

---

# EMC Celerra iSCSI Solutions Microsoft Exchange 2003 Best Practices

*Storage Configuration Guidelines*

---

**Abstract**

This paper covers the storage configuration guidelines and best practices to host Microsoft Exchange 2003 on EMC Celerra systems via iSCSI.

Published September 2005

---

Copyright © 2005 EMC Corporation. All rights reserved.

EMC believes the information in this publication is accurate as of its publication date. The information is subject to change without notice.

THE INFORMATION IN THIS PUBLICATION IS PROVIDED “AS IS.” EMC CORPORATION MAKES NO REPRESENTATIONS OR WARRANTIES OF ANY KIND WITH RESPECT TO THE INFORMATION IN THIS PUBLICATION, AND SPECIFICALLY DISCLAIMS IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE.

Use, copying, and distribution of any EMC software described in this publication requires an applicable software license.

EMC<sup>2</sup>, EMC, EMC ControlCenter, AlphaStor, ApplicationXtender, Catalog Solution, Celerra, CentraStar, CLARAlert, CLARiiON, ClientPak, Connectrix, Co-StandbyServer, Dantz, Direct Matrix Architecture, DiskXtender, Documentum, EmailXtender, EmailXtract, HighRoad, Legato, Legato NetWorker, Navisphere, OpenScale, PowerPath, RepliStor, ResourcePak, Retrospect, Smarts, SnapShotServer, SnapView /IP, SRDF, Symmetrix, TimeFinder, VisualSAN, VSAM Assist, Xtender, Xtender Solutions, and where information lives are registered trademarks and EMC Developers Program, EMC OnCourse, EMC Proven, EMC Snap, EMC Storage Administrator, Access Logix, ArchiveXtender, Authentic Problems, Automated Resource Manager, AutoStart, AutoSwap, AVALONidm, C-Clip, Celerra Replicator, Centera, CLARevent, Codebook Correlation Technology, Common Information Model, CopyCross, CopyPoint, DatabaseXtender, Direct Matrix, DiskXtender 2000, EDM, E-Lab, EmailXaminer, Engenuity, eRoom, FarPoint, FLARE, FullTime, Graphic Visualization, InfoMover, Invista, MirrorView, NetWin, NetWorker, OnAlert, Powerlink, PowerSnap, RepliCare, SafeLine, SAN Advisor, SAN Copy, SAN Manager, SDMS, SnapImage, SnapSure, SnapView, StorageScope, SupportMate, SymmAPI, SymmEnabler, Symmetrix DMX, UltraPoint, Viewlets, VisualSRM, and WebXtender are trademarks of EMC Corporation. All other trademarks used herein are the property of their respective owners.

Part Number H1784

# Table of Contents

- Introduction ..... 4**
- Definition of Terms ..... 4**
- General Environment Recommendations ..... 5**
- Server Recommendations ..... 6**
- Exchange Configuration Recommendations ..... 8**
- Storage Recommendations ..... 8**
  - General Recommendations ..... 8
  - Physical Disk Drive Recommendations ..... 9
  - Storage Configuration Recommendations ..... 9

## Introduction

It is important to plan an Exchange solution that can grow while maintaining optimum performance, high availability, and disaster recovery. This document is meant to be a resource guide for optimizing the performance for Exchange 2003 storage configuration on EMC® Celerra® via iSCSI. It also provides a step-by-step process of planning the storage configuration for Microsoft Exchange 2003 for a single server and multiple servers.

The intended audience for this white paper is IT administrators and system engineers who have an interest in implementing Microsoft Exchange 2003 using EMC Celerra systems. It is assumed that the reader has a general knowledge of Microsoft Exchange, Active Directory, and EMC Celerra features and terminology.

## Definition of Terms

**Active Directory:** An advanced directory service introduced with Windows 2000 Server. It stores information about objects on a network and makes this information available to users and network administrators through a protocol such as LDAP.

**Automatic Volume Management (AVM):** A feature of the Celerra Network Server that creates and manages volumes automatically, without manual volume management by an administrator. AVM organizes volumes into pools of storage that can be allocated to file systems.

**Data Mover:** A Celerra Network Server cabinet component running the data access in real time (DART) operating system that retrieves files from a storage device and makes the files available to a network client.

**Disk volume:** On Celerra systems, a physical storage unit as exported from the storage array. All other volume types are created from disk volumes.

**iSCSI (Internet SCSI):** A protocol for sending SCSI packets over TCP/IP networks.

**iSCSI initiator:** An iSCSI endpoint, identified by a unique iSCSI-recognized name that begins an iSCSI session by issuing a command to the other endpoint (the iSCSI target).

**iSCSI target:** An iSCSI endpoint, identified by a unique, iSCSI-recognized name that executes commands issued by the iSCSI initiator.

**RAID:** Redundant array of independent disks. A method for storing information where the data is stored on multiple disk drives to increase performance and storage capacities and to provide redundancy and fault tolerance.

**RAID 1:** RAID method that provides data integrity by mirroring (copying) data onto another disk. This RAID type provides the greatest assurance of data integrity at the greatest cost in disk space.

**RAID 5:** Data is striped across disks in large stripes. Parity information is stored so data can be reconstructed if needed. One disk can fail without data loss. Performance is good for reads, but slower for writes.

**RAID group:** The EMC CLARiiON® storage-system term for a Celerra disk group.

**SP:** Storage processor on a CLARiiON storage system. On a CLARiiON storage system, a circuit board with memory modules and control logic that manages the storage-system I/O between the host's Fibre Channel adapter and the disk modules.

**SP A:** Storage processor A. A generic term for the first storage processor in a CLARiiON storage system.

**SP B:** Storage processor B. A generic term for the second storage processor in a CLARiiON storage system.

**VSS (Volume Shadow Copy Service):** A Windows service and architecture that coordinates various components to create consistent point-in-time copies of data called shadow copies.

## General Environment Recommendations

**Recommendation #1** Use Microsoft Clustering for high availability and to allow non-disruptive server maintenance and software upgrades.

A cluster is a collection of servers known as nodes that together provide a single, highly available system for hosting applications such as Exchange 2003. The Microsoft Server 2003 service pack 1 Enterprise Edition supports iSCSI clusters of up to eight nodes.

Microsoft clustering offers the following benefits:

- **High availability**  
Microsoft clusters provide a highly available messaging system that can protect against Exchange server failures of hardware, operating systems, device drivers, or applications. If one of the nodes in a cluster is unavailable as a result of failure, another node immediately begins providing service.
- **Nondisruptive messaging service during upgrade or maintenance**  
The Exchange service does not have to be disrupted for server maintenance or software upgrades in a Microsoft cluster environment.
- **Scalability**  
Additional nodes can be added to the cluster when the overall load of the cluster exceeds its capability.
- **Manageability**  
It is easy to inspect the status of all cluster resources and move workloads around onto different nodes.

At the time of release of this document, EMC's Celerra platforms have been qualified by Microsoft to be configured with clusters of two server nodes, as active/passive. For updated information and details on node configuration and hardware that Celerra systems can support, consult the Microsoft support matrix: <http://www.microsoft.com/windows/catalog/server/>.

**Recommendation #2** Use Legato Replistor to replicate to a remote standby server for remote disaster recovery capabilities.

EMC Legato® Replistor® software provides a reliable remote data replication solution with Exchange 2003. In the Replistor configuration there is a source system, which is the production Exchange server containing the data to be protected, and a target system, which is a standby Exchange server where the data is replicated.

The Replistor software can provide automated failover to a remote site in which the processing and identity of the production server is transferred to a standby server. Whereas MSCS is an excellent choice for local server high availability, the prerequisites for geographically dispersed clusters (which are required for Microsoft clusters when the nodes reside in separate locations) are often difficult and costly to meet. Replistor offers a simple alternative—replicating data via any IP WAN or LAN, and not requiring the failover site to have hardware that matches the production environment. When the production server fails, it initiates a series of operations including starting and stopping the Exchange services on production and standby servers, dynamically updating the DNS database for alias creation/modification, and updating Active Directory entries for Exchange specific attributes. Using a LAN and a WAN allows the production system and standby system to reside in two different sites. Because of this capability, Replistor can protect an entire site.

**Recommendation #3** Use Gigabit Ethernet for iSCSI network connection between Exchange server(s) and the Celerra system.

Maintaining optimal network performance is crucial to the deployment of Exchange on iSCSI because there is considerable network traffic generated by the Exchange 2003 server. For optimum network performance, use Gigabit Ethernet cabling, switches, and network interface cards for network connection between Exchange 2003 server(s) and Celerra systems.

**Recommendation #4** Use Replication Manager/SE (RM/SE) for Celerra and VSS to implement instant local recovery capabilities.

Replication Manager/SE (RM/SE) for Celerra has the ability to create point-in-time replicas of databases and file systems residing on Celerra iSCSI virtual LUNs allowing some recovery scenarios to bypass the need for loading data from tape. RM/SE provides easy-to-use, quick backup and restore capabilities, and is integrated with Microsoft's Volume Shadow Copy Services (VSS). With RM/SE, the administrator can take multiple backups, both full/incremental and differential. Each Exchange database backup is checked using Microsoft's ESEUTIL tool, and RMSE integrates with your existing tape backup software.

When RM/SE is used for backup, it has the following effect on performance:

- **Increase in write latency.** The database write latency will increase, but is still well below the Microsoft recommended log and database latency values for good performance.
- **Decrease in user space.** Since the RM/SE stores backups on the same file system as the production iSCSI LUN, storage requirements for a given number of users will depend on the number of backups an Exchange administrator wants to keep. This concept of *space reservation* is common to most snapshot implementations and is designed to ensure that snapshots always have sufficient space to complete, and the worst-case restore scenarios can complete successfully. The formula to calculate the total Celerra file system size that needs to be created in order to keep the iSCSI LUN and backup snaps is:

$$\begin{aligned} \text{TotalFileSystemSize} = & (\text{LUN\_Size} * 2) \\ & + [(\text{No\_Of\_Snaps}) * (\text{LUN\_Size} * \text{Change\_Rate})] \\ & + (\text{N} * \text{LUN\_Size}) \end{aligned}$$

Where:

**LUN\_Size** is the size of the production iSCSI LUN where the production Exchange database or transaction log files will reside.

**No\_Of\_Snaps** is the total number of replicas of the production iSCSI LUN that will be kept at any time.

**Change\_Rate** is the amount of change on the production iSCSI LUN between each replica.

**TotalFileSystemSize** is the size that the file system needs to be to handle the production iSCSI LUN and all of its replicas.

**N** is the number of mounted replicas.

## Server Recommendations

**Recommendation #5** Increase the Microsoft Initiator time-out value to 600 seconds.

By default, the Microsoft iSCSI Initiator time-out is set to 60 seconds. This time-out defines how much time the initiator will hold a request before reporting an iSCSI connection error. This value can be increased in order to accommodate some longer outages, such as Data Mover cluster events. If an iSCSI timeout occurs on an Exchange server that hosts the Exchange database and transaction logs on iSCSI LUNs, it will result in unmounting the database.

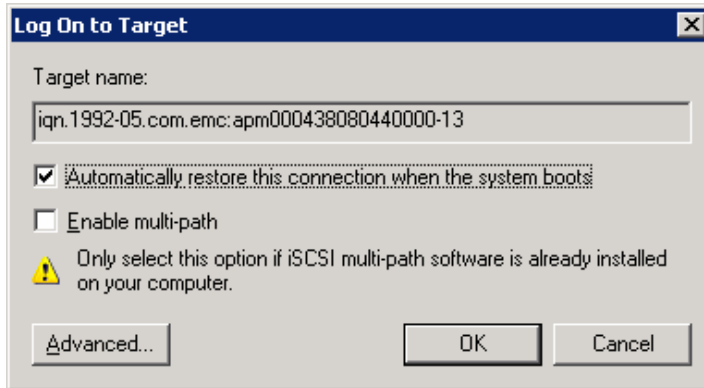
To change the time-out value, search the Windows Registry for the `MaxRequestHoldTime` entry under `HKEY_LOCAL_MACHINE\SYSTEM\CurrentControlSet`, and change the value to 600. The following is an example of the Registry entry in one of the Exchange Servers:

```
HKEY_LOCAL_MACHINE\SYSTEM\CurrentControlSet\Control\Class\ {4D36E97B-
E325-11CE-BFC1-08002BE10318}\0002\Parameters
```

```
MaxRequestHoldTime = 600 (DWORD)
```

**Recommendation #6** Select the **Automatically restore the connection when the system boots** checkbox when configuring the iSCSI Initiator on the Exchange server.

When an Exchange server is rebooted, the iSCSI disks will not be available unless the **Automatically restore the connection when the system boots** checkbox is selected. Select the checkbox from the **Log On to Target** dialog box in the Microsoft iSCSI initiator window.



**Figure 1. Log On to Target Dialog Box with Automatic Restore Option Enabled**

**Recommendation #7** If public folders are heavily used, move them to an iSCSI LUN that is not heavily used by other Exchange mailbox stores.

Exchange public folders can be used as a shared document repository, discussion groups, shared calendars, and several other purposes. These public folder uses are not considered part of the projected number of Exchange mailbox users supported in this paper since some companies use this feature extensively and others ignore it.

If public folders are used extensively, it is important to move the Public folder database onto a Celerra iSCSI LUN for better performance. Before moving the database, it is advisable to monitor the I/O load exerted on the public folder and move it to an iSCSI LUN where its file system is used as a database repository. It is also important to choose a file system that the Exchange database does not heavily use so that it will have the capability to handle the public folder I/O. iSCSI LUN utilization can be monitored by Windows disk performance counters from perfmon.

**Recommendation #8** Move SMTP message queue folder from the Exchange server's local hard disk to a Celerra iSCSI LUN that is not heavily used by Exchange mailbox stores.

By default, SMTP messaging queues are stored on the Exchange server's local hard disk. SMTP queues are used not only for inbound/outbound mail traffic, but also for mail traffic between mailbox stores. SMTP queues are best placed on high-availability storage since the messages in transit will be lost if the SMTP queue storage is lost—a good reason to place them on a Celerra iSCSI LUN. In addition, when Exchange server is used very heavily, it is a good practice to move the SMTP queue folder to Celerra iSCSI LUN for better performance. Choose a file system that is not heavily used by the Exchange database so that it will have the capability to handle the SMTP queue I/O.

## Exchange Configuration Recommendations

**Recommendation #9** Create multiple Exchange storage groups for best performance and fewer Exchange storage groups for effective Exchange server resource utilization.

Any time a user sends or reads a message, or any time a user modifies data stored in his or her mailbox, the change is first committed to the transaction logs. A storage group has its own set of transaction logs and the log I/O operation is sequential. If multiple storage groups are used, the number of parallel log operations will increase since the log operation between storage groups is parallel. On the other hand, multiple storage groups tend to use more server resources than few storage groups.

**Recommendation #10** Create multiple Exchange mailbox stores for quick backup and recovery and fewer Exchange mailbox stores for easy administration.

Having a small number of mailboxes with multiple mailbox stores has advantages of quick backup and recovery as well as minimal mailbox disruption if data corruption occurs. On the other hand, having a large number of mailboxes with fewer mailbox stores is easy to administer since fewer mailbox stores will need to be maintained.

## Storage Recommendations

### *General Recommendations*

Storage design is very important for the Exchange environment because disk subsystem bottlenecks are generally the cause of more performance problems than processor or memory deficiencies.

**Recommendation #11** Plan storage layout for performance, not capacity.

The most common error people make when planning an Exchange server is designing for capacity and not for performance or IOPS (I/O per second). The most important single storage parameter for performance is disk latency. High disk latency is synonymous with slower performance. Microsoft guidelines for good performance are:

- Average read and write latencies below 20 ms
- Maximum read and write latency below 50 ms

In today's disk technology, the increase in storage capacity of a disk drive has outpaced the increase in IOPS. Therefore, the IOPS capacity is the standard to use when planning Exchange storage configurations.

**Recommendation #12** For optimum performance, follow Table 1 to find the number of mailboxes that can be hosted on an Automatic Volume Management (AVM) file system.

The AVM feature of the Celerra Network Server automates volume creation and management. AVM system-defined pools provide a simple way to create and manage file systems through automatic creation and management of volumes. This eliminates the need to manually create stripes, slices, or meta volumes, while supporting high-availability and best performance considerations.

**Table 1. Mailbox Count for Optimum Performance**

Mailbox Profile	Description	Mailbox Size	RAID 1 (User count for each set of 8 spindles)	RAID 5 (User count for each set of 20 spindles)
Light	Infrequent mail access	< 50 MB	4000 Users	6000 Users
Average	Constant mail access	75 –100 MB	2000 Users	3000 Users
Heavy	Active mail access	100 – 200 MB	1000 Users	1500 Users

The AVM creates a file system using eight spindles for a RAID 1 configuration and 20 spindles for a RAID 5 configuration. Table 1 shows the maximum number of mailboxes that can be hosted by such file systems for different mailbox profile users. Since the recommendation is to plan the storage according to IOPS, the mailbox size can be increased for each mailbox profiles without any performance degradation. The number is valid for both NS500 and NS700 systems, but it does not account for the extra capacity that will be required in each file system for snapshots of the mailboxes. For more on snapping iSCSI LUNs containing mailboxes, refer to Recommendation #4.

## Physical Disk Drive Recommendations

**Recommendation #13** Use high-rpm disk drives for best Exchange performance.

Higher-rpm drives provide higher overall random access throughput and shorter response times than slower-rpm drives. For optimum performance, higher-rpm drives are recommended.

**Recommendation #14** Use Fibre Channel disk drives for best Exchange performance.

For best performance, Fibre Channel drives are always recommended for Exchange I/O because of their significantly better performance with random I/O, the dominant I/O pattern for Exchange databases.

**Recommendation #15** Use Advanced Technology-Attached (ATA) drives when storage costs are the primary consideration.

ATA drives have larger storage space, slower response rotational speed, and moderate performance with random I/O. But they are less expensive than Fibre Channel drives. ATA drives are therefore the best option when storage costs are the primary consideration.

## Storage Configuration Recommendations

**Recommendation #16** Use diskpar to align iSCSI LUNs for best performance.

This is the most critical recommendation among all the other recommendations. When Microsoft Disk Manager formats the Celerra iSCSI LUNs, it always creates the partition starting at the 64th sector, therefore misaligning it with the underlying physical disk. Due to this misalignment, Exchange I/O that would have fit evenly on the disks may result in more than one I/O to the physical disk drive. To fix the disk alignment, Microsoft provides a command line tool `Diskpar.exe`. `Diskpar.exe` that comes with the Windows 2000 Resource Kit and it can explicitly set the starting offset in the Master Boot Record

(MBR). This utility is merged with `diskpart.exe` on Windows Server 2003 Service Pack 1 Support Tools.

Exchange Server 2003 writes data in multiples of 4 KB I/O operations (4 KB for the databases and up to 32 KB for streaming files). Since the Celerra file-system block size is 8 KB, which is a multiple of the Exchange I/O of 4 KB, use `diskpar` to set the offset to 16 sectors which equals 8 KB. The disk alignment technique increases the Exchange I/O performance significantly (up to 65 percent) on Celerra iSCSI LUNs.

**Recommendation #17** Use dedicated Celerra file systems for Exchange.

File systems created for Exchange applications should not be used for other I/O operations. This will ensure predictable performance for Exchange.

**Recommendation #18** Use AVM to create performance-optimized file systems.

Some of the advantages of using AVM with system-defined pools to create a file system are as follows:

- Disk volumes are selected from multiple RAID groups.
- Disk volumes are selected from same RAID type.
- Least-utilized disk volumes are used first.
- Disk volumes are chosen to achieve SP balancing.
- Disk volumes are chosen to achieve bus balancing.

**Recommendation #19** Use RAID 1 for best Exchange performance.

When a user sends an e-mail message through Exchange, the system first writes to the transaction log synchronously and then commits the logs later to the Exchange database in an asynchronous *lazy write*. The transaction logs are written sequentially and the Exchange database is read and written randomly. Although the mix varies from customer to customer, roughly 90 percent of the Exchange I/O goes to the database and only 10 percent goes to the transaction logs.

The typical performance bottleneck comes from the Exchange database since it comprises the larger volume of I/O and is very random in nature. For random I/O, RAID 1 outperforms RAID 5 on Celerra with CLARiiON backend. Both RAID 1 and RAID 5 RAID groups perform equally well on random I/O until the write cache becomes saturated. The advantage of using RAID 1 is that it can flush the cache of random writes faster than RAID 5. Additionally, in the event of disk failure, rebuild time of RAID 1 is much faster than rebuild time of RAID 5. Although the RAID 1 storage efficiency is lower than RAID 5, fewer RAID 1 drives are needed overall versus RAID 5 drives to meet the I/O demand for a given number of Exchange users (remember the best practice from earlier—focus on IOPS, not capacity). RAID 1 achieves 65 percent higher IOPS than RAID 5 on Fibre Channel drives for Exchange load. The NS500/NS700 disks can be configured as RAID 1 or RAID 5 to host Exchange database and transaction logs.

**Recommendation #20** Use RAID 5 if RM/SE for Celerra is to be used for instant local recovery capabilities.

RM/SE for Celerra requires that iSCSI LUN and its point-in-time replicas must be on the same file system. This requires that the file systems must be a few times larger than the iSCSI LUN. RAID 5 configurations allow for the creation of larger file systems than RAID 1 configuration for the same number of spindles because RAID 5 has only 20 percent overhead where as RAID 1 has 50 percent overhead. Therefore RAID 5 is more cost effective than RAID 1 in an RM/SE environment.

---

**Recommendation #21** For any two file systems that share the same physical disks, Exchange databases should go in one, while transaction logs should go in the other.

In certain cases, two different file systems might share the same spindles. For highly optimal performance in these instances, be sure that one file system is used for databases, while the other is used for transaction logs.

**Recommendation #22** Do not store both the logs and mailbox stores from the same Exchange storage group on the same physical disks.

This is an added constraint to Recommendation #21. For highly optimal performance, Exchange databases and transaction logs from the same storage group should be in different file systems that do not share spindles.

**Recommendation #23** Keep Jet and streaming databases together on the same iSCSI LUN.

All Exchange Mailbox databases have a Jet database (.edb file) where the content is generated by MAPI clients, and a streaming database (.stm file) where the content is generated by Internet protocol clients. Since both files compose a single Mailbox database, it is advisable to keep them together on the same Celerra iSCSI LUN. In addition, RM/SE for Celerra will not take a replica if both files are not in the same Celerra iSCSI LUN.

**Recommendation #24** Use a dedicated network interface on the Celerra system for iSCSI network traffic to the Exchange server(s).

Using a dedicated network interface on a Celerra Data Mover for iSCSI traffic increases network throughput and reduces interference from other network clients, therefore increasing the Exchange server performance. A single dedicated Celerra network interface can be shared with multiple Exchange servers as long as the only traffic on the interface is iSCSI network traffic.