Microsoft Exchange 2007 Continuous Replication Best Practices Guide

Robert Quimbey, Network Appliance Inc.

June 2007 | TR-3600

TABLE OF CONTENTS

Introduction

Microsoft Exchange 2007 includes new features that increase availability; it runs only on x64; and the disk I/O workload has changed significantly from previous versions. Messaging is a mission-critical service that is consuming an ever-increasing percentage of the information worker's day. Messaging availability can be reduced by logical corruption and by server, network, storage, and site failures. This document explores the Continuous Replication disk I/O workload and provides guidance in the storage design of a Continuous Replication solution on NetApp storage. This guidance applies to both Local Continuous Replication (LCR), and Clustered Continuous Replication (CCR).

See the Microsoft Exchange Server 2007 Best Practices Guide for general Exchange Server 2007 best practices and see the Microsoft document Technical Architecture of Exchange Server 2007 for in-depth Exchange 2007 information.

Intended Audience

This technical report is intended for messaging and storage professionals who design, test, deploy, and manage their corporate messaging infrastructure. For methods and procedures mentioned in this technical report, it is assumed that the reader has working knowledge of the following:

- Microsoft Exchange 2007 architecture
- Exchange storage architecture and administration
- NetApp Data ONTAP®

Continuous Replication Overview

Continuous Replication, or log shipping, is the new High Availability feature in Exchange 2007. It provides storage resiliency and CCR server resiliency. With LCR, the active and target LUNs are hosted on a single server, and the failover is manual, which means that the availability is directly tied to how quickly someone can execute commands on the server. CCR utilizes Microsoft Clustering using nonshared storage in a two-node active-passive configuration, adding automated failover and server resiliency.

During initial setup, the database is copied, or seeded, to the target location. When a log file is filled, it is renamed from the active `Exx.log` to the next log file in the sequence, such as `Exx00000002.log`. The target pulls (or copies) the closed log file, inspects it, and then replays the log file into the target database, keeping it up to date. In the event of a catastrophic failure of the active storage, the target storage generally lacks only the active (`Exx.log`) log file, which in Exchange 2007 has been reduced to 1MB in size (from 5MB) to lower the amount of vulnerable data . In the event of a catastrophic failure of the active storage, the target storage becomes the active, and availability returns very quickly.

The Hub Transport servers in the organization can send any mail lost during the failure from the Hub Transport dumpster on CCR-enabled storage groups, so in most cases the only data at risk is client activity that does not go through transport, such as changing a message property, tasks, or calendaring. The Hub Transport servers cannot be clustered, so in a site disaster, Hub Transport servers that are lost are not available to recover mail from, and that mail may be lost. For more information, see the Microsoft document Working with the Queue Database on Transport Servers. NetApp has both a ReplicatorX™ and a SnapMirror® solution to replicate the Hub Transport server data; see the technical report Exchange 2007 Disaster Recovery Model Using NetApp Solutions.

In Exchange 2007 Database Backup and Restore, Microsoft stresses that Continuous Replication is an availability feature, and not a replacement for backup solutions. Logical corruption and deleted items that have been removed during online maintenance have that activity replicated to the target database. One benefit of Continuous Replication is that Volume Copy Shadow Service (VSS) backups can be performed on the target copy, freeing more time on the active LUNs for online maintenance to run. The disk I/O impact of the VSS backup and integrity check is handled on the target LUNs and does not affect the active LUNs. Unlike many VSS clone strategies, SnapManager® 4.0 for Exchange (SME) can quickly perform a VSS backup on the active, the target, or both nodes in a CCR cluster. The flexibility to defer (and throttle) the integrity check to off-peak hours, and to choose any verified backup for restoration, with log replay, provides an up-to-the-minute restore. Other non-NetApp VSS strategies involve a quick backup, but then the data must be streamed to tape because keeping multiple backups as shadow copies affects performance. With NetApp, the administrator can still choose to stream to tape or to a Virtual Tape Library (VTL). The NetApp storage architecture enables the administrator to perform hundreds of backups with little to no measurable performance impact. Customers that deploy large mailboxes (>1GB) may be more willing to reduce the number of regular full backups when using Continuous Replication and facing 6 to 10TB worth of databases on a single Exchange Server. Microsoft has examples of weekly full and daily incremental backup strategies on the TechNet Web site.

Lost Log Resiliency (LLR)

LLR is a new feature that delays writes to the database until a specified number of logs have been generated. LLR enables database recovery even if one of the most recent log files is lost. Only on the active node in a CCR environment can the number of log files affected be changed, up to a maximum of six. This number can be specified in the Exchange attribute `msExchDataLossForAutoDatabaseMount`.

Transaction Log Roll

The log roll mechanism is used to prevent a large window of vulnerability for storage groups that have very low activity. Basically, the current transaction log file (`Exx.log`) is periodically closed, even if it is not full. This helps minimize data loss in a lossy failover (hard failure with data loss), and can lower the RPO on storage groups with light activity. For more information on LLR and log rolling, see Microsoft TechNet.

Continuous Replication Design Considerations

When designing a storage solution for Exchange 2007 that involves replication, the first step is to outline the service-level agreement (SLA). Two key parts to any SLA are the recovery point objective and the recovery time objective.

Recovery Point Objective (RPO)

The RPO is the amount of data that can be lost. A 35-minute RPO means that in a disaster, up to the last 35 minutes of data can be lost. An RPO of zero (or near zero) tolerates almost no data loss and may require technologies in addition to Continuous Replication, depending on the infrastructure.

Recovery Time Objective (RTO)

The RTO is how long it takes to recover from an outage and return to service. This is where Continuous Replication excels. Because it is designed to increase availability, the service interruption will be very brief.

With an RPO of zero, an asynchronous solution using SyncMirror® can be used to complement CCR. With a larger RTO, single copy clustering (SCC) with SnapManager for Exchange and SnapMirror may be a better solution. Exchange 2007 Disaster Recovery Model Using NetApp Solutions describes two detailed solutions, a 5-minute RPO with a 30-minute RTO, and a 0-minute RPO with a 35-minute RTO.

Active and Target Storage Isolation

The purpose of Continuous Replication is to increase availability and provide storage resiliency by asynchronously copying transaction log files to a target LUN. It is not a best practice for both the active and target LUNs to be placed on the same physical disks, because a catastrophic disk failure would affect both source and target, defeating the purpose of continuous replication. Customers that are considering such a solution would be better served by using single copy clustering (SCC), because it requires half the capacity and disk I/O of continuous replication.

The next point of failure is the storage controller. Placing the active and target LUNs on the same controller, but in separate aggregates, causes the storage controller to become a single point of failure. This may be an acceptable risk, although much of the risk can be alleviated by clustering the storage controller, or by moving the target to an entirely separate storage controller, which is the best practice.

Finally, to provide business continuity in the event of a disaster, the target node and storage must be placed in a different location from the active node. Many customers take advantage of Clustered Continuous Replication (CCR) to provide site resiliency. The passive, or target, node is in another location, or site, and is connected to a completely separate storage array. Due to a limitation with Windows® 2003, both nodes in the cluster must be on the same subnet, which requires additional network design considerations. When designing CCR for site resiliency, the network must both keep up with the cluster heartbeat (<500ms latency) and be able to keep up with the log shipping. If, for example, the Exchange Server is producing 30 log files per second, the throughput from the active node to the target node would need to be in excess of 30MB per second.

Performance

In production, the target database LUNs can cause 2x to 3x more I/O than the active LUNs. (Log replay is I/O intensive.) Even though the target performs more I/O than the active, it is recommended to size both active and passive nodes identically with regard to performance and capacity. Because the passive node can become the active in the event of a failover, its storage must be the same as the active node. When identically provisioned, the passive node keeps up with the log replay, and performs more disk I/O at an increase in disk latency. This increase in disk latency requires care in the LUN design.

Backup with SnapManager 4.0 for Exchange (SME)

SME 4.0 is currently LCR and CCR aware, and is integrated with many new Exchange 2007 features, such as up to 50 storage groups and powershell integration. Powershell is the scripting interface that Exchange 2007 uses to automate tasks. SME 4.0 is tightly integrated with powershell, and a command line reference can be found in the SME 4.0 Installation and Administration Guide. By default, with LCR, running a VSS backup on the target affects the production LCR server CPU when running an integrity check. With SME 4.0, the integrity check can be offloaded to a remote verification server. With CCR, the administrator has the flexibility to back up the active node or the target node. Unlike many software-based VSS backups, SME 4.0 can take 250+ Snapshot™ copies and keep them on disk with negligible performance impact for very quick recovery (minutes). Some solutions take a VSS backup, stream it off to tape, and then destroy the volume shadow copy, because keeping more than a couple of shadow copies adversely affects performance. When these tape solutions must be restored, it will be from tape, a lengthy and painful process. See the SnapManager 4.0 for Microsoft Exchange technical report for more information.

Continuous Replication Disk I/O

Introducing Continuous Replication to a production Exchange server slightly changes the active log workload, because those log files must be read when copied to the target. More interesting is that the target LUN workload varies significantly from the active LUNs. LoadGen 2007 was used to simulate 3,000 Outlook® 2007 online-mode users with 250MB mailboxes. For scenario details, see the Server and Storage Configuration at the end of this technical report.

DATABASE I/O

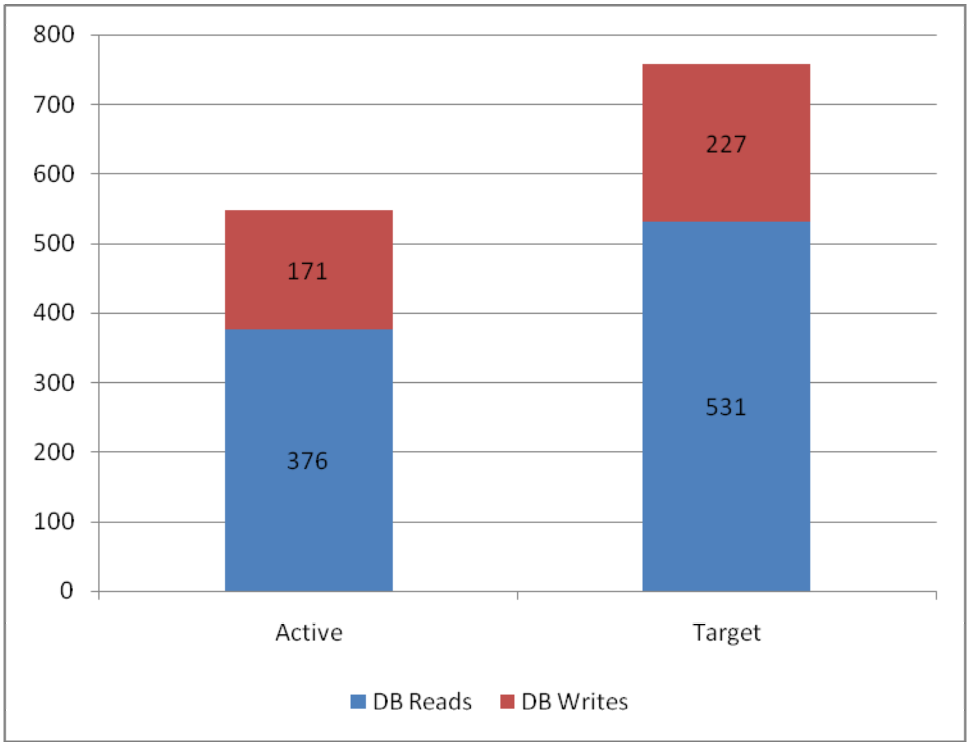


Figure 1) Active versus target DB I/O.

The target database read I/O is 41% higher and the target database write I/O is 33% higher than on the active LUNs. The average total I/O increase is 38.5%. Outlook users in cached mode cause 2x to 3x more database I/O on the passive database LUNs (versus 40% more with Outlook users in online mode).

The read-to-write ratio is virtually unchanged at 68:32 on the active and 70:30 on the target. Each second, 1.24MB of data is transferred on the active, and 1.75MB of data is transferred on the target, a 41% increase. Depending on the user profile, the read-to-write ratio has approached 50:50 in production with some user profiles.

In order to meet capacity demands, more disk performance was available than was required, and the disk latencies reflect that. These latencies were achieved using RAID-DP™, which unlike many double parity solutions delivers awesome performance, and even better reliability than RAID10.

DB Latency	CCR1	CCR2
DB Read	9ms	9ms
DB Write	11ms	1ms
Log Read	<1ms	<1ms
Log Write	<1ms	<1ms

In production, with some user configurations, the target DB I/O can be 2x to 3x higher than the active DB I/O.

LOG I/O

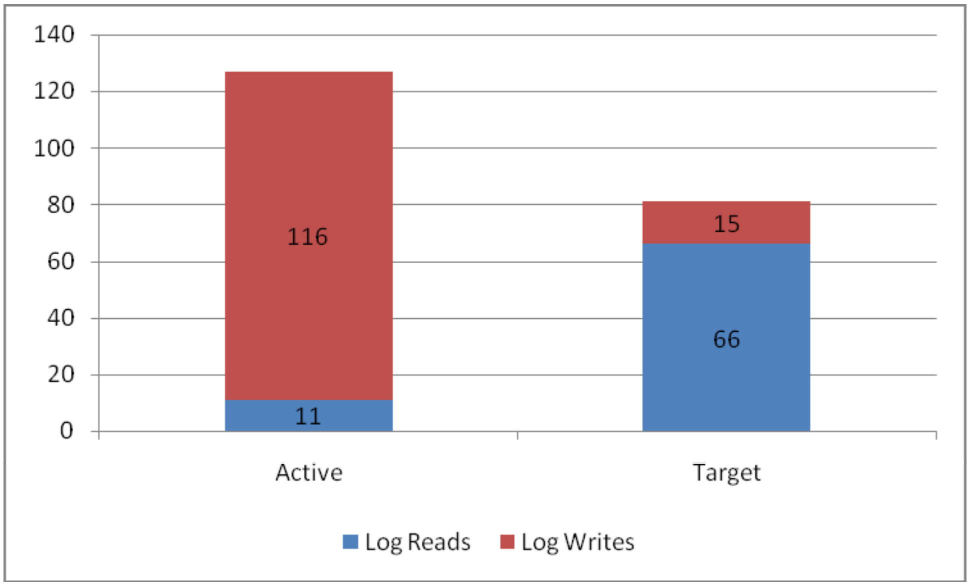


Figure 2) Active versus target log I/O.

The target log writes drop 87%, and the target log reads increase 600%. The total number of I/Os per second on the target LUNs is less than on the active LUNs, yet more bytes are being transferred on the target. The important point to consider with the log I/O decrease is to identically provision both the active and target log LUNs with regard to performance and capacity.

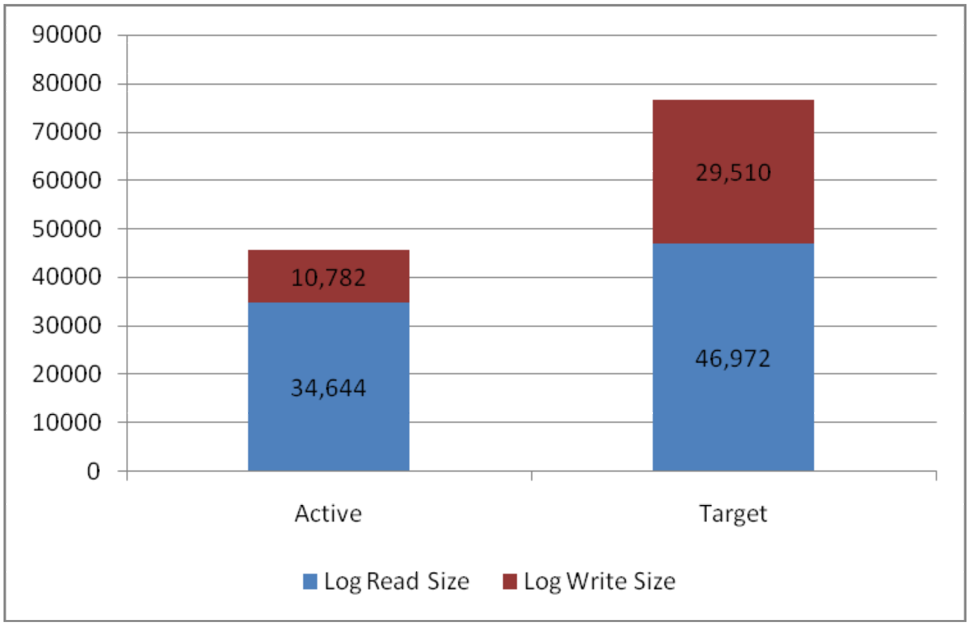


Figure 3) Active versus target log I/O Size (bytes).

Looking at the data from another angle, each write I/O is almost 300% larger on the target. The average read I/O increases 36% on the target.

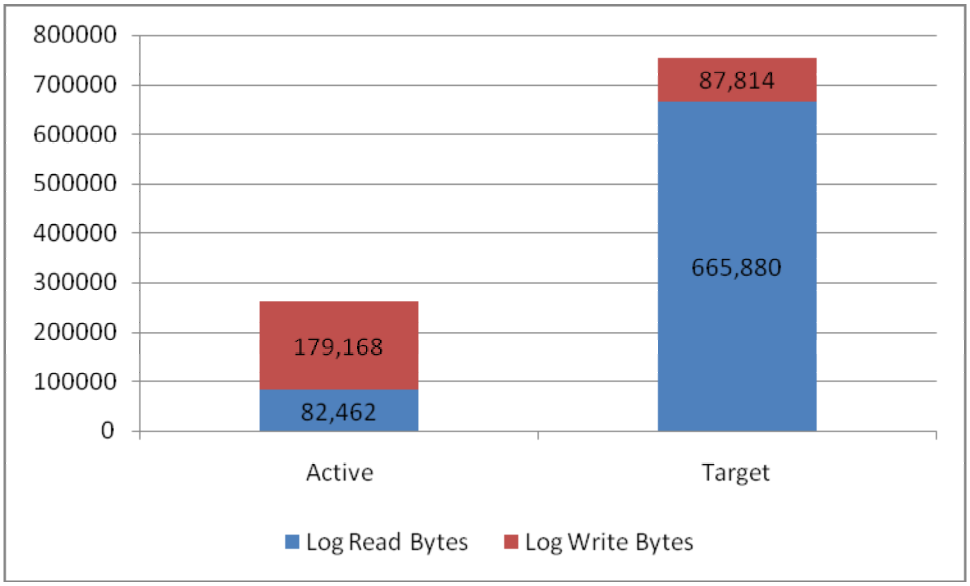


Figure 4) Active versus target log I/O size (bytes).

This graph shows how the log workload is vastly different on the target. The amount of bytes being transferred on the target log LUNs is 288% more! The write bytes are cut in half, while the reads increase eight fold. Although the amount of data being transferred is larger on the target, overall the number of log I/Os per second (disk transfers) is less on the target (131:77 Active:Target).

Exchange 2007 SP1

One of the new features in Exchange 2007 SP1 that is tentatively scheduled to ship at the end of 2007 is Standby Continuous Replication (SCR). SCR is a log shipping target, where the logs are replicated to, inspected, and then played into a standby database. The key with SCR is that the source can be any Exchange 2007 mailbox server—CCR, Single Copy Cluster (SCC), or standalone. Many customers are very successful today with SCC combined with VSS Snapshot copies using SME for very high availability, and for these customers SCR is very compelling. Another benefit of Continuous Replication in SP1 is a vast performance improvement in the target disk I/O. Currently, the target database disk I/O can be 2x to 3x the active database disk I/O, and this will be reduced to less than the active, a huge performance win.

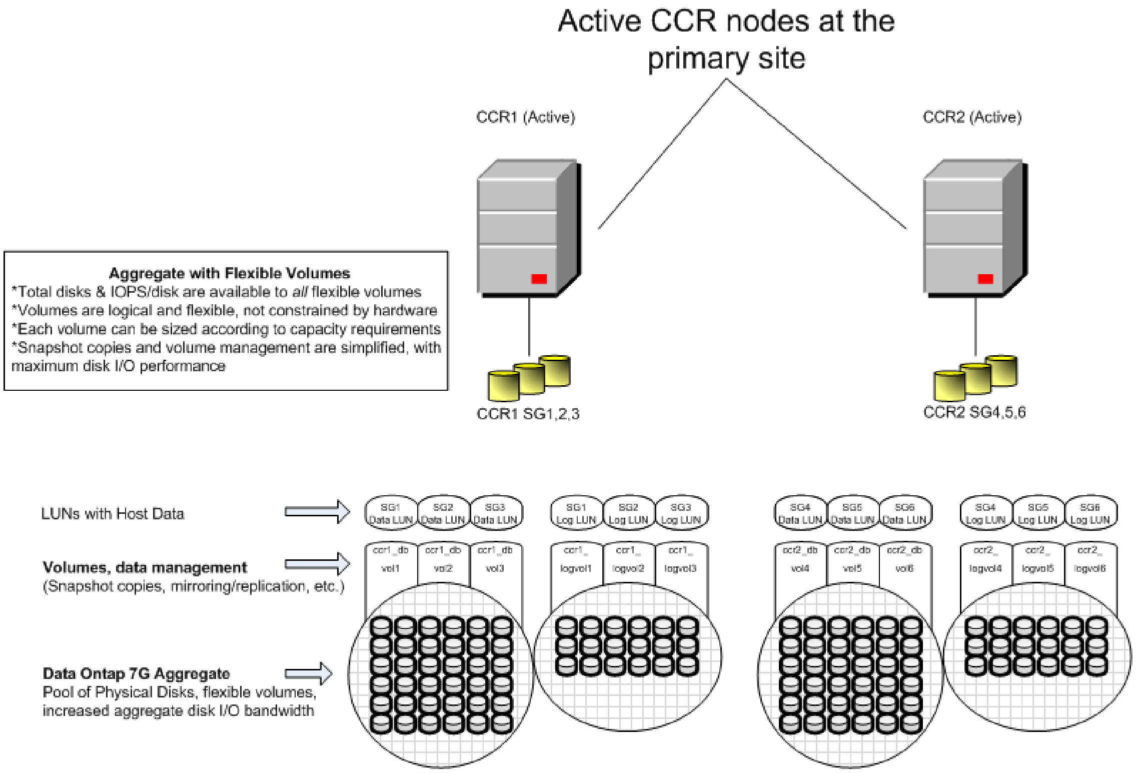
Continuous Replication LUN Design

Creating LUNs that provide adequate performance is only half of the solution. The LUNs must be provisioned with enough capacity as well. For detailed capacity planning information, see Planning Disk Storage.

The first best practice is to ensure that the transaction logs and databases are in separate aggregates. From both a performance (mixing workloads) and reliability perspective, do not place the transaction logs and databases on the same physical disks. The *general* Microsoft Exchange best practice is that like workloads can share spindles, even between Exchange Servers.

When creating volumes inside the aggregate, it is recommended to use FlexVol® volumes (and to create a separate flexible volume for each storage group. FlexVol volumes are equally spread across every disk in the aggregate, and they are the layer where Snapshot copies occur. By creating a separate FlexVol volume for each storage group inside the database aggregate and the log aggregate, the VSS restores do not affect any other storage group. This is a best practice.

With Continuous Replication, it is recommended that the physical disks backing the LUNs are further isolated from other servers. For example, if there are two CCR clusters, create a separate log aggregate for each active node and a separate log aggregate for each target node. Then create a separate database aggregate for each active node and a separate database aggregate for each target node. Figure 5 illustrates the disk configuration for the active nodes on two separate CCR clusters. Notice that each storage group has its own FlexVol volume. This same configuration should be repeated for the target nodes at the DR site on the DR storage.



Aggregate with Flexible Volumes
*Total disks & IOPS/disk are available to *all* flexible volumes
*Volumes are logical and flexible, not constrained by hardware
*Each volume can be sized according to capacity requirements
*Snapshot copies and volume management are simplified, with maximum disk I/O performance

Figure 5) Using Data ONTAP 7G with flexible volumes with multiple CCR clusters.

Clearly, the workloads are different from the active LUNs on the target LUNs. The good news is that if each CCR node is isolated at the disk level, or aggregate level, the target keeps up if it is identically provisioned, although at a higher latency. The alternative is to overprovision each LUN with twice the performance required by the active LUNs if physical disks are shared between CCR clusters to prevent poor performance on active nodes in the event of a failover.

CCR Best Practices

- Isolate each node in a CCR cluster in its own aggregate.
- Separate logs and databases in their own aggregates.
- Create a separate FlexVol volume for each storage group.
- Use RAID-DP™ for superior performance and protection.
- Run SME backup on the target node, and extend the online maintenance window on the active node.
- Be sure to plan for enough capacity and performance; see Planning Disk Storage.
- Provision the active and target LUNs identically with regard to capacity and performance.
- Consider ReplicatorX and/or SnapMirror to achieve a <5 minute RPO.

Summary

Microsoft Exchange is a business-critical application, and there are many solutions available to increase availability. It is important to spend the time to identify the service-level agreement and to then utilize technologies in the storage design to meet those goals. Once the solution is designed, it is critical to test it (see Jetstress and LoadGen in Tools for Exchange Server 2007), and once deployed, to monitor both the Exchange Server and the NetApp storage.

NetApp has proven data protection and disaster recovery tools for Microsoft Exchange. SnapManager for Exchange backup and restore capabilities, combined with SnapMirror technologies, provide a solid and robust solution for protecting and recovering exchange data while meeting stringent RPO and RTO objectives.

Server and Storage Configuration

The CCR cluster was configured so that the nodes were isolated from each other in the following configuration.

- Active and passive CCR nodes were HP DL140G3s with 16GB RAM.
- Each node was connected via iSCSI with 1GbE to a separate FAS3050 controller.
- 3000 Outlook online-mode heavy users were split between two LoadGen 2007 clients.
- Six storage groups with 500 users each with LUNs presented as mount points.
  - Each LUN was aligned (diskpart) and formatted with a 64kB allocation unit size.
- Each 3050 (active and target) had two aggregates using RAID-DP, one log and one DB.
  - Log aggregate 5-144GB FC disks (3 data, 2 parity).
  - DB aggregate 20-144GB FC disks (18 data, 2 parity).
- Each aggregate had six FlexVol volumes, one for each storage group
- Each database was ~134GB in size, with a 30GB content index.