second, but if conversions are constantly being abandoned—this is an event in the event log—this time can be reduced so that the conversions are abandoned earlier and waste less time. Reducing the time may mean that some items do not have their content indexed; the metadata is still indexed and the item archived as normal.

Value	Key	Content
ConversionTimeout	HKEY_LOCAL_MACHINE	DWORD value. Default:
	\SOFTWARE	10 (minutes).s
	\KVS	
	\Enterprise Vault	

Monitoring your system

This section gives some suggestions on what to monitor to check that your system is performing as expected.

Using Performance Monitor

Performance Monitor provides useful information on the overall performance of the system and may show up potential bottlenecks.

The following is a suggestion of what to collect for a full monitoring of an Enterprise Vault system. You can use a file containing the following parameters to set up Performance Monitor.

```
<HTML>
<HEAD>
<META HTTP-EQUIV="Content-Type" CONTENT="text/html;" />
<META NAME="GENERATOR" Content="Microsoft System Monitor" />
</HEAD>
<BODY>
<OBJECT ID="DISystemMonitor1" WIDTH="100%" HEIGHT="100%"
CLASSID="CLSID:C4D2D8E0-D1DD-11CE-940F-008029004347">
<PARAM NAME="_Version" VALUE="196611"/>
<PARAM NAME="LogName" VALUE="WinSys_NET"/>
<PARAM NAME="Comment" VALUE="Standard Enterprise Vault monitors"/>
<PARAM NAME="LogType" VALUE="0"/>
<PARAM NAME="CurrentState" VALUE="0"/>
<PARAM NAME="RealTimeDataSource" VALUE="1"/>
<PARAM NAME="LogFileMaxSize" VALUE="-1"/>
<PARAM NAME="DataStoreAttributes" VALUE="34"/>
<PARAM NAME="LogFileBaseName" VALUE="WinSys_NET"/>
<PARAM NAME="LogFileSerialNumber" VALUE="1"/>
<PARAM NAME="LogFileFolder" VALUE="C:\PerfLogs"/>
<PARAM NAME="Sql Log Base Name" VALUE="SQL:!WinSys_NET"/>
```

```
<PARAM NAME="LogFileAutoFormat" VALUE="1"/>
<PARAM NAME="LogFileType" VALUE="2"/>
<PARAM NAME="StartMode" VALUE="0"/>
<PARAM NAME="StopMode" VALUE="0"/>
<PARAM NAME="RestartMode" VALUE="0"/>
<PARAM NAME="LogFileName" VALUE="C:\PerfLogs\WinSys_NET_000001.blg"/>
<PARAM NAME="EOFCommandFile" VALUE="" />
<PARAM NAME="Counter00001.Path" VALUE="\Cache\*" />
<PARAM NAME="Counter00002.Path" VALUE="\LogicalDisk(*)\*" />
<PARAM NAME="Counter00003.Path" VALUE="\Memory\*" />
<PARAM NAME="Counter00004.Path" VALUE="\MSMQ Queue(*)\*" />
<PARAM NAME="Counter00005.Path" VALUE="\Network Interface(*)\*" />
<PARAM NAME="Counter00006.Path" VALUE="\PhysicalDisk(*)\*" />
<PARAM NAME="Counter00007.Path" VALUE="\Process(*)\*" />
<PARAM NAME="Counter00008.Path" VALUE="\Processor(*)\*" />
<PARAM NAME="Counter00009.Path" VALUE="\Server\*" />
<PARAM NAME="Counter00010.Path" VALUE="\System\*" />
<PARAM NAME="CounterCount" VALUE="10"/>
<PARAM NAME="UpdateInterval" VALUE="60"/>
<PARAM NAME="SampleIntervalUnitType" VALUE="1"/>
<PARAM NAME="SampleIntervalValue" VALUE="60"/>
</OBJECT>
</BODY>
</HTML>
*****
```

The following are counters of particular interest.

■ **\Processor(_Total)\% Processor Time**

When archiving at its maximum rate, a well-tuned Enterprise Vault server is expected to use between 70% and 100% of the total CPU available. If the value is less than this, there may be a bottleneck elsewhere in the system.

■ **\Process(*)\% Processor Time**

The system may be using CPU but not getting the expected archiving rate. A process may be consuming more CPU than expected, or an application other than Enterprise Vault may be consuming CPU.

During archiving, you can expect to see the following processes consuming CPU:

- The archiving task. The name varies according to type of archiving.
- StorageArchive.exe, used when mailbox or journal archiving. For other archiving types, the StorageArchive functions are within the Archiving task.

- DirectoryService.exe. A small amount of CPU is used by the Directory service.
- IndexServer, used to index.
- EVConverterSandbox, used to convert items into HTML.
- StorageCrawler, fetches items from Storage for indexing.

Some processes are recreated frequently such as EVConverterSandbox. This can lead to misleading results from Performance Monitor, with spurious very high values.

In addition, some processes may be active that are concerned with Windows authorization and permission checking and for controlling services.

- Lsass.exe
- Crss.exe
- Svchost.exe

- \LogicalDisk(*)\Avg. Disk Queue Length
\LogicalDisk(*)\Disk Transfers/sec

The local disks may be overloaded. The Transfers/sec counter can tell you whether you have a high transfer rate, whereas the Disk Queue counter can tell you whether the disk has problems servicing the requested transfers. In general, the disk queue should not exceed 2 (minus the number of active spindles) for extended periods.

- \LogicalDisk(*)\Free Megabytes

It is useful to know the rate at which space is being used on disks. For example, by looking at the disk on which indexes or vault stores are held, you can calculate the space taken by each archived item. Very often when the data is not stored on NTFS devices, the NAS device supplies its own Performance Monitor Counters.

- \MSMQ Queue(<server>\private\$\enterprise vault storage archive)\Messages in Queue

When items are archived from an Exchange Server, they are retrieved from the Exchange Server and placed on the Storage Archive queue. See "Setting the number of connections to the Exchange Server" on page 29 for a description of the expected values and actions to take.

Using SQL Performance Monitor counters to measure archiving rate

There are no inbuilt performance counters to show the current archiving rate. It is possible to improvise by using a SQL query and posting the result of the query to one of the SQL user-settable counters. The archiving rate can then be viewed with the other Performance Monitor counters.

The following SQL runs continuously, updating the performance counters every minute with the number of archived items and the hourly archiving rate. Edit the name of the vault store. You can run this as a SQL job. The job can fail in SQL 2005 although it will run as SQL query.

```
DECLARE @CCOUNT INT
DECLARE @CCOUNT_LAST INT
DECLARE @CCOUNT_DIFF INT
use evname_of_vaultstore
WHILE 1=1
BEGIN
select @CCOUNT = (select rows from sysindexes where name='pk_saveset')
EXEC sp_user_counter1 @CCOUNT
SET @CCOUNT_DIFF = ((3600 / 60) * (@CCOUNT - @CCOUNT_LAST))
EXEC sp_user_counter2 @CCOUNT_DIFF
SELECT @CCOUNT_LAST = @CCOUNT
WAITFOR delay '00:01:00.00'
END
```

Useful SQL

The following SQL snippets can either be pasted directly in SQLQuery or created as separate files. All should be run in the context of the Vault Store database.

Hourly archiving rate

This shows for each hour:

- The time.
- The number of items archived in that hour.
- The megabytes archived—original size as reported by MAPI when archiving. The actual original size in Exchange may be larger for very small items.
- The megabytes archived—compressed size.

If only the last 24 hours is required, you can uncomment the commented line.

This can be used to check that the archiving rate is as expected and that the compression rate is as expected.

This query imposes a load on the database and may take a few minutes for Vault Stores with over 50,000,000 archived items.

```
select "Archived Date" = left (convert (varchar, archiveddate,20),14),
"Hourly Rate" = count (*),
"MB (original)" = sum (originalsize)/1024/1024,
"MB (compressed)" = sum (itemsized)/1024
from saveset,savesetproperty
where saveset.savesetidentity = savesetproperty.savesetidentity
--and archiveddate > dateadd("hh", -24, getdate ())
group by
left (convert (varchar, archiveddate,20),14)
order by
"Archived Date"
desc
```

"OriginalSize" is a field introduced in Enterprise Vault 7.0. For earlier versions, the following can be used. This has no information on the original size.

```
select "Archived Date" = left (convert (varchar, archiveddate,20),14),
"Hourly Rate" = count (*),
"MB (compressed)" = sum (itemsized)/1024
from saveset
--where archiveddate > dateadd("hh", -24, getdate ())
group by
left (convert (varchar, archiveddate,20),14)
order by
"Archived Date"
desc
```

Archiving rate/minute

It can sometimes be useful to know the archiving rate per minute. For example, suppose that the archiving rate is high for a few minutes and then low for an extended period. This may indicate that the system is capable of achieving high archiving rate, but some other factor such as a lack of items to archive may give the impression of a slow archiving rate.

```

select "Archived Date" = cast (archiveddate as smalldatetime),
"Minute Rate" = count (*),
"MB (original)" = sum (originalsize)/1024/1024,
"MB (compressed)" = sum (itemsized)/1024
from saveset, savesetproperty
where
saveset.savesetidentity = savesetproperty.savesetidentity
-- and archiveddate > dateadd("hh", -24, getdate ())
group by
cast (archiveddate as smalldatetime)
order by
"Archived Date"
Desc

```

For versions earlier than Enterprise Vault 7.0, use the following SQL without the original item size.

```

select "Archived Date" = cast (archiveddate as smalldatetime),
"Minute Rate" = count (*),
"MB (compressed)" = sum (itemsized)/1024
from saveset
--where archiveddate > dateadd("hh", -24, getdate ())
group by
cast (archiveddate as smalldatetime)
order by
"Archived Date"
desc

```

Archived file sizes

It is often useful to know the size and compression ratio of files being archived. This permits an understanding of the pattern of email. For each file size by KB, the following SQL shows the number of files archived at that file size and the total original and archived size. Because of integer truncation, an original size of 0 means a size between 0 KB and 1 KB, and so on. For small files, the actual original sizes may be larger than reported.

```
select
"Count" = count (*),
"Original Size" = originalsize/1024, "Archived Size" =
sum(itemsize)/count(*),
"KB (orig)" = sum (originalsize)/1024,
"KB (Arch)" = sum (itemsize)
from saveset,savesetproperty
where saveset.savesetidentity = savesetproperty.savesetidentity
group by
originalsize/1024
```

Using Log Parser

Log Parser is a powerful, versatile tool that provides universal query access to text-based data such as log files, XML files and CSV files, as well as key data sources on the Windows operating system such as the event log, registry, file system, and Active Directory. The tool is freely downloadable from the following page on the Microsoft site:

<http://www.microsoft.com/downloads/details.aspx?familyid=890cd06b-abf8-4c25-91b2-f8d975cf8c07>

Log Parser and IIS logs

The following are examples of using queries to extract information from the IIS logs. IIS logs are created daily in the folder

C:\WINDOWS\system32\LogFiles\W3SVC1.

Turn on the following settings in IIS before you run this request:

- Method (cs-method)
- URI Stem (cs-uri-stem)
- URI Query (cs-method)
- Bytes Sent (sc-bytes)
- Bytes Received (cs-bytes)
- Time Taken (time-taken)

Example 1 - Getting daily hits for all IIS requests

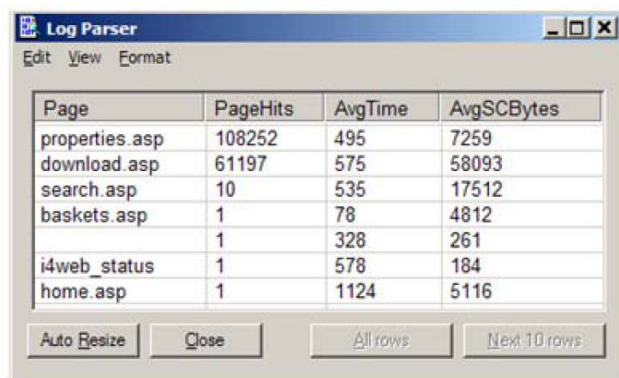
Create a file that contains the following text and name it, for example, Queries1.sql. Substitute the name of the log file as necessary.

```
SELECT extract_filename (cs-uri-stem) AS Page, COUNT(*) AS PageHits,  
avg (time-taken) as AvgTime, avg(sc-bytes) as AvgSCBytes  
FROM  
\\MyServer\c$\windows\system32\logfiles\w3svc1\ex061020.log  
where sc-status = 200  
GROUP BY Page  
ORDER BY PageHits DESC
```

Run this file as follows:

```
Logparser file:Queries1.sql -i:iisw3c -o:datagrid
```

Log Parser displays the results as in the following example:



The screenshot shows the Log Parser application window with a menu bar (Edit, View, Format) and a table of results. The table has four columns: Page, PageHits, AvgTime, and AvgSCBytes. The data is sorted by PageHits in descending order. At the bottom of the window, there are four buttons: Auto Resize, Close, All rows, and Next 10 rows.

Page	PageHits	AvgTime	AvgSCBytes
properties.asp	108252	495	7259
download.asp	61197	575	58093
search.asp	10	535	17512
baskets.asp	1	78	4812
	1	328	261
i4web_status	1	578	184
home.asp	1	1124	5116

For example, this shows that there were 108,252 downloads using `properties.asp` (called when a user downloads an HTML version from the Web Browser). On average each download took 0.495 seconds, and the average length of the data downloaded was 7,259 bytes.

Example 2: Getting download rates

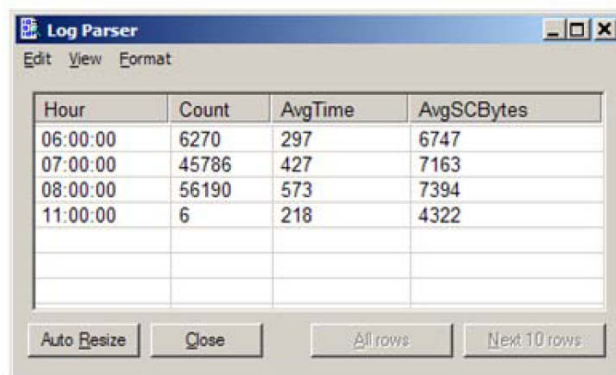
Create a file that contains the following text and name it, for example, Queries2.sql. Substitute the name of the log file as necessary.

```
Select QUANTIZE( Time, 3600 ) as Hour, count (*) as Count,
avg (time-taken ) as AvgTime,avg(sc-bytes) as AvgSCBytes
from \\MyServer\c$\windows\system32\logfiles\w3svc1\ex061020.log
where (cs-uri-stem like '%%properties.asp%%')
and sc-status = 200
group by
QUANTIZE( Time, 3600 )
```

Run this file as follows:

```
logparser file:Queries2.sql -i:iisw3c -o:datagrid
```

Log Parser displays the results as in the following example:



The screenshot shows the Log Parser application window with a menu bar (Edit, View, Format) and a table of results. The table has four columns: Hour, Count, AvgTime, and AvgSCBytes. The data is grouped by hour, showing statistics for 06:00:00, 07:00:00, 08:00:00, and 11:00:00. At the bottom of the window, there are four buttons: Auto Resize, Close, All rows, and Next 10 rows.

Hour	Count	AvgTime	AvgSCBytes
06:00:00	6270	297	6747
07:00:00	45786	427	7163
08:00:00	56190	573	7394
11:00:00	6	218	4322

For example, this shows that between 7 a.m. and 8 a.m. there were 45,786 downloads using `properties.asp`. The average time per download was 0.427 seconds, and the average size was 7193 bytes.

Use the results from the first query and substitute the other download or search methods for `properties.asp` in the query. For example:

- `Download.asp` - download from the integrated search.
- `Viewmessage.asp` - download from Archive Explorer.
- `Search.asp` - Browser Search.
- `Searcho2k.asp` - Integrated Search.

Getting index statistics

You can use Log Parser to obtain indexing statistics from `updates.log`, which is a new file in Enterprise Vault 7.0. `updates.log` is created in the indexing folder for every archive.

Example 3: Getting indexing rates for an archive

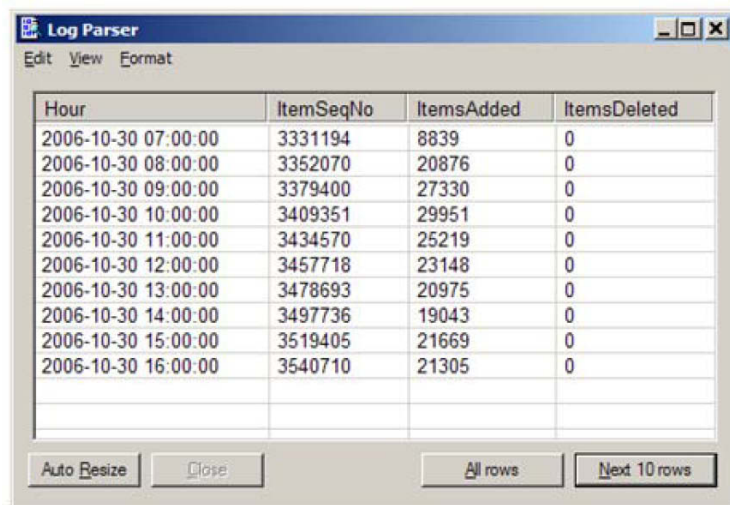
Create a file that contains the following text and name it, for example, `Queries3.sql`. Substitute the name of the log file as necessary.

```
select QUANTIZE( Field1, 3600 ) as Hour,
max(to_int(extract_token(Field5,1,'='))) as ItemSeqNo,
sum (to_int(extract_token(Field7,1,'='))) as ItemsAdded,
sum (to_int(extract_token(Field9,1,'='))) as ItemsDeleted
from
'Indexes\1D8BE8BF16B78D74989E0A47F61E4A40C1110000server.dom.com\updates
.log'
where (Field2 = 'M')
group by Hour
```

Run the query as follows:

```
logparser file:Queries3.sql -i:tsv -o:datagrid -headerrow:off
```

Log Parser displays the results as in the following example:



The screenshot shows the Log Parser application window with a menu bar (Edit, View, Format) and a table of results. The table has four columns: Hour, ItemSeqNo, ItemsAdded, and ItemsDeleted. The data shows hourly statistics for the date 2006-10-30 from 07:00:00 to 16:00:00. The ItemSeqNo values increase by 1000 each hour, while ItemsAdded and ItemsDeleted values fluctuate. At the bottom of the window, there are buttons for 'Auto Resize', 'Close', 'All rows', and 'Next 10 rows'.

Hour	ItemSeqNo	ItemsAdded	ItemsDeleted
2006-10-30 07:00:00	3331194	8839	0
2006-10-30 08:00:00	3352070	20876	0
2006-10-30 09:00:00	3379400	27330	0
2006-10-30 10:00:00	3409351	29951	0
2006-10-30 11:00:00	3434570	25219	0
2006-10-30 12:00:00	3457718	23148	0
2006-10-30 13:00:00	3478693	20975	0
2006-10-30 14:00:00	3497736	19043	0
2006-10-30 15:00:00	3519405	21669	0
2006-10-30 16:00:00	3540710	21305	0

This provides the following information for this index for each hour in updates.log:

- The index sequence number
- Items added in the hour
- Items deleted in the hour

Miscellaneous

VMware ESX Server

Virtual Servers offer a great deal of convenience when running Enterprise Vault. Enterprise Vault has been tested against VMware 3.5 and VMware 4.

Both versions of VMware offer similar performance when running Enterprise Vault. They have been shown to meet the performance rates specified elsewhere in this guide.

To achieve the best performance, you must properly specify and configure a system that is running VMware. This requires some expertise in VMware, which is outside the scope of this guide.

NTFS collections

After archiving, items may be collected into CAB files of 10 MB, allowing quicker backups or transfer to tertiary storage. The collection process is fairly lightweight and normally runs during the day. There are more files to collect in Enterprise Vault 8.0 but items are collected at least as fast as they are ingested.

Note that when a new item is added to an item that has been collected then the item is not retrieved to add the new sharer, but the file or its indexable content will be retrieved for indexing. This can be an expensive process so it is advisable to defer the collection until all items in the collection that are referenced by items in other Vault Store partitions have been archived. For example, if the policy is to archive all items older than four weeks, defer collections on journaled items that are archived immediately until after four weeks. If not done, more CPU is used on the Enterprise Vault server and more data is transferred from the Vault Store partition containing the CAB files.

Note that NTFS collections are not connected with Centera collections.

Export to PSTs

On a server with four cores, items may be exported to PSTs at the rate of 30,000 items an hour. If a higher rate is required, five export processes can be started, giving a maximum rate of 100,000 items an hour.

Storage expiry

Items eventually pass their retention category and must be deleted from Enterprise Vault. Storage Expiry can be concurrent with archiving, but this slows down both archiving and deletion of items. On a server with four cores, items are expired at the rate of 100,000 items an hour. The rate is higher when items are stored in Centera collections where it is up to 200,000 items an hour. This figure applies to items that were archived in versions before Enterprise Vault 8.0. For items that are archived in Enterprise Vault 8.0, assume that they are deleted at least as fast as they are archived.

Provisioning

The Exchange Server Provisioning task completes its work in one or two minutes only. On a site where there are 100,000 users, provisioning takes about 20 minutes.

The Domino Provisioning task is slower taking about an hour for every 10,000 users but in this case provisioning will run concurrently with archiving.

Index rebuild

On a server with four cores, an index on a fast storage device will be rebuilt at a rate of 100,000 items an hour.

When the Enterprise Vault System is actively archiving, this rate falls to 60,000 items an hour.

Backtrace

Introduced in Enterprise Vault 9.0.3, the Backtrace feature collects a limited volume of trace into memory for each Enterprise Vault process and outputs that trace to a log file when a (filterable) event log warning or error occurs. This is a complementary mechanism to DTrace, and you can use it in all scenarios where you can capture DTrace.

Backtrace is not enabled by default, but Symantec Support may advise you to enable it. You may also choose to enable Backtrace yourself so that, if a problem arises, you can include the trace with the description of the problem when you contact Support. This will usually lead to a quicker diagnosis and resolution of the problem.

There is an overhead when you enable Backtrace. CPU usage can increase by 5% to 15%, and this may have an effect on throughput rates. If the CPUs have spare capacity, there is little or no impact on perceived performance. However, if the CPUs are fully loaded, there can be up to a 15% drop in throughput, most notably when ingesting.

32-bit and 64-bit platforms

In general, the main difference between 32-bit and 64-bit Windows is that the latter provides access to larger capacity memory. This enables software to take advantage of the faster physical RAM during processing over slower storage such as disk, which can help to improve performance.

The Intel Architecture-64 technology also introduces support for parallel instruction processing designed to increase performance, but requires software specifically designed to support this technology.

32-bit Windows

The x86 platform (also referred to as “IA-32”, “Intel Architecture 32”, and “x86-32”) consists of 32-bit processors that are able to address 32-bit memory (4 GB of RAM). Pentium Pro processors onwards have been extended to be capable of addressing 36-bit memory (64GB) via “Physical Address Extensions” which the operating system needs to support. Windows 2000 onwards supports Physical Address Extensions which can be enabled with the boot flag /PAE. However, the application software needs to have been written to use the Address Windowing Extensions (AWE) API in order to access the extended memory. The extended memory also has drawbacks on performance and system memory (only half the system page table entries are available).

The operating system has access to 2 GB of virtual address space for its purposes.

An individual process can address 2 GB which can be extended to 3 GB by enabling operating system support through the /3GB boot flag. However, the process must be compiled with LargeAddressAware flags. Using /3GB also reduces the space available to the operating system by 1 GB and, importantly, it reduces the space for system page table entries. This limits the total system memory to 16 GB. 16-bit applications are still supported on the 32-bit Windows.

Enterprise Vault and Accelerators on 32-bit Windows

Enterprise Vault and the Accelerators have primarily been developed upon 32-bit Windows. The Enterprise Vault binaries are compiled as 32-bit and Accelerators use .NET technology. These are both fully supported as per the Certification tables.

64-bit Windows: Windows x64

The x64 platform (generally referred to as “64-bit” but also “AMD64”, “x86-64”, “Intel 64”, “EM64T”, and “IA-32e”) consists of a 64-bit extended processor capable of executing instructions in a 32-bit or 64-bit “mode”. Current processors can address 40-bit memory (up to 1 TB of RAM). The 64-bit mode instruction set is basically the x86 32-bit instruction set with some adjustments to certain instruction encodings to support new registers and instructions.

The processor can support either 32-bit operating systems (which will run as per x86-32, including 16-bit applications) or the x64 operating systems. When running an x64 operating system (which runs in 64-bit mode), all 32-bit applications run through a 32-bit handling layer known as Windows32 On Windows64 (WOW64). This performs certain file and registry mappings, and also handles switching the processor between 32-bit and 64-bit “modes” (the 32-bit software is still running 32-bit instructions natively). 16-bit applications are no longer supported on Windows x64.

A 32-bit process can only load 32-bit DLLs and still has 2GB memory limit (which is extended to 4GB if the process was compiled with LargeAddressAware flags) and applications which use the 36-bit AWE interface are also supported.

A 64-bit process can only load 64-bit DLLs and can access 8TB of virtual address space. The 32/64-bit Windows APIs remain very similar, and a 32-bit process can communicate with a 64-bit process through the normal inter-process methods and COM.

The operating system now has access to 8TB virtual address space for its purposes.

The .NET compilers generally generate microcode that is portable between the 32-bit and 64-bit .NET Common Language Runtime (CLR). On Windows x64 the .NET Framework installs both 32-bit and 64-bit versions of the CLR. By default, .NET managed code starts in the 64-bit CLR unless the code has been compiled with target platform flags specifying 32-bit. By default, the ASP.NET environment uses the 64-bit CLR, but it can be switched to the 32-bit environment (for the whole Web server) through the ASP.NET configuration scripts.

Enterprise Vault, Accelerators and SQL on Windows x64

As Enterprise Vault is compiled as 32-bit binaries, it needs to run on Windows x64 within the WOW64 handling layer. Enterprise Vault is supported on Windows x64 using the WOW64 layer as per the *Compatibility Charts*.

The Accelerators have been written using .NET technology, but they have limited support.

The Compliance and Discovery .NET managed code loads external 32-bit DLLs, so it must be run within the 32-bit CLR (and the 32-bit ASP.NET). See the *Compatibility Charts* for current support information.

Using a remote SQL Server running on x64 is supported as per the *Compatibility Charts*. The x64 editions provide performance benefits due to memory efficiency.

64-bit Windows: Windows Itanium

The true 64-bit Windows platform (normally referred to as “Intel Architecture 64”, “Itanium”, “64-bit”, or “IA-64”) consists of a 64-bit processor using a totally different 64-bit RISC instruction set. At this time just one main processor is available with Windows support. This is the Intel Itanium, which can address 44-bit memory (16 TB). The Intel Itanium 2 is soon to arrive and can address 50-bit (1 PB), which this has many complexities.

The Windows operating system and application software is compiled to use the 64-bit instruction set (different to x64 and x86) and the compiler is specialized to support parallel instruction execution (Itanium can execute 6 instructions in parallel). All 32-bit software runs through the 32-bit WOW64 handling layer which performs file and registry mappings. On Itanium, a 32-bit extension layer must also convert 32-bit x86 instructions to 64-bit IA-64 instructions, which can degrade performance or prevent proper running. 16-bit applications are not supported.

A 32-bit process can only load 32-bit DLLs and has a 2 GB memory limit (4 GB if compiled with LargeAddressAware flags). The 36-bit AWE API is not supported.

A 64-bit process can only load 64-bit DLLs and can access 7 TB of virtual address space. The 32/64 bit Windows APIs remain very similar, and a 32-bit process can communicate with a 64-bit process through the normal inter-process methods and COM.

The operating system has access to 7 TB of virtual address space for its purposes.

The .NET compilers generally generate microcode that is portable between the 32-bit and 64-bit .NET CLR. On Windows Itanium, the .NET Framework installs both 32-bit and 64-bit versions of the CLR. By default, .NET managed code starts in the 64-bit CLR unless the code has been compiled with target platform flags specifying 32-bit. By default, the ASP.NET environment uses the 64-bit CLR, but it can be switched to the 32-bit environment (for the whole Web server) through the ASP.NET configuration scripts.

Enterprise Vault, Accelerators and SQL on Windows Itanium

At the time of publication, Enterprise Vault and the Accelerators are not supported on this platform. Similarly, using a remote SQL Server that is running on Windows Itanium is not supported.

See the *Compatibility Charts* for the current information.