Implementing Oracle® Database 11g Running with Direct NFS Client on Network Appliance™ MetroCluster

TR-3614 | September 2007
Amarnath Rampratap, Niranjan Mohapotra | Network Appliance, Inc.

Table of Contents

Introduction

Most of today's business applications are data-centric, requiring fast and reliable access to intelligent information architectures that can often be provided by a high-performance relational database system. Oracle is one among the relational database systems that provides such a back-end data store for mission-critical line-of-business applications.

The latest release, Oracle Database 11*g*, offers significant architectural enhancements in performance and scalability. With the proliferation of e-business, an enterprise today operates in an extremely complex and a highly networked global economy and is more susceptible to interruptions than in the past. The cost of interruptions or downtimes varies across industries and companies, and this cost can be as much as millions of dollars in an hour. While these numbers are staggering, the reasons are quite obvious. A business can suffer downtimes that can be unplanned and planned. Unplanned downtimes may be caused by hardware or system failures, data storage failures, human errors, software glitches, natural disasters, and so on. NetApp along with Oracle has extended its reach to increase productivity and keep information close to hand, flexible enough to meet your organization's administrative model.

Purpose and Scope

The purpose of this technical report is to provide reference architecture for Oracle Database 11*g* Release 1 running with DNFS on Network Appliance MetroCluster solutions that is designed to achieve a highly available relational database management system environment.

Intended Audience

This technical report is intended for Information technology professionals, storage professionals, Oracle DBAs, and business continuity professionals responsible for database management infrastructure. For methods and procedures in this technical report, it is assumed that the reader has knowledge of the following:

Oracle Database system architecture:

- Oracle storage architecture and database administration
- Oracle Real Application Clusters (RAC)
- Service-level expertise of Oracle Server recovery options

Working knowledge on NetApp solutions, including the following:

- Data ONTAP®
- NetApp MetroCluster

Assumptions

Throughout this document it is assumed that we have two physical sites, "Site A" and "Site B". These are separated by 10.5 km. Naming of all components will clearly show whether they are physically located at "Site A" or "Site B".

High-Level Topology Diagram

The overall solution uses NetApp MetroCluster as a back end for storage availability and two extended Oracle RACs, each with two active nodes spanning across sites as front end for application availability. The nodes are running Oracle Enterprise Linux® Update 5 hosting two instances and one database per cluster accessing their storage via DNFS (Direct Network File System) Client. In a normal situation, one database is active on "Site A" and accessing their storage via DNFS at "Site A," and the other database is active on "Site B" and accessing their storage via DNFS at "Site B."

The general layouts of components used in this reference architecture are shown below:

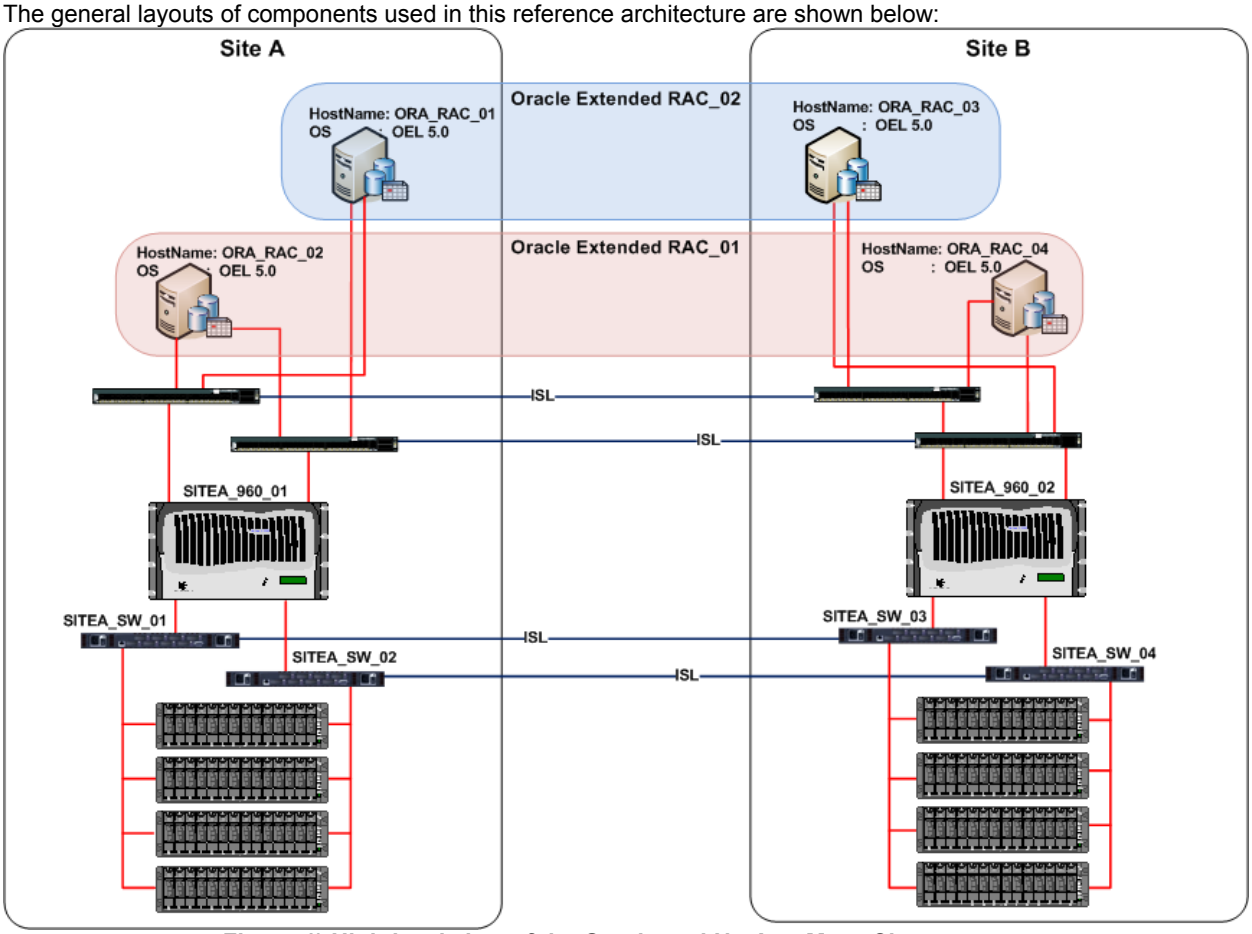

Figure 1) High-level view of the Oracle and NetApp MetroCluster setup.

Material List

Hardware	Vendor	Name	Version	Description
Storage	NetApp	FAS960AA	N/A	Storage Controller
Hosts	IBM	Two Oracle RAC (Two Nodes) on each site 2 X IBM eServer XSeries 445 (2.8 Ghz/4GB RAM) 2 X IBM eServer XSeries 206 (2.8 Ghz/4GB RAM)	N/A	Hosts Server
Front-End Network	Cisco	4948 (4)	IOS 12.1	48 Port Ethernet Switch
Back-End SAN (MetroCluster)	Brocade	200E (4)	5.1.0	16 Port FC Switch
Software	**Vendor**	**Name**	**Version**	**Description**
Storage	NetApp	SyncMirror®	7.2.3	Replication
	NetApp	Data ONTAP	7.2.3	Operating System
	NetApp	Cluster_Remote	7.2.3	Failover
Hosts	Oracle	Enterprise Linux	5.0	Operating System
	Oracle	Oracle Database	11*g* R1	Database
	Oracle	Oracle CRS	11*g* R1	Oracle Cluster Service

Platform Specification

FAS Storage Controller

The controller and back-end Fibre Channel switches were configured using the instructions in the Data ONTAP 7.2.3 Active/Active Configuration Guide, and the current firmware levels and other notes found on the NOW™ (NetApp on the Web) site.

- Data ONTAP release: 7.2.3
- Brocade firmware: v5.1.0

Two FAS960 series controllers (each with four DS14mk2-HA shelves full of 66GB 15kdrives) connected with the VI-MC interconnect, and four Brocade 200E switches were used in this test. The controllers were named SITEA_960_01 and SITEB_960_02, and the switches were named SITEA_SW01, SITEA_SW02, SITEB_SW03, and SITEB_SW04.

Slot Assignments

The controllers are configured identically in terms of hardware with the following cards/slot assignments:

Slot #	Card	Purpose
1	X3300A: Remote management card	Remote monitoring/management
5	X2050A: Dual optical Fibre Channel for mirroring	Disk connection
6	X1922A: VI-MetroCluster	Cluster interconnect
7	X3140A: NVRAM4	NVRAM card
8	X2050A: Dual optical Fibre Channel for mirroring	Disk connection
11	X2050A: Dual optical Fibre Channel for target interconnect	Target card

Network Settings

SITEA_960_01

Interface	IP Address	Purpose
E0	172.17.149.20, Partner e0	LAN

SITEB_960_02

Interface	IP Address	Purpose
E0	172.17.149.25, Partner e0	LAN

Aggregate Layout

Controller	Aggregate Name	Options	# of Disks	Purpose
SITEA_960_01	ORADATA_A	RAID_DP™, aggr Mirrored	14	Oracle DB for ORCL_SITEA
SITEA_960_01	ORALOGS_A	RAID_DP, aggr Mirrored	14	Oracle Archive Logs
SITEA_960_01	ORA11G_A	RAID_DP, aggr Mirrored	3	Oracle OCR file and CSS file
SITEA_960_02	ORADATA_B	RAID_DP, aggr Mirrored	14	Oracle Database for NTAP_SITEB
SITEA_960_01	ORAOGS_B	RAID_DP, aggr Mirrored	14	Oracle Archive Logs
SITEA_960_01	ORA11G_B	RAID_DP, aggr Mirrored	3	Oracle OCR file and CSS file

Volume Layout

Controller	Volume Name	Options	Size (GB)	Purpose
SITEA_960_01	Vol0	RAID_DP, flex mirrored, create_ucode=on, convert_ucode=on	42.6	Root Volume
SITEA_960_01	ORADATA_A	RAID_DP, flex mirrored, create_ucode=on, convert_ucode=on	200	Oracle Database & Redo Logs
SITEA_960_01	ORALOGS_A	RAID_DP, flex mirrored, create_ucode=on, convert_ucode=on	150	Oracle Archive Logs
SITEA_960_01	ORA11G_A	RAID_DP, flex mirrored, create_ucode=on, convert_ucode=on	20	Oracle CRS and CSS files

SITEA_960_02	Vol0	RAID_DP, flex mirrored, create_ucode=on, convert_ucode=on	42.6	Root Volume
SITEA_960_02	ORADATA_B	RAID_DP, flex mirrored, create_ucode=on, convert_ucode=on	200	Oracle Database & Redo Logs
SITEA_960_02	ORALOGS_B	RAID_DP, flex mirrored, create_ucode=on, convert_ucode=on	150	Oracle Archive Logs
SITEA_960_02	ORA11G_B	RAID_DP, flex mirrored, create_ucode=on, convert_ucode=on	20	Oracle CRS and CSS files

MetroCluster Configuration

Switch Configuration

The back-end FC switches in a MetroCluster environment must be set up in a specific manner for the solution to function properly. In the sections below, the switch and port connections are detailed and should be implemented exactly as documented.

SITEA_SW_01
IP Address: 172.17.149.235
Domain ID: 1

Port	Bank/Pool	Connected with	Purpose
0	1/0	SITEA_960_01, 5a	SITEA_960_01 FC HBA
1	1/0	SITEA_960_01, 8a	SITEA_960_01 FC HBA
2	1/0		
3	1/0		
4	1/1		
5	1/1	SITEA_960_02, Pool1 Shelf 3B	
6	1/1		
7	1/1		
8	2/0		
9	2/0	SITEA_960_02, Pool0, Shelf 1B	
10	2/0		
11	2/0		
12	2/1	SITEA_960_01, FCVI, 6a	Cluster Interconnect
13	2/1	SITEB_SW_03, Port 5	ISL
14	2/1		
15	2/1		

SITEA_SW_02
IP Address: 172.17.149.236
Domain ID: 2

Port	Bank/Pool	Connected with	Purpose
0	1/0	SITEA_960_01, 5a	DISK HBA for BANK 2 Shelves
1	1/0	SITEA_960_01, 8a	DISK HBA for BANK 2 Shelves
2	1/0		
3	1/0		
4	1/1		
5	1/1	SITEA_960_02, Pool1 Shelf 3A	
6	1/1		
7	1/1	SITEA_960_01 FCVI, 6b	Cluster Interconnect
8	2/0		
9	2/0	SITEA_960_01, Pool0, Shelf 1A	
10	2/0		
11	2/0		
12	2/1		
13	2/1	SITEB_SW_04, Port4	ISL
14	2/1		
15	2/1		

SITEB_SW_03
IP Address: 172.17.149.237
Domain ID: 3

Port	Bank/Pool	Connected with	Purpose
0	1/0	SITEA_960_02, Pool1 Shelf 3B	
1	1/0		
2	1/0		
3	1/0	SITEA_960_01, FCVI, 6a	Cluster Interconnect
4	1/1		
5	1/1	SITEA_SW_01, Port 13	ISL
6	1/1		
7	1/1		
8	2/0	SITEA_960_02, 5a	Disk HBA Bank 2 Shelves
9	2/0	SITEA_960_02, 8a	Disk HBA Bank 2 Shelves
10	2/0		
11	2/0		
12	2/1	SITEB_960_02, Pool 0, Shelf 1B	
13	2/1		
14	2/1		
15	2/1		

SITEB_SW_04
IP Address: 172.17.149.238
Domain ID: 4

Port	Bank/Pool	Connected with	Purpose
0	1/0	SITEA_960_01, Pool 1, Shelf 3A	
1	1/0		
2	1/0		
3	1/0		
4	1/1	SITEA_SW_02, Port 13	ISL
5	1/1		
6	1/1		
7	1/1		
8	2/0	SITEB_960_02, 5b	Disk HBA Bank 2 Shelves
9	2/0	SITEB_960_02, 8b	Disk HBA Bank 2 Shelves
10	2/0		
11	2/0		
12	2/1	SITEB_960_02, Pool 0, Shelf 1A	
13	2/1	SITEB_960_02 FCVI, 6b	Cluster Interconnect
14	2/1		
15	2/1		

Host Servers

Host Configuration

Two Oracle extended Real Application Clusters (RAC) were set up spanning across sites with two nodes (IBM eServer xSeries with four Intel® xeon CPU and 4GB RAM) as members in each RAC. The hosts in the RACs are named as ORA_RAC_01, ORA_RAC_02, ORA_RAC_03, and ORA_RAC_04.

Software Configuration

The hosts in the Oracle RAC are installed according to vendor-supplied procedures with:
Oracle Enterprise Linux 5.0

Oracle Cluster Ready Service (CRS) 11*g* Release 1
Oracle Database 11*g* Release 1

Network Configuration

The following table details the network settings for various Oracle hosts.

Hostname	IP Address	Purpose
ORA_RAC_01	172.17.148.216	LAN
ORA_RAC_01	192.168.73.1	Heartbeat
ORA_RAC_01	172.17.148.218	Virtual IP
ORA_RAC_02	172.17.148.219	LAN
ORA_RAC_02	192.168.73.2	Heartbeat
ORA_RAC_02	172.17.148.221	Virtual IP
ORA_RAC_03	172.17.148.225	LAN
ORA_RAC_03	192.168.73.3	Heartbeat
ORA_RAC_03	172.17.148.227	Virtual IP
ORA_RAC_04	172.17.148.228	LAN
ORA_RAC_04	192.168.73.4	Heartbeat
ORA_RAC_04	172.17.148.230	Virtual IP

Oracle RAC/Database Configuration

After a complete installation of Oracle Enterprise Linux 5.0 the following configurations were done before installing Oracle CRS and Database:

Step 1: Set up the kernel parameters as follows:
Add the following entries into the /etc/sysctl.conf file as a root user:
kernel.shmmni = 4096
# semaphores: semmsl, semmns, semopm, semmni
kernel.sem = 250 32000 100 128
net.ipv4.ip_local_port_range = 1024 65000
net.core.rmem_default=4194304
net.core.rmem_max=4194304
net.core.wmem_default=262144
net.core.wmem_max=262144

Step 2: Set up the user limits for Oracle user as follows:
Add the following entries into the /etc/security/limits.conf file:
oracle soft nproc 2047
oracle hard nproc 16384
oracle soft nofile 1024
oracle hard nofile 65536

Step 3: Install the following packages for Enterprise Linux 5.0:
From Enterprise Linux 5.0 Disk 1:
rpm -Uvh binutils-2.*
rpm -Uvh elfutils-libelf-0.*
rpm -Uvh glibc-2.*
rpm -Uvh glibc-common-2.*
rpm -Uvh libaio-0.*
rpm -Uvh libgcc-4.*
rpm -Uvh libstdc++-4.*
rpm -Uvh make-3.*

From Enterprise Linux 5.0 Disk 2:
rpm -Uvh compat-libstdc++-33*
rpm -Uvh elfutils-libelf-devel-0.*
rpm -Uvh glibc-devel-2.*
rpm -Uvh gcc-4.*
rpm -Uvh gcc-c++-4.*
rpm -Uvh libaio-devel-0.*

rpm -Uvh libstdc++-devel-4.*
rpm -Uvh unixODBC-2.*
rpm -Uvh unixODBC-devel-2.*

From Enterprise Linux 5.0 Disk 3:
rpm -Uvh sysstat-7.*

Step 4: Enable rsh access without password for Oracle user and root user to all the nodes in the Oracle RAC before installing Oracle CRS and Oracle Database.

Note: OCR file and CSS file were hosted on the storage using NFS.

Step 5: After Installation of Oracle Database 11*g*, the following steps were performed to configure the DNFS client:
Log in as 'Oracle' user (the user who owns the Oracle binary) and do the following steps.
Prompt > cd $ORACLE_HOME/lib
Prompt > mv libodm11.so libodm11.so_stub
Prompt > ln –s libnfsodm11.so libodm11.so

Note: The above steps were performed to host the data files, redo log files and the control files on NetApp storage using direct NFS.

Data Configuration
Benchmark factory was used to generate TPCC workload on the Oracle Database during the test scenarios.

Functional Tests

The functional test results will be broken into two levels:

- Basic system operation: operation of the system, in the various failure scenarios outlines later in this report
- Oracle operation: the reaction of Oracle to the various failure scenarios outlined later in this report

Test Scenarios

The following subsections describe the various test scenarios that were executed upon successful build of the solution discussed earlier in this document. The test scenarios include various component failures, including server hardware, network, storage system, etc. Unless mentioned otherwise, prior to the execution of each test, the environment was reset to the normal running state. The normal running state had all the Oracle RAC nodes operational and the ORCL_SITEA database active on SITE A and NTAP_SITEB database active on Site B. Additionally, benchmark factory was configured to perform typical user transactions on both databases, ORCL_SITEA and NTAP_SITEB.

Complete Loss of Power to Disk Shelf

No single point of failure should exist in the scenario. Therefore, the loss of an entire disk shelf was tested. This test was accomplished by simply turning off both power supplies while a load was applied.

Task	Power off the shelf "IBMX445-WHQL02 Pool0," observe results, and then power it back on.
Expected/Observed Results	Relevant disks go offline, plex is broken, but service to clients (availability and performance) is unaffected. When power is returned to the shelf the disks will be detected and a resync of the plexes will occur without any manual action.
Oracle Results	Oracle service was unaffected; no failure of Oracle resource or CRS service occurred.

Loss of One Link on One Disk Loop

No single point of failure should exist in the scenatio. Therefore, the loss of one disk loop was tested. This test was accomplished by removing a fiber patch lead from one of the shelves.

Task	Remove fiber entering IBMX445-WHQL01 Pool0, ESH A, observe results, and then reconnect the fiber.
Expected/Observed Results	Controller messages that some disks are connected to only one switch will be displayed, but service to clients (availability and performance) will be unaffected. When the fiber is reconnected controller messages that disks are now connected to two switches will be displayed.
Oracle Results	Oracle service was unaffected; no failure of Oracle resource or CRS service occurred.

Loss of a Brocade Switch

No single point of failure should exist in this scenario. Therefore, the loss of an entire Brocade switch was tested. This test was accomplished by simply removing the power cord from the switch while the load was running.

Task	Power off the Fibre Channel switch "SITA-SW_04," observe results, and then power it back on.
Expected/Observed Results	Controller messages that some disks are connected to only one switch and that one of the cluster interconnects is down are displayed, but service to clients (availability and performance) is unaffected. When the power is restored and the switch completes its boot process, controller messages are displayed to indicate that the second cluster interconnect is again active.
Oracle Results	Oracle service was unaffected; no failure of Oracle resource or CRS service occurred.

Loss of One ISL

No single point of failure should exist in the solution. Therefore, the loss of one of the interswitch links (ISLs) was tested. This test was accomplished by simply removing the fiber connection between two of the switches while a load was applied.

Task	Remove the fiber between SITA_SW_01 and SITA_SW_03.
Expected/Observed Results	Controller messages that some disks are connected to only one switch and that one of the cluster interconnects is down will be displayed, but service to clients (availability and performance) will be unaffected. When ISL is reconnected controller messages will be displayed to indicate that the disks are now connected to two switches and that the second cluster interconnect is again active.
Oracle Results	Oracle service was unaffected; no failure of Oracle resource or CRS service occurred.

Failure of Controller

No single point of failure should exist in the solution. Therefore, the loss of one of the controllers itself was tested.

Task	Power off the running SITEA_960_01 controller.
Expected/Observed Results	As a result of the change of processing from one controller to the other, host interruption should be minimal if any, because the failover is masked by the "disk time out" value in Oracle, which is set to 200 secs by default. Failure of database or CRS should not occur.
Oracle Results	Oracle service was unaffected; no failure of Oracle resource or CRS service occurred.

Failback of Controller

As a follow-up to the previous test, the data serving must be failed back to the previously failed controller to return to the normal operating state. This test was accomplished by issuing a command on the surviving controller to request that processing be returned to the previously failed controller.

Task	Power on SITEA_960_01. Issue a cf giveback command on SITEB_960_02 to cause the failback to occur.
Expected/Observed Results	As a result of the change of processing from one controller to the other, host interruption should be minimal if any, because the failover is masked by the "disk time out" value in Oracle, which is set to 200 secs by default. Failure of database or CRS should not occur.
Oracle Results	Oracle service was unaffected; no failure of Oracle resource or CRS service occurred.

Crash of a Node in the Oracle RAC

To test the availability of the Oracle RAC setup we will intentionally power off one of the nodes in the Oracle RAC.

Task	Power off ORA_RAC_01.
Expected/Observed Results	Host should crash and the remaining nodes in the Oracle RAC should realize what happened. But the database should not be affected and the crs_stat –t command should reflect the failure of the failed node and the database and the instance on the surviving node should be online.
Oracle Results	Oracle service was unaffected, crs_stat –t command showed that one node has failed and the database and the instance on the surviving nodes as online.

Failure of LAN Connection on Node in Oracle RAC

To test availability of the Oracle RAC solution we will remove the LAN (public) interface from a node on the Oracle RAC.

Task	Remove the LAN cable from the LAN (public) interface on Node ORA_RAC_01.
Expected/Observed Results	crs_stat –t should show that the instance on the node with the failed LAN connection is offline and the database should be online on the instance on the surviving node.
Oracle Results	Oracle service was unaffected; GSD and ONS resources and the instance of the node with failed LAN connection were offline.

Failure of Both Oracle Nodes on the Primary Site

To test the availability of the Oracle RAC solution, we will fail both the Oracle Enterprise Linux nodes on the primary site.

Task	Remove power to ORA_RAC_01 and ORA_RAC_02.
Expected/Observed Results	One node in each of the RAC should fail and the service to clients should not be affected and the database should be online on the instances on the surviving nodes.
Oracle Results	Oracle service was unaffected; failed nodes were offline in crs_stat -t.

Loss if Entire Site Declared

To test availability of the overall solution we will simulate loss of an entire site.

Task	Test the failure of the "SITE A" site by interrupting the following components in this order, in rapid succession: **Step 1: Simulate Failure** Remove both ISLs. Remove power to all nodes. Remove power from SITEA_960_01. **Step 2: Recovery** Declare the disaster and perform a takeover at the surviving site, SITE B. Issue the following command on SITEB_960_02: **`SITEB_960_02> cf forcetakeover -d`** Note: It is important to shut down the database and the CRS service, as the disktimeout value in Oracle is set to 200 seconds by default; if the failover process exceeds 200 seconds, this can result in a CRS reboot. Use the partner command on SITEB_960_02 to access SITEA_960_01 (now running on the same controller as SITEB_960_02). Start Oracle CRS on both the nodes in Site B.
Expected/Observed Results	One node in each of the Oracle RACs should fail and the service to clients should not be affected and the database should be online on the instances on the surviving nodes.
Oracle Results	Oracle service was affected when the CRS was shut down; the failover process (including declaring a disaster and takeover) completed in 130 seconds; failed nodes were offline in crs_stat -t.

Restore of Entire Site/Recover from Disaster

To test availability of the overall solution we will simulate recovery after loss of an entire site.

Task	Test the recovery/restore of SITE A by bringing back the following components in this order: Step 1: Power on the controller (SITEA_960_01). Step 2: Reconnect the ISLs between sites so that SITEB_960_02 can see the disk shelves from SITEA_960_01. After the connection, the SITEB_960_02 pool 1 volumes automatically begin to resync. Step 3: The following steps should be followed on the Oracle nodes in SITE A: Individually power on each of the Oracle nodes. Verify that the CRS service starts correctly and the node has become onlne using crs_stat –t command line. Step 4: When all the nodes are online, issue the cf status command at SITEA_960_01 console to verify if the giveback is possible and use cf giveback command to fail back to SITE A.
Expected/Observed Results	On the cluster giveback to the SITE A controller, the results should be similar to the normal giveback. There should be no downtime during the failback process.
Oracle Results	Oracle service was unaffected, failed nodes booted successfully after restoring power, and the CRS service started on the nodes in SITE A and was back online.

Conclusion

Documentation of the configuration, tests performed, and the results observed were used to update this technical report. Failure scenarios on the storage side in a MetroCluster environment did not affect Oracle Database/RAC functionality and performance.