

Arctera Enterprise Vault™

Performance Guide

15.2

Legal Notice

Copyright ©2025 Arctera US LLC. All rights reserved.

Arctera and the Arctera Logo are trademarks or registered trademarks of Arctera US LLC or its affiliates in the U.S. and other countries. Other names may be trademarks of their respective owners. This product may contain third-party software for which Arctera 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. Refer to the Third-party Legal Notices document accompanying this Arctera product or available at:

<https://www.arctera.io/license-agreements>

The product described in this document is distributed under licenses restricting its use, copying, distribution, and de-compilation/reverse engineering. No part of this document may be reproduced in any form by any means without prior written authorization of Arctera US LLC 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. ARCTERA US LLC 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, et seq." Commercial Computer Software and Commercial Computer Software Documentation," as applicable, and any successor regulations, whether delivered by Arctera as on premises or hosted services. Any use, modification, reproduction release, performance, display or disclosure of the

Contents

| | | |
|------------------|---|-----------|
| Chapter 1 | Introduction | 9 |
| | What's new in this guide..... | 9 |
| | Supporting documents | 10 |
| Chapter 2 | Indexing engine..... | 13 |
| Chapter 3 | Storage Queue..... | 15 |
| | Choosing a location for the Storage Queue | 15 |
| | Location of the Storage Queue when it is not used for safety copies | 15 |
| | Location of the Storage Queue when it is used for safety copies | 16 |
| Chapter 4 | Metadata store (MDS) | 17 |
| | Updating existing archives to use MDS..... | 17 |
| Chapter 5 | Storage Compliance Sampling | 19 |
| Chapter 6 | Storage Classification | 20 |
| | Sizing a system..... | 20 |
| | Impact of classification | 21 |
| | Exporting records..... | 22 |
| Chapter 7 | Hardware | 25 |
| | Recommended processors and memory | 25 |
| | VMware ESX Server..... | 26 |
| | Recommended memory | 26 |
| | Hyperthreading | 26 |
| | Storage | 26 |
| | Enterprise vault store | 26 |
| | Indexes | 27 |
| | Local disks | 27 |
| | Network | 27 |
| Chapter 8 | Exchange mailbox archiving | 29 |
| | Archiving from Exchange 2013 or 2016 | 29 |
| | Number of physical cores | 30 |

| | |
|---|-----------|
| Calculating disk space | 30 |
| Disk space used by vault store partitions | 31 |
| Disk space used by indexes | 32 |
| Network usage | 32 |
| Communicating with and copying data from Exchange servers | 32 |
| Communicating with SQL | 33 |
| Writing to the vault store partition..... | 33 |
| Reading and writing indexes | 33 |
| Effect on the Exchange server..... | 33 |
| Tuning parameters for Exchange mailbox and journaling..... | 34 |
| Setting the number of connections to the Exchange server..... | 34 |
| Changing the distribution list cache size..... | 34 |
| Chapter 9 Exchange journaling..... | 37 |
| Introduction | 37 |
| Number of physical cores | 37 |
| Calculating disk space | 38 |
| Disk space used by vault store partitions | 39 |
| Disk space used by indexes | 40 |
| Network usage | 40 |
| Communicating with the Exchange servers..... | 40 |
| Communicating with SQL | 41 |
| Writing to the storage medium..... | 41 |
| Reading and writing indexes | 41 |
| The impact of journal report decryption on journaling..... | 41 |
| Tuning parameters for Exchange mailbox and journaling..... | 42 |
| Chapter 10 PST migration..... | 43 |
| Introduction | 43 |
| Choice of CPU..... | 43 |
| Location and collection | 44 |
| Increasing the number of concurrent migrations | 45 |
| Changing the sample size for PST ownership identification..... | 46 |
| Calculating disk space | 46 |
| Disk space used by vault store partitions | 47 |
| Disk space used by indexes | 48 |
| Chapter 11 Domino mailbox archiving..... | 49 |
| Number of physical cores..... | 49 |
| Adjusting the number of threads | 50 |
| Calculating disk space | 51 |

| | |
|---|-----------|
| Disk space used by vault store partitions..... | 51 |
| Disk space used by indexes..... | 52 |
| Retrieving archived items..... | 52 |
| Chapter 12 Domino journaling..... | 55 |
| Number of cores..... | 55 |
| Number of concurrent connections to the Domino server..... | 56 |
| Calculating disk space..... | 56 |
| Disk space used by vault stores..... | 57 |
| Disk space used by indexes..... | 58 |
| Chapter 13 SMTP archiving..... | 59 |
| Number of physical cores..... | 59 |
| Calculating disk space..... | 60 |
| Disk space used by vault store partitions..... | 60 |
| Disk space used by indexes..... | 62 |
| Disk space used by holding folder..... | 62 |
| Disk space used by message tracking log folder..... | 63 |
| Network usage..... | 63 |
| Communicating with and receiving data from SMTP source..... | 63 |
| Communicating with SQL..... | 63 |
| Writing to the vault store partition..... | 64 |
| Reading and writing indexes..... | 64 |
| The impact of SSL/TLS..... | 64 |
| Microsoft Purview Information Protection (MPIP)..... | 64 |
| Chapter 14 Microsoft Teams Archiving..... | 65 |
| About Enterprise Vault Microsoft Teams Archiving..... | 65 |
| Microsoft Teams Archiving Overview..... | 65 |
| Disk Planning..... | 66 |
| Deployment for Performance Measurements..... | 67 |
| Performance Numbers..... | 67 |
| Chapter 15 File System Archiving..... | 69 |
| Number of cores..... | 69 |
| Calculating disk space..... | 70 |
| Disk space used by vault stores..... | 70 |
| Disk space used by indexes..... | 71 |
| Network usage..... | 71 |
| Communicating with the file server..... | 72 |
| Communicating with the SQL database..... | 72 |

| | |
|--|-----------|
| Transfer of data to the storage medium and retrieval for indexing | 72 |
| Reading and writing indexes | 72 |
| File types | 72 |
| File system folder links to Enterprise Vault Search | 73 |
| Chapter 16 SharePoint | 75 |
| Introduction | 75 |
| Number of cores | 75 |
| Calculating disk space | 76 |
| Disk space used by vault stores | 76 |
| Disk space used by indexes | 77 |
| Retrieving items | 77 |
| Chapter 17 Enterprise Vault Extensions | 79 |
| Introduction | 79 |
| Number of cores | 80 |
| Calculating disk space | 80 |
| Disk space used by vault stores | 81 |
| Disk space used by indexes | 81 |
| Chapter 18 Archiving to Centera | 83 |
| Archiving with and without Centera collections | 83 |
| Centera sharing model | 84 |
| Choice of Enterprise Vault server | 84 |
| Centera settings | 85 |
| Centera limits | 85 |
| Self-healing | 86 |
| NTFS to Centera migration | 86 |
| Chapter 19 Archiving to a storage device through the Storage Streamer API | 87 |
| Choice of Enterprise Vault server | 87 |
| Chapter 20 Backtrace | 89 |
| Chapter 21 Move Archive | 91 |
| Overview | 91 |
| Setting Move Archive parameters | 91 |
| Moving small number of users | 92 |
| Moving large number of users | 92 |
| General notes | 94 |

| | |
|---|-----|
| Chapter 22 Combined activity | 95 |
| Chapter 23 Backup of indexes | 97 |
| Chapter 24 Document conversion | 99 |
| IFilters and Optical Character Recognition of image files | 99 |
| Converting to HTML or text..... | 101 |
| Excluding files from conversion | 101 |
| Conversion timeout | 102 |
| Chapter 25 Amazon Web Services (AWS) Cloud..... | 103 |
| Deployment for Performance Measurements..... | 103 |
| Performance Numbers | 105 |
| Chapter 26 Microsoft Azure Cloud | 107 |
| Deployment for Performance Measurements..... | 107 |
| Performance Numbers | 108 |

Introduction

This document provides guidelines on expected performance when running Arctera 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 a result of further performance investigations and feedback. In most cases, the performance of Enterprise Vault 15.2 is equivalent to that of previous versions.

Supporting documents

Use this guide in conjunction with the following documents:

- Enterprise Vault Compatibility Charts, which is available from the following page of the Arctera Support website:
<http://www.veritas.com/docs/000097605>
- Arctera eDiscovery 15.2 Best Practices Guide, which is available from the following page on the Arctera Support website:
<http://www.veritas.com/docs/000081985>
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 Surveillance as well.
- Arctera Enterprise Vault™ Best Practice for Implementing Enterprise Vault in AWS and Microsoft Azure Cloud, which is available from the following page on the Arctera Support website:
[Enterprise Vault™ Best Practice for Implementing Enterprise Vault on AWS and Microsoft Azure Cloud.](#)
- Exchange Server documentation, which is available at the following location:
[https://technet.microsoft.com/en-s/library/aa996058\(v=exchg.150\).aspx](https://technet.microsoft.com/en-s/library/aa996058(v=exchg.150).aspx)
- Enterprise Vault 15.2: SQL Best Practices Guide, which is available at the following location:
<http://www.veritas.com/docs/100012617>
- Enterprise Vault Best Practices Guide - Implementing Enterprise Vault on VMware, which is available at the following location:
https://www.veritas.com/support/en_US/article.100023811
- Enterprise Vault Best Practices Guide - Enterprise Vault indexing, which covers details of the EV indexing system. The document is available from the following page of the Arctera Support website:
https://www.veritas.com/support/en_US/doc/EV_Indexing_BP_143
- IMAP Access Client Configuration Guides, which are available at the following location:
<http://www.veritas.com/docs/100040609>
- Best Practices for Deploying SMTP Archiving, which is available at the following location:
https://www.veritas.com/content/support/en_US/doc/ev_12_bp_SMTP_00

- Migrating from the Legacy SMTP Archiving Solution, which is available at the following location:

https://www.veritas.com/support/en_US/doc-viewer.118344599-118344609-0.index

Indexing engine

Enterprise Vault 14.2 and later introduced a new 64-bit Elasticsearch indexing engine. The new features and functionality of this engine need to be considered by both existing customers while upgrading and new customers installing a fresh deployment. Enterprise Vault offers backwards compatibility with existing 32-bit Alta-Vista indexes, 64-bit Velocity indexes seamlessly federating searches between old and new data indexed in new 64-bit Elasticsearch indexing engine. The recommended number of cores has been increased to 16 and the recommended memory has been increased to 32 GB.

With the hardware upgraded to the new recommended level, users will see faster response times from searches from the new 64-bit Elasticsearch indexing engine when compared with the old Velocity 64-bit indexing 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 Arctera Support website at the following location:

https://www.veritas.com/support/en_US/doc/EV_Indexing_BP_143

Storage Queue

Enterprise Vault 11.0 introduced a new mechanism for ingesting items: Storage Queue. For Exchange archiving, the Storage Queue replaces some of the functions of Microsoft Message Queuing (MSMQ). For other agents, it allows asynchronous ingest. The agent inserts items into the queue, and the Storage service then takes them from the queue and archives them into the Vault Store. The Storage Queue can also be used to hold safety copies of items so that, when items are archived, they can be removed from Exchange without waiting for a backup of the Vault Store partitions.

This new mechanism is more efficient than the older ones. The Storage Queue uses fewer resources than MSMQ, and, for the other agents, there is a better balance between the processes that fetch items from the agents and the processes that store items in the Vault Stores.

Choosing a location for the Storage Queue

The main factor to consider when choosing a location for the Storage Queue is whether it will be used to hold safety copies of items.

Location of the Storage Queue when it is not used for safety copies

The Storage Queue location should be on a fast-local disk on a fault tolerant device (RAID 1 or higher). The disk should be large enough to hold a backlog of items. The default backlog that can be held before agent services pause to allow the Storage Service to reduce the queue is 50,000 items. Note that when the disk reaches 90% capacity, ingest pauses until some items have been removed from the queue. Therefore, you do not have to allow for 50,000 items, and, in most circumstances, you should not see a backlog accumulating. A minimum disk size of 25 GB is recommended.

The IOPS characteristics of the Storage Queue are comparable with or slightly better than those of the Message Queue doing the same task. So, if you are satisfied with the performance of the Message Queue disk, this will be suitable for the Storage Queue. (However, you should allow for the extra

Choosing a location for the Storage Queue

space required.) In testing, IOPS per 100 KB item ingested were observed to be in the range of 20 to 40.

Do not locate the queue on any of the following:

- A network device. This is likely to slow down ingest and could more than double the network traffic that Enterprise Vault generates.
- The system drive. There will be contention for IOPS and space.
- An otherwise active drive. Other activity on the disk may slow down ingest, which may stop if the available space falls below the threshold.

You may place the queue on the drive that currently holds MSMQ. The residual MSMQ activity will create very little load on the disk.

Location of the Storage Queue when it is used for safety copies

If you use the Storage Queue to hold the safety copy of items until the Vault Store is backed up, you need a disk that is specified as above, but you must also provision enough space to hold all the data between backups. Allow for missed backups and times when extra data is generated; if the disk becomes full, Enterprise Vault stops ingesting items. The reliability of the disk is paramount, as this is where the safety copies are retained in the event that the data in the Vault Store partition is lost before backup.

By default, the data is held uncompressed in the Storage Queue. For Exchange archiving, the space taken up by each item on the Storage Queue is about twice the size of that item in Exchange. For example, if the average size of an item in Exchange is 100 KB and you archive 1,000,000 items between backups, the space used on the disk is $2 \times 100 \times 1,000,000$ KB, or 190 GB. To this you must add a safety factor to allow for missed backups or unexpected surges in items to be archived.

You have the option to hold the data in compressed form. If items are held compressed, they take up the same space as in Exchange. This halves the space required on the Storage Queue. The extra resources used by compression reduce the ingest rate by up to 10%.

If you are using the Storage Queue to hold safety copies, items are removed from Exchange or replaced with shortcuts immediately after archive. This means that Enterprise Vault is doing more work during ingest than if it delayed the post-processing until later. So, ingest is slower but post-processing is faster, and the overall process is faster. On average, the ingest rate is reduced by 10% but post-processing is four times as fast. Post-processing now consists of checking that the item is successfully stored and has been backed up. It does not need to remove the original item or replace it with a shortcut.

Metadata store (MDS)

Enterprise Vault 11.0 introduced a new metadata store (MDS) technology, which enables a “Fast Browse” feature. Fast Browse-enabled archives can list the contents of an archived folder more quickly than was previously the case. This performance improvement is particularly noticeable in IMAP connections and the new Enterprise Vault Search application.

If you plan to implement the MDS technology, you must take into account the impact on SQL and the space that the databases require. This is described in the Enterprise Vault *SQL Best Practices Guide* at <http://www.veritas.com/docs/100012617>.

If you intend to enable IMAP connections, you must plan the roll-out carefully and follow the guidance in the *IMAP Access Client Configuration Guides* at <http://www.veritas.com/docs/100040609>.

Updating existing archives to use MDS

When upgrading from a previous version of Enterprise Vault, you must update your existing archives to take advantage of the MDS technology. To calculate the time taken to update the archives, allow one hour for every 3,000 archives and one hour for every 2,500,000 items. The sum of these two values is the expected time to update on an otherwise idle system with eight cores of 2.2 GHz. That is:

$$(\text{Archives}/3000 + \text{Items}/2500000) \times (\text{Cores}/8)$$

The update process consumes CPU on both the Enterprise Vault server and SQL server. You can expect 50% of the CPU to be used on the Enterprise Vault server and 30% on a similarly specified SQL Server. This may slow down other Enterprise Vault activity, resulting in a general slowdown until the update is complete. For this reason it may be inadvisable to update all users to MDS at the same time. If you are enabling users to allow fast searching, you can let them initiate their own updates when they first search. On the other hand, if you want your users to benefit immediately from using Enterprise Vault Search, or you want to enable them for IMAP, you will want to update them before their first use.

Storage Compliance Sampling

Enterprise Vault 11.0.1 introduced Compliance Sampling in storage, which enables centralized sampling of all journal types by Surveillance (previously known as Compliance Accelerator) 11.0.1 and later. The new architecture can also provide overall performance improvements over the previous connector-based sampling.

The new sampling may increase the size of the vault store databases during archiving, and it increases the load at each Storage service.

If you plan to implement the Compliance Sampling technology, you must take into account the impact on SQL and the space that the databases require. This is described in the Enterprise Vault *SQL Best Practices Guide* at <http://www.veritas.com/docs/100012617>.

Storage Classification

Enterprise Vault 12 introduced classification services within storage, which enables centralized classification of all data types without the need for external or agent-based classification products and enables reclassification. The integration into storage also provides considerable performance benefits over external agents.

Storage Classification analyzes all content and helps to determine the retention strategy for all archived items.

Enterprise Vault 12.1 extends classification and retention to provide records management by marking items with a record type, and provides PowerShell cmdlets to export required records.

Enterprise Vault 12.2 enhances classification with the introduction of far more comprehensive policy and classification management with the new Arctera Insight Classification.

Enterprise Vault 12.3 introduces smart partitions, which store data based upon its classification.

Sizing a system

Classification is integrated into the Storage service, which works in conjunction with either the Microsoft File Classification Infrastructure or the Arctera Insight Classification to provide classification of items during archiving or indexing and deletion/expiry. Items can be re-classified by rebuilding indexes after policy or rule changes. However, if smart partitions are used, re-classified items are not moved between storage partitions.

For the basic classification offering provided by FCI, the classification rules are defined through the Microsoft File Server Resource Manager (FSRM), and processed using an appropriate classifier method (which includes the basic Arctera Insight Classification). For more information, see the Enterprise Vault *Classification using FCI* guide.

You can configure the extensive policy-based classification management provided with the new Arctera Insight Classification using the Vault Administration Console. For more information, see the Enterprise Vault *Classification using VIC* guide.

Note: We recommend that you use or migrate to using the Arctera Insight Classification rather than FCI, as this provides far greater control through comprehensive policy-based classification management. For more information, see the Enterprise Vault *Classification migrating to VIC* guide.

If FCI is used, it is recommended the classifier method used is the “Arctera Insight Classification”, as it has been expressly designed to process rules in the most efficient way.

Impact of classification

Extra processing is required on the Enterprise Vault servers when Storage Classification is enabled, which can impact performance in various circumstances.

Exactly when Enterprise Vault classifies the items is determined by whether you are archiving the classified items to smart partitions rather than standard vault store partitions, as follows:

- If you have chosen to use smart partitions, Enterprise Vault classifies the items at archiving time, which may extend the time to archive items.
- If you have not chosen to use smart partitions, Enterprise Vault classifies the items at indexing time, which may extend the time to index items.

Either way, the time required to archive and index all items should be similar.

The number of items that a Storage server can process depends on three main factors:

- The total number of enabled policies/rules and their complexity.
- The total number of CPU cores.
- The disk performance of the Enterprise Vault cache location.

The performance metrics shown below assume that the Enterprise Vault servers are running with recommended specification. 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.

VIC classification

Enabling VIC classification would impact Indexing performance in terms of throughput and size. The overall impact depends on multiple factors, such as:

- The total number of classification policies used.
- Complexity of custom classification policies.
- Average item size.
- Number of matches or classification tags added.
- Underlying hardware.

Rebuilding indexes with 200 rules that require classification reduces Indexing throughput by 25%.

Enabling the **Retrieve classification policy match information** setting at the site level Advanced setting classification section would have minimal impact on Indexing throughput. However, the Index size would increase by 8-10%.

Exporting records

The records management system that you have adopted may require you to export items from your archives for long-term retention elsewhere. For example, in the Capstone approach to records management, it is customary to make periodic transfers of permanent records to the U.S. National Archives and Records Administration (NARA). NARA does not have an interest in temporary records or non-records, so there is no need to transfer them.

Enterprise Vault 12.1 provides PowerShell cmdlets with which you can export selected items from an archive. See the *Administrators Guide* and *PowerShell Cmdlets* guide for more information.

The item export rate that a Storage server can deliver depends on several factors:

- The total amount of RAM
- The total number of CPU cores
- The export disk location performance

The following table shows sample export rates for the PowerShell cmdlet `Export-EVNARAArchive` from a Storage server with 8 cores and 16 GB of RAM. 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 export rate may be achieved on physical servers.

| Export type | Hourly export item rate | Hourly export data rate |
|-------------------------|--------------------------------|--------------------------------|
| Exchange – Native (MSG) | 360,000 | 79 GB/hr |
| Exchange – PST | 160,000 | 35 GB/hr |
| Domino – Native (EML) | 170,000 | 37 GB/hr |
| SMTP – Native (EML) | 200,000 | 44 GB/hr |
| FSA – Native | 240,000 | 190 GB/hr |

Exporting data from vault stores that employ storage collections can typically reduce the throughput by 50%.

The default of 16 export threads provides the optimum performance and resource utilization. However, the `Export-EVNARAArchive` cmdlet accepts a `-MaxThreads` parameter with which you can adjust the number of threads. You may find that increasing the number may yield a small increase in throughput at the expense of resource utilization.

There is no performance benefit in running multiple PowerShell export cmdlets concurrently rather than one cmdlet that is exporting multiple archives sequentially. Errors can result if you run too many PowerShell cmdlets concurrently.

Exporting to a network share in PST format can significantly impact the throughput. Such exports are typically seven times slower than normal.

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, that is 8 core and 16 GB of RAM. For the server hosting the Indexing service, 16 core and 32 GB of memory is required.

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 consider 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.veritas.com/docs/100023811>

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

32 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 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 eDiscovery (previously known as Discovery Accelerator) or Surveillance (previously known as 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 as follows:

- The creation of temporary files used when archiving and conversion.
- IOs that MSMQ or Storage Queue has generated.
- IOs to the Enterprise Vault cache locations.

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. In Enterprise Vault 11.0, the Storage Queue has replaced many of the functions of MSMQ, and it is less important to isolate the MSMQ IOs. See "Storage Queue" on page 15 for information on the location of the Storage Queue.

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.

Archiving from Exchange 2013 or 2016

Depending on your Exchange Server environment, you may see a slowdown when ingesting from Exchange 2013 or 2016. 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 and 2016, 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.

These are all background activities that do not affect the ingest rate or user activity. However, 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,

Number of physical cores

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 165 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 (165 KB) |
|-----------------|-----------------------------|
| 16 | 45,000 |

The average size of mail messages has an effect on the throughput.

As a minimum, the rate at which Enterprise Vault deletes items that are ready for expiry matches the ingest rate, but it may exceed this rate.

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 new 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 Enterprise Vault *SQL Best Practices Guide*, which is available from:

<http://www.veritas.com/docs/100012617>

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. 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 different sharing model. See “Archiving to Centera” on page 83.

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 |
|---------------|------------|
| Full | 9% |

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.

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

Tuning parameters for Exchange mailbox and journaling

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 Enterprise Vault archives items depends mainly on the specification of the system; that is, the processing power, IO capacity, and network. There are some parameters that can be changed for specific problems. It is not suggested that any of the following are used as a matter of course.

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.

Tuning parameters for Exchange mailbox and journaling

| Value | Key | Content |
|-------------|---|---|
| DLCacheSize | HKEY_LOCAL_MACHINE \SOFTWARE \Wow6432Node \KVS \Enterprise Vault \Agents | DWORD value set to an integer value. Default is now 500. Recommend to remove any previous change. |

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.

The ingest rate may be affected when ingesting from a single Exchange 2013 or 2016 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 165 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 (165 KB) |
|-----------------|-----------------------------|
| 16 | 45,000 |

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

As a minimum, the rate at which Enterprise Vault deletes items that are ready for expiry matches the ingest rate, but it may exceed this rate.

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 Enterprise Vault *SQL Best Practices Guide*, which is available at the following location:

<http://www.veritas.com/docs/100012617>

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 are the files that are not shared.
- 2 Take 60% of the size of attachments. 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 different sharing model. See “Archiving to Centera” on page 83.

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 |
|---------------|------------|
| Full | 9% |

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.

The impact of journal report decryption on journaling

If journal report decryption is configured on Exchange Server, 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

Tuning parameters for Exchange mailbox and journaling

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.

Location and collection

- **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 |
|-----------------|----------------------------------|-----------------------------|
| 16 | 15,000 | 90,000 |

As a minimum, the rate at which Enterprise Vault deletes items that are ready for expiry matches the ingest rate, but it may exceed this rate.

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 |

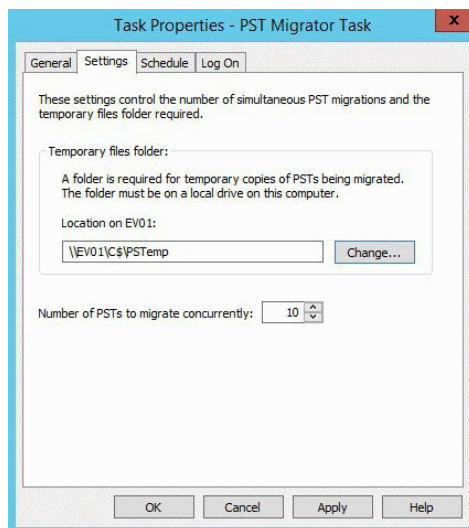
By default, the PST Collector task copies 100 PSTs to a holding area. As PSTs are migrated, more PSTs are collected to keep the number of PSTs in the holding area constant. This is a continual background process during migration and ensures that there are always files ready for migration.

Increasing the number of concurrent migrations

In Enterprise Vault 11.0.1, the PST Locator task can sample messages within each PST to identify the most suitable owner. However, the additional processing required to open each PST and analyze the content can significantly impact the throughput.

Increasing the number of concurrent migrations

The default number of PST files to migrate concurrently is 10. You may find that the throughput rate is improved by increasing this to 15. This leads to higher resource usage.

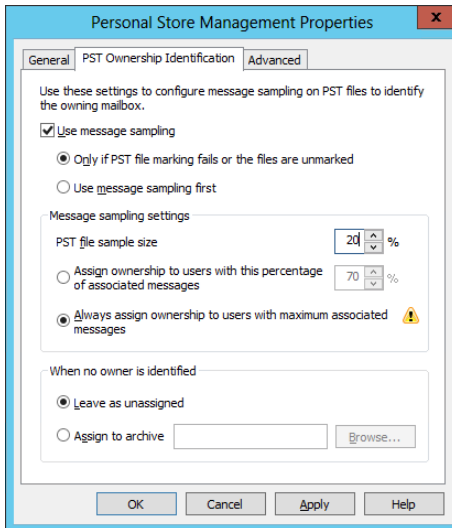


You can also set the number of concurrent tasks in an EVPM script, as follows:

```
ConcurrentMigrations = 15
```

Changing the sample size for PST ownership identification

The default sample size for PST ownership identification is 80%. If you are confident with the selectivity of authors within typical PSTs, you may find that you can improve the throughput rate by reducing the sample size to a smaller sample, such as 20%.



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 Enterprise Vault *SQL Best Practices Guide*, which is available at the following location:

<http://www.veritas.com/docs/100012617>

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 consider some of the complexities.

You should also note that PST files may come from diverse sources with few shared parts. This must 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. 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 different sharing model. See “Archiving to Centera” on page 83.

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 |
|---------------|------------|
| Full | 8% |

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) |
|-----------------|----------------------------|
| 16 | 90,000 |

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

Adjusting the number of threads

As a minimum, the rate at which Enterprise Vault deletes items that are ready for expiry matches the ingest rate, but it may exceed this rate.

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.

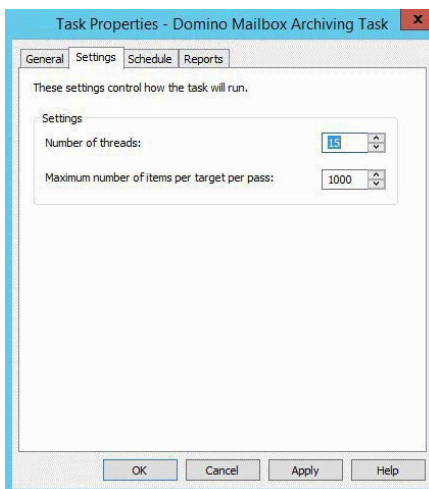
Domino servers are found on many different architectures. Check the Enterprise Vault [Compatibility Charts](#).

Adjusting the number of threads

It is important to set up the optimum number of threads for Domino archiving. The factors to consider 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 10 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 Enterprise Vault *SQL Best Practices Guide*, which is available at the following location:

<http://www.veritas.com/docs/100012617>

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 consider 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. 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.

Retrieving archived items

- 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 83 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 |
|---------------|------------|
| Full | 8% |

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.

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 can 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) |
|-----------------|----------------------------|
| 16 | 90,000 |

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

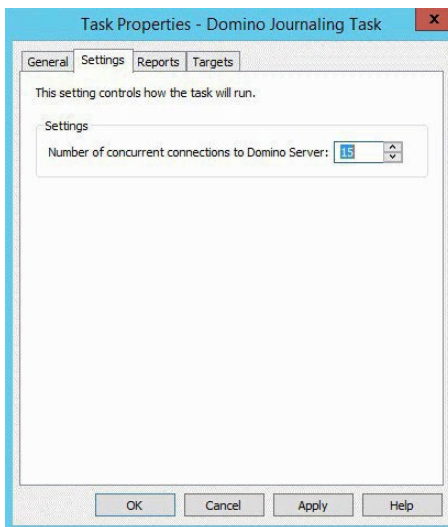
As a minimum, the rate at which Enterprise Vault deletes items that are ready for expiry matches the ingest rate, but it may exceed this rate.

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 consider 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 Enterprise Vault *SQL Best Practices Guide*, which is available at the following location:

<http://www.veritas.com/docs/100012617>

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. 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 83 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 |
|---------------|------------|
| Full | 8% |

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.

SMTP archiving

Enterprise Vault 11.0.1 introduced a new SMTP agent and archiving service that replaces the previous file-based archiving solution. SMTP archiving provides a convenient method of journal archiving from any SMTP source.

For migration guidance, see the technical note [Migrating from the Legacy SMTP Archiving Solution](#).

For further guidance and sizing, see the [Best Practices White Paper for Enterprise Vault SMTP Archiving](#) white paper.

In most cases, when you are choosing servers for email archiving, the most critical factor is the ingest rate. For SMTP archiving, there is normally a continuous stream of messages, and everything must be archived as soon as possible. This stream fits in between daily usage, backups, defragmentation, and all the other background activities that take place on the server.

Archiving is slower if there is other concurrent activity.

Number of physical cores

The following table shows the expected ingest rates with all services running on a single server for numbers of physical cores where the average message size including attachments is 165 KB.

For further guidance for distributing the workload, see the [Best Practices White Paper for Enterprise Vault SMTP Archiving](#) white paper.

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 | 165 KB |
| 16 | 90,000 to 100,000 |

As a minimum, the rate at which Enterprise Vault deletes items that are ready for expiry matches the ingest rate, but it may exceed this rate.

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, and two areas of temporary 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 new 64bit Elasticsearch indexes.
- The SMTP holding folder, which contains the SMTP stream.
- If enabled, the SMTP message tracking log folder.
- The Storage Queue locations, which for SMTP are recommended to be configured to retain safety copies. For guidelines on how to size the Storage Queue locations, see “Storage Queue” on page 15.
- 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 Enterprise Vault *SQL Best Practices Guide*, which is available at the following location:

<http://www.veritas.com/docs/100012617>

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 12 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. 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.
- 4 If archiving SMTP journal streams from Exchange servers and there is more than one journal target, there can be 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 targets that participate in sharing within the vault store group. This is the size of the DVSSP files after sharing.

However, the use of the SMTP scalability features can eliminate this fan-out factor using a single target that can be distributed across archives. For more information, see the [Best Practices for Deploying SMTP Archiving](#) white paper.

- 5 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 83 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 |
|---------------|------------|
| Full | 7% |

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.

Disk space used by holding folder

The holding folder location should be on a fast-local disk on a fault-tolerant device (RAID 1 or higher). The disk should be large enough to hold a backlog of items. The holding folder should be provisioned with sufficient capacity to accommodate 48 hours of incoming data to allow for missed backups and times when extra data is generated; if the disk becomes full, Enterprise Vault stops ingesting items. For example, if the average size of an item is 120 KB and you archive 1,000,000 items per day, the recommended holding folder capacity is $2 \times 120 \times 1,000,000$ KB, or 230 GB.

The reliability of the disk is paramount, as this area contains the only copy of incoming items until they are added to the Storage Queue. Once the items are added to the queue the only copy is then present on the Storage Queue, and therefore it is recommended to use Storage Queue safety copies.

In testing, IOPS per 100 KB item ingested were observed to be in the range of 20 to 40. Therefore, the holding folder typically averages around 300 IOPS under normal load.

Do not locate the holding folder on any of the following:

- A network device. This is likely to slow down ingest and could more than double the network traffic that Enterprise Vault generates.
- The system drive. There will be contention for IOPS and space.
- An otherwise active drive. Other activity on the disk may slow down ingest, which may stop altogether if the available space falls below the threshold.

Disk space used by message tracking log folder

Enabling SMTP message tracking logging does not impact the performance, although depending on throughput will rapidly consume disk capacity.

The SMTP Message Tracking log folder should be located on locally attached storage.

Provision approximately 250MB for every million messages received.

In testing, IOPS per million items ingested were observed to be in the range of 5. Therefore, the message tracking folder typically averages around 1 IOPS under normal load.

Network usage

The network is used for the following purposes while ingesting items from SMTP sources:

- Communicating with and receiving data from the SMTP sources.
- 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, and so on.

Communicating with and receiving data from SMTP source

Assume that the network traffic between the SMTP source and the Enterprise Vault SMTP server is equal to twice 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.

The impact of SSL/TLS

Enabling SSL/TLS connection to the SMTP agent can typically reduce the throughput by up to 40%.

Microsoft Purview Information Protection (MPIP)

With the MPIP feature enabled for Office 365 emails, archiving and indexing might be impacted and run slower. Performance affecting factors include the number of encrypted emails, encrypted attachments, nested emails, etc.

Microsoft Teams Archiving

About Enterprise Vault Microsoft Teams Archiving

Enterprise Vault 14.2 and later introduced capability to archive Microsoft Teams instant messaging (IM) and channel communications using Arctera Merge1 in rich format.

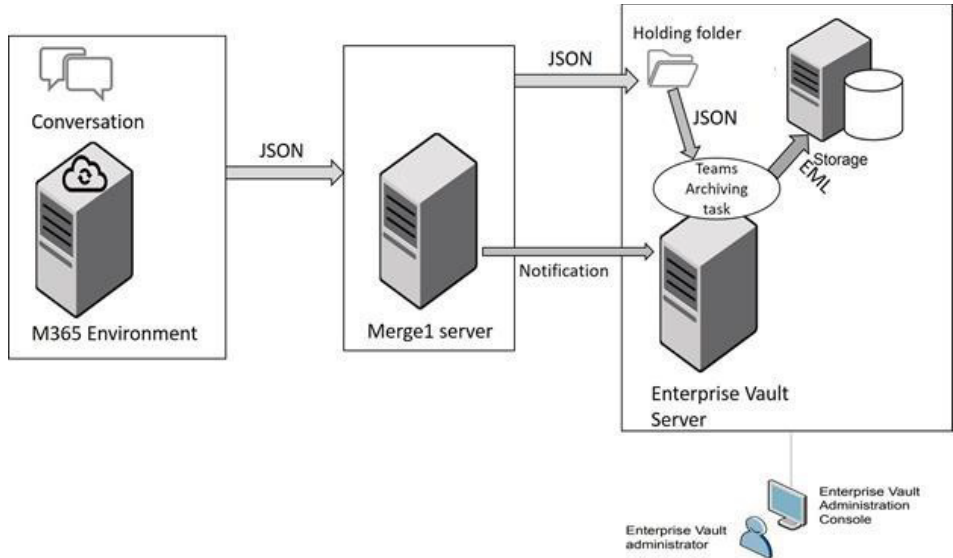
For more details, refer to the *Setting Up Microsoft Teams Archiving Guide*.

Microsoft Teams Archiving Overview

Microsoft Teams archiving with Enterprise vault links the Microsoft 365 environment and Enterprise vault using Merge1 server. Enterprise Vault assumes that archiving is enabled for individual or all users in the M365 Domain, and appropriate archiving policy have been applied. Enterprise Vault requires M365 and Merge1 environment details to be configured to actively archive conversation to a file server. These conversations are then archived in Enterprise Vault environment as EML files.

Figure 1 illustrates the process when archiving conversation from Microsoft Teams.

Figure 1: Microsoft Teams archiving overview



The archiving process is as follows:

- A conversation takes place between Microsoft Teams users.
- Merge1 fetch the conversation JSON to Chat holding folder on Enterprise Vault server.
- Merge1 notifies Enterprise Vault after every 50 messages.
- Enterprise Vault segregates messages by date and chat room id in chat segregated folder.
- Teams archiving task processes files in chat segregated folder and creates EML files for each room ID and day.
- Teams archiving task archives the item in a Teams archive. During processing, the task applies the retention category or retention plan that is specified in the target properties.

Disk Planning

To reduce the number of IOPS on same disk, keep **ChatHolding** folder and **ConvertedChatHolding** on separate disks.

You can update the properties using Enterprise Vault Administration Console.

- Path for the **ChatHolding** folder can be changed from **Importer > Properties > Connection and Folders > Holding Folder**.

- Path for **ConvertedChatHolding** folder can be changed from **Teams archiving task > Properties > Advanced**.

Deployment for Performance Measurements

We have the following deployment for taking the performance measurements of Enterprise Vault Archiving rates for Microsoft Teams content once it has been downloaded using Arctera Merge1.

Hardware:

- Enterprise Vault Server with Indexing Services: 16 Core & 32 GB RAM
- SQL Server: 8 Core & 16 GB RAM
- Merge1 Server: 8 Core & 16 GB RAM

Dataset:

150,000 unique conversations having 15,00,000 messages downloaded locally in the ChatHolding folder. For every message 1 JSON file is downloaded. So, 15,00,000 unique JSON files.

Storage:

The ChatHolding and ConvertedChatHolding folders configured on the local NTFS Partitions.

Performance Numbers

The following table shows the expected ingestion rates for the number of unique conversations getting converted to EML and archived. The average number of messages getting exchanged per conversation is 10. As such, the number of messages number is 10 times the number of conversations getting archived.

Enterprise Vault 14.2

| | Unique Conversations/hour | Unique messages/hour |
|----------------|---------------------------|----------------------|
| Ingestion Rate | 4000* | 40,000* |

The ingestion rate would be dependent on certain factors:

- Number of files shared as part of Teams conversations.
- Size of messages/files being shared.

Performance Numbers

In Enterprise Vault 14.2.1, we made some performance-related adjustments and saw a considerable improvement, as seen in the table below.

Enterprise Vault 14.2.1

| | Unique Conversations/hour | Unique messages/hour |
|----------------|----------------------------------|-----------------------------|
| Ingestion Rate | 180,000* | 1,800,000 (1.8 million)* |

* The ingestion rate would be dependent on certain factors:

- Number of files shared as part of Teams conversations.
- Size of messages/files being shared.

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) |
|-----------------|-----------------------------|
| 16 | 160,000 |

The above table shows the ingest rate for a possible scenario. The figures apply to both NTFS and NAS volumes.

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.

As a minimum, the rate at which Enterprise Vault deletes items that are ready for expiry matches the ingest rate, but it may exceed this rate.

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 Enterprise Vault *SQL Best Practices Guide*, which is available at the following location:

<http://www.veritas.com/docs/100012617>

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 83 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 |
|---------------|------------|
| Full | 6% |

The percentage for Full will be less if there is little indexable content.

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.

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 folder links to Enterprise Vault Search

Enterprise Vault 12.1 and later provides the ability to create a folder shortcut URL file that links to the archived folder in Enterprise Vault Search. Enabling this feature does not impact the overall archiving rate.

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) |
|-----------------|-----------------------------|
| 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.

As a minimum, the rate at which Enterprise Vault deletes items that are ready for expiry matches the ingest rate, but it may exceed this rate.

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 Enterprise Vault *SQL Best Practices Guide*, which is available at the following location:

<http://www.veritas.com/docs/100012617>

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 83 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 |
|---------------|------------|
| Full | 7% |

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.

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 was introduced 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, Arctera 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.

Number of cores

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 sizes
- The file types

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) |
|-----------------|------------------------------|
| 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 (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. 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 Enterprise Vault *SQL Best Practices Guide*, which is available at the following location:

<http://www.veritas.com/docs/100012617>

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 83 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 |
|---------------|------------|
| 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.

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 Arctera 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.

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 if 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 eDiscovery, but it does affect the following:

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

Choice of Enterprise Vault server

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 |

Moving large number of users

| Move type | Comparison with original ingest |
|--------------------------------|--|
| 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.

Backup of indexes

Enterprise Vault 14.2 introduces Elasticsearch as new indexing engine. Elasticsearch engine does not support file-system level backup of Elasticsearch indexes. It supports snapshot mechanism to backup indexes.

Note: After upgrade to Enterprise Vault 14.2 or later, the backup strategy of the legacy 32-bit and 64-bit non-Elasticsearch indexes will remain unchanged.

For details, refer to the white paper for backup of indexes

https://www.veritas.com/support/en_US/doc/EV_WHITEPAPER_BACKINGUP_EV.

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 advanced site settings that control conversion. Changes should be made with care.

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

IFilters and Optical Character Recognition of image files

Enterprise Vault 12 introduces support for Windows IFilters and indexing text content within certain image files using Optical Character Recognition (OCR), including GIF, JPG/JPEG, PNG, and TIF/TIFF.

Enterprise Vault 12.2 extends the OCR support to processing images embedded within documents, and 12.3 provides additional granularity to control which types of documents are processed for embedded images.

Enterprise Vault works in conjunction with the Windows IFilter to perform the OCR processing of image files and retrieve all text content for indexing.

This additional processing can be processor intensive, so it may potentially impact the archiving and indexing throughput.

With embedded image processing disabled, for typical distribution of office type data, the impact is usually negligible. However, with embedded image OCR processing enabled the archiving throughput can be reduced significantly, typically around 20% - 50%.

If image OCR processing is not required for certain image types, you can save processing power by removing the image file types from conversion.

IFilters and Optical Character Recognition of image files

| Site Advanced content conversion setting name | Value |
|---|--|
| File types for OCR conversion | String value containing list of file types. The list format is: <code>.filetype[.filetype]</code> Each file type must be prefixed by a period. Default: .GIF.JPG.JPEG.PNG.TIF.TIFF Only the above file types are currently supported. |

If OCR processing of embedded images is not required, you can increase archiving throughput and save processing power by disabling embedded image processing.

| Site Advanced content conversion setting name | Value |
|---|----------------|
| OCR Conversion of embedded images | Off (default). |

If the embedded image OCR processing is only required for certain document types, you can save processing power by specifying individual file types for embedded image processing.

| Site Advanced content conversion setting name | Value |
|--|--|
| File types for OCR conversion of embedded images | String value containing list of file types. The list format is: <code>.filetype[.filetype]</code> Each file type must be prefixed by a period. Default: * This is a list of the container documents that will be processed for embedded images. |

Alternatively, processing of scanned pages within PDF documents can be enabled without enabling the full embedded image OCR processing by disabling embedded image processing as above, and enabling PDF scanned pages as below.

| Site Advanced content conversion setting name | Value |
|---|----------------------|
| OCR Conversion of scanned pages | On (default is Off). |

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. 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. However, the general formatting will be lost for previewing items in Enterprise Vault Search. There are Advanced Site Settings that you can change to force this.

Prior to Enterprise Vault 11.0.1, Excel files were converted to text by default because of the expense of converting. With the 64-bit converters in Enterprise Vault 11.0.1 and later, Excel files are converted to HTML by default.

| Site Advanced content conversion setting name | Value |
|---|--|
| File types converted to text | <p>String value containing list of file types.</p> <p>The list format is:</p> <p><code>.filetype[.filetype]</code></p> <p>Each file type must be prefixed by a period. For example:</p> <p><code>.DOC.XLS.XLSX.XSLM</code></p> <p>All file types can be converted to text by using the * wildcard character.</p> |

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 audio or video files.

Conversion timeout

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.

| Site Advanced content conversion setting name | Value |
|---|--|
| File types excluded from conversion | String value containing list of file types. The list format is: <code>.filetype[.filetype]</code> Prefix each file type with a period. For example: <code>.WAV.WMA</code> |

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.

| Site Advanced content conversion setting name | Value |
|---|-----------------------|
| Conversion Timeout | Default: 10 (minutes) |

Amazon Web Services (AWS) Cloud

Enterprise Vault now supports Amazon Simple Storage Service (S3) as primary storage, letting you store primary archived data in the AWS public cloud. Enterprise Vault now supports Amazon Commercial Cloud Services (C2S) to store primary archived data in the AWS Government cloud for US Federal Agencies.

Arctera recommends deploying Enterprise Vault in AWS to get all the benefits that AWS provides for IaaS by using Cloud Native services.

For hybrid environments where Enterprise Vault is deployed on-premise and primary storage resides in the cloud, there would be a significant impact on archiving, indexing, and retrieval performance. Additionally, any retrieval will incur higher costs making the hybrid environment less optimal for Supervision and Discovery.

When you create a partition for Amazon Simple Storage Service (S3), the data is stored in non-WORM mode by default. You can choose to create the partition for Amazon Simple Storage Service (S3) in WORM mode using S3 Object Lock feature of Amazon. In this case, you need to set additional permissions for the IAM Role and STS Assume Role authentication methods. For details, refer to [Using Amazon Simple Storage Service \(S3\) as a primary storage for Enterprise Vault](#).

Deployment for Performance Measurements

Enterprise Vault performance varies based on the Amazon environment used, such as the storage class and the authentication method used. The Instance type c5.2xlarge has been used to measure the performance, and it has been observed that the performance matches the Enterprise Vault recommendations. The c5.2xlarge instance is well suited for high-

Deployment for Performance Measurements

performance computing and any other similar tasks. Deploying a more performance-oriented instance yields higher performance.

For best practices on deploying Amazon S3, see [Best Practice for Implementing Enterprise Vault in AWS and Microsoft Azure Cloud](#).

The following deployments are used for measuring the performance:

- Instance Type: c5.2xlarge
- Enterprise Vault Partition: Amazon S3 Partition with the following:
 - Authentication = IAM Role
 - Storage Class = Standard
 - Encryption = None
- Enterprise Vault 14.1
 - Content Source = SMTP
 - Number of items = 15,00,000

Note: For EV 14.2 and later we recommend using instance type C5 4xLarge or higher. The minimum recommended configuration for EV 14.2 and later is 16core and 32GB RAM.

Performance Numbers

Enterprise Vault observed the following performance metrics for a deployment in Amazon with EC2 instance and S3 Standard as storage.

| | Items per hour | Size per item in KB |
|----------------|-----------------|---------------------|
| Archiving time | 85000- 95000 | 167 |
| Indexing time | 50000 – 60000 | 167 |
| Ingestion rate | 80000 – 90000 | 167 |
| Expiry time | 220000 – 240000 | 167 |

Note: Performance numbers for archiving and storage expiry will be slightly on the lower side when the partition is created in WORM mode using the S3 Object Lock feature as compared with the partition created in the default non-WORM mode.

Microsoft Azure Cloud

Enterprise Vault now supports Microsoft Azure Blob Storage as primary storage, letting you store primary archived data in the Azure public cloud. Enterprise Vault also supports Microsoft Azure Blob Storage as primary storage, letting you store primary archived data in the Azure Government cloud for US Government Agencies.

You can use Hot and Cool access tier to store and access data. You can use this primary partition to archive, restore, search, and delete the data when Enterprise Vault is hosted in the on-premise and cloud network.

Arctera recommends deploying Enterprise Vault in Microsoft Azure Cloud to get all the benefits that Azure Cloud provides for IaaS by using Cloud Native services.

For hybrid environments where Enterprise Vault is deployed on-premise and primary storage resides in the cloud, there would be a significant impact on archiving, indexing, and retrieval performance. Additionally, any retrieval will incur higher costs making the hybrid environment less optimal for Supervision and Discovery.

Deployment for Performance Measurements

Enterprise Vault performance varies based on the Azure environment used, such as the access tier and the authentication method used. The Instance type F8s_v2 has been used to measure the performance, and it has been observed that the performance matches the Enterprise Vault recommendations.

For best practices on deploying Microsoft Azure Government Cloud or Microsoft Azure Blob Storage, [Enterprise Vault™ Best Practice for Implementing Enterprise Vault on AWS and Microsoft Azure Cloud](#).

The following deployments are used for measuring the performance:

- Instance Type: F8s V2

- Enterprise Vault Partition: Microsoft Azure Blob Storage Partition with the following:
 - Authentication = Standard
 - Access tier = Default
 - Encryption = Microsoft-managed (enabled by default)
- Enterprise Vault 14.1
 - Content Source = SMTP
 - Number of items = 15,00,000

Note: For EV 14.2 and later we recommend using instance type F16s v2 or higher. The minimum recommended configuration for EV 14.2 and later is 16core and 32GB RAM.

Performance Numbers

Enterprise Vault observed the following performance metrics for a deployment in Microsoft Azure Blob Storage.

| | Items per hour | Size per item in KB |
|----------------|----------------|---------------------|
| Archiving time | 90000 – 100000 | 167 |
| Indexing time | 40000 – 50000 | 167 |
| Ingestion rate | 90000 – 100000 | 167 |
| Expiry time | 90000 – 100000 | 167 |