

# Symantec Enterprise Vault™

## Performance Guide

10.0

*Last updated: June 27, 2013*



# 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: 2013-06-27.

## Legal Notice

Copyright © 2013 Symantec Corporation. All rights reserved.

Symantec, the Symantec Logo, the Checkmark Logo, 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

<http://www.symantec.com>

# Technical Support

Symantec Technical Support maintains support centers globally. Technical Support's primary role is to respond to specific queries about product features and functionality. The Technical Support group also creates content for our online Knowledge Base. The Technical Support group works collaboratively with the other functional areas within Symantec to answer your questions in a timely fashion. For example, the Technical Support group works with Product Engineering and Symantec Security Response to provide alerting services and virus definition updates.

Symantec's support offerings include the following:

- A range of support options that give you the flexibility to select the right amount of service for any size organization
- Telephone and/or Web-based support that provides rapid response and up-to-the-minute information
- Upgrade assurance that delivers software upgrades
- Global support purchased on a regional business hours or 24 hours a day, 7 days a week basis
- Premium service offerings that include Account Management Services

For information about Symantec's support offerings, you can visit our Web site at the following URL:

[www.symantec.com/business/support/](http://www.symantec.com/business/support/)

All support services will be delivered in accordance with your support agreement and the then-current enterprise technical support policy.

## Contacting Technical Support

Customers with a current support agreement may access Technical Support information at the following URL:

[www.symantec.com/business/support/](http://www.symantec.com/business/support/)

Before contacting Technical Support, make sure you have satisfied the system requirements that are listed in your product documentation. Also, you should be at the computer on which the problem occurred, in case it is necessary to replicate the problem.

When you contact Technical Support, please have the following information available:

- Product release level
- Hardware information

- Available memory, disk space, and NIC information
- Operating system
- Version and patch level
- Network topology
- Router, gateway, and IP address information
- Problem description:
  - Error messages and log files
  - Troubleshooting that was performed before contacting Symantec
  - Recent software configuration changes and network changes

## Licensing and registration

If your Symantec product requires registration or a license key, access our technical support Web page at the following URL:

[www.symantec.com/business/support/](http://www.symantec.com/business/support/)

## Customer service

Customer service information is available at the following URL:

[www.symantec.com/business/support/](http://www.symantec.com/business/support/)

Customer Service is available to assist with non-technical questions, such as the following types of issues:

- Questions regarding product licensing or serialization
- Product registration updates, such as address or name changes
- General product information (features, language availability, local dealers)
- Latest information about product updates and upgrades
- Information about upgrade assurance and support contracts
- Information about the Symantec Buying Programs
- Advice about Symantec's technical support options
- Nontechnical presales questions
- Issues that are related to CD-ROMs or manuals

## Support agreement resources

If you want to contact Symantec regarding an existing support agreement, please contact the support agreement administration team for your region as follows:

Asia-Pacific and Japan [customercare\\_apac@symantec.com](mailto:customercare_apac@symantec.com)

Europe, Middle-East, and Africa [semea@symantec.com](mailto:semea@symantec.com)

North America and Latin America [supportsolutions@symantec.com](mailto:supportsolutions@symantec.com)



# Contents

Chapter 1	Introduction.....	11
	What's new in this guide.....	11
	Supporting documents.....	11
Chapter 2	New indexing engine .....	13
Chapter 3	Hardware.....	15
	Recommended processors and memory.....	15
	VMware ESX Server.....	16
	Recommended memory.....	16
	Hyperthreading.....	16
	Storage.....	16
	Enterprise vault store.....	16
	Indexes.....	17
	Local disks.....	17
	Network.....	17
Chapter 4	Exchange mailbox archiving .....	19
	Archiving from Exchange 2013.....	19
	Number of physical cores.....	20
	Calculating disk space.....	20
	Disk space used by vault store partitions.....	21
	Disk space used by indexes.....	22
	Network usage.....	22
	Communicating with and copying data from Exchange servers.....	22
	Communicating with SQL.....	22
	Writing to the vault store partition.....	23
	Reading and writing indexes.....	23
	Summary.....	23
	Effect on the Exchange server.....	24
	Tuning parameters for Exchange mailbox and journaling.....	24
Chapter 5	Exchange journaling .....	31
	Introduction.....	31
	Number of physical cores.....	32

	Calculating disk space.....	32
	Disk space used by vault store partitions .....	33
	Disk space used by indexes.....	34
	Network usage.....	34
	Communicating with the Exchange servers .....	34
	Communicating with SQL.....	34
	Writing to the storage medium .....	35
	Reading and writing indexes.....	35
	Summary .....	35
	The impact of journal report decryption on journaling .....	36
	Tuning parameters for Exchange mailbox and journaling .....	36
Chapter 6	PST migration.....	37
	Introduction .....	37
	Choice of CPU.....	37
	Location and collection.....	38
	Increasing the number of concurrent migrations.....	39
	Calculating disk space.....	39
	Disk space used by vault store partitions .....	40
	Disk space used by indexes.....	41
Chapter 7	Domino mailbox archiving .....	43
	Number of physical cores .....	43
	Adjusting the number of threads .....	44
	Calculating disk space.....	45
	Disk space used by vault store partitions .....	45
	Disk space used by indexes.....	46
	Network traffic.....	47
	Retrieving archived items .....	47
Chapter 8	Domino journaling .....	49
	Number of cores.....	49
	Number of concurrent connections to the Domino server .....	50
	Calculating disk space.....	50
	Disk space used by vault stores .....	51
	Disk space used by indexes.....	52
	Network traffic.....	52
Chapter 9	NSF migration.....	53
	Introduction .....	53
	Number of cores.....	54



	Calculating disk space .....	54
	Disk space used by vault stores.....	54
	Disk space used by indexes.....	55
	Network traffic .....	56
<b>Chapter 10</b>	<b>File System Archiving.....</b>	<b>57</b>
	Number of cores .....	57
	Calculating disk space .....	58
	Disk space used by vault stores.....	59
	Disk space used by indexes.....	59
	Network usage .....	60
	Communicating with the file server.....	60
	Communicating with the SQL database.....	60
	Transfer of data to the storage medium and retrieval for indexing ...	60
	Reading and writing indexes .....	60
	Summary .....	61
	File types .....	61
	File System Archiving and SMTP archiving .....	61
<b>Chapter 11</b>	<b>SharePoint .....</b>	<b>63</b>
	Introduction.....	63
	Number of cores .....	63
	Calculating disk space .....	64
	Disk space used by vault stores.....	64
	Disk space used by indexes.....	65
	Network traffic .....	65
	Retrieving items .....	66
<b>Chapter 12</b>	<b>Enterprise Vault Extensions.....</b>	<b>67</b>
	Introduction.....	67
	Number of cores .....	68
	Calculating disk space .....	68
	Disk space used by vault stores .....	68
	Disk space used by indexes.....	69
<b>Chapter 13</b>	<b>Archiving to Centera.....</b>	<b>71</b>
	Archiving with and without Centera collections .....	71
	Centera sharing model.....	72
	Choice of Enterprise Vault server .....	72
	Centera settings .....	73
	Centera limits .....	73

	Self-healing.....	74
	NTFS to Centera migration .....	74
Chapter 14	Archiving to a storage device through the Storage Streamer API ...	75
	Choice of Enterprise Vault server .....	75
Chapter 15	Data Classification Services.....	77
	Sizing a system .....	77
	Effect on Enterprise Vault servers.....	78
	Network usage.....	79
Chapter 16	User activity .....	81
	Average user response times .....	81
Chapter 17	Virtual Vault .....	83
	Overview .....	83
	Initial synchronization .....	84
	Incremental synchronization .....	85
Chapter 18	Backtrace.....	87
Chapter 19	Move Archive .....	89
	Overview .....	89
	Setting Move Archive parameters .....	89
	Moving small number of users .....	90
	Moving large number of users.....	90
	General notes .....	92
Chapter 20	Combined activity .....	93
Chapter 21	Rules of thumb for IOPS.....	95
Chapter 22	Backup of indexes .....	97
Chapter 23	Document conversion.....	99
	Converting to HTML or text.....	99
	Excluding files from conversion.....	100
	Conversion timeout .....	101

# 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.

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

## What's new in this guide

This guide has been updated from the previous version as the result of further performance investigations and feedback. The most notable additions in this latest version are as follows:

- A new chapter on the Enterprise Vault Extensions feature.  
See “Enterprise Vault Extensions” on page 67.
- Information on archiving from Exchange 2013 CU1.  
See “Archiving from Exchange 2013 CU1” on page 19.

## Supporting documents

Use this guide in conjunction with the following documents:

- *Best practices white paper for Enterprise Vault indexing*, which covers details of the new indexing system. The document is available from the following page of the Symantec Support website:

<http://www.symantec.com/docs/DOC4250>

- *Enterprise Vault Compatibility Charts*, which is available from the following page of the Symantec Support website:  
<http://www.symantec.com/docs/TECH38537>
- *Symantec Discovery Accelerator 10.0 Best Practices Guide*, which is available from the following page on the Symantec Support website:  
<http://www.symantec.com/docs/TECH159520>  
The guide discusses the different aspects that you need to consider during sizing, and recommends best practices for implementation. Most of the advice in the guide applies to Compliance Accelerator as well.
- *Symantec Enterprise Vault Data Classification Services - Implementation Guide*, which is available from the following page of the Symantec Support website:  
<http://www.symantec.com/docs/DOC4121>
- Centera and Celerra best practice documentation: *Symantec Enterprise Vault and EMC Storage Applied Best Practices*, which is available at the following location:  
<http://www.emc.com/collateral/hardware/white-papers/h6790-symantec-enterprise-vault-centera-wp.pdf>
- *Exchange Server 2010 documentation*, which is available at the following location:  
<http://technet.microsoft.com/en-gb/library/bb124558.aspx>
- *Symantec Enterprise Vault 10.0: SQL Best Practices Guide*, which is available at the following location:  
<http://www.symantec.com/docs/DOC5365>
- *Enterprise Vault Best Practice Guide - Implementing Enterprise Vault on VMware*, which is available at the following location:  
<http://www.symantec.com/docs/TECH180094>
- *Virtual Vault Best Practices Guide*, which is available at the following location:  
<http://www.symantec.com/docs/TECH124700>

# New indexing engine

Enterprise Vault 10.0 introduces a new indexing engine. 32-bit technology is replaced with 64-bit technology. To take advantage of the new technology, Enterprise Vault requires more processing power and more memory than previous versions. The recommended number of cores has been increased to 8 and the recommended memory has been increased to 16 GB.

With the hardware upgraded to the new recommended level, users will see faster response times from searches from the new 64-bit indexing engine when compared with the old 32-bit engine, especially with the system under load. Enterprise Vault also becomes far more scalable especially when used with the Accelerator products.

This guide assumes that the Indexing service runs on the same server as other Enterprise Vault services on an Enterprise Vault server. You do not have to install the Indexing service on every Enterprise Vault server. For example, in larger deployments of Enterprise Vault the Indexing and Storage services can be located on more powerful computers to optimize search and retrieve performance. Associated Storage and Indexing services can reside on different computers.

For further guidance and sizing, see the *Best practices white paper for Enterprise Vault indexing*. This is available on the Symantec Support website at the following location:

<http://www.symantec.com/docs/DOC4250>



# 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.

## Recommended processors and memory

This guide gives the recommended number of cores for each archiving source to achieve a given level of throughput. The other components of the system, such as the disks, memory, and network, need to match this power.

Eight cores or more are recommended. Either a combination of multi-core or multi-CPU can be used, but server sizing should not be based on hyper-threading.

If the Indexing services are located on a separate server from the other Enterprise Vault services, the base Enterprise Vault server may be sized as in previous versions of Enterprise Vault, but 8 GB of memory or more are recommended. For the server hosting the Indexing service, eight cores and 16 GB of memory are recommended.

Throughput rates are given in relation to the total number of cores. The type and power of the processor or core is also important. The figures assume a processor of 2.7 GHz or similar – for example, an Intel® Xeon® processor. The figures may be adjusted to take into account processors with higher or lower specifications, but you should be aware that the raw processing power may not accurately reflect its ability to do work.

The figures in this guide are meant to allow you to size systems correctly, with realistic throughput figures that can be easily achieved. They are derived from benchmarks on typical customer loads based on current customers. The figures are adjusted to allow for factors that may apply to working systems but are not covered in a benchmark and to allow some leeway if the original estimates of throughput were too low.

If these figures are not reached, it is probable that some other factor apart from processing power is causing a bottleneck. Before you upgrade the processors, check whether they are running at full capacity or whether some other element of the system is causing a bottleneck. With some tuning, you may be able to increase the throughput.

## VMware ESX Server

All the figures in this guide assume that you are running Enterprise Vault in a virtual environment and that you have followed the recommendations in the *Enterprise Vault Best Practice Guide - Implementing Enterprise Vault on VMware*. This guide is available at the following location:

<http://www.symantec.com/docs/TECH180094>

If you move a system from a physical environment to a virtualized environment, you may experience a degradation in performance. However, you should still attain the throughput figures in this guide.

## Recommended memory

16 GB of memory or more is recommended. The Enterprise Vault servers should be capable of easy upgrades to memory.

## Hyperthreading

Hyperthreading does not provide benefit to Enterprise Vault, and in some circumstances it may impact performance. Therefore, hyperthreading should be not be used.

## Storage

### Enterprise vault store

One of the benefits 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 and Centera devices 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 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 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.

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.

## Local disks

Archiving generates IOs on local disks. The primary causes of these are

- The creation of temporary files used when archiving and conversion
- IOs generated by MSMQ
- IOs to the Enterprise Vault cache location.

To isolate the IOs that MSMQ and Enterprise Vault cache cause, place the MSMQ files and the Enterprise Vault cache on fast local disks separate from the system disk and from each other. 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 IO loads.

## 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. 100BASE-T is normally sufficient, but 1000Base-t is recommended.

See also the sections on network usage for the various archiving agents to calculate what bandwidth is required.

## 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. Archiving is slower if there is other concurrent activity.

When archiving from Exchange 2010, throughput is best when the recommended ratio of Mailbox servers to Client Access Server (CAS) servers is maintained. The current recommendation is 4:3 (Mailbox : CAS).

For more information, see the Microsoft Exchange TechCenter at <http://technet.microsoft.com/en-gb/library/bb124558.aspx>. If the number of CAS servers is less than the recommended number, the ingest rate is affected.

### Archiving from Exchange 2013 CU1

Depending on your Exchange Server environment, you may see a slowdown when ingesting from Exchange 2013 CU1. This will vary depending on your Exchange system, but the ingest rate from a single Exchange Server could be up to 30% slower. However, if you are ingesting from two or more servers, the total slowdown should be marginal.

We have also noticed that some activities are slower on Exchange 2013, but the effect varies from system to system. The activities that are slower are as follows:

- Enabling mailboxes.
- Post-processing of user mailbox items (but not journal items). This is when items in Exchange are replaced by shortcuts after the vault store partitions have been backed up.
- Synchronization of mailboxes.

**Number of physical cores**

These are all background activities that do not affect the ingest rate or user activity. However, you should expect that these activities may take up to twice their current time.

It is difficult to be precise but, in general, the more complex the Exchange environment, the slower the ingest rate and other activities. For example, this is especially likely to be the case in an environment with many Exchange Servers at different versions in a forest that contains many domains.

## Number of physical cores

The following table shows the expected ingest rates for numbers of physical cores where the average message size including attachments is 70 KB. It is assumed that the system is running on VMware and that CPU and memory resources are dedicated (reserved) to the Enterprise Vault server, and not shared with other virtual machines on the host. A 10% to 20% higher ingest rate may be achieved on physical servers.

Number of cores	Hourly ingest rate (70 KB)
8	60,000
16	90,000

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, which holds the 64-bit indexes.
- The SQL databases that are used to hold the Directory, vault store, and fingerprint databases. For guidelines on how to size the databases, see the *Symantec Enterprise Vault 10.0: SQL Best Practices Guide*, which is available at the following location:

<http://www.symantec.com/docs/DOC5365>

## Disk space used by vault store partitions

The single instance model works in the following way:

- 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 steps provide some rules for estimating the amount of space used. This is a simple calculation that does not take into account some of the complexities.

### To estimate the amount of space used

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 or later 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 71 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	4%
Full	12%

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

## Network usage

The network is used for the following purposes while ingesting items from Exchange user mailboxes and 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 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 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 vault store partition

Data is written in to the vault store partition 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 twice the size of the original item for every item indexed.

## Summary

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

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

	Hourly ingest rate	
	60,000	90,000
Enterprise Vault server ↔ Exchange server	19,000	28,500
Enterprise Vault server ↔ SQL server (vault store)	3,000	4,500
Enterprise Vault server ↔ SQL server (Directory)	1,350	2,000
Enterprise Vault server ↔ SQL server (fingerprint)	170	250

	Hourly ingest rate	
	60,000	90,000
Enterprise Vault server ↔ Storage medium	7,200	11,000
Enterprise Vault server ↔ Index location	19,200	30,000

## 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.

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.

## 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.



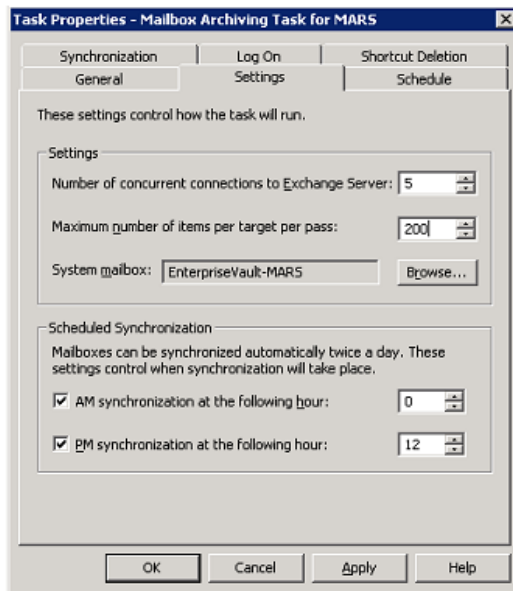
Items are passed between these processes by message queues. The queues used by mailbox archiving are:

- 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.

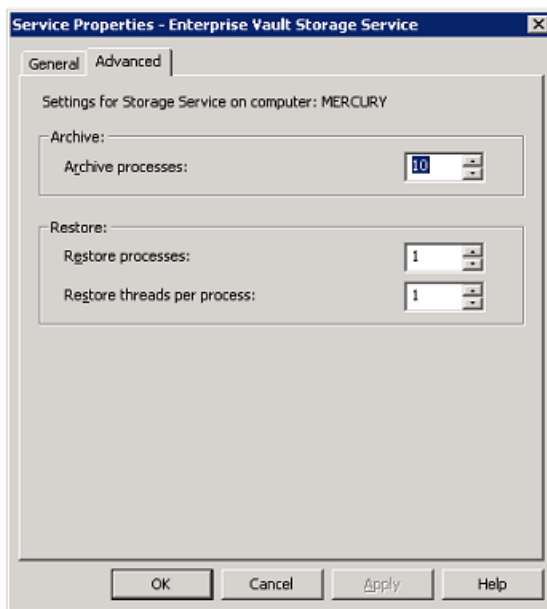


## Setting the number of storage archive processes

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 Enterprise Vault time to clear the **Storage Archive** 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 \Wow6432Node \KVS \Enterprise Vault \Storage	DWORD value set to an integer value. Default is 1500. Raise this to 5000.
ThrottleUpperThreshold	HKEY_LOCAL_MACHINE \SOFTWARE \Wow6432Node \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, 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.

If this is the case, increase the number of concurrent connections from 5 to 10.

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 \Wow6432Node \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 mailbox 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:

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

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

The ingest rate may be affected when ingesting from a single Exchange 2013 server. A suggested rate is included in the table below, but the rate may vary from one system to another. The ingest rate is less affected as the number of Exchange Journaling tasks increases.

## Number of physical cores

The following table shows the expected ingest rate for numbers of physical cores where the average message size including attachments is 70 KB. It is assumed that the system is running on VMware and that CPU and memory resources are dedicated (reserved) to the Enterprise Vault server, and not shared with other virtual machines on the host. A 10% to 20% higher ingest rate may be achieved on physical servers.

Number of cores	Hourly ingest rate (70 KB)	Hourly ingest rate (70 KB) from single Exchange 2013 server
8	60,000	45,000
16	90,000	45,000

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 SQL databases that are used to hold the Directory, vault store, and fingerprint databases. For guidelines on how to size the databases, see the *Symantec Enterprise Vault 10.0: SQL Best Practices Guide*, which is available at the following location:

<http://www.symantec.com/docs/DOC5365>



## Disk space used by vault store partitions

The single instance model works in the following way:

- 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 steps provide some rules for estimating the amount of space used. This is a simple calculation that does not take into account some of the complexities.

### To estimate the amount of space used

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 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 71 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	4%
Full	12%

The percentage for 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.

## 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

## Communicating with 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. 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 twice 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	
	60,000	90,000
Enterprise Vault server ↔ Exchange server	19,000	28,500
Enterprise Vault server ↔ SQL server (vault store)	3,000	4,500
Enterprise Vault server ↔ SQL server (Directory)	670	1,000
Enterprise Vault server ↔ SQL server (fingerprint)	170	250
Enterprise Vault server ↔ Storage medium	7,200	11,000
Enterprise Vault server ↔ Index location	19,200	30,000

## 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 2,000 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
8	15,000	60,000
16	15,000	90,000

## Location and collection

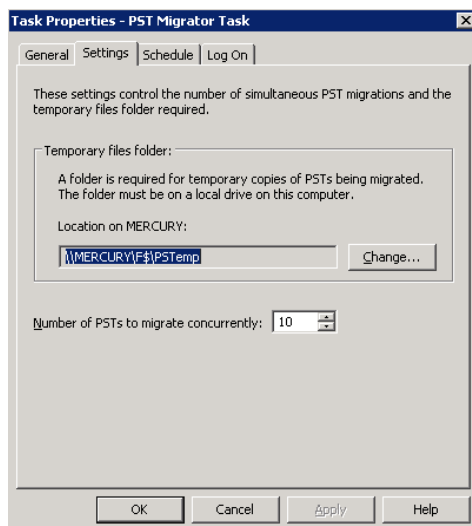
In the locate and migrate method, the PSTs are located before migration. The time to locate the PSTs is not predictable and may require many domains and servers to be searched. The following table shows one example.

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 SQL databases that are used to hold the Directory, vault store, and fingerprint databases. For guidelines on how to size the databases, see the *Symantec Enterprise Vault 10.0: SQL Best Practices Guide*, which is available at the following location:

<http://www.symantec.com/docs/DOC5365>

## Disk space used by vault store partitions

The single instance model works in the following way:

- 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 steps give 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 store more messages with attachments in PSTs, and you should allow for this by adjusting the percentage of messages with attachments.

### To calculate the amount of space used

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 71 for more details.

---



## Disk space used by indexes

To calculate the expected index size

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

Indexing type	Percentage
Brief	4%
Full	12%

The percentage for 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.



# 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 Domino servers, and the tasks on multiple Enterprise Vault servers may archive mailboxes on a single Domino server.

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

## Number of physical cores

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. It is assumed that the system is running on VMware and that CPU and memory resources are dedicated (reserved) to the Enterprise Vault server, and not shared with other virtual machines on the host. A 10% to 20% higher ingest rate may be achieved on physical servers.

Number of cores	Hourly ingest rate (70 KB)
8	60,000
16	90,000

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 Domino 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. Check the *Enterprise Vault Compatibility Charts* which is available on the Symantec Support website at the following location:

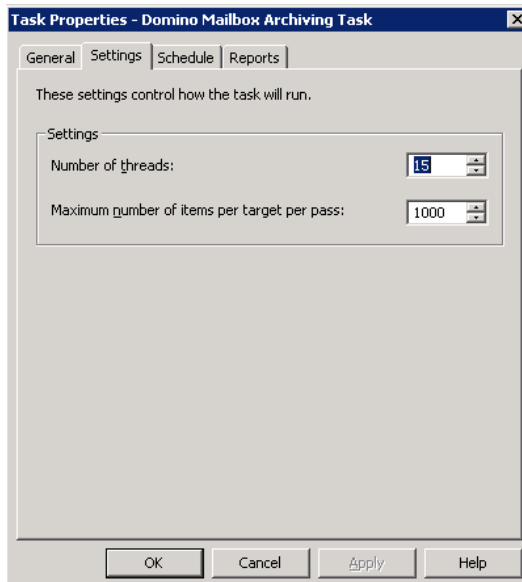
<http://www.symantec.com/docs/TECH38537>

## 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 SQL databases that are used to hold the Directory, vault store, and fingerprint databases. For guidelines on how to size the databases, see the *Symantec Enterprise Vault 10.0: SQL Best Practices Guide*, which is available at the following location:

<http://www.symantec.com/docs/DOC5365>

### Disk space used by vault store partitions

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 works in the following way

- 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 steps provide some rules for estimating the amount of space used. This is a simple calculation that does not take into account some of the complexities.

#### To estimate the amount of space used

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 71 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	4%
Full	12%

The percentage for 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.

## 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	
	60,000	90,000
Enterprise Vault server ↔ Domino server	20,500	31,000
Enterprise Vault server ↔ SQL server (vault store)	3,600	5,400
Enterprise Vault server ↔ SQL server (Directory)	1,500	2,200
Enterprise Vault server ↔ SQL Server (fingerprint)	300	450
Enterprise Vault server ↔ Storage medium	24,000	36,000
Enterprise Vault server ↔ Index location	20,400	30,600

## 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 (EVDG) 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 EVDG, 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 EVDG.

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 IOs 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.

## Number of cores

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. It is assumed that the system is running on VMware and that CPU and memory resources are dedicated (reserved) to the Enterprise Vault server, and not shared with other virtual machines on the host. A 10% to 20% higher ingest rate may be achieved on physical servers.

Number of cores	Hourly ingest rate (70 KB)
8	60,000
16	90,000

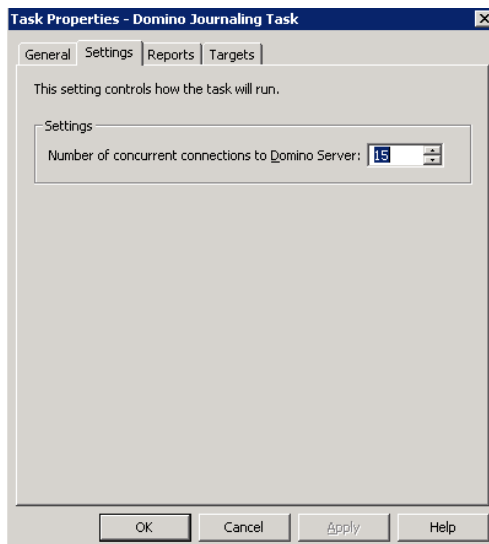
The average size of mail messages has an effect on the throughput. The observed effect is that when the average message size is doubled, throughput is reduced by one third.

## 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 SQL databases that are used to hold the Directory, vault store, and fingerprint databases. For guidelines on how to size the databases, see the *Symantec Enterprise Vault 10.0: SQL Best Practices Guide*, which is available at the following location:  
<http://www.symantec.com/docs/DOC5365>

## Disk space used by vault stores

The single instance model changed in Enterprise Vault 8.0. The principal changes were 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 steps provide some rules for estimating the amount of space used. This is a simple calculation that does not take into account some of the complexities.

### To estimate the amount of space used

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 71 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	4%
Full	12%

The percentage for 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.

## 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 for different archiving rates.

	Hourly ingest rate	
	60,000	90,000
Enterprise Vault server ↔ Domino server	20,500	31,000
Enterprise Vault server ↔ SQL server (vault store)	3,600	5,400
Enterprise Vault server ↔ SQL server (Directory)	1,500	2,200
Enterprise Vault server ↔ SQL Server (fingerprint)	300	450
Enterprise Vault server ↔ Storage medium	24,000	36,000
Enterprise Vault server ↔ Index location	20,400	30,600

# 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	SHKEY_LOCAL_MACHINE \SOFTWARE \Wow6432Node \KVS \Enterprise Vault \Agents	DWORD value set to an integer value. Default is 5. Raise this to 15.

## Number of cores

The following table shows the archiving rates per hour assuming 15 threads and where the average message size including attachments is 70 KB. It is assumed that the system is running on VMware and that CPU and memory resources are dedicated (reserved) to the Enterprise Vault server, and not shared with other virtual machines on the host. A 10% to 20% higher ingest rate may be achieved on physical servers.

Number of cores	Hourly ingest rate (70 KB)
8	60,000
16	90,000

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 SQL databases that are used to hold the Directory, vault store, and fingerprint databases. For guidelines on how to size the databases, see the *Symantec Enterprise Vault 10.0: SQL Best Practices Guide*, which is available at the following location:

<http://www.symantec.com/docs/DOC5365>

## 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 changed in Enterprise Vault 8.0. The principal changes were 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 steps provide some rules for estimating the amount of space used. This is a simple calculation that does not take into account some of the complexities.

#### To estimate the amount of space used

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 71 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	4%
Full	12%

The percentage for 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.

## 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 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	
	60,000	90,000
Enterprise Vault server ↔ NSF file location	42,000	63,000
Enterprise Vault server ↔ SQL server (vault store)	3,600	5,400
Enterprise Vault server ↔ SQL server (directory)	1,500	2,200
Enterprise Vault server ↔ SQL server (fingerprint)	300	450
Enterprise Vault server ↔ Storage medium	24,000	36,000
Enterprise Vault server ↔ Index location	20,400	30,600



# 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 IO to the disks containing the files to be archived.

## Number of cores

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. It is assumed that the system is running on VMware and that CPU and memory resources are dedicated (reserved) to the Enterprise Vault server, and not shared with other virtual machines on the host. A 10% to 20% higher ingest rate may be achieved on physical servers.

Number of cores	Hourly ingest rate (100 KB)
8	60,000
16	90,000

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

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 8 Core	Text	54000	3900
100 KB 8 Core	Mixed Office	60000	6000
200 KB 8 Core	Mixed Office	48000	9000
540 KB 8 Core	Big PDF	5400	2900
3000 KB 8 Core	JPEG only	13200	37200

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.

## 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 SQL databases that are used to hold the Directory, vault store, and fingerprint databases. For guidelines on how to size the databases, see the *Symantec Enterprise Vault 10.0: SQL Best Practices Guide*, which is available at the following location:

<http://www.symantec.com/docs/DOC5365>

## 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 71 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 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 documents

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

Indexing type	Percentage
Brief	2%
Full	6%

The percentage for 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 2%.

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

## 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.

### Transfer of data to the storage medium and retrieval 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 twice 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	
	60,000	90,000
Enterprise Vault server ↔ File server	18,000	27,000
Enterprise Vault server ↔ SQL server (vault store)	3,000	4,500
Enterprise Vault server ↔ SQL server (Directory)	850	1,300
Enterprise Vault server ↔ SQL server (fingerprint)	300	450
Enterprise Vault server ↔ Storage medium	36,000	54,000
Enterprise Vault server ↔ Index location	29,000	44,000

## 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.



# SharePoint

## Introduction

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

## Number of cores

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. It is assumed that the system is running on VMware and that CPU and memory resources are dedicated (reserved) to the Enterprise Vault server, and not shared with other virtual machines on the host. A 10% to 20% higher ingest rate may be achieved on physical servers.

Number of cores	Hourly ingest rate (200 KB)
8	60,000
16	90,000

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 SQL databases that are used to hold the Directory, vault store, and fingerprint databases. For guidelines on how to size the databases, see the *Symantec Enterprise Vault 10.0: SQL Best Practices Guide*, which is available at the following location:

<http://www.symantec.com/docs/DOC5365>

### 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 71 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 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 documents

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

Indexing type	Percentage
Brief	2%
Full	6%

The percentage for 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 2%.

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

## Network traffic

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

	Hourly ingest rate	
	60,000	90,000
Enterprise Vault server ↔ SharePoint server	18,000	27,000
Enterprise Vault server ↔ SQL server (vault store)	3,000	4,500
Enterprise Vault server ↔ SQL server (Directory)	850	1,300
Enterprise Vault server ↔ SQL server (fingerprint)	300	450
Enterprise Vault server ↔ Storage medium	36,000	54,000
Enterprise Vault server ↔ Index location	29,000	44,000

## 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 Extensions

## Introduction

The Enterprise Vault Extensions feature is new in Enterprise Vault 10.0.4. This feature enables partners to develop new archiving applications to extend the types of source items that Enterprise Vault can archive. For example, these items can include instant messages, voice messages, and large images with OCR. Third-party archiving applications that ingest into Enterprise Vault are visible in the Enterprise Vault Administration Console, and their contribution to archived data is reportable through the Enterprise Vault reporting functionality.

Developers are free to develop third-party archiving applications in any way that they want. Such applications may have their own performance characteristics that affect resource usage and the ingest rates into vault stores. Therefore, Symantec can only give general guidance on performance.

A third-party archiving application can run on an Enterprise Vault server or a dedicated server, so a number of factors affect the ingest rate. These include:

- The resource usage of the third-party archiving application.
- The hardware resources of the server that hosts the third-party agent.
- The nature and location of the storage that holds the source data to be ingested.

There are many types of data sources that you can ingest through Enterprise Vault Extensions, so it is not possible or relevant to provide accurate, expected ingest rates. It is reasonable to assume that a third-party archiving application has similar ingest rates and sizing requirements to those of the established archiving agents. For example, an application that ingests mail items from a mail system should have similar ingest rates to those for Exchange and Domino mailbox archiving. Similarly, an application that ingests files (or items of a similar size to files) should have an ingest rate that is similar to that of File System Archiving.

There are many possible item types that this chapter does not cover. These range from very small instant messages to very large medical images. Some additional information is included below to help in sizing small items.

## Number of cores

The choice of CPU depends on three factors:

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

The table below shows the ingest rates for small text files. It assumes the system is running on VMware and that CPU and memory resources are dedicated (reserved) to the Enterprise Vault server; they are not shared with other virtual machines on the host. A 10% to 20% higher ingest rate may be achieved on physical servers.

Number of cores	Hourly ingest rate (< 10 KB)
8	120,000
16	180,000

## 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 (saveaset), DVSSP (saveaset shared part) and DVSCC (saveaset 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. When Enterprise Vault ingests small files, it creates DVS files only.
- The index area.
- The SQL databases that are used to hold the Directory, vault store, and fingerprint databases. For guidelines on how to size the databases, see the *Symantec Enterprise Vault 10.0: SQL Best Practices Guide*, which is available at the following location:

<http://www.symantec.com/docs/DOC5365>

## Disk space used by vault stores

When Enterprise Vault archives an item, it first compresses the item and then adds metadata to it. The compression ratio depends on the type of file. For example, the compression ratio for small text files is lower than that for large, non-text files. In addition, small items do not have components that are eligible for sharing.

**To estimate the amount of required storage**

- Multiply the number of items to be archived by 10 KB (or by 12 KB if you need to take account of the default allocation unit size on NTFS).

The result is the total size of the DVS files.

---

**Note:** These recommendations do not apply to Centera, which uses a completely different sharing model. See “Archiving to Centera” on page 71 for more details.

---

## Disk space used by indexes

Small text files create comparatively large indexes. Pure text data generates more indexable content than Office documents or image files, and the metadata is a higher proportion of the index size.

**To calculate the expected index size**

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

Indexing type	Percentage
Brief	25%
Full	100%



# 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.

## Centera settings

Writes to the Centera are slightly slower than to other devices, but many IOs 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 if 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 for an Enterprise Vault server with the recommended configuration.

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

# Archiving to a storage device through the 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 introduced 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.

## Choice of Enterprise Vault server

On devices tested so far, there is no substantial difference in performance between archiving to a device through the Storage Streamer API and archiving to CIFS/SMB and Centera storage devices. Refer to the tables for each archiving type for the throughput rate.

76 | Archiving to a storage device through the Storage Streamer API

**Choice of Enterprise Vault server**

# Data Classification Services

Data Classification Services (DCS) uses various components of Symantec Enterprise Vault and Symantec Data Loss Prevention to analyze Microsoft Exchange email content and help to determine the archiving and retention strategy for all messages.

DCS consists of the following components:

- An Oracle database
- An Enforce server
- One or more Data Classification servers

## Sizing a system

A DCS system normally consists of separate Enforce/Oracle servers and DCS servers. These servers may be set up as physical or virtual servers.

Enforce and Oracle may be installed on the same server. This server should have at least 4 cores and 8 GB of memory. See the *Symantec Enterprise Vault Data Classification Services - Implementation Guide* which is available on the Symantec Support website:

<http://www.symantec.com/docs/DOC4121>

For all but the smallest environments, it is recommended that the DCS server is installed on a separate server from the Enforce/Oracle server and that there is more than one DCS server to provide resilience. See the *DCS Implementation Guide* for server sizing guidelines for disk and memory requirements.

The number of items that a DCS server can process depends on two factors:

- The total number of rules in enabled policies
- The total number of cores

The following table shows the total number of items that will be processed per hour by the DCS servers with cores running at 2 GHz. It is assumed that the system is running on VMware and that CPU and memory resources are dedicated (reserved) to

the DCS server, and not shared with other virtual machines on the host. A 10% to 20% higher ingest rate may be achieved on physical servers.

	Number of rules					
	0	15	30	100	200	300
4 cores	210,000	95,000	65,000	25,000	15,000	10,000
8 cores	420,000	195,000	125,000	50,000	25,000	15,000
12 cores	635,000	290,000	190,000	70,000	40,000	25,000
16 cores	845,000	390,000	255,000	95,000	50,000	35,000

The number of items processed is the number of items that are submitted for classification.

Note the following:

- The processing rates are not affected by the number of items that match a rule or rules.
- The processing rates are not affected by the action taken whether it is include/exclude from review, to change the retention category or “Do not Archive”. The only exception is when the action is “Do not Archive” and the item is moved to the Deleted item folder. The number of items that can be processed is reduced by a maximum of 10%, depending on how many items match the rule.
- The processing cores may be distributed among one or more DCS servers. More than one DCS server is recommended for resilience. The load is distributed evenly among all DCS servers. For example one 8-core server is equivalent to two 4-core servers. All DCS servers should be of the same specification.

## Effect on Enterprise Vault servers

Extra processing is required on the Enterprise Vault Servers when Data Classification is enabled. Ingest rates are reduced by 30%.

It may be beneficial to increase the threads for each Journal task to 10.

## Network usage

There will be network traffic between the DCS servers and the Enterprise Vault servers. A rule of thumb is that the network traffic is twice the size of the original item.

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	
	50,000	80,000
Enterprise Vault server ↔ Data Classification Server	17,000	27,000

**Network usage**



## User activity

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 and Virtual Vault. 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.

### Average user response times

The following assumes a normal pattern of Enterprise Vault usage on a server with 4,000 users on an Enterprise Vault server with 8 cores and 16 GB of memory.

The following table shows the average response times in seconds when users are simultaneously interacting with Enterprise Vault in three different ways:

- Using Archive Explorer (AE)
- Using browser or integrated Search (Search)
- Using Virtual Vault (VV)

Additionally, Virtual Vault performs background incremental synchronizations on the users.

Transaction	Transactions/Hour	Average response time (seconds)
AE click on a folder	800	2
AE enter Archive Explorer	600	1
AE expand a node	800	1
AE list archives	800	1

## Average user response times

Transaction	Transactions/Hour	Average response time (seconds)
AE search a folder or subfolder	800	1
AE search the current folder	400	1
AE view a message	800	1
AE view a random page	20	1
Search open browser	400	1
Search	2200	3
Search view item online	3000	1
VV move 1 to 3 items between folders	800	1
VV move item into archive	800	1
VV view a message	800	1
VV background synchronizations	3000	-

Search response times will be slower for the first search for each user as the user's index is brought into memory. The first search may take up from 5 to 30 seconds to complete.

Searches and some other user activities that query the index will be slower when the indexes are located on NAS devices.

Response times will be slower when the client is remote and not connected by a fast network especially when viewing large messages.

Moves on the Virtual Vault are done locally on the client and are equivalent to performing the same actions on a PST file. Virtual Vault synchronizations are background jobs and therefore no response time is given.

# 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 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 or Virtual Vault is enabled on existing system, 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	1,000 users	3,000 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.

	1,000 users	3,000 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	1,000 users	3,000 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.

See the *Virtual Vault Best Practices Guide* for more information on how to set up Virtual Vault. The guide is available from the following location:

<http://www.symantec.com/docs/TECH124700>



# Backtrace

Backtrace lets you obtain log files of tracing information from Enterprise Vault processes in which the logging starts before a problem occurs. Unlike the DTrace utility log files, a Backtrace log contains tracing information from a single process.

Backtrace retains tracing information in memory until a previously defined trigger event occurs. It then writes a limited amount of DTrace information to a log file. This file contains DTrace information from before and after the trigger event occurred. This is useful when submitting support cases and diagnosing problems.

The impact of Backtrace on performance has not been tested on all areas of the product, but it can have a noticeable impact on some areas. Therefore, it should not be enabled automatically, and, if it is enabled, you should check that you are still getting the desired throughput.

Turning on Backtrace for the following has little impact on throughput and performance provided that Enterprise Vault is running on the recommended specification and the CPU on the Enterprise Vault server is less than 80%:

- Exchange mailbox archiving
- Exchange journaling
- Domino mailbox archiving
- Domino journaling

Turning on Backtrace for the following has a noticeable impact but may be done provided that you can accept a drop in throughput or increase in CPU usage. In these cases, a drop in throughput of 20% to 30% is expected:

- File System Archiving
- SharePoint archiving

Backtrace does not have a noticeable effect on the search and acceptance areas of Discovery Accelerator, but it does affect the following:

- Export rates are reduced by 30%.
- Analytic throughput rates are reduced by 50%.

So, it is not advisable to use Backtrace if you plan to enable analytics on any cases.





# 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 Move Archive task's **Task Properties: Settings tab**. 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. The extra steps taken when moving an archive (most notably the verification phase), mean that 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 the copy phase of the 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 copy phase of 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 copy 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** in 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 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.

## Combined activity

Enterprise Vault may be running several activities simultaneously. For example all these activities may be concurrent:

- User activity including:
  - Searching
  - Downloading or retrieving archived items
  - Uploading items
- Virtual Vault synchronization
- Management tasks including:
  - Move Archive
  - Re-indexing
  - Export Archive
  - PST Migration
- Scheduled tasks including:
  - Mailbox archiving
  - Journal archiving.

In this situation, user activities are least impacted with response times minimally affected.

Virtual Vault incremental synchronization should also complete within the required time frame.

Any management tasks or scheduled tasks will run slowly. With concurrent user activities, management tasks and scheduled tasks are severely impacted:

- Management tasks are reduced to 30% of their normal processing speed
- Scheduled tasks are reduced to 10% of their normal processing speed

Careful attention therefore should be paid to the scheduling of management tasks. They may be scheduled concurrently with user activity but will run slower. This may be what is required and speed will increase as user activity decreases. However, user

activities must be taken into account when planning large scale migrations or re-indexing.

Scheduled tasks may run at the same time but it is unlikely that the required archiving rates will be achieved. In particular it is worth considering locating journal archiving on a separate server from mailbox archiving so that daytime journal archiving does not conflict with user activity.

It should be noted that the case here used is where users are actively using Enterprise Vault. In many sites the use of Enterprise Vault by users may be much lighter and the effect on other activities much smaller.

# Rules of thumb for IOPS

Because of the range of activities within Enterprise Vault, it is difficult to give figures for IOPS that will meet all situations. These figures are given as guidelines but there are many factors that can influence the IOPS.

Generally the peak time for IOPS is during ingest and if the system is sized to handle the ingest load it should also be able to handle other tasks and daily user activity.

The following figures are for an Enterprise Vault server ingesting items from Exchange mailboxes with 4000 users and with an ingest rate of 50,000 items an hour. The figures should be multiplied up for larger systems

Component	Estimated typical sustained IOPS
<b>SQL Server</b>	<b>Data/Log</b>
Directory	10 to 100 /10 to 100
Vault store	300/150
Fingerprint	25 /20
<b>Enterprise Vault server</b>	
Index server	500 1000+ if also using DA (see DA best practice guide)
Vault store partition	450

Storage systems should be specified for at least twice these values to cover for peaks and unexpected usage.





# Backup of indexes

Because of the change in the indexing engine in Enterprise Vault 10.0 there may be changes in the time to back up indexes. This section looks at some of the considerations when performing an incremental backup of indexes.

- The amount of data that will be backed up in an incremental backup is much larger than the amount of data indexed. This is because individual items are not backed up but the containing files which will of course change even after the addition of a single item.
- More data per item is backed up for user indexes than for journal indexes. This is because, up to a point, the addition of a few items will have the same effect as the addition of more items, a larger file will be changed and will be marked for back up.
- There is a great variation in the amount of data that requires backup in the journal indexes. This is because some journal indexes have reached a point when all the index files need consolidating where others are consolidating only the smaller files.

Given the huge variation in the types and sizes of archives backed up, the following is a rough “rule of thumb” for the size of data that needs backing up in an incremental backup. The actual speed of the backup depends on many factors such as the backup destination and network speeds.

For indexes where a few items are added during each archiving session such as those for users, the size of the data to be backed up will be 20 times the size of data added to the index.

For indexes where a large number of items are added during each archiving session such as those for journals, the size of the data to be backed up will be 10 times the size of data added to the index.

The size of data added to the index will normally be 12% of the original size of the data where full indexing is enabled.



# 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.

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

Value	Key	Content
ConvertWordToText	HKEY_LOCAL_MACHINE \SOFTWARE \Wow6432Node \KVS \Enterprise Vault	String value 0 (default) - Convert to HTML 1 - Convert to text.
ConvertExcelToText	HKEY_LOCAL_MACHINE \SOFTWARE \Wow6432Node \KVS \Enterprise Vault	String value 0 - Convert to HTML 1 (default) - Convert to text
TextConversionFileTypes	HKEY_LOCAL_MACHINE \SOFTWARE \Wow6432Node \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

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. The list format is: .filetype[.filetype]. Prefix each file type with a period and end the list with a period. For example: .GIF.JPG.
FromConversion	\SOFTWARE \Wow6432Node \KVS \Enterprise Vault	

## 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.
	\SOFTWARE \Wow6432Node \KVS \Enterprise Vault	Default: 10 (minutes)

