

EMC CLARiiON Storage Solutions: Microsoft Exchange 2007

Best Practices Planning

Abstract

This white paper presents the latest storage configuration guidelines and best practices for Microsoft Exchange Server 2007 on EMC[®] CLARiiON[®] storage systems.

August 2008

Copyright © 2007, 2008 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.

For the most up-to-date listing of EMC product names, see EMC Corporation Trademarks on EMC.com

All other trademarks used herein are the property of their respective owners.

Part Number H4060.1

EMC CLARiiON Storage Solutions:
Microsoft Exchange 2007
Best Practices Planning

Table of Contents

Executive summary	5
Introduction	5
Audience	5
Microsoft Exchange Server 2007	5
Environmental parameters for storage design.....	6
User community information	6
Mailbox count and size	6
User activity	6
Other user profile characteristics.....	6
Backup/recovery requirements	7
Other organizational requirements or constraints.....	7
Sample measurements	7
Planning storage for the Exchange production data.....	8
Exchange storage groups	8
How many databases per storage group?	8
How many storage groups per server?	8
How many storage groups per database/log LUN pair?	9
Calculating the base I/O per user requirement.....	9
Calculating the IOPS requirement for an Exchange environment.....	10
RAID types and the read/write ratio	10
Other factors that may impact I/O	11
Calculating the capacity requirement for database LUNs	12
Offline defragmentation	13
Large mailboxes	13
Choosing a RAID and disk type	15
Comparing RAID 1/0 to RAID 5.....	15
Comparing 10k rpm to 15k rpm drives	16
Comparing 73 GB, 146 GB, and 300 GB drives	16
Charting for the optimal disk and RAID type	17
Summary	18
MetaLUNs	19
Building blocks	20
Log LUN configuration	20
Additional storage considerations for the Exchange production data	22
Other Exchange server roles.....	22
Public folders	22
Mixing database and log LUNs on the same drives.....	23
Planning storage for local recovery	23
SnapView for disk-based replication.....	23
Clone-based replication.....	24
Snapshot-based replication.....	25
RecoverPoint CDP	25
Exchange continuous replication	25
Online backup to disk.....	26
Planning storage for local message archiving.....	27

Additional infrastructure considerations.....	27
Storage system considerations.....	27
Storage system tuning	29
Windows file-system alignment.....	29
Windows allocation unit size.....	30
Network connectivity	30
Fibre Channel SAN	30
iSCSI SAN	31
Direct attach	32
Putting it all together	32
Consider site-specific constraints.....	32
Configure the cleanest looking layout diagram	32
Validate the design.....	32
Conclusion	33
References	33
EMC	33
Microsoft.....	33

Executive summary

When you implement Microsoft Exchange 2007, your messaging environment is likely to be very different from what it was in an earlier version. The Exchange architecture has changed, and the user profile characteristics, connectivity choices, storage technology, and options for managing and protecting your data are all evolving.

However, the basic principles for designing storage as part of a robust, reliable messaging system remain largely the same. It is still important to design to meet your peak load requirements, and to remember that Exchange is sensitive to I/O latency. When you maintain balanced storage resources, you provide the best possible performance for your users, and make optimum use of your storage investment. By architecting for high availability, your design has the best chance of meeting or exceeding your SLA requirements.

Introduction

This white paper describes the step-by-step process for planning the layout for Exchange data on an EMC® CLARiiON® storage system. It also provides considerations and the latest best practice recommendations. The approach taken is to design a layout that meets the following goals:

- Optimal performance — During peak periods, user response times are still acceptable and there is no buildup of mail queues.
- Efficient backup and rapid recovery — Backups complete within the allotted window, with an acceptable impact on the production environment. Local recovery meets the service level agreement (SLA) requirement.
- Simplicity of design — The resulting configuration is straightforward to implement and easy to manage and expand.

In addition to recommendations for production data storage layout, the paper includes considerations for configuring CLARiiON storage for Exchange backup, local replication, and archiving.

Audience

The intended audience for this white paper is customers – including IT planners, storage architects, and administrators - and EMC technical staff and partners.

The reader should have a general knowledge of Microsoft Exchange and Windows technology, as well as an understanding of basic CLARiiON features and terminology.

Microsoft Exchange Server 2007

There are a number of changes in Exchange 2007 that affect storage configuration. These include:

- 64-bit architecture — This allows the use of significantly more server memory, which in turn caches more I/O that would otherwise have gone to disk.
- Local and cluster continuous replication (LCR/CCR) — These new log shipping capabilities provide additional protection options.
- Storage groups — The maximum number of Exchange storage groups has increased from four to 50; the maximum number of databases is also 50, which offers increased design flexibility.
- Log file size — Log files have changed from 5 MB to 1 MB, allowing a finer level of granularity for log shipping.
- Page size — The size of the database page has changed from 4 KB to 8 KB, and now matches the default CLARiiON cache page size.
- Content indexing — Indexing has been reworked to provide significantly improved performance.
- Unified messaging — Exchange 2007 includes improved voice mail integration.

-
- Hub transport role — All Exchange mail is routed through a hub transport server, even mail within the same Exchange database.
 - Deleted item retention — This Exchange setting is the number of days that users' deleted messages are kept in the database after they have been removed from the Deleted Items folder. The default period has increased from 7 to 14 days.
 - Large mailboxes — Exchange 2007 includes improved handling of very large mailboxes.

Environmental parameters for storage design

There are several factors to consider when designing a storage design for an Exchange environment. Before beginning your design, you should have a comprehensive understanding of both how your organization utilizes its current messaging system and what the requirements for the new system will be.

This section describes data that you can gather to help you understand your current system. Whenever possible, it is valuable to support this data with empirical measurements (concurrent users, messages sent/received, log files/day, and so on) from the current environment.

User community information

This information will help you understand the I/O profile for a set of users over time, and what their storage requirements are. The section “Sample measurements” in this paper provides some recommended Windows System Monitor counters for quantifying some of the following information.

Mailbox count and size

- How many total users do you have today?
- What is your anticipated growth over the next few years?
- How many mailboxes are not associated with an individual user (such as a central help desk mailbox)?
- What are the mailbox size limits?
- What is the deleted items retention period?

User activity

- What is the typical working day?
- When are the peak activity periods?
- Is there geographic dispersal of the users across time zones?
- How many concurrent users are there during the peak period?
- What is the Exchange activity level of the users? This includes:
 - Categorization of the user types helps estimate base IOPS demand (see Table 1).
 - Measured I/O in the existing environment gives the best starting point.

Other user profile characteristics

- What mail clients are used and in what proportion? For example:
 - Outlook (2003/2007, cached or online)
 - Outlook Web Access
 - Mobile devices (for example, BlackBerry)
- Are there special category users—with different security, performance, or backup/recovery requirements?
- Is there anything else pertinent that helps to describe the user profile for this organization?
 - Is there a heavy use of personal folders?

-
- Do users often send large documents?
 - Is there a lot of use of Outlook 2003 shared folders?
 - What are the characteristics of public folder usage? These include:
 - Size of the public store
 - Replication activity among public stores

Backup/recovery requirements

The backup and recovery method you choose is an important factor in your storage design. Once again, measuring the existing environment provides the best starting point for your new design:

- Is there a predetermined backup method?
- What is the timing of online maintenance and various backup activities, and what is the backup window in which they must complete?
- What are the requirements (service level agreement) for recovery?
 - What is the recovery point objective (RPO), or acceptable amount of data loss?
 - What is the recovery time objective (RTO), or acceptable amount of downtime?
- Is a distance replication or disaster recovery (DR) solution planned?
 - What is the distance between the primary site and DR site?
 - What type of network connection links the two sites?
 - What is the available bandwidth for Exchange replication?
 - Are the sites on different subnets?

Other organizational requirements or constraints

This category covers any additional pertinent factors that have already been decided, or have been added as a requirement, including:

- What are the type, number, and location of Exchange servers?
- Are the Exchange servers clustered?
- What is the planned Exchange server role layout?
- What is the SAN/network structure?
- Will any portions of the Exchange environment operate within a virtual machine?
- Are there existing CLARiiON storage systems in place?
- Will **Content Indexing** be enabled?
- What other software will be operating in the Exchange environment? Other software might include:
 - Antivirus software
 - Email archiving solutions
 - Applications integrated with Exchange (for workflow and so forth)
 - Exchange-integrated third-party tools (for mailbox recovery, enhanced monitoring, and so forth)

Sample measurements

You can use the Windows System Monitor tool (PerfMon) to measure a variety of performance characteristics in your current Exchange environment. Select a production server with a user load matching your target community and make sure that the run includes a 1- to 2-hour period during peak user activity. Following are three key sample measurements:

-
- IOPS per user — Use the System Monitor counter **Physical Disk\Disk Transfers/sec** on the Exchange database drive and the following formula:
IOPS per User = (Disk transfers/sec) / (Number of users)
 - Read/Write Ratio— During the peak activity period, the ratio is measured on the database drive as **Disk Reads/sec : Disk Writes/sec**.
 - I/O Latency — You can measure the read or write latency to the database or log drives using the counters **Physical Disk\Avg Disk sec/Read** and **Avg Disk sec/Write**.

Planning storage for the Exchange production data

This section provides guidelines for configuring storage to handle the production Exchange data. When designing a storage configuration for Exchange, the two key measurements to consider are:

- The I/O operations per second (IOPS) that the Exchange environment requires to handle the peak load
- The storage capacity requirement

It is important that the resulting design meets both of these needs.

Exchange storage groups

The Exchange storage group (ESG) is the fundamental unit for layout planning. When backing up Exchange, the elements of an ESG should be treated together. The following are a few common questions about storage groups.

How many databases per storage group?

There can be up to five databases in each Exchange 2007 storage group. For Exchange 2003, the recommendation has typically been to *go wide*, by using the maximum number of storage groups (four), before going *deeper* by configuring more databases within a particular storage group. This is because each storage group has its own associated set of log files, which spreads the risk and offers more granular recoverability. Because the maximum number of databases in Exchange 2007 now matches the maximum number of storage groups (50), usually the best choice is to place a single database in each storage group.

If you are implementing Exchange cluster continuous replication (CCR), there is a limitation of one database per storage group.

How many storage groups per server?

Recommendations for how many storage groups to have on each server are less clear. Considerations include:

- Maximum database size — EMC and Microsoft recommend the following maximum database sizes:
 - 200 GB, when using an EMC rapid recovery solution or Exchange CCR
(This is not a technical limitation. In some configurations with very large mailboxes it may be preferable to use somewhat larger databases rather than a higher number of storage groups.)
 - 100 GB otherwise
- The backup window size and organizational service level agreement (SLA) — These may require a smaller database size.
- The backup/recovery method chosen.
- The tolerance for complexity — Increasing the number of storage groups increases the administrative complexity.

- The desire or need for flexibility — Increasing the number of storage groups offers the most granular management. This allows maintenance and recovery operations that are more compartmentalized and affect the least amount of users possible. It also allows for separate treatment of a set of users with different performance, security, or backup/recovery requirements.

Increasing the number of storage groups may provide performance benefits. Using more ESGs results in more parallel log operations, which makes more user information accessible from memory. However it also diminishes the benefit of single instance storage since that is maintained on a per-database basis.

Choosing 50 storage groups gives you maximum granularity for recovery, and there can also be performance advantages. However, if you have 1,000 users on a server with mailboxes of 200 MB, configuring 50 storage groups – with 20 users per ESG – is likely to be unnecessarily complex.

Then again, for 4,000 or more users with very large mailboxes on a server, you may be required to use a large number of storage groups due to the maximum database size allowed by your SLA.

Much of the testing at EMC has been with eight or 16 ESGs, which EMC feels offers a good balance of performance and manageability for a large percentage of implementations.

How many storage groups per database/log LUN pair?

In most cases, two LUNs should be allocated for each ESG—one for the transaction logs, and one for the database file. Rapid recovery VSS-compliant solutions typically recover an entire LUN at once. By placing each ESG on its own LUN pair, recovery is faster and affects only users in that storage group.

However, especially on a server with a large number of storage groups, there are cases where it may be more practical to use the same database/log LUN pair to hold more than one storage group:

- If the backup/recovery method is at a file level (such as streaming backup or EMC RepliStor[®] replication) rather than a LUN level, it is reasonable to configure multiple storage groups, each in their own folders, on the same LUN. As with all cases of data placement, consider your backup and replication schedules, and combine storage groups on LUNs that result in balanced I/O across the physical drives.
- When multiple ESGs reside on the same LUN, they can share the same free space for growth or offline defragmentation.

Calculating the base I/O per user requirement

The best way to provide enough I/O for your application, especially in a large Exchange environment, is to know your users' usage profiles. Sizing of the storage infrastructure should be based on a careful analysis of the number of current and anticipated users, and their messaging habits and patterns. The fundamental calculation concerns I/Os per second (IOPS) per user.

Table 1 describes four Exchange user categories and provides an estimate of the IOPS demand and recommended server cache per user for each¹.

Table 1. Exchange user profiles

Typical user profile	Messages per day (~50 KB message size)	Expected I/Os per second	Database cache per user
Light	5 sent / 20 received	0.11	2 MB
Average	10 sent / 40 received	0.18	3.5 MB
Heavy	20 sent / 80 received	0.32	5 MB
Very Heavy	30 sent / 120 received	0.48	5 MB

¹ Values referenced from the Microsoft TechNet article *Exchange 2007 – Planning Storage Configurations*

Once you have defined user profiles, you can calculate the total IOPS required by multiplying each user by their predicted use of IOPS. It's best to measure what the user IOPS are now. If you have measured the IOPS per user for an Exchange 2003 environment, you can expect a rough estimate of a 25 percent to 30 percent decrease in IOPS when moving to Exchange 2007 on the same server with 4 GB of memory. The decrease will be greater if you increase the server cache by the recommended amount. However, before designing your storage based on expected Exchange 2007 I/O savings, consider the following:

- It is a best practice to design to handle the I/O requirements during peak periods. This may call for limited dependency on extra server memory, which will be lost in the event of a server crash or server failover. I/O to the storage decreases as the cache gets fully utilized and it can take hours for the cache to get rebuilt.
- Larger mailboxes generate more I/O. If the migration to Exchange 2007 includes significantly larger mailboxes for the users, this will offset some of the I/O savings of moving to the new version.
- Exchange online maintenance often represents the highest I/O demand on the system.
- Other changes made as part of the move to Exchange 2007 will also affect the I/O load. These may include:
 - An increase in the number of users connecting through mobile clients. Count each BlackBerry user as the equivalent of three typical users.
 - Content indexing.
 - A change in the process used for data protection — For example, the use of Exchange CCR requires twice the I/O capacity on the passive node. To support failover in both directions, this calls for configuring double I/O capacity on both sides.

A typical read/write ratio for Exchange 2003 is from 2:1 to 3:1. With cache fully utilized, the read/write ratio for Exchange 2007 is closer to 1:1 because the I/O savings are almost entirely in reads. If your system has a higher write percentage, the type of RAID configuration becomes a more important factor. This is discussed in the “RAID types and the read/write ratio” section of this document.

Organizations typically use many mailboxes in Exchange that are not associated with an individual user. These mailboxes often serve as the central contact point for a group, conference rooms, or integrated applications. Depending on the number of these unassociated mailboxes (it can be significant—10 percent or more) and their activity level (which can vary significantly), it may be appropriate to factor these in to the IOPS calculation. By default, treat these mailboxes as equivalent to the typical user mailboxes, and always include them in the capacity calculation.

Calculating the IOPS requirement for an Exchange environment

This is a key step in configuring CLARiiON storage for good Exchange performance. It is a best practice to configure dedicated disk drives for the Exchange databases. Calculate the periods of highest I/O demand during the day by looking at the anticipated cumulative effect of user activity, system activity (virus checkers), and background activity (local or remote replication). Balance the I/O where possible with scheduling (backup during off-peak hours) and an even distribution of users across ESGs. Then, plan a design that will handle the resulting peak I/O load.

Start with a measurement or an estimate of the I/O profile of the Exchange users in the organization. Plan for the peak user load time—these are typically mid-morning on Monday.

RAID types and the read/write ratio

Depending on the particular organizational requirements, there are two possible RAID type options that can be appropriate for production Exchange database LUNs:

- RAID 1/0 — This offers the best performance with high protection, but only 50 percent of the RAID group capacity is usable. It is frequently recommended because it provides sufficient space across the number of spindles required for handling peak I/O load with today's larger disk drives. On a RAID 1/0

LUN, there are two physical I/O operations for each write requested (a write to each mirrored disk); this is described as a write penalty of two.

- RAID 5 — This configuration offers a higher usable capacity per RAID group than RAID 1/0. It can be effective for environments with very large mailboxes and/or lower IOPS requirements. However, in a RAID 5 group there are four physical I/O operations for each write requested (two reads to calculate parity, one write for data, and one write for parity).

Regardless of the RAID type chosen, it is important to configure enough drives to handle the I/O demand. For further detail, see the section “Choosing a RAID and disk type” in this paper.

Table 2. Write penalty for RAID types

RAID type	Write penalty
RAID 1/0 (striping + mirroring)	2
RAID 5 (striping + parity)	4

Other factors that may impact I/O

There are other administrative operations that may impact I/O. Several of these should be scheduled to take place only during off-peak times (see the following examples). You must account for this I/O activity by increasing the IOPS capacity by some amount over the RAID-adjusted IOPS requirement. To estimate the amount of I/O overhead these additional activities cost, it is best to perform tests in an environment that matches the target production environment as closely as possible. The less sure you are of this overhead, the more capacity you should assign to assure good performance.

Examples of background I/O activity that cannot be scheduled to off-peak times

- High load on the server — The more active mailboxes a server is managing and the less memory it has, the less likely it is that any particular user’s mailbox will be cached.
- Server-based antivirus protection — In addition to the extra read operations, antivirus software can add 20 percent or more to the CPU utilization of the Exchange server.
- Integrated features and applications — The impact here depends on the number and type of any integrated features (such as content indexing) or applications (such as workflow), and the amount of their use.
- Synchronous or asynchronous mirroring — If a mirroring solution is used for distance replication or disaster recovery, the resulting impact on I/O latency must be factored in.
- Log shipping — Exchange LCR and CCR include ongoing shipment of log files. It typically has little impact on a production Exchange server, but on a heavily loaded server log shipping can have an impact on I/O latency.

The total I/O demand during peak user load can be calculated by adding the cumulative overhead of the background activity to the RAID-adjusted IOPS requirement. This overhead is often calculated simply as an added percentage, but some activities (such as virus checkers) involve primarily reads, in which case it’s more accurate to ignore the write penalty in the calculation.

There are other schedulable activities that can significantly increase I/O demand on the Exchange LUNs. You should understand their impact. With an ever-increasing amount of data to back up and a shrinking backup window, it is not uncommon for the peak I/O period for an Exchange environment to take place overnight. If there is any concurrent user activity, it is important to pay attention to response times during this period.

Examples of schedulable activities

- Online backup to disk or tape — This places heavy read activity on the production LUNs. There is added overhead if the Exchange server is used to manage the backup.
- Local clone-based replication — Clone-based replication using an application such as EMC Replication Manager involves synchronizing all source Exchange LUNs in the ESG to their clones.

During the incremental synchronization part of the process there is heavy back-end read activity against the production LUNS.

When the incremental update on the clones is complete, Eseutil performs an integrity check on each page of the database replicas. The validated copy may then also be archived or copied (via SAN Copy™) to a remote site. These checking and archiving functions cause high read activity, but it is to the clones rather than to the production LUNS. As long as best practices are followed and the clone LUNS are on separate drives than the production LUNS, this heavy read activity comes into calculations only in terms of backup scheduling and use of overall array resources.

- Local snap-based replication — The eseutil validation associated with VSS-compliant snapshots will have a more significant impact on production Exchange LUNS than clone-based replication. If snaps are chosen as a backup method with eseutil checking, it is important to factor in the associated impact. This is discussed further in the section “SnapView for disk-based replication.”
- Exchange online maintenance — By default, Exchange schedules online database maintenance for a 4-hour period nightly to perform functions that include online defragmentation and clearing out deleted mailboxes and deleted items that have gone past their retention period. The timing and duration for this maintenance can be adjusted.

Because of the heavy I/O that online maintenance adds, schedule it to take place during the period of lightest activity. At the beginning of online maintenance, Exchange performs an Active Directory lookup for each user in the database. Slightly offsetting the online maintenance start times of the databases reduces the impact of these searches on the Active Directory.

You cannot perform a backup on a database while it is undergoing online maintenance. Maintenance will pause until the backup job for that database completes, but it will not extend the operation past its allotted time window. Take this into consideration to assure that online maintenance gets enough time each day, as Microsoft recommends databases complete a full maintenance pass at least once a week.

In summary, take additional I/O activity into consideration when calculating the anticipated demand. Then design to accommodate the peak I/O load with that overhead factored in.

Calculating the capacity requirement for database LUNS

The capacity calculation for an Exchange 2007 database LUN includes a number of parameters to consider:

- Add the following factors together to estimate the size of the Exchange database:
 - User Count x Mailbox Size — For each category of users in the storage group, multiply the maximum allowed mailbox size by the number of users in that category.
 - Planned growth — If you plan to add users to the storage group or to increase the mailbox maximum at a later date, factor that in.
 - Deleted item retention — You can roughly estimate the space taken up by these items by calculating the total size of the messages received and messages sent by users in the ESG in a day, and multiplying that by the desired deleted item retention period.
 - White space — When online maintenance runs, it removes deleted items that have gone past the retention period and the emptied space is consolidated. The amount of white space will vary, but a rule-of-thumb estimate is to use 10 percent of the Exchange database size.
- Content index — If content indexing is enabled, the index is maintained on the same LUN as the Exchange database. Estimate the size of the content index to be 5 percent of the database size.
- Free space — Allow an additional buffer of 15 percent to 20 percent free space on the LUN for growth and to assure that you don’t run out of capacity. If you plan to perform offline defragmentation on this LUN, allow for a free space equivalent to 110 percent the size of the database (see the “Offline defragmentation” section for additional considerations).

For example, with 400 users in one storage group (no space for offline defrag):

100 heavy users @ 400 MB Mailbox	→ 40 GB	
300 typical users @ 200 MB Mailbox	→ 60 GB	
Sum requirement for mailboxes		→ 100 GB
14 days deleted item retention (2 weeks @ 5 MB per user)	→ 2 x 25 MB x 400 = 20 GB	
White space (5 MB per user)	→ 5 MB x 400 = 2 GB	
Total database size		→ 122 GB
Content index (5% of mailbox total space)	→ 5GB	
LUN free space (~20% of DB size)	→ 25 GB	
Final LUN size		→ 152 GB

When one database is planned for each storage group on its own LUN, deleted item retention is typical, content indexing is enabled, and space is not required for offline defragmentation, you can perform a quick capacity calculation for the LUN by simply adding 50 percent to 60 percent additional space beyond the mailbox total (100 GB + 50% = 150 GB).

Offline defragmentation

In most cases it is no longer necessary to perform offline defragmentation of the databases. Normal online maintenance will defragment the database, but it will not compact the size of the file. The only way to actually shrink the database size is to perform offline defragmentation, where the database is dismounted and Eseutil is used to rebuild a new copy. To have the shortest time offline, this rebuild is performed on the same LUN as the source. There must be free space equaling at least 110 percent of the size of the database for the rebuild. If multiple databases are stored on the same LUN, enough space must be allowed to handle the rebuild of the database with the largest size.

Additional considerations:

- It's possible that a complete offline rebuild of an aged database will result in some improved performance.
- If you have allocated space for a recovery storage group, it's possible to share this space for offline defragmentation.
- Gradually moving mailboxes from one storage group to a new one is potentially an alternative method to renew a database, with lower user downtime. It can also be used to rebalance the activity among storage groups. Don't forget to allocate sufficient capacity in the log LUN of the target storage group (see the section "Log LUN configuration").

Large mailboxes

Exchange mailbox sizes have been trending larger. More than the decreasing cost per gigabyte of storage, this is being caused by:

- Larger messages — The use of HTML and increasing use of bitmap files in signatures increases the average message size by three or more times.
- Larger attachments — This is due to higher resolution graphical content and increasing use of rich media.
- Better search capability — Because Exchange 2007 includes content indexing that is significantly improved, users will be willing to maintain larger mailboxes because they can search for their older mail more efficiently.

- Removal of personal folders (PSTs) — This may be the largest factor of all. Stored on users' PCs and in various file shares, personal folders present problems for protecting the organization's intellectual property and for supporting compliance requirements. Organizations are increasingly absorbing the content of personal folders into the Exchange database and limiting their use.

Exchange 2007 improves handling of very large mailboxes — up to 2 GB and more. However, there are some side effects to consider before deploying very large mailboxes.

- Increased I/O — Although the increase in I/O activity with large mailboxes is much less than it is in Exchange 2003, there is still a considerable cost of moving from, say, a 200 MB mailbox to a 2 GB mailbox.
- Impact on the backup process — Increasing the maximum mailbox size significantly may force a change in the backup process to allow your backup to remain within the nightly backup window.
- Server de-consolidation — With the 50 database maximum per server and the recommended maximum database size of 200 GB (with a rapid recovery solution), there is potential for very large mailboxes to limit the number of users you can place on a server.

For example, 1,000 users with a 1 GB mailbox will require six storage groups, even at the maximum database size (1,100 GB total, including an estimated 100 GB for deleted item retention, database white space, and other internal database structures). Without a rapid recovery solution in place, the recommended database maximum is 100 GB. In that case, 12 storage groups would be required for these 1,000 users. Particularly if your SLA requires smaller databases, deploying very large mailboxes could restrict a server to a smaller number of users than it could otherwise handle. (see Table 3).

Table 3. Storage groups and large mailboxes

SLA	Number of storage groups for 1,000 users			
Database max	250 MB mailbox (320 GB total)	500 MB mailbox (580 GB total)	1 GB mailbox (1,100 GB total)	2 GB mailbox (2,150 GB total)
50 GB	7	12	23	43
100 GB	4	6	12	22
200 GB	2	3	6	11

Email archiving offers an alternative to large mailboxes, providing users with access to older mail while keeping the size of the production Exchange databases manageable. EmailXtender[®] is EMC's email archiving product. It provides:

- Cost-effectiveness — Most older mail can be compressed and maintained on less expensive storage
- Space savings — The archived mail is compressed, with enhanced single-instance storage. This savings is multiplied by the number of Exchange database replicas kept.
- Shorter backup/recovery times — With smaller/fewer Exchange databases, backups and recoveries are faster, and there is flexibility for a larger variety of protection options that fit within the backup window.
- PST assimilation — The content of personal folders can be extracted into the email archive
- Extensive cross-server search capability — This allows timely user- or administrator-based searches.
- Offline access — Like cached Outlook mail and personal folders, a copy of archived email can be kept on users' PCs for offline use.
- Compliance and legal discovery features.

Choosing a RAID and disk type

Regardless of RAID type, a physical disk can handle a certain number of Exchange style IOPS. I/O performance improvements on new disk models have not kept pace with the increase in storage capacity. Consequently, it is particularly important to evaluate Exchange 2007 disk configurations by the environment's I/O requirements as well as its storage capacity requirements.

There is not consistent agreement on the IOPS capability of a disk drive. Although some sequential I/O tests have indicated that a CLARiiON 10k rpm drive can perform at a speed greater than 300 IOPS, we have determined that a more practical value for Exchange 8 KB random I/Os on CLARiiON is in the vicinity of 140 IOPS. Similarly, a practical value to use for 15k rpm drives is 180 IOPS.

Table 4. IOPS per spindle

Disk rpm	CLARiiON disk I/O capacity with Exchange databases
10k	140 IOPS
15k	180 IOPS

Using these values, you can divide into the RAID-adjusted IOPS requirement to determine the number of drives needed to handle the I/O demand. Round the number up to meet RAID 1/0 or RAID 5 requirements.

Comparing RAID 1/0 to RAID 5

It is generally accepted that RAID 1/0 is a better choice for random-write environments like Exchange. The effect is somewhat subtle; since all writes hit the write cache, RAID 1/0 and RAID 5 RAID groups perform equally well until the storage system is sufficiently busy and the write cache becomes saturated. The advantage of using RAID 1/0 rather than RAID 5 with Exchange is that RAID 1/0 groups can flush the cache of Exchange's random write load about 15 percent to 30 percent faster than RAID 5 groups. This equates to cache-speed performance at a higher random-write load. Additionally, rebuild times and rebuild impact are reduced with RAID 1/0 in the event of disk failures.

Table 5. RAID types and relative performance in failure scenarios

RAID type	Rebuild IOPS loss	Rebuild time	Impact of second failure during rebuild
RAID 5	50%	15% to 50% slower than RAID 1/0	Loss of data
RAID 1/0	20% - 25%	15% to 50% faster than RAID 5	Loss of data 14% of time in an eight-disk group (1/[n-1])

The downside may be cost. RAID 1/0 requires more drives for a given capacity, but that extra capacity may not be valuable when the spindle count is determined by I/O throughput. Consider the following calculations to determine the number of users that one physical drive can handle. The first calculation determines the user-level IOPS expected from a drive after taking the RAID adjustment into consideration.

$$\text{User IOPS per drive} = \text{back-end IOPS per drive} \times \text{percentage remaining after RAID adjustment}$$

The percentage remaining after RAID adjustment depends on the read/write ratio and the RAID type. The easiest way to get this figure is to add the left and right sides of the read/write ratio and divide by the (read ratio + write ratio x write penalty). The following are two examples for different RAID types.

- RAID 1/0, User IOPS per drive with a 1:1 read/write ratio on a 10k drive:

$$140 \times (1 + 1) / (1 + 1 \times 2) = 140 \times 2 / 3 = 93 \text{ IOPS}$$

- RAID 5, User IOPS per drive with a 1:1 read/write ratio on a 10k drive:

$$140 \times (1 + 1) / (1 + 1 \times 4) = 140 \times 2 / 5 = 56 \text{ IOPS}$$

The second calculation determines the number of users that the I/O capacity of the drive can handle, followed by an extension of the two previous examples.

$$\text{Users per drive} = \text{User IOPS per drive} / \text{IOPS per user}$$

- RAID 1/0, .5 IOPS per user:
93 / .5 = 186 Users per drive
- RAID 5, .5 IOPS per user:
56 / .5 = 112 Users per drive

Table 6 uses these calculations for six drives, configured as RAID 1/0 and RAID 5, with a few different mailbox sizes, allowing 50 percent additional space on the LUN for deleted item retention and white space within the database, plus indexing and growth.

Table 6. Capacity comparison for six 300 GB drives by RAID type and mailbox size

	Usable capacity	Users @ .5 IOPS	Required capacity		
			200 MB mailboxes	500 MB mailboxes	1 GB mailboxes
RAID 1/0 (3+3) 10k	804 GB (268x3)	1116 (186x6)	335 GB	837 GB	1674 GB
RAID 5 (5+1) 10k	1340 GB (268x5)	672 (112x6)	202 GB	504 GB	1008 GB

Note that once the mailbox size approaches 500 MB, there is not enough space on the RAID 1/0 group to accommodate the number of users whose I/O it can handle. This does not take into consideration the fact that the increased mailbox size causes the IOPS per user to increase.

RAID 5 becomes more appropriate when the capacity requirements are high, relative to the I/O demand, such as when the I/O demand is low (less than .4 IOPS) and the mailbox limit is high (greater than 500 MB).

Comparing 10k rpm to 15k rpm drives

The 15k rpm drive offers up to 30 percent better performance over 10k rpm drives in the kind of random-access case that Exchange presents. The increased speed ensures that the write cache avoids saturation and keeps writes going at cache speeds.

Comparing 73 GB, 146 GB, and 300 GB drives

Smaller drives offer more performance per gigabyte. Larger drive sizes become more appropriate as the ratio of mailbox size versus I/O requirements increases.

Table 7 compares the usable capacity of 10 drives configured as RAID 1/0 and as RAID 5.

Table 7. Usable capacity of 10 drives by disk size and RAID type

Raw capacity	Usable capacity per drive	10 spindles usable capacity — 5+5 R1/0	10 spindles usable capacity — two 4+1 R5
73 GB	66	330 (66x5)	528 (66x8)
146 GB	134	670	1072
300 GB	268	1340	2144

Referring to Table 7, if the required storage for the ESG went beyond 670 GB, the RAID 1/0 spindle count of 10 for 73 GB and 146 GB drives would not be adequate. Options would be to increase the drive count, move up to 300 GB drives, or switch to RAID 5.

Charting for the optimal disk and RAID type

Using calculations like this, and factoring in increased I/O for larger mailboxes, you can estimate the number of spindles required to provide sufficient storage and I/O capacity for various disk configurations. The graph in Figure 1 shows a line estimating the number of spindles needed to meet the IOPS requirement for 4,000 heavy users as the mailbox size grows. A second line shows the number of spindles needed for storage capacity. The higher of the two lines on the graph represents the number of spindles needed to meet the I/O and capacity requirements at a certain mailbox size, and the circle indicates the crossover point where storage capacity starts determining the drive count rather than the I/O requirement. It's at this point where you can take best advantage of both the I/O and storage capabilities of the disk drives. Usually the further away that your design maps from this crossover point, the more you should consider alternative disk configurations.

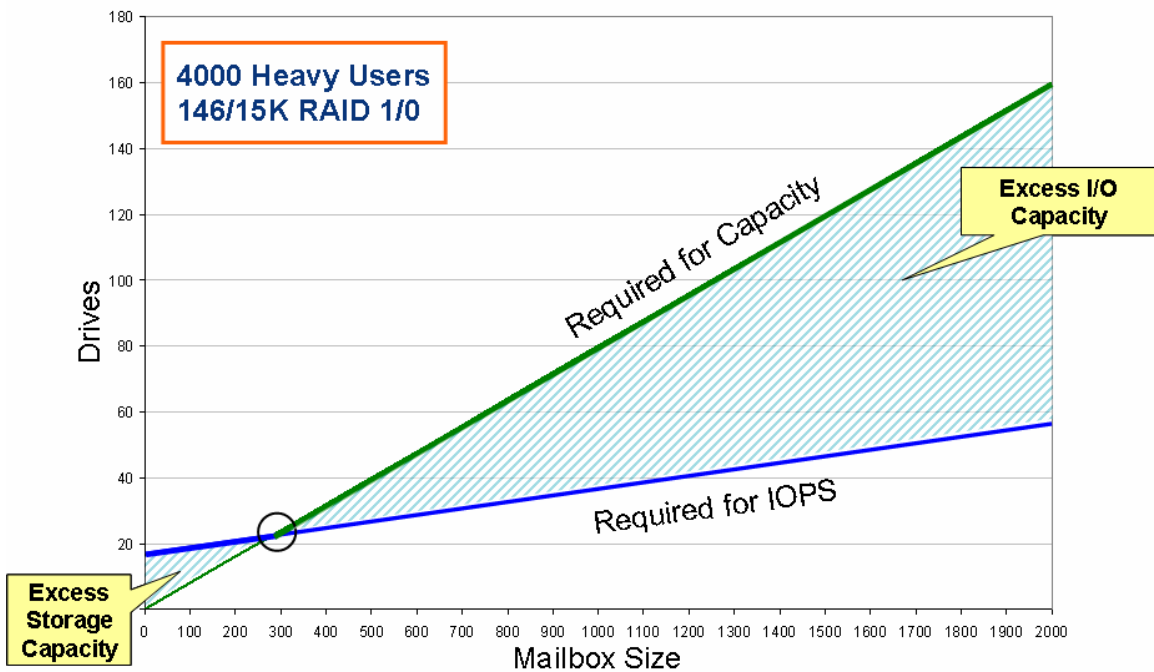


Figure 1. Determining an optimal disk and RAID type

Figure 2 adds the requirement of a 1.2 GB mailbox for the 4,000 users in the example above. The combined drive requirement for both I/O and storage capacity is charted for two additional drive configurations. The circles once again represent the crossover point where the spindle count begins to be determined by the storage capacity need.

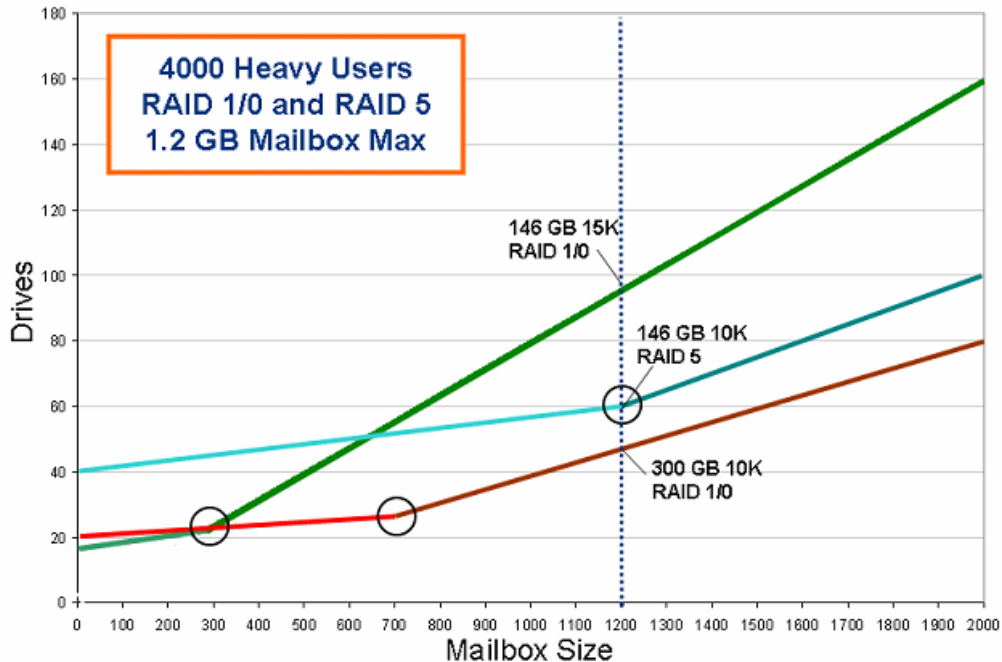


Figure 2. Determining an optimal disk and RAID type

Note that in this example the RAID 5 configuration of 146 GB 10k rpm drives takes best advantage of the storage for this user profile. However, there are other important considerations to factor in before making a final decision:

- Remember that RAID 5 groups take longer to rebuild than RAID 1/0 groups. To some degree, this can be factored into a decision by adding extra IOPS to the user requirements.
- Reasons for choosing a configuration with fewer (potentially more expensive) drives include overall savings on power and floor space, and possibly staying within the drive limit of the array.
- A configuration with faster drives (15k rpm) can get better performance with fewer spindles, which can shorten backup windows. Activities such as clone synchronizations, Exchange online maintenance, and backup to disk/tape will complete more quickly.
- In some cases, by choosing four Gigabit drives your array-to-array replication will complete in less time.

Summary

Although I/O demand has decreased with Exchange 2007 while mailbox sizes have been trending higher, IOPS capacity remains an important factor when determining the required number of drives. EMC recommends RAID 1/0 as the default choice for Exchange data volumes because:

- Under certain I/O loads, a RAID 1/0 group can perform better than a RAID 5 group with the same number of spindles.
- Rebuilds of a disk in a RAID 1/0 group will affect fewer spindles and complete more quickly.
- The smaller usable capacity of a RAID 1/0 group does not matter in the many cases where the spindle count must be determined by the I/O requirements.

RAID 5 is appropriate for some Exchange deployments depending on I/O load, mailbox size, and cost considerations.

The decision to choose between 10k or 15k drives will probably come down to cost. When IOPS are the determining factor in how many disks you need, the number of 15k drives required will usually be less than

the 10k drive requirement. Balance the additional cost per drive against savings on additional power, drives, DAEs, and possibly cabinets.

MetaLUNs

CLARiiON storage systems can combine multiple LUNs into a larger *metaLUN* that spans multiple RAID groups. MetaLUNs offer two primary advantages: I/O load balancing and expandability. Usually you will configure an ESG for its maximum anticipated size at the start, but it is possible to use metaLUN technology to handle gradual growth in capacity, performance, or both.

The main advantage is the ability to distribute I/O over many spindles without resorting to host striping. Striped volumes are particularly advantageous with workloads (such as Exchange) that are random and bursty, and metaLUNs make using striped volumes simple. Suppose that you have planned two ESGs to reside on their own 3+3 RAID 1/0 groups, for a total of 12 spindles. An alternative design would be to create a metaLUN for each of the ESGs—both spanning the two 3+3 groups. The same 12 spindles would still be handling the I/O of the two ESGs, but in this case if the I/O demand of one ESG is higher than the other, the combined load will be balanced across all of the drives. This can help to avoid an I/O bottleneck for a particular ESG. The two storage groups would also share the cost of a disk rebuild. By spanning two RAID groups, you double the risk of them being affected by a rebuild, but a rebuild will affect only half the disks of the metaLUN.

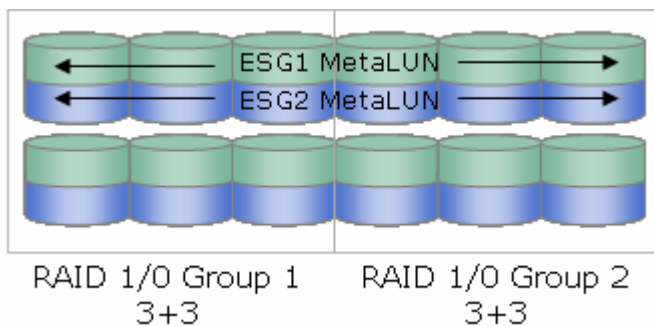


Figure 3. MetaLUNs sharing two RAID groups

Another choice is a single 6+6 RAID group. It yields the same performance but with less growth potential. To accommodate growth, additional free space on the existing RAID group set can be used by concatenating the metaLUN. RAID group expansion also provides room for metaLUN growth (by concatenating the metaLUN), along with added performance.

For optimal performance load balancing across a RAID group set, you can interleave the data of multiple storage groups on the RAID set by creating metaLUNs for each ESG in the following order (illustrated in Figure 4):

- Stripe the first component of ESG1
- Stripe the first component of ESG2
- Concatenate the next component of ESG1
- Concatenate the next component of ESG2

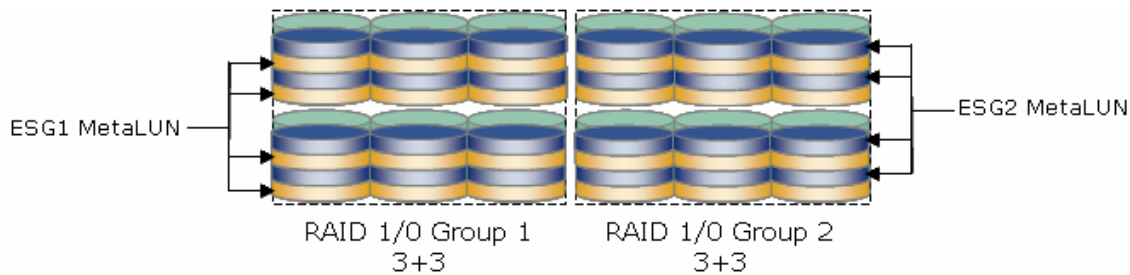


Figure 4. Interleaving metaLUNs

It is possible to overuse metaLUNs. While offering the benefits of flexibility and load balancing, metaLUNs residing on the same RAID group all share the risk of a physical disk failure in that group. You should also pay attention to I/O-intensive activities (such as Eseutil, online maintenance, and clone synchronizations) taking place on multiple metaLUNs sharing the same drives. Typically, two or three ESG metaLUNs sharing a RAID group set become a practical limit.

Building blocks

In practice, Exchange configurations perform well and are easier to manage if you choose from a limited number of proven RAID group types and sizes. These groups serve as flexible building blocks that will be easy to multiply out in the full Exchange storage design.

Following are recommended building block elements that are small enough to be flexible, allow for growth, and have been proven to perform well:

- RAID 10: 2+2, 3+3, 4+4, 5+5
- RAID 5: 4+1

EMC constructs and tests a wide range of Exchange 2007 configurations based on a set of storage building blocks. They serve as best practice examples and are validated with performance testing that is published as part of the Microsoft Exchange Solution Review Program (ESRP).

Log LUN configuration

When configuring disks for Exchange, most attention is paid to the database LUNs because they typically represent the highest risk of performance bottleneck. But the database performance depends on the log response time. Database transactions are gated by the completion of the associated log writes.

When choosing a RAID type for log file LUNs, I/O performance and data protection (rather than capacity) are the overriding factors; therefore RAID 1/0 is the best RAID type to use for log LUNs. It provides better response time than RAID 5 in degraded situations. In the case of a disk failure, RAID 1/0 rebuilds complete more quickly than RAID 5 rebuilds. The longer the rebuild period is, the more vulnerability there is to data loss. Data loss always occurs if a second drive is lost during rebuild of a RAID 5 group (see Table 6).

Although writes to the log LUN are sequential, performance tests have shown that you can take best advantage of a set of drives by sharing a set of log LUNs on them, because the storage array write cache preserves the individual sequential log streams to the back end. A rough rule of thumb to use for log drives is to allocate one-fourth the number of spindles (of matching I/O capacity) you have allocated for the databases, rounding up for RAID 1/0. For example, if you calculated that you need 32 drives to handle the databases of eight Exchange storage groups on a server, allocate eight drives in a RAID 1/0 4+4 group to handle the eight transaction log LUNs for that server.

To avoid added recovery complications in the unlikely case of the physical loss of a RAID group, store the log files for servers on separate drives.

Here are other important considerations in the design of the disk layout for the Exchange transaction logs:

- There is one set of transaction logs for each storage group.
- For manageability and flexibility, each set of log files should reside on their own LUN.
- An ESG transaction log file should always be stored on a different LUN from its associated database. Even on small systems, the LUN for the ESG log file and the LUN for the associated database should never use the same spindles. This is for protection rather than performance. If something should happen to the database, the log files are essential to recover transactions since the last backup. If those log files reside on the same physical disk and that disk is damaged, this option is lost.
- With the exception of recovery operations and CCR, log I/Os are 100 percent writes. They are typically from 512 bytes to 12 KB, but can be significantly larger.
- In Exchange 2007, the number of host writes to the log LUN is approximately half the number of host writes to the database LUN. In Exchange 2003, the number of writes to the log LUN is typically less than one-third the writes to the database LUN.
- The new size of each log file in Exchange 2007 is 1 MB.
- Most online Exchange backup processes delete log files whose transactions have been committed to the database. It is important to confirm that committed log files are being deleted to avoid running out of space on the log LUN.
- Circular logging is an Exchange feature that causes log files to be deleted after their transactions have been committed to the database. Only a handful of the logs are maintained at any time to save space. However, this sacrifices the capability to recover a database up to the minute. This feature is **OFF** by default and should not be turned **ON**.

Calculating log LUN storage capacity requirements

Any change to an Exchange database is first recorded in a transaction log file. This includes user activity and administrative activity such as online maintenance, mailbox moves, and email shortcutting into an archive. You can estimate storage capacity requirements for a log LUN by considering:

- The number of log files generated per user per day — At EMC there are roughly 10 MB of logs generated per user each day, including online maintenance activity.
- The number of days' worth of logs to maintain — Typical full, online Exchange backups, which are run nightly, prune log files. Differential backups do *not* prune the log files. It is important to have the capacity to store several days' worth of log files, in case there is a problem with the backup process, and to provide protection for some disaster recovery solutions.
- Amount of storage needed for mailbox moves — When a mailbox is moved from one ESG to another ESG, transaction logs are generated in the target log LUN that are roughly the same size as the mailbox being moved.
- Storage for email shortcutting — Email archiving products can remove email messages from the Exchange database, leaving just a small stub that points to the actual message in the archived container storage. As with all other transactions, all shortcutting is logged and therefore must be taken into consideration – especially during its initial implementation.

For example, an Exchange server contains storage groups with the following characteristics and requirements:

- 400 users with mailboxes of 500 MB
- 10 MB of logs per user per day (including online maintenance)
- Need to maintain 7 days of log files
- Need to be able to move up to 20 mailboxes at a time

In this case the log LUN requirements are:

```
400 users x 10 MB x 7 days
plus 20 mailboxes x 500 MB
→ 28 GB + 10 GB = 38 GB
```

Additional storage considerations for the Exchange production data

The previous configuration guidelines have covered design recommendations for database and log LUNs for Exchange implementations. This section describes some additional considerations.

Other Exchange server roles

Exchange 2007 divides its functionality into *server roles*. These roles provide flexibility when configuring server(s), and allow you to place a role, or a combination of roles, on one or more physical servers, depending on the size and activity of your environment. This white paper discusses configuration guidelines for the *mailbox* server role. There are two other required server roles in Exchange 2007:

- Hub Transport – All internal mail, even within a server, passes through the hub transport server. It is similar to the Exchange 2003 bridgehead server.
- Client Access – Mail from all clients (Outlook, Outlook Web Access, mobile devices and so forth), passes through the client access server before it is sent to the appropriate mailbox server. It is similar to the Exchange 2003 front-end server. If users access their inbox by using any client other than Microsoft Outlook, you must then install the CAS role in your Exchange environment.

There are also two optional server roles:

- Edge Transport – This is installed on a separate standalone server at the edge of an organization's network. It serves as a protective filter for the internal Exchange environment, with functions including antispam and antivirus.
- Unified Messaging – This server has the capability to merge an organization's voice-over-IP infrastructure with the Exchange environment, allowing the integration of voice and fax into the messaging system.

A more detailed description of these roles and their storage requirements is included in the Microsoft document on Planning Disk Storage for Exchange 2007.

Public folders

It is difficult to provide general guidelines for configuring public folder storage because their usage varies so widely. Some organizations do not use Exchange's public folders, while other organizations use them extensively for shared document repositories, discussion groups, shared calendars, and several other purposes. By default, the public storage is contained in a dedicated database in the first storage group. However, when it is actively utilized it is often configured on its own Exchange server with at least one replicated copy to another Exchange server.

The best starting point for planning public folder storage for a newly migrated Exchange 2007 environment is to examine the current I/O and storage usage and growth rate for public folders in the current

environment. If necessary, adjust these measurements based on any planned changes to the public folder use policy (such as adding a new integrated application that makes significant use of public folders, or switching shared documents to file shares). Then, deploy (using the principles described in this paper) as with any other Exchange storage group.

Note that public folder replication is not allowed to or from Exchange 2007 mailbox servers configured with cluster continuous replication (CCR). There can either be a single non-replicated public folder database on a CCR clustered mailbox server, or there can be a replicated public folder database residing on servers that are outside of the CCR environment.

Mixing database and log LUNs on the same drives

For proper Exchange data protection, transaction log files and the database files for an Exchange storage group should never be stored on the same physical drives. In general, it is a best practice to maintain log LUNs on different drives than database LUNs. However, in some cases it is appropriate to configure a RAID group with the database LUN of one storage group combined with the log LUN of a different storage group. Considerations include:

- Usually the driving factor for combining database and log LUNs on one RAID group is a shortage of drives and the need to provide as much I/O capacity as possible to the database LUNs to satisfy performance requirements. This most often occurs with small configurations.
- If there are only two RAID groups available for Exchange data, the users can be split across two storage groups, and the logs placed with the databases for the alternate ESG. However, you must then factor in the log I/O requirement when calculating the spindle count.
- If you combine database and log LUNs onto one RAID group, you must remember to factor in the log I/O requirement when calculating the spindle count. Note also that log I/O in Exchange 2007 represents a higher percentage of total Exchange I/O than in Exchange 2003.

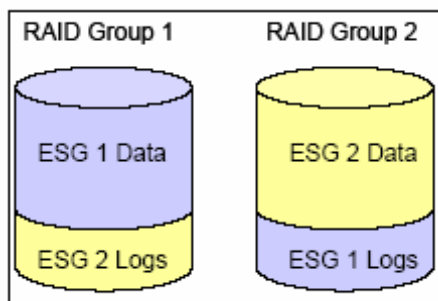


Figure 5. Sharing a database LUN with the logs from an alternate storage group

Planning storage for local recovery

This section provides guidelines for configuring storage to handle local replicas and disk-based backups of Exchange production data. CLARiiON storage systems also support a full range of solutions for remote replication and disaster recovery, details of which are outside the scope of this paper.

SnapView for disk-based replication

SnapView™ is an optional software package for the EMC CLARiiON storage system. Using SnapView, you can create a point-in-time view—or multiple views—of a LUN, which you can make accessible to another server, or simply hold as a point-in-time copy for possible restoration. For instance, a system administrator can make the SnapView replica accessible to a backup server so that the production server can continue processing without the downtime traditionally associated with backups. In the event of a data corruption on the source LUN, SnapView replicas can be used to restore the contents of a corrupted LUN to the point-in-time creation of the replica. SnapView can create replicas using either clones or snapshots.

There are currently two products sold by EMC that facilitate the management and integration with Exchange 2007 and the Windows Volume Shadow Copy Service (VSS) to create verified SnapView replicas of an ESG, ready to be restored immediately:

- Replication Manager
- EMC NetWorker®

Clone-based replication

Clone-based replication provides very rapid recovery of an Exchange database. SnapView clones provide users the ability to create fully populated copies of LUNs within a single array. Once synchronized, clones can be fractured from their source, and then presented to a secondary server for read and write access. Following the initial synchronization, clones can be *incrementally resynchronized*, where only the data that has changed on the source since the clone was fractured is copied to the clone. In the event of a data corruption on the source LUN, clones can be used to restore the source LUN via a *reverse synchronization* operation. This returns the source LUN to the point-in-time view of the source when the clone was fractured. In the unlikely event of a hardware error on the LUN (for instance, a multiple-drive failure), the clone can be repurposed as the production LUN. Thus, clones provide protection against both software errors and hardware errors.

EMC Replication Manager allows you to configure up to eight clones for each of the Exchange production LUNs.

Consider the following when choosing clone-based replication as the backup method:

- Know the characteristics of the clone backups you'll be using.
 - Number of clones for each production LUN — Each extra clone maintains a validated backup copy for a different point in time.
 - Timing and frequency of the backup operation — Spread out backups of the storage groups, and time them for low usage periods to take best advantage of array resources.
- Fibre drives are recommended for clones. Their performance is well suited to clone resynchronizations. ATA drives are not recommended for clone LUNs in an active Exchange environment because the clone resynchronization operation to an ATA drive is considerably (up to several times) slower. The likelihood of affecting the production environment during this time increases. Additionally, with an IOPS rate of less than half that of Fibre drives, the backup window will also be extended on ATA drives by an Eseutil check that takes longer to complete.
- Clone LUNs can be bound on drives that differ in size, speed, RAID geometry, or even drive type. For example, you can use RAID 1/0 for production LUNs and RAID 5 for clones. EMC recommends RAID 5 for clones because clones do not have the same IOPS requirements, and provide greater capacity. Additionally, you can put production data on 15k rpm drives, and use 10k rpm drives for their clones.
- Eseutil is usually the most I/O intensive activity on database clones. Consider Eseutil throughput requirements when determining the spindle count here.
- When using RAID 5 with clones, take modified RAID 3 (MR3) support into consideration. RAID 5 4+1 sets offer a good balance of rebuild time with MR3 support².
- 300 GB drives may be more appropriate with clones. The combination of RAID 5 and 300 GB drives allow you to configure sufficient extra capacity to store multiple clones on the same RAID group. Since clone synch operations are scheduled, the I/O capacity for a set of drives can be shared to handle multiple backups occurring at different times.
- Do not place clone LUNs in the same RAID group that contains their source LUN.

² For information on MR3 writes, refer to the “CLARiiON RAID 5 optimizations” section of the *EMC CLARiiON Fibre Channel Storage Fundamentals* white paper.

-
- Plan clone layouts to avoid backups occurring simultaneously on different LUNs configured on the same RAID group. LUN resynchronizations and Eseutil integrity checks both involve a heavy amount of I/O. Pairing two or more of these activities at the same time slows them down and may affect user response times.
 - Consider using metaLUNs for clones to provide more spindles to improve Eseutil performance, but balance that with the advisability of separating concurrent high-I/O on separate drive sets.
 - Keep in mind the discussion in the “Building blocks” section. For simplicity, try to select your RAID group from the recommended choices, and expand using this as a base.
 - To add some extra protection, if you are configuring multiple clones for an Exchange LUN, it’s best to alternate the clones on separate spindle sets.

Snapshot-based replication

Replication Manager can create a VSS shadow copy via a SnapView snapshot. As with clones, snapshots provide a readable and writable LUN replica. However, snapshots are not fully populated copies. They use a pointer-and-copy-based design, where pointers map to data regions on the source LUN until they are changed, at which point the original data is copied to a reserved area and the pointers are redirected accordingly. In this way, you only have to allocate sufficient disk space to accommodate the changes to the source LUN.

The process of allocating the pointers according to a particular point in time is referred to as *starting a SnapView session*. To see the contents of a particular session, you can *activate a snapshot* to the session. The *reserved LUN* is the private LUN that contains the original source data, and the process of copying that data is referred to as *copy on first write* (since it must only occur on the initial change to the source LUN).

As with clones, SnapView session data can be used to restore a corrupted source LUN. Much like the *reverse synchronization* that clones offers, SnapView sessions can be *rolled back* to the point-in-time view of when the session was started.

RecoverPoint CDP

EMC RecoverPoint provides network-based continuous data protection with on-demand local recovery to any point in time. It maintains a replica of the Exchange databases and also a set of journal volumes that hold tracked changes to the Exchange data. Consider the following when designing storage for local Exchange data protection with RecoverPoint:

- Because the Exchange replicas are used in production if there is failure, configure these with storage and I/O capacity similar to the production database and log volumes.
- The formula to calculate the amount of journal storage needed is:

$$\text{journal size} = \frac{\text{data size} \times \text{daily change rate} \times \text{recovery window}}{.80}$$

where:

data size represents the size of the Exchange replica

daily change rate represents the average daily percentage of change to the Exchange data and **recovery window** represents the number of days’ worth of changes you need to maintain.

For more detailed information on sizing and configuring storage, see the RecoverPoint documentation.

Exchange continuous replication

Local continuous replication (LCR) and cluster continuous replication (CCR) are features of Exchange 2007. LCR enables log replication on a single standalone (non-clustered) server. If the active copy of the database or logs fails, the administrator can quickly (manually) activate the passive copy. LCR offers

protection from a single server from a storage volume or system failure. There is no protection in the case of a server failure.

CCR functions in a similar way but provides added protection against a server failure. It is based on the use of a majority node set (MNS) cluster with one active node, one passive node, and a third node (which can be a file share) to serve as the witness to help determine the need for a failover. CCR is included in this section because with its current requirement of all nodes being in the same subnet, it is seen as most appropriate for same-site or campus environments.

Storage configuration considerations for continuous replication include the following:

- Passive node storage for LCR is typically allocated on the same storage system as the active node.
- Passive node storage for CCR is typically allocated on a separate storage system from the active node
- The amount of disk I/O to the passive LUNs in a continuous replication environment can be double to triple the I/O to the equivalent active LUNs. (This overhead is expected to be less in Exchange 2007 SP1.) To function equivalently in production on either the active or the passive storage, with bidirectional failover capability, you should configure both sides to meet the I/O requirements of the passive side.
- Particularly for LCR, where both active and passive LUNs are associated with the same server on the same storage system, pay attention to the potentially high number of LUNs in relation to the storage system limits. With 50 storage groups, each with its own database/log LUN pair, there would be 200 LUNs associated with just one Exchange server.

Service Pack 1 for Exchange 2007 adds Standby Continuous Replication (SCR). SCR supports log shipping and replay at the storage group level to a standby server or cluster (which may be on a different subnet). This provides another recoverable copy of the Exchange data at a separate location. Because it is intended to be able to take over a production role, the SCR storage should be configured to provide I/O capacity and storage capacity equivalent to its production Exchange volumes.

Online backup to disk

Disk-based Exchange online backup has become a competitive alternative to tape backup, offering faster recovery and higher reliability. Consider the following when online backup to disk is chosen as the primary backup method:

- Capacity requirements are determined by the size and frequency of the backups and the number of copies maintained before archiving to tape.
- Online backups perform sequential writes. A configuration that optimizes this type of I/O performs best.
- CLARiiON ATA disk drives are effective for backup to disk. They perform well with sequential I/O operations. They are also most competitive with the cost of tape backup.
- If keeping more than one backup of an ESG on disk, you can alternate copies across two different RAID groups for added protection.
- Most testing has been done on RAID 5 4+1. It has been determined to be the sweet spot for performance with backup to disk.
- Multiple streams within the same LUN often result in better overall throughput, although the per-stream performance will understandably decline. If the streams are each sent to a separate LUN on the RAID group, the data in a particular stream will be more contiguous on the disk and thus restore somewhat faster (by 5 MB/s to 10 MB/s).
- For the most rapid recovery and the least user impact, individual Exchange databases should be as small as possible with the smallest possible number of users. If the backup method will be online backup to disk, you have added incentive to use the maximum number of storage groups and databases within those ESGs.

-
- If online backup is performed during a period of active production, be sure to take this additional activity into account when calculating I/O requirements, also paying attention to overall array limitations.
 - EMC Disk Library is integrated with a number of software applications that perform Exchange online backups. It offers additional features such as compression and remote replication.

Planning storage for local message archiving

Message archiving on CLARiiON storage is a growing component of new messaging system designs. The archiving implementation generally involves adding one or more servers to the environment. These servers manage the transfer of the content of older messages out of the standard Exchange database structure. The archiving servers also manage the near-line retrieval of these messages when a user calls for one.

EMC EmailXtender software provides this archiving capability. The EmailXtender manual provides a formula to estimate the amount of data to be archived:

- Container File Storage (disk space for archived messages) is the product of the following factors:
 - Number of users
 - Messages per user per day
 - Days per work week
 - Number of weeks mail is retained
 - Average message size (KB)
 - .000001 (to convert the result to GB)

For example:

$$3200 \times 20 \times 5 \times 250 \times 50 \times .000001 = 4000 \text{ GB}$$

EmailXtender also estimates an approximately 20 percent overhead for the installation and Message Center drive, plus .5 GB for queuing. The space required on the installation LUN in this case would be:

$$4000 \times .2 + .5 = 800.5 \text{ GB}$$

Because a very large quantity of data is usually archived, and the access to this data is less time critical, this is an appropriate application for CLARiiON ATA drives, and for EMC Centera[®] content addressable storage.

Additional infrastructure considerations

Storage system considerations

After you determine the I/O requirements of the new messaging system and settle upon an appropriate number and type of disk drives, the next step is to consider the throughput and features of the storage system.

Storage system designs that balance array resources, by considering both the layout of the data and the timing of schedulable activities, provide the best throughput and data protection. Consider the following when planning your disk layout:

- Avoid configuring Exchange database LUNs on the CLARiiON persistent storage manager (PSM) drives (drives 0-2). However, these drives may be suitable for the lower I/O requirements of a properly configured set of log LUNs.
- It is not required to have the log LUN and database LUN for an ESG managed by opposite CLARiiON storage processors (SPs). It is more important to ensure that the overall I/O demand is balanced across the two SPs.

- You gain a small performance advantage (3 to 4 percent) by binding the RAID 1/0 primaries and secondaries on different back-end buses. The main advantage to this approach is that the administrator does not have to know which LUN is on which back-end bus. The back end is balanced by virtue of the RAID group layout.
- Don't forget to include hot spares in the configuration. Typically, you should add one hot spare for every 30 drives.
- Clone LUNs must be assigned to the same SP as their source. (You can configure clones on the opposite SP to its source, but the clone would be trespassed for synchronizations in this case.) Ensure that both the current SP owner and preferred owner are the same for both the source and target LUNs.
- Some activities within the Exchange environment place heavy use of the CLARiiON SP. These include performing an Eseutil check against a database, and running CLARiiON layered applications – particularly when acting upon several LUNs at once. When configuring an array for Exchange and scheduling Exchange administrative operations, it is important to consider the limits of the SP CPU resources.

Table 8 lists useful specifications for current CLARiiON CX models (based on FLARE[®] 26).

Table 8. CLARiiON CX series storage systems feature summary

Feature	CX3-80	CX3-40	CX3-40c	CX3-20	CX3-20c	Cx3-10c
Maximum disks	480	240	240	120	120	60
Storage processors (SP)	2	2	2	2	2	2
CPUs/SP	2@3.6GHz	2@2.8GHz	2@2.8GHz	1@2.8GHz	1@2.8GHz	1@1.8GHz
Front-end FC ports/SP – Fibre Channel	4@4 Gb	4@4 Gb**	2@4 Gb	2@4 Gb**	2@4 Gb	2@4 Gb
Front-end FC ports/SP – iSCSI	-	-	4@1 Gb	-	4@1 Gb	2@1 Gb
Back-end FC ports/SP	4@4 Gb	2@4 Gb**	2@4 Gb	1@4 Gb	1@4 Gb	1@4 Gb
Array cache (total and write in MB)	6728 / 3072	3016 / 2500	3016 / 2500	1053 / 1053	1053 / 1053	310 / 310
Highly available hosts (2 HBAs each)	256	128	128	128	128	64
Maximum LUNs	2048	2048	2048	1024	1024	512
RAID groups	240	120	120	60	60	30
MirrorView images (Total primary + secondary)	200	200	200	100	100	100
Snapshot reserved LUNs	512	256	256	128	128	64
Clone sources per array	512	256	256	128	128	64
Clone images per array	1024	512	512	256	256	128
SAN Copy concurrent sessions	16	8	8	8	8	4
Exchange users (Fibre Channel) **	24,000	18,000	18,000	12,000	12,000	6,000

* The CX3-20f has six front-end Fibre Channel ports per SP and the CX3-40f has four front-end Fibre Channel ports and 4 back-end ports per SP.

** As described in this paper, the number of Exchange users an array can handle varies greatly. The number in this table is provided primarily for comparison between the models. It refers to a storage system dedicated to Exchange users (4 IOPs/user) with mailboxes of 200 MB and local replication to clones.

Remember these specifications when determining the appropriate storage system(s) for the new Exchange storage design. For example, you can create up to 128 clone sources on a CLARiiON CX3-20, and each of these clone sources can have two clones. If the local replication plan calls for more than two clones per clone source, the maximum total number of clones (targets) on the CX3-20 must still be no more than 256. Note that if you are implementing clone-based replication with each Exchange storage group on its own database/log LUN pair, a single Exchange server with 50 ESGs will have 100 attached production LUNs (clone sources) and 100 clones.

Once the CLARiiON storage system is in place, ready to be configured, follow the tuning guidelines in this section.

Storage system tuning

The CLARiiON storage system is well behaved and high performing when you use the default parameters. Some settings must be set at installation. For most systems, follow these guidelines:

- Read and write cache: **ON** for all LUNs
- LUN stripe element size: 128 blocks (default)
- Pre-fetch settings: Default settings
- Cache page size: 8 KB (default)
- Cache settings:

Table 9. Recommended CLARiiON cache settings

	CX3-80	CX3-40	CX3-20	CX3-10
Array Cache (Total and Write in MB) (not including SP memory)	6728 / 3072	3016 / 2500	1053 / 1053	310 / 310
Recommended Write Cache	3072 (max)	2500 (max)	975	270
Recommended Read Cache (remainder per SP after write allocation)	3656	516	78	40

Exchange servers

The following server tuning and environment recommendations are included here because the problems that they prevent are often mistaken for disk issues:

- The location and speed of domain controllers, DNS servers, and the global catalog are vital to Exchange performance. The supporting servers should be local enough to the Exchange servers to provide a fast connection. There should also be enough of them to support the Exchange environment. Microsoft recommends that you use at least one CPU processor for a global catalog server for each of the four processors dedicated to Exchange. For example, there should be one dual-CPU global catalog server to handle two four-processor Exchange servers.
- Do not use file-level scanning antivirus software on any Exchange files. This can cause corruption to the files, including database, log, and checkpoint files. For details, refer to Microsoft Knowledge Base article 328841, "Exchange and antivirus software."

<http://support.microsoft.com/?id=328841>

Windows file-system alignment

Windows has an internal structure called the master boot record (MBR) that inhabits the beginning of a physical device. The MBR uses hidden sectors on the drive. This value is defaulted to 63. The result is that on a CLARiiON LUN, Windows always creates the first partition on that disk starting at the 64th sector, thus misaligning it with the underlying RAID stripe. This causes disk crossings for a percentage of small I/O (typical of Exchange), resulting in slightly lower performance.

The solution is to modify the master boot record's hidden sectors from 63 to the value matching the stripe element size. We recommend a stripe element size of 128, which is the default size. `Diskpart.exe` is a command-line utility provided by Microsoft with Windows Server 2003. It can explicitly set the starting offset of the master boot record by using the **align=64** parameter when creating the primary partition. (For more information about using Diskpart, see the Microsoft Windows Help.)

For Exchange, setting the offset with Diskpart is preferred to setting the offset in Navisphere during the binding of a LUN. Aligning after the LUN is bound offers these advantages:

- Any clone or SAN Copy of this disk will automatically include the alignment.

-
- If a change in the offset is required after implementation, data will still be lost but you do not have to rebind the LUN.
 - This is consistent with the recommended method for Symmetrix systems.

Windows allocation unit size

The default allocation unit size for Windows disks is 4 KB. When formatting a new Windows disk for use with Exchange data (database or logs), EMC recommends that you format the disk with an allocation unit size of 64 KB. This is not likely to affect the performance of normal Exchange activity. However, performance of Eseutil and other larger-I/O activities may improve 20 percent or more.

Network connectivity

This section discusses considerations for three types of network connections between the Exchange servers and CLARiiON storage.

Fibre Channel SAN

Storage area networks (SANs) have become an industry standard, particularly for an organization's critical applications. Until recently Fibre Channel was the only practical option for SAN deployment, and it remains the choice for highest performing networks. Having storage centralized on a SAN offers many advantages, which are outlined in the following sections.

With SAN centralized management:

- You can manage all storage from a central point using a single console.
- It provides centralized control of an email life cycle for multiple servers:
 - Growth.
 - Local and remote protection.
 - Archiving and compliance.
- You can remotely assign new disks to a server rather than physically installing them when needed.
- It allows for better integration of a messaging system with other business applications.
- It enables simpler reimaging, patching, and disk replacement. In addition, you can use CLARiiON NQM for throttling some activities and giving priority to others.

SAN provides more flexibility, including:

- Storage on demand and the ability to easily assign storage for added server requirements.
- New storage groups.
- Unified messaging and the ability to easily assign/remove storage for temporary or periodic functions.
- Offline defragmentation.
- Mailbox moves.
- Online expandability including:
 - The capacity of on-the-fly for more/larger mailboxes or added I/O demand such as mobile users.
 - Increasing the size of log LUNs.
- Online LUN migration for changing I/O requirements including:
 - Transparently moving data to a new disk set (RAID type/disk type)
 - Useful for updated tiering of archived email.
- More options for backup/recovery with zero/low impact on the Exchange server.
- Facilitating server virtualization. A server role can be transported to a new physical server while remaining attached to the same central storage.

-
- The ability of servers to boot from the SAN to further consolidate storage and make a server more easily replaceable
 - Database portability, a well-liked new feature in Exchange 2007 that allows an IT administrator to move an Exchange database from one server to another. In a SAN environment this process is completed by simply reassigning that capacity to a new server.

SAN optimizes resource utilization because:

- A high volume of traffic is shifted to the SAN, allowing better LAN performance.
- More disk space is “stranded” when associated with a single server.
- It allows for common space available across multiple servers for offline defrag or recovery storage groups.
- Facilities can share tape units.
- There is the potential for decreased servers and other equipment, with better storage utilization, resulting in consolidation, and lower licensing costs.

SAN offers improved data protection with:

- A highly reliable storage platform (approaching five 9s).
- Time-tested VSS-compliant rapid recovery solutions, including:
 - EMC protected restore
 - Efficient array-level replication
- Improved disaster recovery and continuity options.
- Automatic hot replacement of disks.
- RAID protection of all disks (can include boot disk).
- The ability to quickly reassign disks to another server in case of server failure.

The total cost of ownership may be lower. Although the initial deployment cost of a SAN is higher, there are cost benefits in the long run due to:

- Labor savings from centralized management.
- A payback in TCO of centralized storage over time (even though upfront costs of server-specific storage may be less).
- Lower costs associated with more efficient disk utilization including floor space and power.
- Less costly downtime.
- Decreased costs because you are maintaining fewer servers.
- Shared storage utilization.
- Access to storage over distance.
- Management efficiencies compared to traditional DAS.
- Consolidation that brings a more efficient approach to storage deployment and management through rationalizing storage resources, increasing capacity utilization and centralizing storage management.
- Simpler storage management
- Disk-to-disk backup

iSCSI SAN

Deploying storage on an iSCSI SAN yields most of the same benefits as a Fibre Channel SAN, often with lower deployment costs. Storage is centralized for high flexibility, manageability, and utilization.

It is also important to note that an iSCSI SAN has more potential to limit the throughput of high I/O demand activities such as Eseutil, backup to disk/tape, or array-to-array replication.

Direct attach

With the typically lower I/O requirements of Exchange 2007 and its added replication options, there is more consideration of deploying on less expensive direct-attached storage (DAS). However, in most medium-to-large Exchange environments, the long-term benefits of SAN deployments outweigh the initial savings of DAS.

Considerations include:

- Direct-attach deployments are most feasible for smaller organizations, particularly where the storage requirements are relatively stable.
- There are fewer backup options.
- This can lead to server proliferation and “stranded” storage.
- It is more difficult to repurpose servers and storage, which leads to added equipment.
- With DAS you are physically moving or recabing a disk system to achieve database portability.

CLARiiON storage systems can be configured as direct-attached storage (see Table 8). By connecting up to 12 Exchange servers directly to a CLARiiON storage system (total of six with HA connections to each SP), you can take advantage of the CLARiiON data protection strengths and some of the centralized storage features among the attached servers.

Putting it all together

This section presents some final suggestions for designing a successful Exchange storage design.

Consider site-specific constraints

The resulting storage design must take into account any requirements or restrictions in the environment where the messaging system will be implemented. For example, an organization may have an existing CX700 available that they want to use before purchasing an additional array. They may not be able to dedicate an entire array to Exchange use, or they may have a certain type of drive already in place for the Exchange data. There are likely to be many other decisions made already that will affect the storage design, such as number and location of Exchange servers, number of ESGs per server, network capacity, and so forth.

Regardless of the constraints, the core requirement of providing enough drives to meet peak I/O demand remains, as does the strong recommendation to keep log and database LUNs for the same ESG on separate spindles.

Configure the cleanest looking layout diagram

Using a building block style, draw a clean-looking storage layout diagram. Assign names to the LUNs that are organized and meaningful. This will ease understanding of the design, help identify possible weaknesses, and aid in the storage administration of the implementation.

Validate the design

Before committing to a particular design it is a good idea to conduct a peer review, and if possible compare it with known good configurations, such as the EMC building block designs reported in the Microsoft ESRP program (see the “References” section for additional information).

Where possible, the configuration should be built and tested with performance tools such as Microsoft JetStress, LoadGen, and Performance Monitor, and EMC Navisphere Analyzer.

Anticipate unforeseen issues to arise early in a rollout and be prepared to address them.

Conclusion

Built for organizations that consider their messaging systems highly important, CLARiiON storage systems offer a dependable five 9s platform for Microsoft Exchange deployment. They offer a great deal of flexibility, which suits the wide range and changing nature of Exchange environments. As new versions of Exchange are released, and as CLARiiON array technology advances, best practice information for Exchange storage design is constantly evolving. This paper presents a current snapshot of these latest guidelines.

References

EMC

- *EMC CLARiiON Fibre Channel Storage Fundamentals* white paper
- *EMC CLARiiON Best Practices for Fibre Channel Storage* white paper
- [Microsoft Exchange Solution Reviewed Program \(ESRP\)](#) page on EMC.com

Microsoft

- [Microsoft Exchange Solution Reviewed Program \(ESRP\) for Exchange 2007](#) on Microsoft TechNet
- [Exchange 2007 – Planning Storage Configurations](#) on Microsoft TechNet