Technical Report

Providing Zero Downtime for Enterprise Applications Using Oracle Real Application Cluster on Extended Distance Cluster, Oracle Automatic Storage Management Normal Redundancy, and NetApp MetroCluster

Antonio Jose Rodrigues Neto, Jeffrey Steiner, Jim Lanson, Karthikeyan Nagalingam, Lou Lydiksen, and Neil Gerren, NetApp

February 2010 | TR-3816

EXECUTIVE SUMMARY

Whether you have a single data center, a campus, or a metropolitan-wide environment, a combined solution using NetApp® and Oracle® technologies is a cost-effective alternative that can provide continuous data availability for your mission-critical enterprise applications. NetApp MetroCluster is an industry-leading solution that combines NetApp storage array–based clustering and synchronous mirroring to help deliver continuous availability and minimal data loss. NetApp Fabric MetroCluster combined with Oracle Real Applications Clusters (RAC) on Extended Distance Clusters, and Oracle Automated Storage Management (ASM) Normal Redundancy offers transparent recovery from failures so mission-critical Oracle databases can continue uninterrupted with zero downtime.

TABLE OF CONTENTS

1 INTRODUCTION

1.1 PURPOSE

This document provides a configuration design for Oracle Databases on NetApp storage requiring high availability and zero downtime for mission-critical applications. This solution leverages technologies from Oracle and NetApp, including MetroCluster, ASM, and RAC. It also provides detailed architectural and configuration diagrams and tables.

1.2 OVERVIEW OF THE NETAPP TECHNOLOGY USED

METROCLUSTER

MetroCluster is a unique solution that combines array-based clustering with synchronous mirroring, implemented at the RAID level, to deliver continuous availability and zero data loss at the lowest cost. As a self-contained solution at the array level, NetApp MetroCluster in conjunction with Oracle ASM Normal Redundancy provides transparent recovery for site failures so business-critical Oracle databases continue uninterrupted. This eliminates repetitive change management activities while reducing the risk of human error and administrative overhead.

You can now benefit from new MetroCluster enhancements:

- Nondisruptive upgrades to minimize planned downtime
- Testing with Oracle RAC Extended Cluster to achieve continuous availability in a critical environment
- Integration with other NetApp technologies to gain additional storage efficiencies

MetroCluster configurations consist of a pair of active-active NetApp storage controllers configured with mirrored aggregates and extended distance capabilities to create a high-availability solution. The primary benefits include:

- Higher availability with geographic protection
- Minimized risk of lost data
- Simplified management and recovery
- Reduced system downtime
- Quicker recovery when a disaster occurs
- Minimal disruption to users and client applications

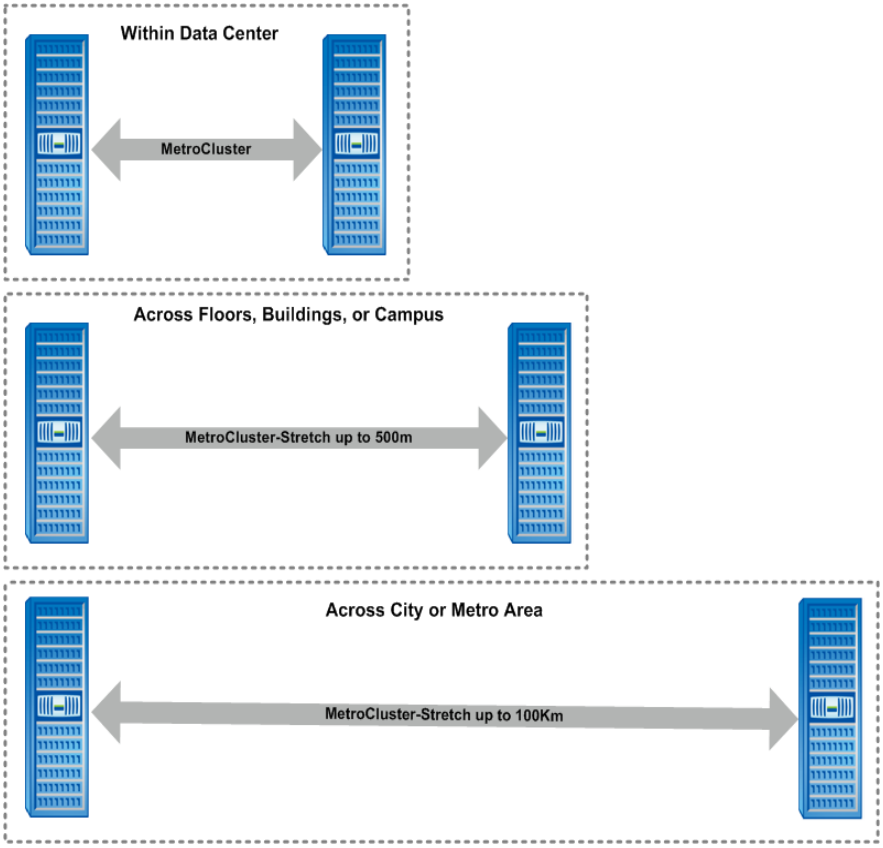

Figure 1) MetroCluster.

METROCLUSTER TYPES

Stretch MetroCluster (sometimes referred to as nonswitched) is simply an active-active configuration that can extend up to 500m depending on speed and cable type. It also includes RAID-level synchronous mirroring (SyncMirror®) and the ability to do a site failover with a single command. Additional resiliency can be provided through the use of multipathing. Further information on multipathing can be found in the *Data ONTAP Active-Active Configuration Guide* located on the NOW™ site.

Fabric MetroCluster (switched) uses four Fibre Channel switches in a dual-fabric configuration and a separate cluster interconnect card to achieve clustering over greater distances (up to 100km depending on speed and cable type) between node A and node B locations.

V-Series MetroCluster is simply either of the above configurations with a NetApp V-Series system. Because of the architectural differences between V-Series and a standard active-active configuration, V-Series MetroCluster has additional flexibility when it comes to the maximum number of disk spindles and Fibre Channel switches. For more information, see the V-Series documentation and the V-Series Compatibility Guide on NOW.

OPERATION

NetApp MetroCluster behaves in most ways like an active-active configuration. All protection provided by core NetApp technology (RAID-DP®, Snapshot™ copies, automatic controller failover) also exists in a MetroCluster configuration. However, MetroCluster adds complete synchronous mirroring at an aggregate level along with the ability to perform a complete site failover from a storage perspective with a single command.

MIRRORING

NetApp SyncMirror, an integral part of MetroCluster, combines the disk-mirroring protection of RAID 1 with industry-leading NetApp RAID 6 and RAID-DP technology. In the event of an outage—whether due to a disk problem, cable break, or host bus adapter (HBA) failure—SyncMirror can instantly access the mirrored data without any operator intervention or disruption to client applications. SyncMirror maintains strict physical separation between two copies of your mirrored data. Each copy is referred to as a plex. Each controller's data has its "mirror" at the other location (see Figure 1).

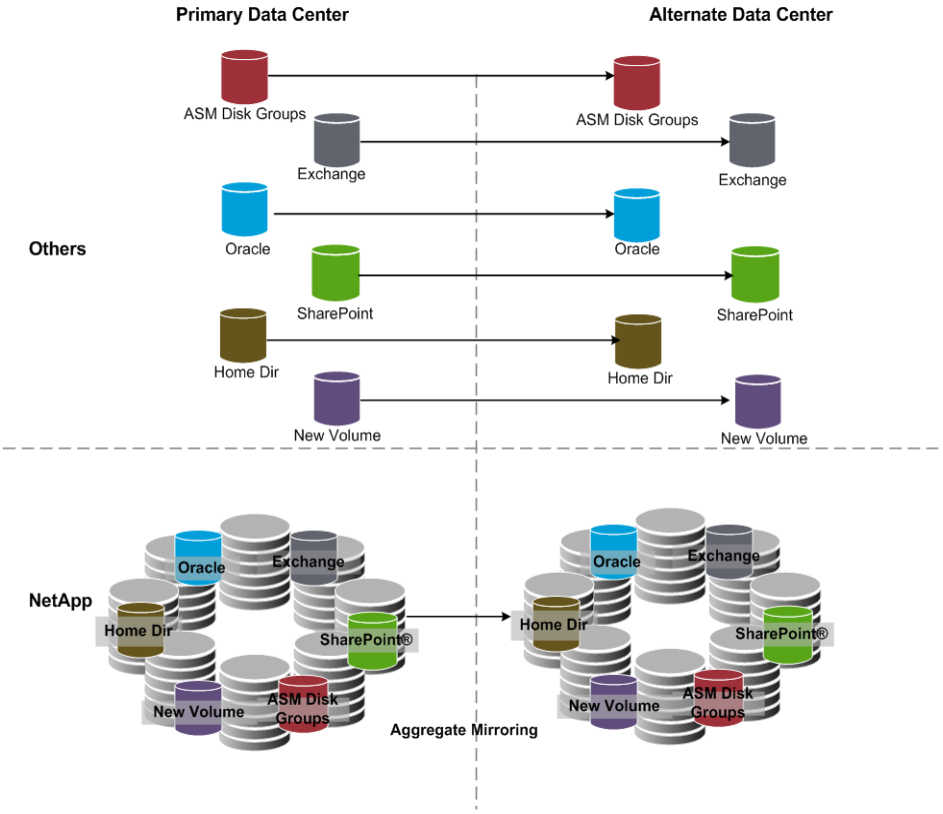

Figure 2) SyncMirror pools and plexes.

Notice in Figure 1 that other protection solutions operate at an individual volume level. This means that to protect all the volumes (which could be hundreds), there must be some type of replication relationship created after each source and destination volume is created. With MetroCluster, all mirroring is performed at an aggregate level so that all volumes are automatically protected with one simple replication relationship.

RECOVERING FROM A DISASTER

As mentioned earlier, recovery from single component failures (controller, disk shelf, and so on) is automatic and transparent to the user. In the event of a complete site failure, such as the loss of the primary data center, the MetroCluster approach manages the failover and completes faster than other approaches. With volume-based synchronous mirroring solutions, in order to bring up data services at the alternate data center after the primary fails, each replication relationships must be "broken" and the volumes placed online. For environments with many volumes this takes time, even with scripts. MetroCluster simplifies this process by allowing the administrator to bring all data services online with a single command. This failover can be further automated using the NetApp MetroCluster Tiebreaker service.

1.3 OVERVIEW OF THE ORACLE TECHNOLOGY USED

ORACLE REAL APPLICATION CLUSTER

Oracle's Real Application Clusters (RAC) option supports the deployment of a single database across a set of clustered servers, providing fault tolerance from hardware failures or planned outages. Oracle RAC running on extended host clusters provide the highest level of Oracle capability in terms of availability, scalability, and low-cost computing. Oracle RAC supports mainstream business applications of all kinds. This includes OLTP, DSS, and the unique Oracle ability to effectively support mixed OLTP/DSS environments. This also includes popular packaged products such as SAP®, PeopleSoft, Siebel, and Oracle E*Business Suite, as well as custom applications.

Oracle RAC provides a single image installation and management. The DBA has a single point of control to install and manage a RAC cluster from the GUI interface or command line.

Oracle Database 10*g* and later includes Oracle Clusterware, a complete, integrated cluster management solution available on all Oracle Database platforms. This clustering functionality includes mechanisms for cluster messaging, locking, failure detection, and recovery. For most platforms, no third-party cluster management software is required. Oracle will, however, continue to support select third-party clustering products on specified platforms.

Oracle Clusterware includes a high-availability API for applications. Oracle Clusterware can be used to monitor, relocate, and restart your applications. With Real Application Clusters, Oracle Clusterware automatically manages all Oracle processes.

ORACLE AUTOMATIC STORAGE MANAGEMENT

Oracle Automatic Storage Management (ASM) is a feature in Oracle Database 10*g* and later that provides the database administrator with a storage management interface that is consistent across all storage and server types.

Figure 2 illustrates ASM using normal redundancy.

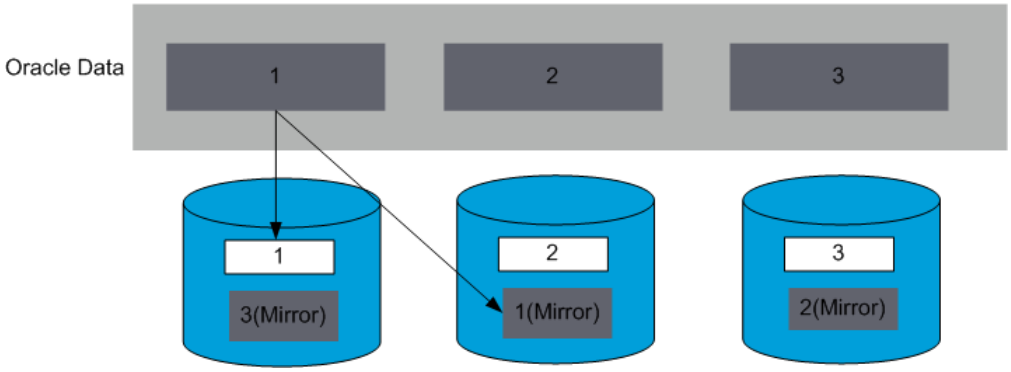

Figure 3) ASM using normal redundancy.

ASM is a volume manager and a file system for Oracle Database files that supports both single-instance Oracle Database and Oracle Real Application Clusters (Oracle RAC) configurations.

In a normal redundancy disk group configuration, by default, Automatic Storage Management uses two-way mirroring for data files and three-way mirroring for control files to increase performance and reliability. NetApp recommends that these ASM failure groups be on different NetApp controllers.

Alternatively, you can use two-way mirroring or no mirroring. A normal redundancy disk group requires a minimum of two failure groups (or two disk devices) for two-way mirroring. The effective disk space in a normal redundancy disk group is half the sum of the disk space in all of its devices. For most installations, Oracle recommends that you use normal redundancy disk groups.

For more information on Oracle ASM, see www.oracle.com/technology/products/database/asm.

2 DISASTER RECOVERY PLANNING CHALLENGES

Disaster recovery is defined as the processes, policies, and procedures related to preparing for recovery or continuation of technical infrastructure critical to an organization after a natural (for example, flood, tornado, volcano eruption, earthquake, or landslide) or human-induced disaster (for example, a threat having an element of human intent, negligence, or error, or involving a failure of a human-made system).

Disaster recovery planning is a subset of a larger process known as business continuity planning and should include planning for resumption of applications, data, hardware, communications (such as networking) and other IT infrastructure. A business continuity plan (BCP) includes planning for non-IT related aspects such as key personnel, facilities, crisis communication, and reputation protection and should refer to the disaster recovery plan (DRP) for IT-related infrastructure recovery/continuity.

Generically, a disaster can be classified as logical or physical disasters, which are addressed with high-availability, recovery processing, and/or disaster recovery processes.

- **Logical Disasters**

  Logical disasters include, but are not limited to data corruption by users or technical infrastructure. Technical infrastructure disasters can result from file system corruption, kernel panics, or even system viruses introduced by employees.

- **Physical Disasters**

  A failure of any of the storage components on site 1 or site 2 that supersedes the resiliency features of a high-availability pair of NetApp controllers not based on MetroCluster or SyncMirror, which would normally result in downtime or data loss.

In certain cases, there are mission-critical applications that should not be stopped even in a disaster. By leveraging Oracle RAC Extended Clusters and NetApp storage technology, it is possible to address those failure scenarios, providing a robust deployment for critical database environments and applications.

3 VALUE PROPOSITION

Mission-critical applications typically need to be implemented using:

RPO = 0

(Recovery point objective = zero): This is a requirement where data loss is unacceptable in the event of any type of any failure.

RTO ~= 0

(Recovery time objective as close to zero as possible): This is a requirement where the time to recovery from a disaster scenario should be as close to 0 minutes as possible.

Oracle (RAC) on Extended Distance Clusters (using ASM normal redundancy) combined with NetApp technology (NetApp fabric MetroCluster + Snapshot technology) meets these RPO requirements.

Neither NetApp MetroCluster nor Oracle ASM normal redundancy, by themselves, will meet all the requirements of truly mission-critical applications. For more information on Oracle ASM mirroring and disk group redundancy, see http://download.oracle.com/docs/cd/E11882_01/server.112/e10500/asmdiskgrps.htm#CHDHDGDI.

Oracle ASM normal redundancy used in conjunction with multiple NetApp resiliency features such as MetroCluster will make sure of zero downtime for all the following failure scenarios:

- Any kind of Oracle Database instance crash
- Any kind of Oracle Database corruption
- A triple-disk failure in a RAID-DP group
- A double FCAL or SAS loop failure
- Shelf failures that cause triple-disk failures in a RAID-DP group
- FAS controller failure
- Switch failure
- Multipathing failure
- Lost writes
- SATA spasm errors
- Complete site failure

Logical Disasters

- Database crashes

Physical Disasters

- Shelf failure
- Controller failure
- Switch failure
- Multipath failure

Upon a complete site disaster, a situation known as "split brain" could occur.

Split Brain

Unless proper fencing technology is employed, any shared-data clustering technology can experience a condition commonly called "split brain."

This is a situation where cluster nodes cannot communicate with each other, but can still write to shared-data resources causing data between sites to fall out of synchronization.

In an Oracle RAC environment, RAC Cluster Synchronization Services use voting disks to avoid split brain situations.

In a two-site scenario where you do not want to give one site priority over the other, you have to choose one of the following:

- Manually determine which is the surviving site
- Have a third voting disk at a third site
- Have monitoring software at a third site with "tie-breaker" rules which determines the truly failed site, for a proper site failover.

In this document we chose to employ a third voting disk accessed via NFS. For more information, see http://www.oracle.com/technology/products/database/clusterware/pdf/grid_infra_thirdvoteonnfs.pdf

In a failure scenario where communication between the two sites is lost, both the Oracle and NetApp solutions alone provide RPOs and RTOs close to zero.

By default NetApp MetroCluster will wait for the storage administrator to manually determining the surviving site.

Without ASM normal redundancy, the Oracle instance would pause or crash upon a site failure.

With ASM normal redundancy and using a third voting disk, the surviving Oracle RAC nodes will be able access the shared-data and a majority of the voting disks.

In the case of the split brain, NetApp Fabric MetroCluster will not take any action. It will wait for a disaster declaration. ASM normal redundancy will provide transparent access to disks in this situation, providing zero downtime for applications.

Based on this, more than 90% of the failure scenarios will be handled by NetApp technologies without any effect to the Oracle Database servers. Only a split brain scenario (not typical) must be handled by Oracle ASM normal redundancy.

In an ASM normal redundancy environment without MetroCluster, if an ASM disk is dropped due to the ASM instance being unable to communicate with it before the DISK_REPAIR_TIME is exceeded, a rebalance (full base-line copy) of the ASM disk must occur. This can be very time consuming due to the locking messages that must be coordinated between ASM instances.

In Oracle ASM 11g, there is a feature called ASM Fast Rebalance. This greatly shortens the time required to do a rebalance by placing the disk group in a restricted mode where only one ASM instance has access to it. However, this prevents any RDMBS instance from accessing the disk group being rebalanced, thus causing downtime.

With MetroCluster, even in a complete site failure, the ASM instances will always have access to all of their ASM disk groups without interruption.

Because of this, none of the ASM instances detect a failure, and the “resyncing” or “rebalancing” occurs at the NetApp storage controller level without any effect to the Oracle Database servers.

4 EXTENDED DISTANCE RAC CLUSTER WITH ASM REDUNDANCY

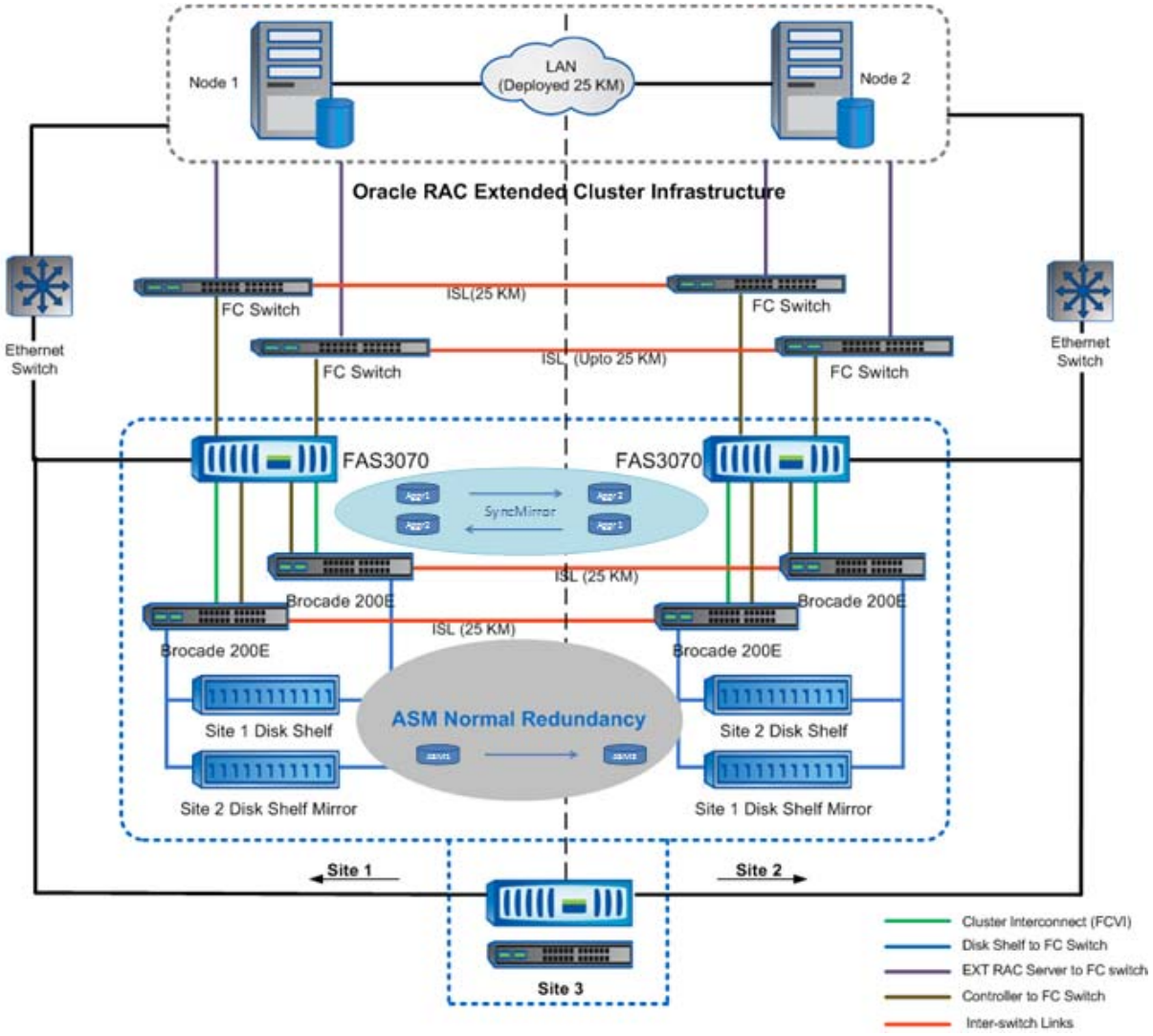

Figure 4) ASM normal redundancy.

Figure 3 illustrates an RAC cluster with ASM redundancy. In this architecture:

1. Oracle Real Application Clusters are distributed across two sites.
2. Oracle ASM is used with the normal redundancy option.
3. Site 1 volumes such as DATA1, LOG1, and ARCH1 are the primary groups in the ASM disk group and their mirror copies are DATA2, LOG2, and ARCH2 volumes in site 2. ASM normal redundancy mirrors data between these two sites.
4. NetApp Fabric MetroCluster using SyncMirror creates a mirrored copy of the aggregate, which contains the volumes where the site 1 aggregate is (sync) mirrored to site 2, and the site 2 aggregate is (sync) mirrored to site 1.

5. Multipathing: In the Oracle RAC extended clusters, multipathing is one of the important configurations, where we are using Asymmetric Logical Unit Access (ALUA) on hosts running RedHat Enterprise Linux®. ALUA is also known as Target Port Group Support (TPGS).

DM-Multipath works with ALUA to find the primary and secondary paths to be used for failover. For more information, see the appendix for `multipath.conf` and ALUA settings used.

5 HIGH-LEVEL ARCHITECTURE AND DEPLOYMENT DETAILS

Figure 4 illustrates the architecture of this configuration. WAN Simulator provides the 25KM distance simulation for the private and public network between Oracle RAC nodes. Latency is 10 ms for 25km distance simulation. “btcppe110” NetApp storage controller provides the third voting disk through NFS for Oracle extended RAC cluster members through public network.

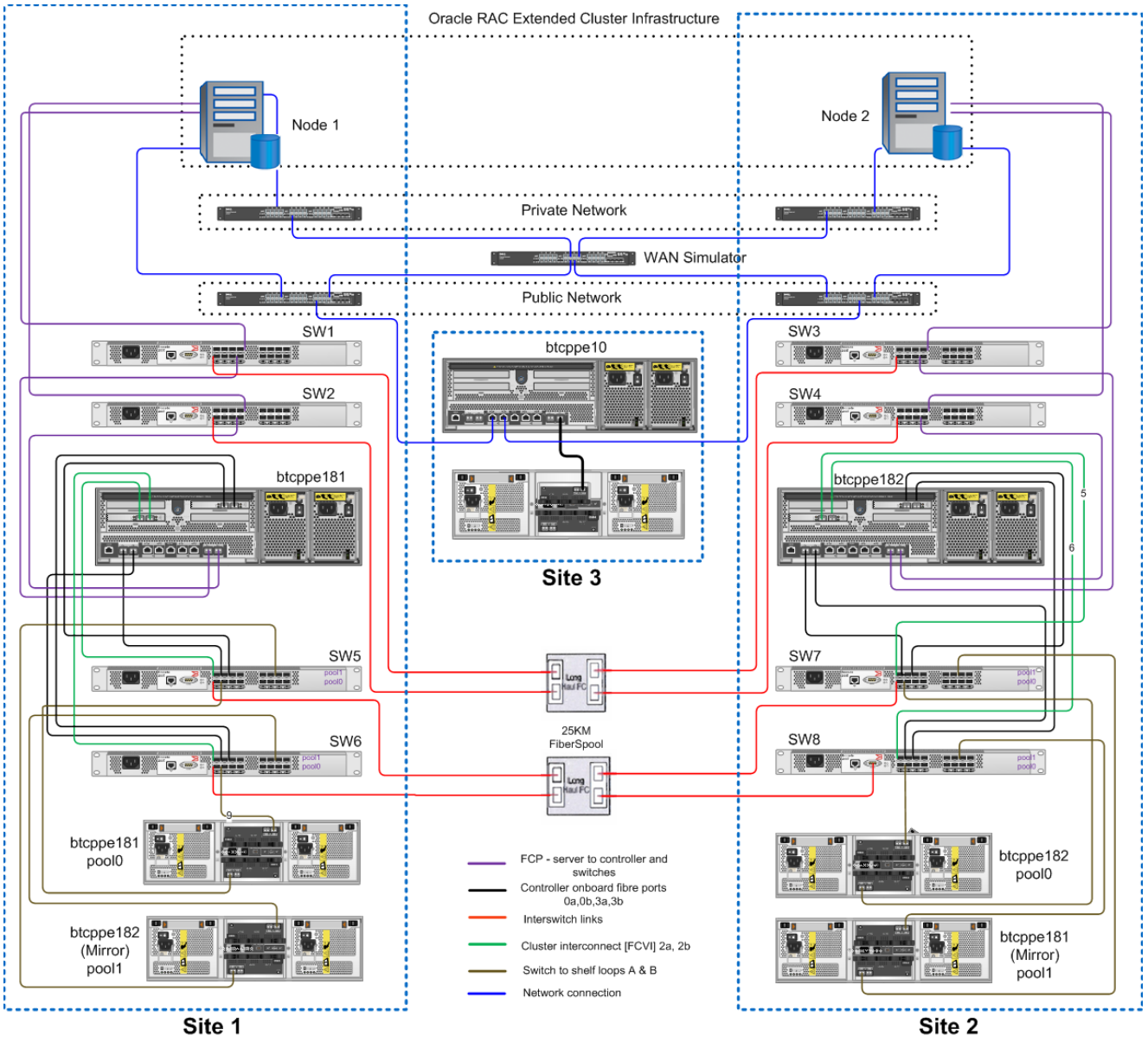

Figure 5) High-level architecture.

The following tables list the deployment details of this architecture.

Table 1) Oracle hosts specification.

Oracle Hosts	
Server	Two IBM servers, one per site
Operating Systems	RedHat Linux 5.3
NIC (Qty and type)	Quad port, 1Gbps
HBA (Qty and type)	Fibre Channel, Dual port per server
Host Attach Kit/version	FCP/5.0
Multipathing	Yes (native)
SAN Switches/Models/Firmware	Eight Brocade 200E, 5.3, switches; Four switches per site [two for NetApp MetroCluster and two for Oracle RAC nodes (frontend)],
Local Storage Used	For Oracle binaries

Table 2) Oracle specification.

Oracle	
Version	Oracle Database 10*g* R2 (10.2.0.4)
ASM?	Yes with normal redundancy
RAC?	Yes
Oracle CRS	Two OCRs files and 3 voting disks (2 raw devices and 1 NFS – third voting)

Table 3) NetApp storage specification.

NetApp Storage	
Model	FAS3070 cluster pair, split across the sites (btcppe181, btcppe182)
Operating Systems	Data ONTAP® 7.3.1.1
NIC (Qty and type)	4 per storage controller
HBA (Qty and type)	Fibre Channel
Back-End switches	Brocade 200e x 8 Numbers
Software	Fabric MetroCluster
	SyncMirror
	FCP, NFS, FlexClone®, and SnapRestore®

Table 4) Data layout.

	Data Layout				
Storage Controller	**Aggregate**	**Volume**	**LUN**	**ASM and OS Mapping**	**ASM Disk Group**
btcppe181	racaggr	oradata1	/vol/oradata1/oradata1.lun	ORCL:ORADATA1	DATA
btcppe182	extrac_B	oradata2	/vol/oradata2/oradata2.lun	ORCL:ORADATA2	
btcppe181	racaggr	oralog1	/vol/oralog1/oralog1.lun	ORCL:ORALOG1	LOG
btcppe182	extrac_B	oradata2	/vol/oralog2/oralog2.lun	ORCL:ORALOG2	
btcppe181	racaggr	oraarch1	/vol/oraarch1/oraarch1.lun	ORCL:ORAARCH1	ARCH
btcppe182	extrac_B	oraarch2	/vol/oraarch2/oraarch2.lun	ORCL:ORAARCH2	
btcppe181	racaggr	ocr1	/vol/ocr1/orc1.lun	/dev/raw/raw1	-
btcppe182	extrac_B	ocr2	/vol/ocr2/orc2.lun	/dev/raw/raw2	-
btcppe181	racaggr	votedisk1	/vol/votedisk1/votedisk1.lun	/dev/raw/raw4	-
btcppe182	extrac_B	votedisk2	/vol/votedisk2/votedisk2.lun	/dev/raw/raw5	-
btcppe110	aggr1	votedisk3	/vol/votedisk3	/votedisk3/votedisk3.crs	-

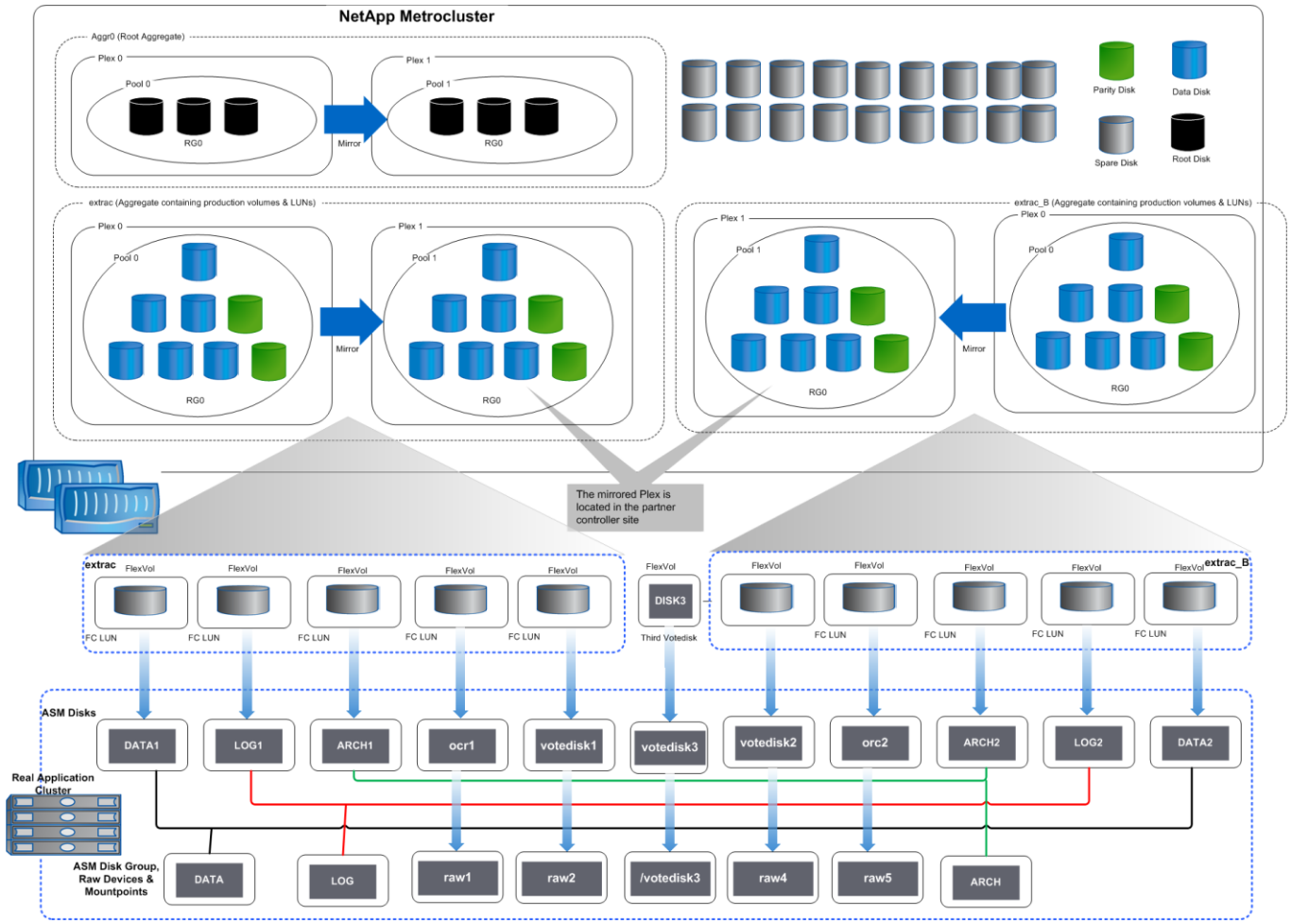

Figure 6) Physical and logical storage configuration of NetApp FAS controllers in MetroCluster setup.

6 FAILURE SCENARIOS AND RECOVERY

This section describes the test scenarios that were executed after the successful implementation of the environment described in this document. Test scenarios include various component failures. Unless stated otherwise, before the execution of each test, the environment was reset to the "normal" running state with Oracle RAC running on the servers with generating disk activity.

For each of the following test scenarios, both the volumes (LUN and NFS) are verified.

Table 5) Failure scenarios.

Logical Failures	
Failure Scenario	**Description**
1	Database restore (using SnapRestore)
2	Database clones
Physical Failures	
Failure Scenario	**Description**
3	Loss of an Oracle Node
4	Loss of an Oracle Host HBA
5	Loss of Disk(s)
6	Loss of Complete Disk Shelf
7	Loss of NetApp Storage Controller
8	Loss of Fibre Channel Switch
9	Loss of One ISL
10	Loss of an Entire site
11	Restore an Entire site (Recover from Disaster)

6.1 FAILURE SCENARIO 1: DATABASE RESTORE USING SNAPRESTORE

Table 6) Database restore using SnapRestore.

Description	Restore the site 1 database using site 2 Snapshot copy.
Task(s)	1. Create a Snapshot copy based on the RTO and RPO required. 2. Corrupt the database from the database volume. 3. Make a full restore or partial restore from the ASM disk using the Snapshot copy.
Expected Results	The database is restored and functions properly.
Results	The database is restored to its original state and functions properly.
Time to Restore?	Seconds

6.2 FAILURE SCENARIO 2: DATABASE CLONES

Table 7) Database clones.

Description	Create a dev/test environment production environment data.
Task	1. Create a backup of the production database. 2. Protect the backup using NetApp SnapMirror® or SnapVault®. 3. Transfer the backup to site 2. 4. At site 2, create a clone of the production database using the transferred Snapshot copy. You can customize the parameter as per your requirement. 5. Check the functionality and space utilization of the cloned database.
Expected Results	The dev/test environment created and cloned from the data set.
Results	A clone is created at site 2.
Disruptive?	No

6.3 FAILURE SCENARIO 3: LOSS OF ORACLE NODE

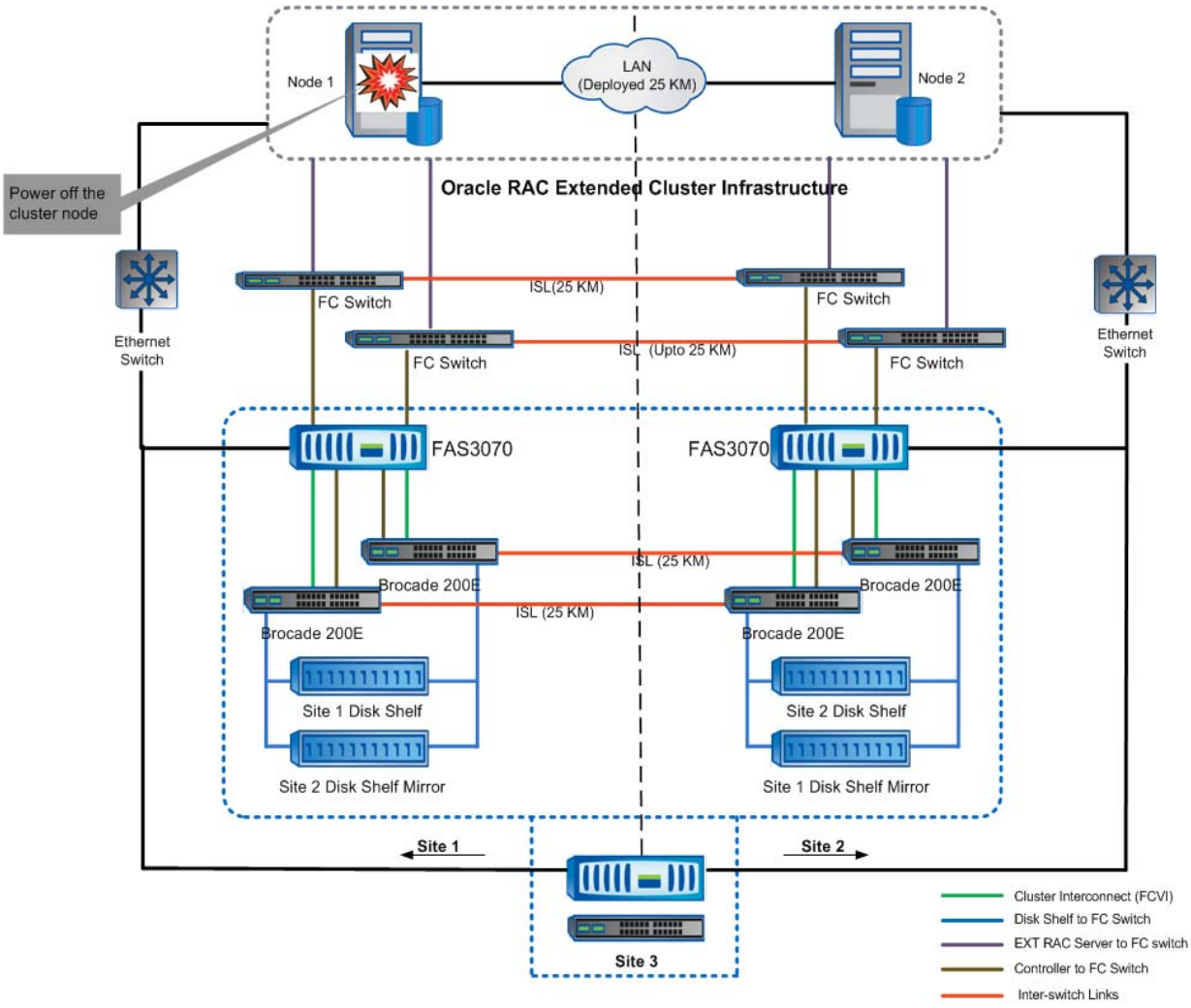

Figure 7) Loss of an Oracle node.

Table 8) Loss of an Oracle node.

Description	No single point of failure should exist in the solution. Therefore, the loss of one of the Oracle hosts in the cluster is tested. This test is accomplished by halting an appropriate host in the cluster while running a query.
Task	On an Oracle client running a simple query (sqlplus), power off one of the cluster members.
Expected Results	The query and load of the failed node should move to the active node.
Results	The actual results are consistent with the expected results.
Disruptive?	No

6.4 FAILURE SCENARIO 4: LOSS OF AN ORACLE HOST HBA

Table 9) Loss of an Oracle host HBA.

Description	No single point of failure should exist in the solution. Therefore, the loss of one of the host HBAs on one of the Oracle nodes was tested. This test is accomplished by disconnecting the Fibre Channel cable while running a query.
Task	On an Oracle client running a simple query (sqlplus), remove one of the HBA cables from the dual HBA card.
Expected Results	Multipathing will address the workload with no interruption. The Oracle client remains unaffected by the HBA failure.
Results	The actual results are consistent with the expected results.
Disruptive?	No

6.5 FAILURE SCENARIO 5: LOSS OF DISK(S)

Table 10) Loss of disk(s).

Description	No single point of failure should exist in the solution. Therefore, the loss of a single disk was tested. This test is accomplished by removing a drive from the aggregate containing the data files while running a query.
Task	On an Oracle client running a simple query (sqlplus), remove the SW6:5.24 disk from the shelf (id: 1), which is a member of the oradata volume.
Expected Results	Oracle Database is able to serve to the Oracle client uninterrupted, where the transactions (that is, select, insert) continues without any issues.
Results	The actual results are consistent with the expected results.
Disruptive?	No

6.6 FAILURE SCENARIO 6: LOSS OF COMPLETE DISK SHELF

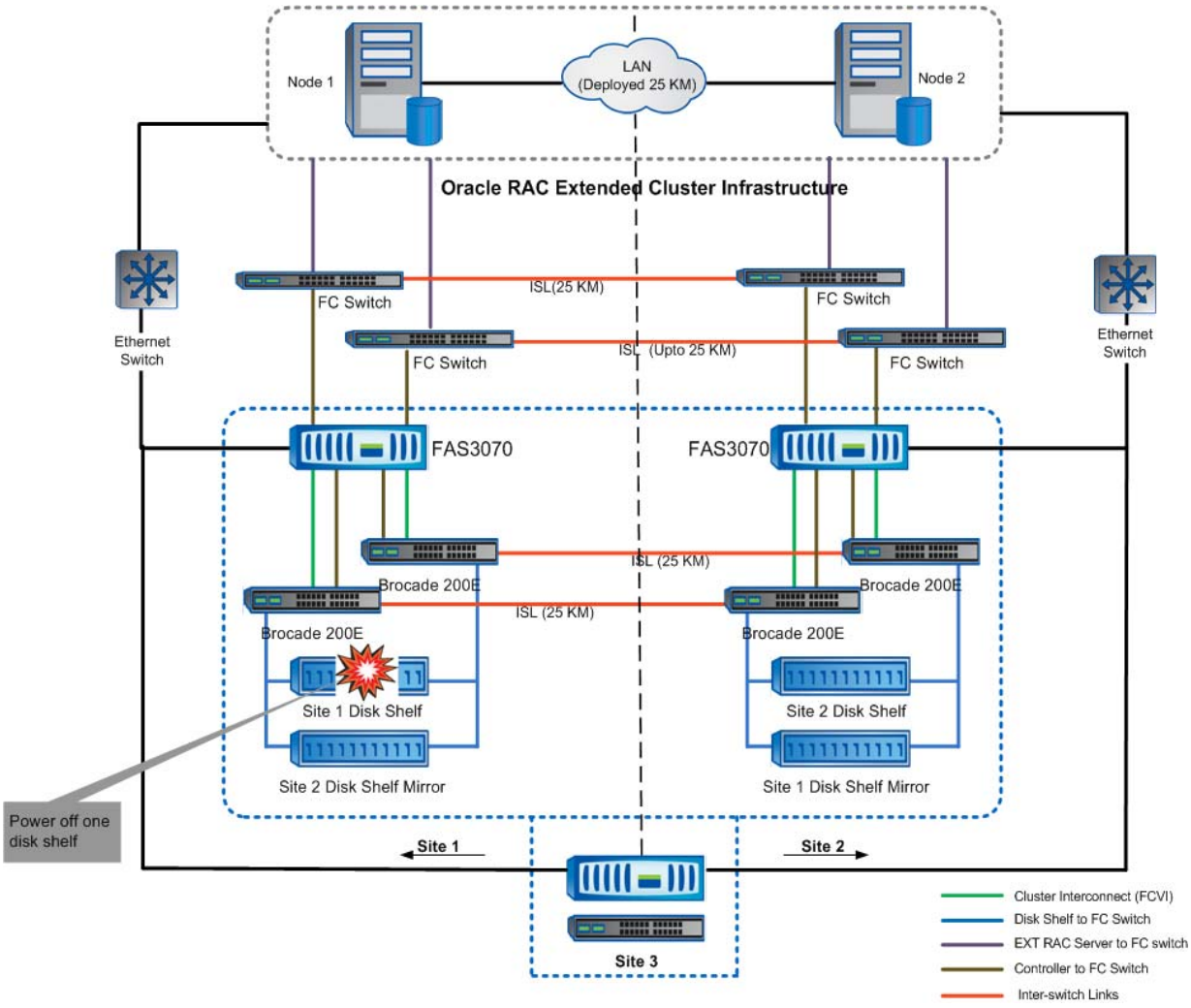

Figure 8) Loss of complete disk shelf.

Table 11) Loss of complete disk shelf.

Description	No single point of failure should exist in the solution. Therefore, the loss of an entire shelf was tested. This test is accomplished by turning off both power supplies in the shelf while running the deduplication process.
Task	1. On an Oracle client running a simple query (sqlplus), power off the SITE 1 Pool0 shelf. 2. Observe the results, and then power it back on.
Expected Results	Relevant disks will go offline. The plex will be broken, but service to clients (availability and performance) will remain unaffected. When the shelf is powered on, the disks will be detected and a resync of the plexes will occur without any manual intervention.
Results	The actual results are consistent with the expected results.
Disruptive?	No

6.7 FAILURE SCENARIO 7: LOSS OF STORAGE CONTROLLER: FAILOVER AND GIVEBACK

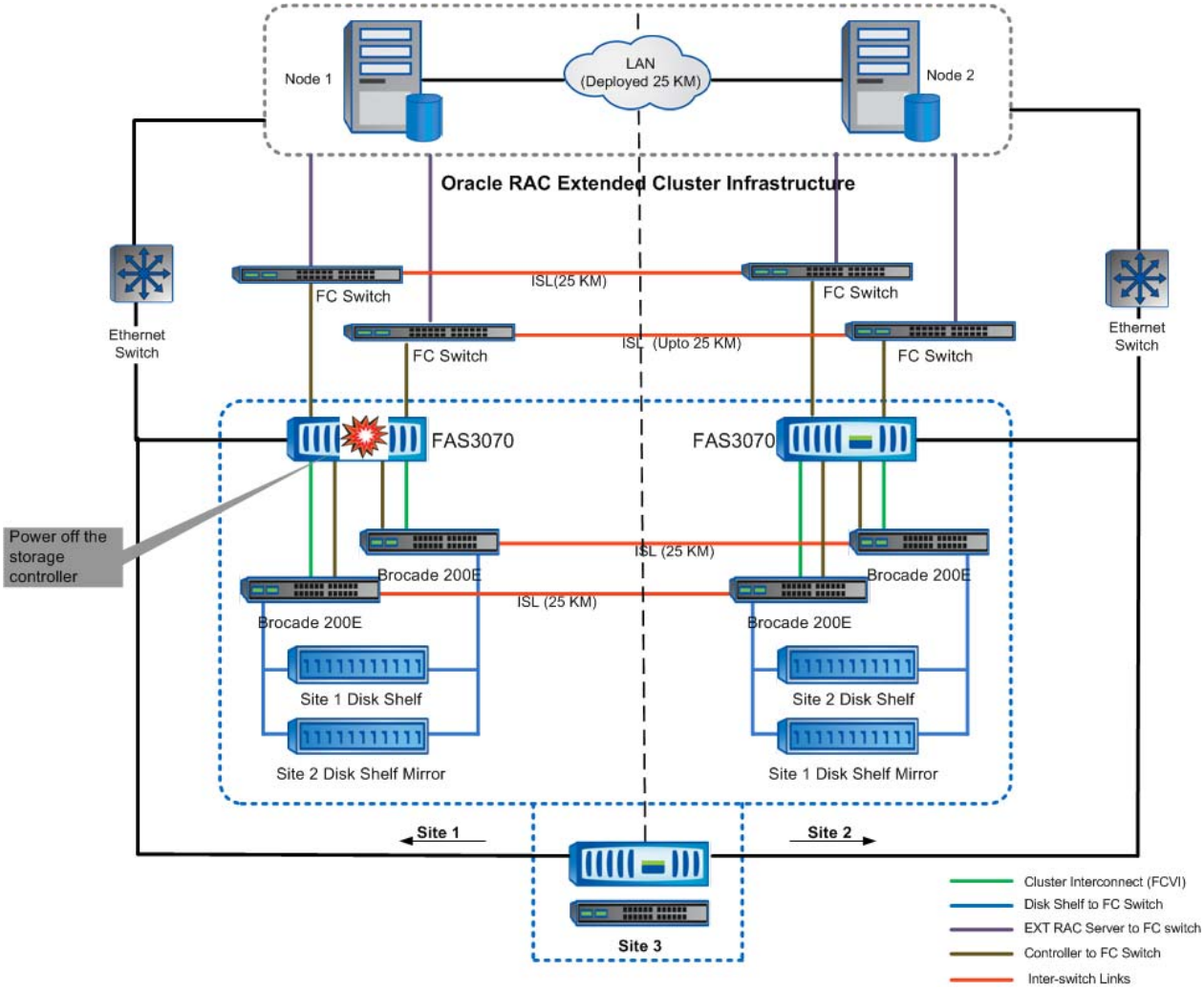

Figure 9) Loss of storage controller.

Table 12) Loss of a storage controller.

Description	No single point of failure should exist in the solution. Therefore, the loss of one of the controllers is tested.
Task	On an Oracle client running a simple query (sqlplus), power off the site 1 controller by turning off both power supplies.
Expected Results	The site 2 controller will take over. The Oracle client running a simple query (sqlplus) will continue to access the Oracle Database RAC without any errors.
Results	The actual results are consistent with the expected results. btcppe182 takes over the btcppe181 controller and is able to serve the clients without any errors or interruption.
Disruptive?	No

6.8 FAILURE SCENARIO 8: LOSS OF FIBRE CHANNEL SWITCH

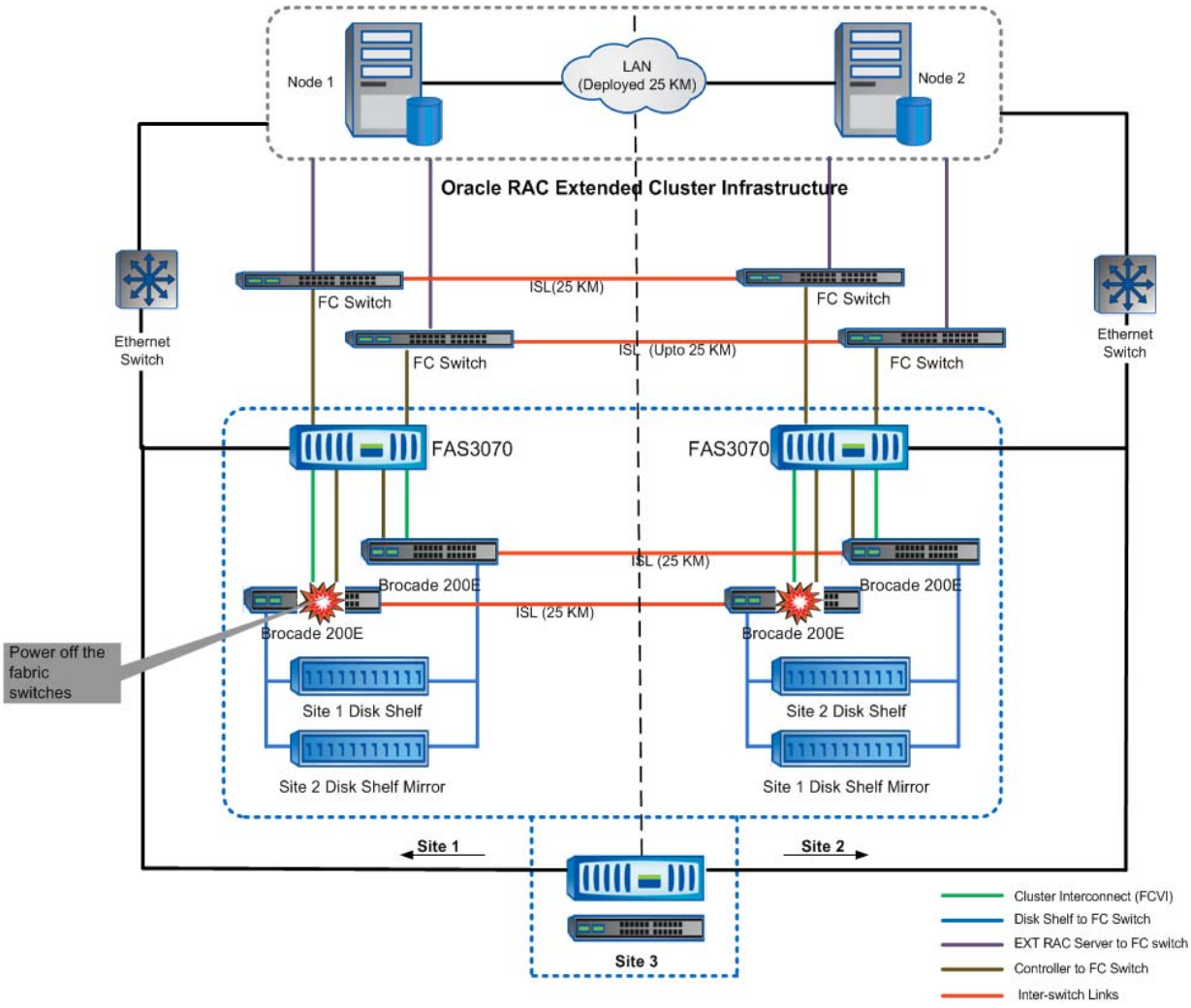

Figure 10) Loss of fibre channel switch.

Table 13) Loss of fibre channel switch.

Description	No single point of failure should exist in the solution. Therefore, the loss of an entire Brocade switch was tested. This test was accomplished by simply removing the power cord from the switch while a load is applied.
Task	1. On an Oracle client running a simple query (sqlplus), power off the Fibre Channel switches in site 1-SW6 and site 2-SW8. 2. Observe the results, and then power it back on.
Expected Results	The controller should display a message indicating that some disks are connected to only one switch and that one of the clusters interconnects is down, but service to clients (availability and performance) is unaffected. When power is restored and the switch completes its boot process, the controller should display messages to indicate that the disks are now connected to two switches and that the second cluster interconnects is again active.
Results	Switch multipathing works and Oracle clients able to complete their transactions without I/O errors.

```
btcppe181> aggr status racaggr -r
Aggregate racaggr (online, raid_dp, mirrored) (block checksums)
Plex /racaggr/plex4 (online, normal, active, pool1)
RAID group /racaggr/plex4/rg0 (normal)

RAID Disk Device          HA  SHELF BAY CHAN Pool Type  RPM  Used (MB/blks)    Phys (MB/blks)
--------- ------          ------------ ---- ---- ---- ----- --------------    --------------
dparity   SW7:10.77       0a    4   13  FC:B   1  FCAL 10000 272000/557056000  280104/573653840
parity    SW7:10.72       3d    4   8   FC:B   1  FCAL 10000 272000/557056000  280104/573653840
data      SW8:10.65       0b    4   1   FC:A   1  FCAL 10000 272000/557056000  280104/573653840
data      SW7:10.66       0a    4   2   FC:B   1  FCAL 10000 272000/557056000  280104/573653840
data      SW7:10.67       0a    4   3   FC:B   1  FCAL 10000 272000/557056000  280104/573653840
data      SW8:10.68       0b    4   4   FC:A   1  FCAL 10000 272000/557056000  280104/573653840
data      SW8:10.69       0b    4   5   FC:A   1  FCAL 10000 272000/557056000  280104/573653840
data      SW7:10.73       3d    4   9   FC:B   1  FCAL 10000 272000/557056000  280104/573653840

Plex /racaggr/plex15 (online, normal, active, pool0)
RAID group /racaggr/plex15/rg0 (normal)

RAID Disk Device          HA  SHELF BAY CHAN Pool Type  RPM  Used (MB/blks)    Phys (MB/blks)
--------- ------          ------------ ---- ---- ---- ----- --------------    --------------
dparity   SW6:9.112       3c    7   0   FC:B   0  FCAL 10000 272000/557056000  280104/573653840
parity    SW5:5.21        3d    1   5   FC:B   0  FCAL 10000 272000/557056000  280104/573653840
data      SW6:5.29        3c    1   13  FC:A   0  FCAL 10000 272000/557056000  280104/573653840
data      SW5:5.25        3d    1   9   FC:B   0  FCAL 10000 272000/557056000  280104/573653840
data      SW6:5.26        3c    1   10  FC:A   0  FCAL 10000 272000/557056000  280104/573653840
data      SW6:5.27        3c    1   11  FC:A   0  FCAL 10000 272000/557056000  280104/573653840
data      SW5:5.19        0a    1   3   FC:B   0  FCAL 10000 272000/557056000  280104/573653840
data      SW6:9.118       0b    7   6   FC:B   0  FCAL 10000 272000/557056000  280104/573653840

btcppe181>
```

Disruptive?	No

6.9 FAILURE SCENARIO 9: LOSS OF ONE ISL

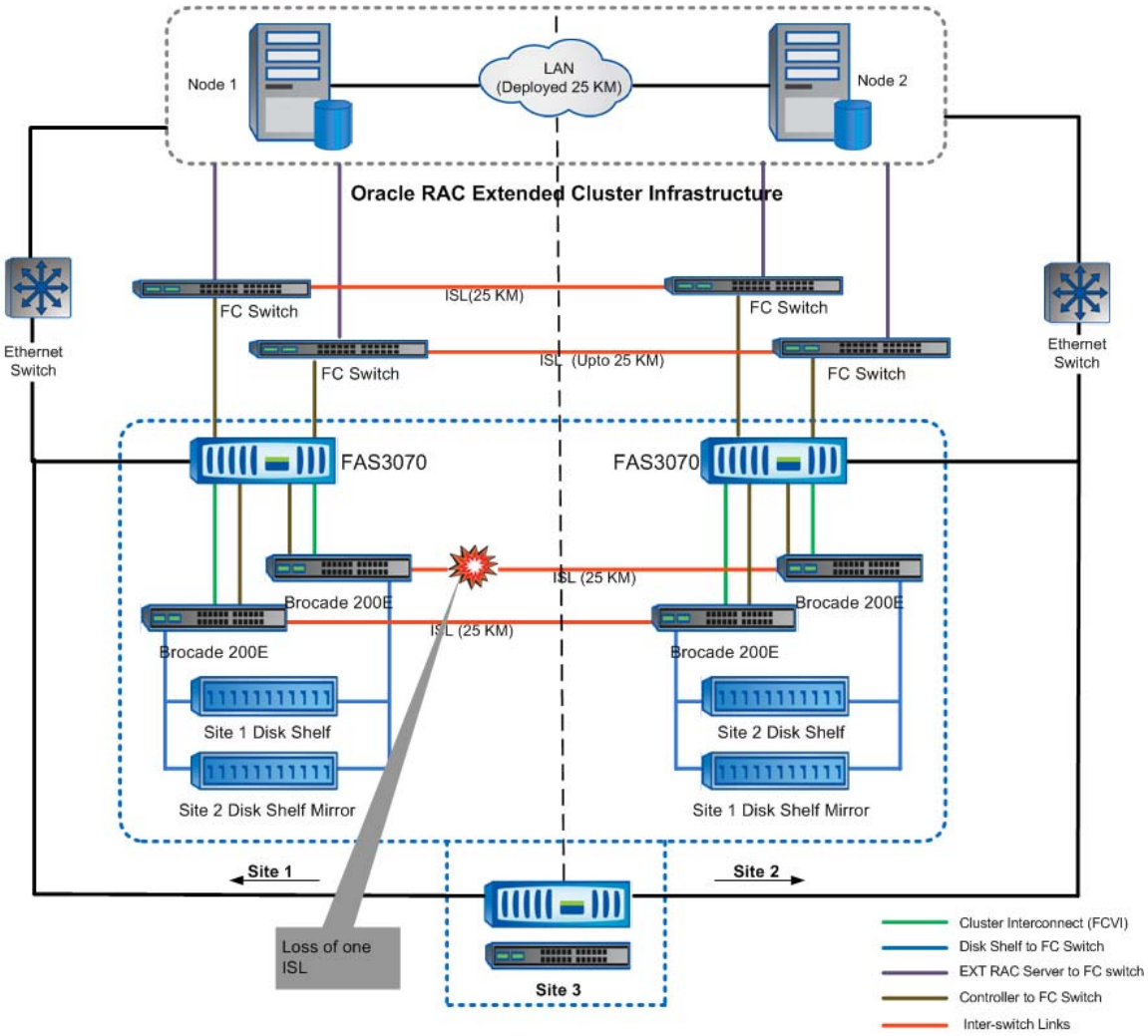

Figure 11) Loss of one ISL.

Table 14) Loss of one ISL.

Description	No single point of failure should exist in the solution. Therefore, the loss of one of the interswitch links (ISLs) is tested. This test was accomplished by removing the Fibre Channel cable between two switches while a load is applied.
Task	On an Oracle client running a simple query (sqlplus), remove the Fibre Channel cable between SITE1-SW5 and SITE2-SW7.
Expected Results	The controller displays the messages that some disks are connected to only one switch and that one of the cluster interconnects is down, but service to clients (availability and performance) is unaffected. When the ISL is reconnected, the controller displays a message that the disks are now connected to two switches and that the second cluster interconnect is again active.

Results	1. Disconnect the one ISL cable from SW5 fabric switch:
	`btcppe181> Fri Nov 20 19:12:30 IST [btcppe181: cf.nm.nicTransitionDown:warning]: Cluster Interconnect link 0 is DOWN`
	`btcppe181> cf status`
	`Cluster enabled, btcppe182 is up.`
	`VIA Interconnect is up (link 0 down, link 1 up).`
	2. Then plug the cable back:
	`btcppe182> cf status`
	`Cluster enabled, btcppe181 is up.`
	`VIA Interconnect is up (link 0 down, link 1 up).`
	`btcppe182> Fri Nov 20 19:19:58 IST [btcppe182: cf.nm.nicTransitionUp:info]: Interconnect link 0 is UP`
	`btcppe182> cf status`
	`Cluster enabled, btcppe181 is up.`

Disruptive?	No

6.10 LOSS OF ONE LINK IN ONE DISK LOOP

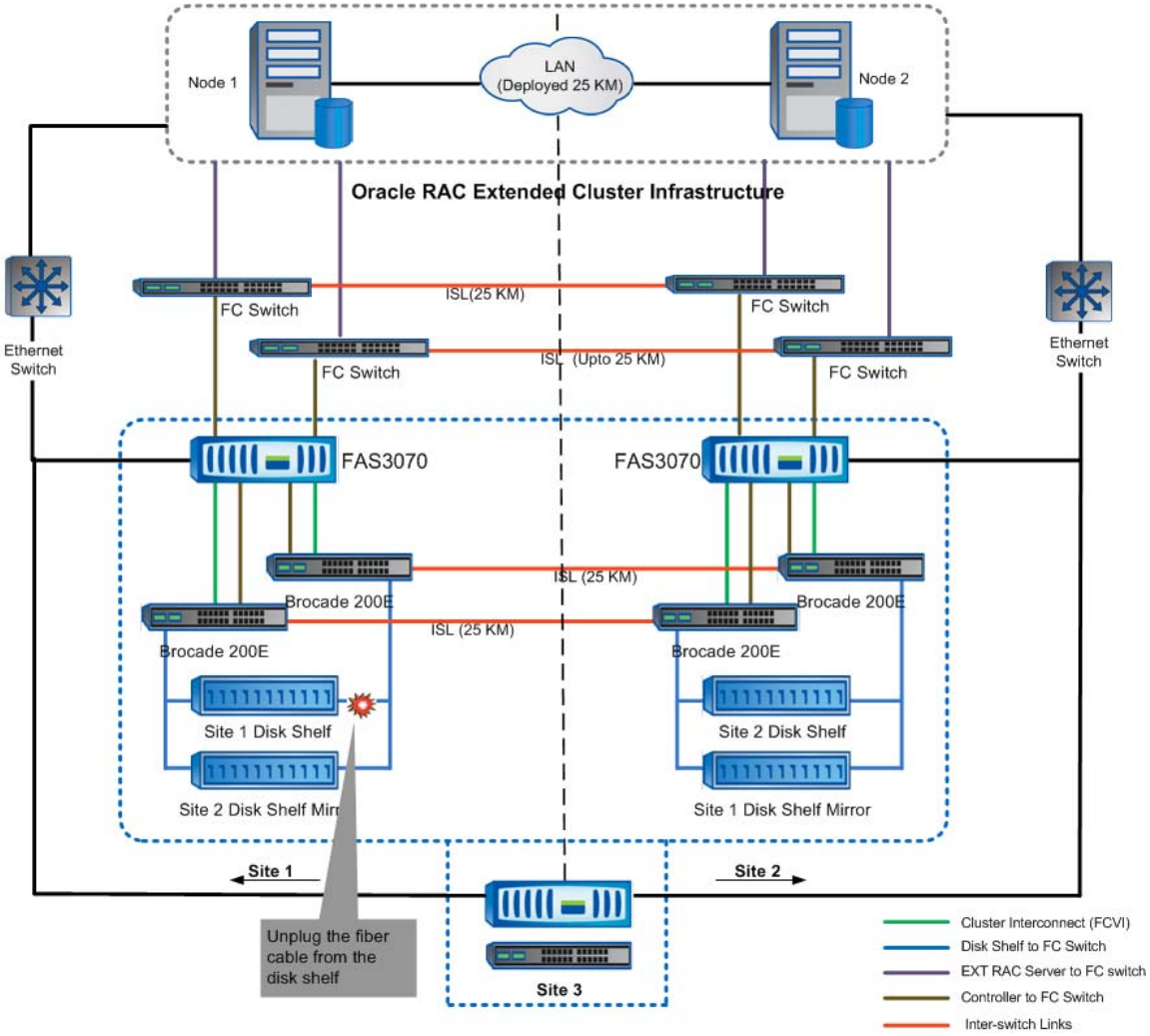

Figure 12) Loss of one link in one disk loop.

Table 15) Loss of one link in one disk loop.

Description	No single point of failure should exist in the solution. Therefore, the loss of one of the paths to shelf was tested.
Task	1. On an Oracle client running a simple query (sqlplus), disconnect the Fibre Channel cable in one of the disk shelves. 2. Observe the results, and then reconnect the Fibre Channel cable.
Expected Results	No disruption to data availability: • The controller displays the message that some disks are connected to only one switch. • No change detected in the Oracle Database server level, and Oracle clients run without any interruption. • When the Fibre Channel is reconnected, the controller displays the messages that disks are now connected to two switches.
Results	The actual results are consistent with the expected results.
Disruptive?	No

6.11 FAILURE SCENARIO 10: LOSS OF AN ENTIRE SITE

In case of complete site disaster, all physical components of the extended RAC and fabric MetroCluster such as interswitch links (ISL), Oracle node, storage controller, fabric switches for server and storage, pool0 and pool1 disk shelves of one site all become unavailable simultaneously, a manual force failover of the NetApp MetroCluster needs to be performed to declare the disaster due a split brain scenario.

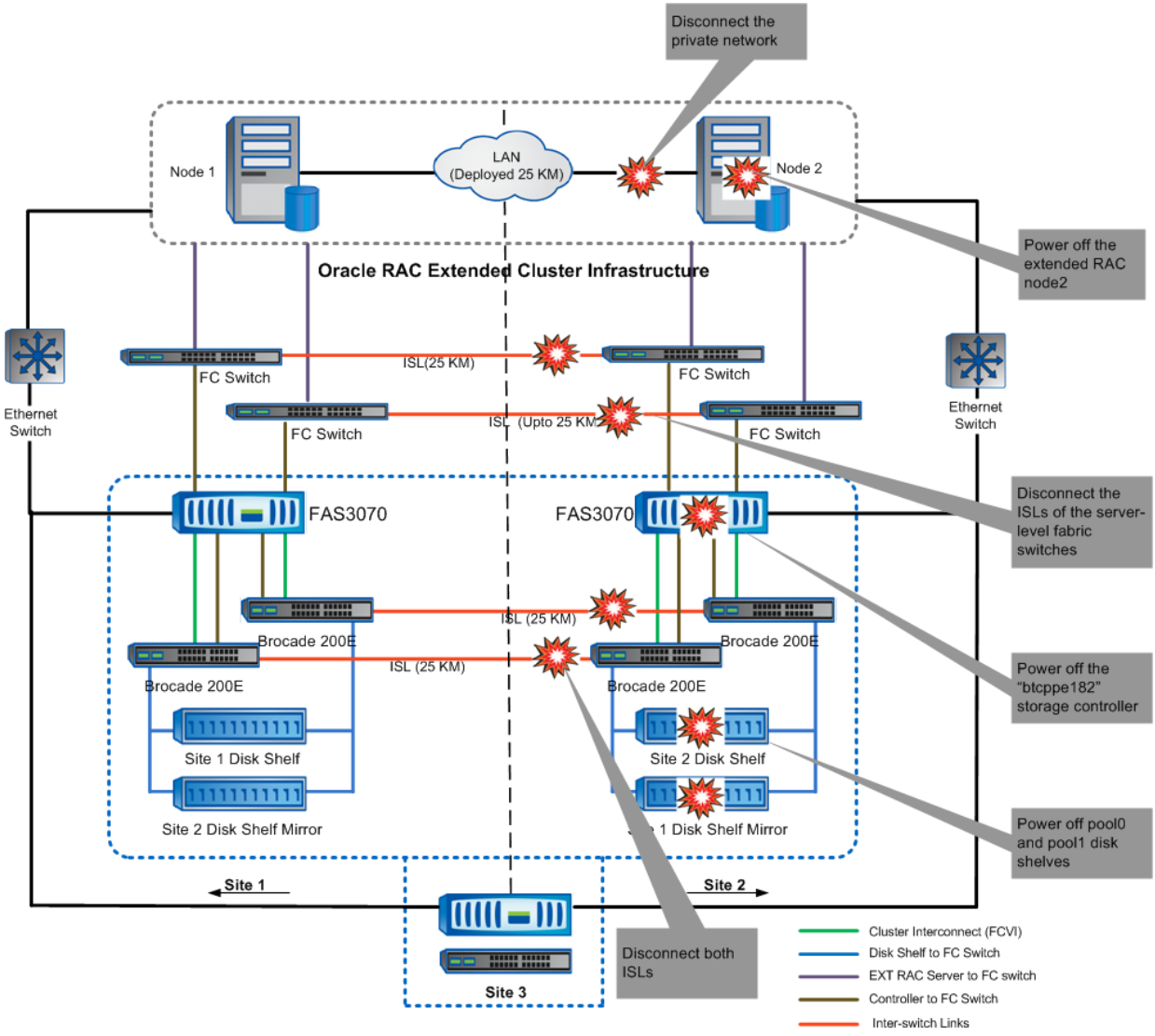

Figure 13) Loss of an entire site.

One way to simulate a real-world site disaster in the lab is to interrupt following components of the lab setup in the order given, in rapid succession (so that the partner site component is unable to automatically detect any failure) as follows:

1. Disconnect the ISLs from the SW7 and SW8 switches.
2. Disconnect both ISLs from the SW3 and SW4 switches.
3. Power off all the disk shelves in site 2.
4. Disconnect the private link from the extended Oracle RAC server (node 2).
5. Power off the NetApp storage controller (btcppe182).
6. Power off the extended Oracle Database servers (node2).

To simulate the loss of a site, do as follows:

1. Check the ASM disks and their status using custom `asm.sql` query as follows:

```
[oracle@nodel ~]$ cat asm.sql
set lines 500
prompt

prompt ASM disks


prompt =========

col "Group"             form 999
col "Disk"              form 999
col "Header"            form a9
col "Mode"              form a8
col "Failure Group"     form a10
col "Path"              form a19

select  group_number "Group"
,       disk_number "Disk"
,       header_status "Header"
,       mode_status     "Mode"
,       state           "State"
,       total_mb        "Total MB"
,       free_mb         "Free MB"
,       name            "Disk Name"
,       failgroup       "Failure Group"
,       path            "Path"
from    v$asm_disk
order by group_number, disk_number
/


[oracle@nodel ~]$
```

```
SQL> @asm.sql

ASM disks
=========

Group Disk Header    Mode     State      Total MB    Free MB Disk Name                      Failure Gr Path
----- ---- --------- -------- -------- ---------- ---------- ------------------------------ ---------- ----------------
---
    0    0 FOREIGN   ONLINE   NORMAL         3065          0                                           /dev/raw/raw5
    0    1 FOREIGN   ONLINE   NORMAL         3065          0                                           /dev/raw/raw4
    0    2 FOREIGN   ONLINE   NORMAL         3065          0                                           /dev/raw/raw2
    0    3 FOREIGN   ONLINE   NORMAL         3065          0                                           /dev/raw/raw1
    1    0 MEMBER    ONLINE   NORMAL       158724     146454 ARCH1                          ARCH1      ORCL:ARCH1
    1    1 MEMBER    ONLINE   NORMAL       158724     146454 ARCH2                          ARCH2      ORCL:ARCH2
    2    1 MEMBER    ONLINE   NORMAL       389157     350723 DATA2                          DATA2      ORCL:DATA2
    2    2 MEMBER    ONLINE   NORMAL       389157     350723 DATA181                        DATA1      ORCL:DATA1
    3    1 MEMBER    ONLINE   NORMAL        76796      76458 LOG2                           LOG2       ORCL:LOG2
    3    2 MEMBER    ONLINE   NORMAL        76796      76458 LOG181                         LOG1       ORCL:LOG1

10 rows selected.

SQL>
```

   Note that the members `ARCH1`, `ARCH2`, `DATA1`, `DATA2`, `LOG1`, and `LOG3` show the state as `ONLINE` and `NORMAL`.

2. Verify the site-level disaster.
3. Disconnect the ISLs from the storage-side fabric switches site 2 (SW7 and SW8).

   The site 1 storage controller displays the message that the site-level disaster occurred in site 2.

The storage controller in surviving site 1 displays that its partner node is down. As mentioned previously, during an entire site failure, an automated cluster takeover will not be initiated by the surviving storage controller node.

```
btcppe181> cf status
btcppe182 may be down, takeover disabled because of reason (partner mailbox disks not accessible or invalid)
btcppe181 has disabled takeover by btcppe182 (unsynchronized log)
VIA Interconnect is down (link 0 down, link 1 down).
The DR partner site might be dead.
 To take it over, power it down or isolate it as described in the Data Protection Guide manual, and then use
cf forcetakeover -d.
btcppe181>
```

Note: You do not have to perform a manual force takeover when using ASM normal redundancy with NetApp fabric MetroCluster.

4. Check the ASM disk status after the site crash simulation as follows:

```
SQL> @asm.sql

ASM disks
*********

Group Disk Header    Mode     State      Total MB    Free MB Disk Name                      Failure Gr Path
----- ---- --------- -------- -------- ---------- ---------- ------------------------------ ---------- ----------
-
    0    0 UNKNOWN   ONLINE   NORMAL         3065          0                                            /dev/raw/raw5
    0    1 FOREIGN   ONLINE   NORMAL         3065          0                                            /dev/raw/raw4
    0    2 UNKNOWN   ONLINE   NORMAL         3065          0                                            /dev/raw/raw2
    0    3 FOREIGN   ONLINE   NORMAL         3065          0                                            /dev/raw/raw1
    0    4 MEMBER    ONLINE   NORMAL       158724          0                                            ORCL:ARCH1
    0    5 UNKNOWN   ONLINE   NORMAL       158724          0                                            ORCL:ARCH2
    0    6 UNKNOWN   ONLINE   NORMAL        76796          0                                            ORCL:LOG2
    0    7 UNKNOWN   ONLINE   NORMAL       389157          0                                            ORCL:DATA2
    2    1 UNKNOWN   OFFLINE  HUNG         389157     350723 DATA2                          DATA2
    2    2 MEMBER    ONLINE   NORMAL       389157     350723 DATA181                        DATA1      ORCL:DATA1
    3    1 UNKNOWN   OFFLINE  HUNG          76796      76458 LOG2                           LOG2

Group Disk Header    Mode     State      Total MB    Free MB Disk Name                      Failure Gr Path
----- ---- --------- -------- -------- ---------- ---------- ------------------------------ ---------- ----------
-
    3    2 MEMBER    ONLINE   NORMAL        76796      76458 LOG181                         LOG1       ORCL:LOG1

12 rows selected.

SQL>
```

The state of the DATA2, LOG2, and ARCH2 are now shown as UNKNOWN or HUNG.

Despite this, the Oracle client is able to access the database and instances through ASM mirror copies of those disks such as DATA1, LOG1, ARCH1 and their status are NORMAL.

In this solution during site failure there is **no single point of failure**, the Oracle clients continue to access the database with the help of “ASM normal redundancy” mirror copy.

Therefore, combining Oracle ASM with NetApp fabric MetroCluster we can provide a zero-downtime disaster recovery resolution.

PERFORMING A GIVEBACK WHEN THE FAILED SITE IS ONLINE

To perform giveback once the failed site is back online, do as follows:

1. Switch on the disk shelves connected to the site 2.
2. Reconnect the ISL between sites.

   The storage controller for site 1 (btcppe181) will be able to access the disk shelves at site 1. Then the disks will automatically sync to each other using SyncMirror.
3. Power on the Node2, which is in the site 2.
4. Power on the site 2 storage controller (btcppe182).
5. Make sure that the aggregates for site 2 are mirrored before giveback, which can be checked from the partner storage controller using the **`partner`** and **`aggr status`** commands. If this is not the case, do the following:

   a. Suppose if the aggr status is "out-of-date," offline the mirrored aggregate and reinitialize the mirror for the source aggregate using **`aggr mirror racaggr –v racaggr(1).`**

   b. Continue the above step for all the aggregate before "cf giveback"; otherwise storage controller cannot come back if the root volume aggregate pool0 not recovered properly.
6. Check the cluster status and verify that a giveback is possible and all mirror resyncs are completed before performing **`cf giveback.`**

   a. Site 2 storage controller (btcppe182) waiting for giveback:

```
NetApp Release 7.3.1.1: Mon Apr 20 22:45:56 PDT 2009
Copyright (c) 1992-2008 NetApp.
Starting boot on Fri Nov 27 09:01:24 GMT 2009
Fri Nov 27 09:02:55 GMT [scsi.path.excessiveErrors:error]: Excessive errors encountered by adapter 0a on disk
 device SW5:10.73.
Fri Nov 27 09:02:55 GMT [diskown.isEnabled:info]: software ownership has been enabled for this system
Fri Nov 27 09:02:57 GMT [monitor.chassisPower.degraded:notice]: Chassis power is degraded: sensor PSU2 AC IN
Waiting for giveback
Reservation confFrlict found on thi is node's disks!No
27 09:02:59 GMT [ses.giveback.wait:info]: Enclosure Services will be unavailable while waiting for giveback.

Press Ctrl-C for Maintenance menu to release disks.
Waiting for giveback
Waiting for giveback
```

   b. Check the "cf status" in site 1 storage controller (btcppe181):

```
btcppe181(takeover)>
btcppe181(takeover)> cf status
btcppe181 has taken over btcppe182.
btcppe182 is ready for giveback.
btcppe181(takeover)>
```

7. Issue the **`cf giveback`** command in site 1 storage controller (btcppe181):

```
btcppe181(takeover)>
btcppe181(takeover)> cf giveback
please make sure you have rejoined your aggregates before giveback.
Do you wish to continue [y/n] ?? y
btcppe181(takeover)> Fri Nov 27 09:05:07 GMT [btcppe181 (takeover): cf.misc.operatorGiveback:info]: Cluster m
onitor: giveback initiated by operator
Fri Nov 27 09:05:07 GMT [btcppe181: cf.fm.givebackStarted:warning]: Cluster monitor: giveback started
Fri Nov 27 09:05:08 GMT [btcppe182/btcppe181: iscsi.service.shutdown:info]: iSCSI service shutdown
Fri Nov 27 09:05:08 GMT [btcppe182/btcppe181: fcp.service.shutdown:info]: FCP service shutdown
```

8. In the site 2 storage controller (btcppe182), check that the DATA2, LOG2, and ARCH2 disks show their states as `MEMBER`, `ONLINE,` and `NORMAL`.

```
[oracle@node2 ~]$ sqlplus / as sysdba

SQL*Plus: Release 10.2.0.4.0 - Production on Mon Nov 30 12:13:55 2009

Copyright (c) 1982, 2007, Oracle.  All Rights Reserved.


Connected to:
Oracle Database 10g Enterprise Edition Release 10.2.0.4.0 - Production
With the Partitioning, Real Application Clusters, OLAP, Data Mining
and Real Application Testing options

SQL> @asm.sql

ASM disks
*********

Group Disk Header    Mode     State      Total MB    Free MB Disk Name                      Failure Gr Path
----- ---- --------- -------- -------- ---------- ---------- ------------------------------ ---------- ------------------
    0    0 FOREIGN   ONLINE   NORMAL         3065          0                                            /dev/raw/raw5
    0    1 FOREIGN   ONLINE   NORMAL         3065          0                                            /dev/raw/raw4
    0    2 FOREIGN   ONLINE   NORMAL         3065          0                                            /dev/raw/raw2
    0    3 FOREIGN   ONLINE   NORMAL         3065          0                                            /dev/raw/raw1
    0    4 MEMBER    ONLINE   NORMAL       158724          0                                            ORCL:ARCH2
    0    5 MEMBER    ONLINE   NORMAL       389157          0                                            ORCL:DATA2
    0    6 MEMBER    ONLINE   NORMAL        76796          0                                            ORCL:LOG2
    1    0 MEMBER    ONLINE   NORMAL       158724     146176 ARCH1                          ARCH1      ORCL:ARCH1
    1    1 UNKNOWN   OFFLINE  HUNG         158724     146436 ARCH2                          ARCH2
    2    1 UNKNOWN   OFFLINE  HUNG         389157     350723 DATA2                          DATA2
    2    2 MEMBER    ONLINE   NORMAL       389157     350723 DATA181                        DATA1      ORCL:DATA1

Group Disk Header    Mode     State      Total MB    Free MB Disk Name                      Failure Gr Path
----- ---- --------- -------- -------- ---------- ---------- ------------------------------ ---------- ------------------
    3    1 UNKNOWN   OFFLINE  HUNG          76796      76458 LOG2                           LOG2
    3    2 MEMBER    ONLINE   NORMAL        76796      76458 LOG181                         LOG1       ORCL:LOG1

13 rows selected.

SQL> select name,state from v$asm_diskgroup;

NAME                           STATE
------------------------------ -----------
ARCH                           MOUNTED
DATA                           MOUNTED
LOG                            MOUNTED
```

9. Add the disks such as “DATA2,” “LOG2,” “ARCH2” back to the “DATA,” “LOG,” “ARCH” ASM disk group. **Note:** When adding the disks back, use different names, for example, “DATA182,” “LOG182,” “ARCH182.”

```
SQL> alter diskgroup ARCH add failgroup ARCH2 disk 'ORCL:ARCH2' name ARCH182 force;

Diskgroup altered.

SQL> select * from v$asm_operation;

GROUP_NUMBER OPERA STAT      POWER     ACTUAL      SOFAR   EST_WORK   EST_RATE EST_MINUTES
------------ ----- ---- ---------- ---------- ---------- ---------- ---------- -----------
           1 REBAL RUN           1          1          6      12023         34         353

SQL> alter diskgroup ARCH rebalance power 11;

Diskgroup altered.

SQL> select * from v$asm_operation;

GROUP_NUMBER OPERA STAT      POWER     ACTUAL      SOFAR   EST_WORK   EST_RATE EST_MINUTES
------------ ----- ---- ---------- ---------- ---------- ---------- ---------- -----------
           1 REBAL RUN          11          0          0      12592          0           0

SQL> alter diskgroup DATA add failgroup DATA2 disk 'ORCL:DATA2' name DATA182 force;

Diskgroup altered.

SQL> alter diskgroup DATA rebalance power 11;

Diskgroup altered.

SQL> alter diskgroup LOG add failgroup LOG2 disk 'ORCL:LOG2' name LOG182 force;

Diskgroup altered.

SQL> alter diskgroup LOG rebalance power 11;

Diskgroup altered.
```

10. Check the ASM disk group status to confirm that the disks were added to the disk group and are ONLINE:

```
[oracle@node1 ~]$ export ORACLE_SID=+ASM1
[oracle@node1 ~]$ sqlplus / as sysdba

SQL*Plus: Release 10.2.0.4.0 - Production on Tue Dec 1 13:34:38 2009

Copyright (c) 1982, 2007, Oracle.  All Rights Reserved.


Connected to:
Oracle Database 10g Enterprise Edition Release 10.2.0.4.0 - Production
With the Partitioning, Real Application Clusters, OLAP, Data Mining
and Real Application Testing options

SQL> @asm.sql

ASM disks
*********

Group Disk Header    Mode     State      Total MB    Free MB Disk Name                      Failure Gr Path
----- ---- --------- -------- -------- ---------- ---------- ------------------------------ ---------- ------------------
    0    0 FOREIGN   ONLINE   NORMAL         3065          0                                           /dev/raw/raw4
    0    1 FOREIGN   ONLINE   NORMAL         3065          0                                           /dev/raw/raw5
    0    2 FOREIGN   ONLINE   NORMAL         3065          0                                           /dev/raw/raw1
    0    3 FOREIGN   ONLINE   NORMAL         3065          0                                           /dev/raw/raw2
    1    0 MEMBER    ONLINE   NORMAL       158724     146088 ARCH1                          ARCH1      ORCL:ARCH1
    1    2 MEMBER    ONLINE   NORMAL       158724     146088 ARCH182                        ARCH2      ORCL:ARCH2
    2    0 MEMBER    ONLINE   NORMAL       389157     350723 DATA182                        DATA2      ORCL:DATA2
    2    2 MEMBER    ONLINE   NORMAL       389157     350723 DATA181                        DATA1      ORCL:DATA1
    3    0 MEMBER    ONLINE   NORMAL        76796      76458 LOG182                         LOG2       ORCL:LOG2
    3    2 MEMBER    ONLINE   NORMAL        76796      76458 LOG181                         LOG1       ORCL:LOG1

10 rows selected.

SQL>
```

6.12 FAILURE SCENARIO 11: RESTORATION OF THE ORIGINAL SITE

Table 16) Restoration of the original site.

Description	To test the availability of the overall solution, recovery after the loss of an entire site is simulated.
Task	1. Power on only the site 2 disk shelves. 2. Reconnect the ISL between sites so that site1 can see the disk shelves from site 2. After connection, the site2 pool1 volumes automatically begin to resync. 3. In partner mode on site2, reestablish the mirrors in accordance with the active-active and *MetroCluster Configuration Guide* located on NOW. 4. Using the `command aggr` status, make sure that all mirror resynchronization is complete before proceeding. 5. Power on the site 2 controller. Use the `cf status` command to verify that a giveback is possible and use `cf giveback` to fail back.
Expected Results	The resync of volumes is completed successfully. On cluster giveback to the site 2 controller, the results are similar to a normal giveback, as tested previously. This is a maintenance operation involving a small interruption.
Results	Results were as expected. It is important to note that until the `cf giveback` command was issued, there was absolutely no disruption.

6.13 COMBINATION TEST (SIMULTANEOUS FAILURES IN BOTH SITES)

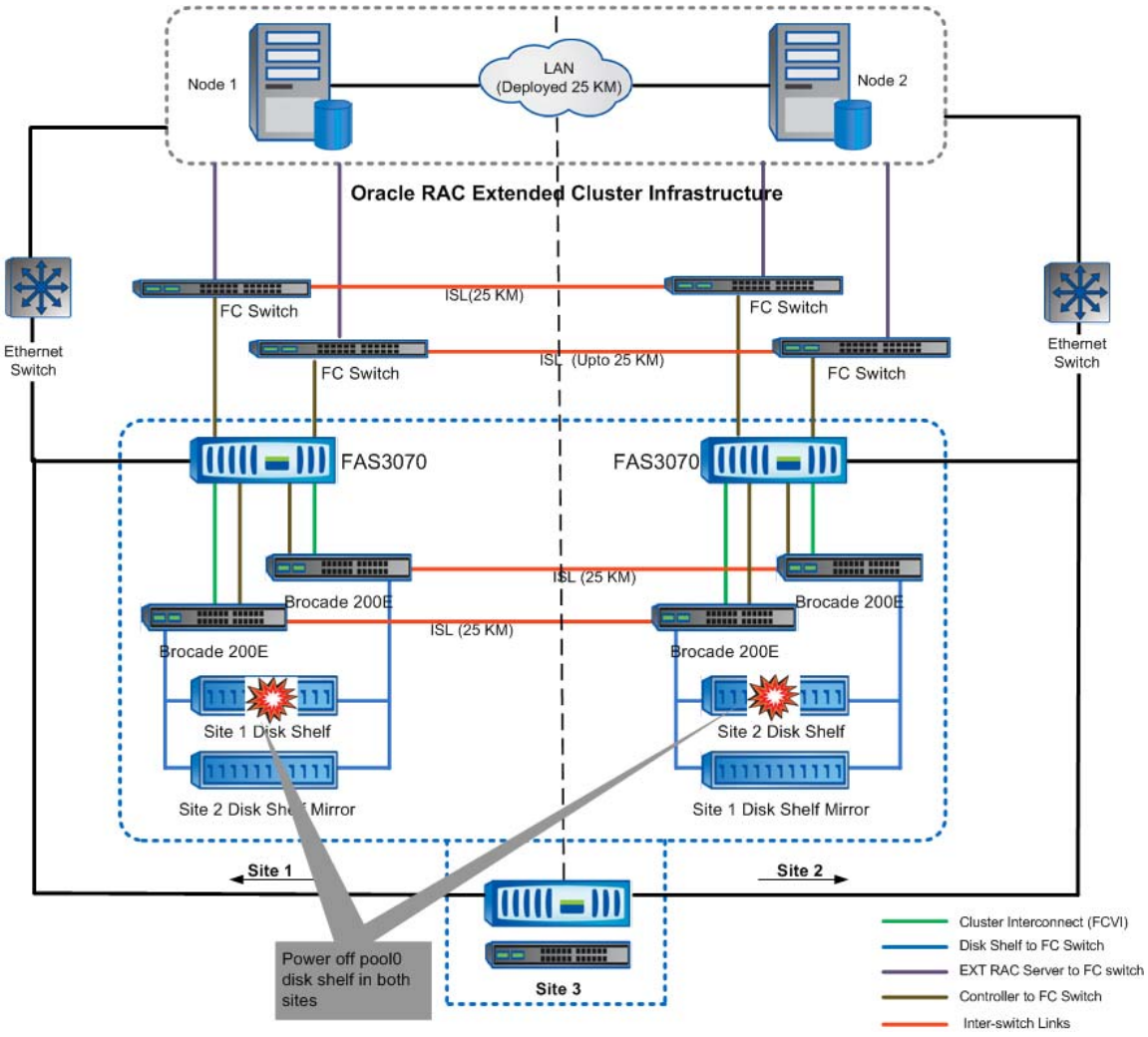

Figure 14) Failures in both sites (1).

Table17) Combination tests (simultaneous failures in both sites).

Tests Performed	Power off disk pool 0 in site 1. Power off disk pool 0 in site 2.
Expected Results	Oracle clients and applications should not see any change and continue to operate normally.
Actual Results	Actual results were in line with the expected behavior, and the tests passed as expected.
Oracle RAC HA behavior	No event
Disruptive?	No

6.14 COMBINATION TEST (SIMULTANEOUS FAILURES IN BOTH SITES)

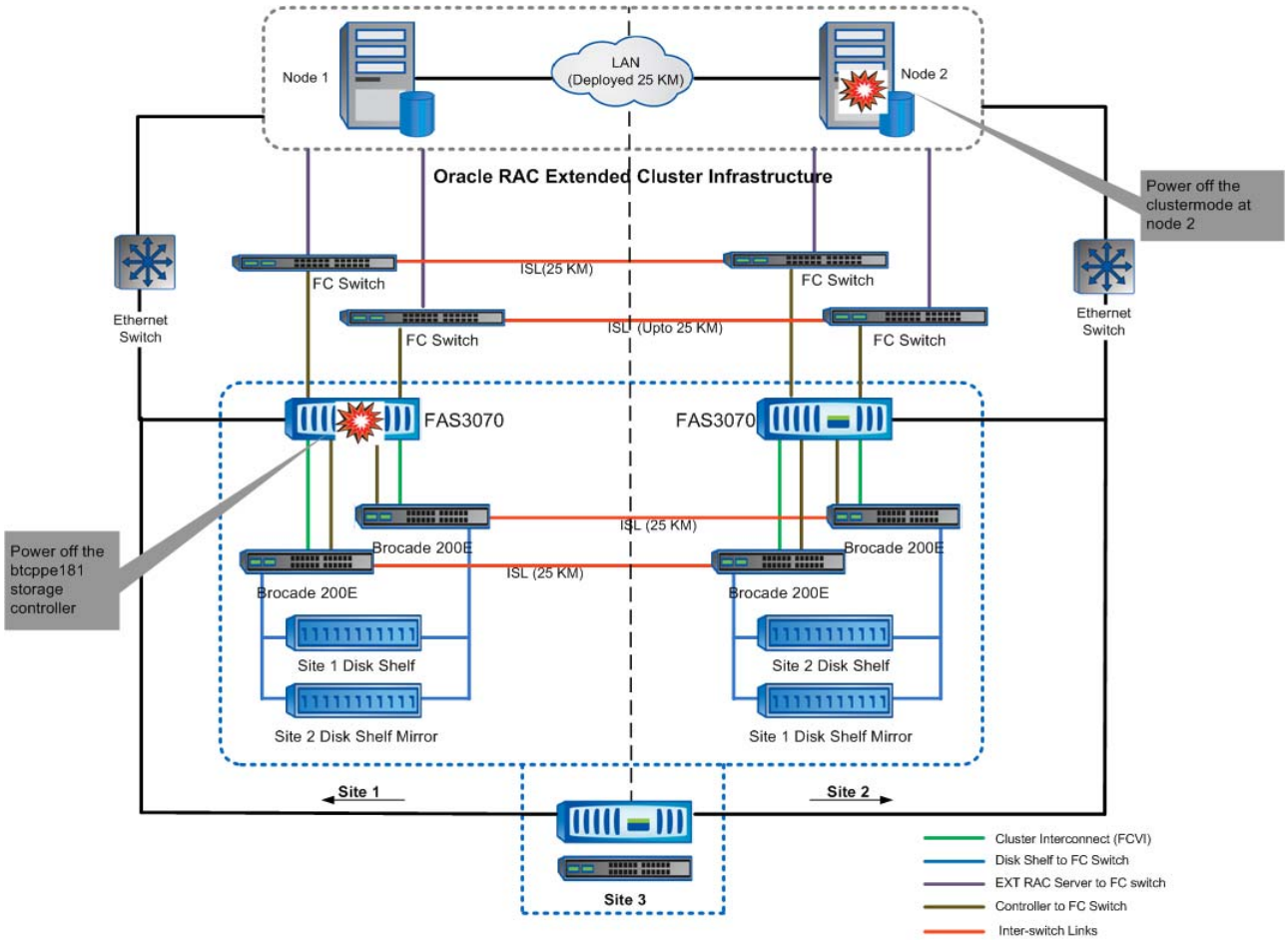

Figure 15) Failures in both sites (2).

Table 18) Combination tests (simultaneous failures in both sites).

Tests Performed	1. Power off the Oracle Database server in one site 2 (node2) 2. Power off the storage controller in site 1 (btcppe181)
Expected Results	• Oracle instance in the Oracle RAC Server automatically migrate to the surviving Oracle RAC Server. • The surviving storage controller automatically takes over the powered off controller.
Actual Results	Actual results were in line with the expected behavior, and the tests passed as expected.
MetroCluster Behavior	Surviving storage controller performs automatic takeover; there is no disruption of data access to either site.
Oracle RAC HA behavior	The instance is automatically switched to the surviving node. The load is also transferred to the surviving node.
Effect to Data Availability	Applications or data residing on the Oracle instance on the Oracle Database server will be available in the surviving node of the Extended RAC setup.

6.15 COMBINATION TESTS (SIMULTANEOUS FAILURES IN BOTH SITES)

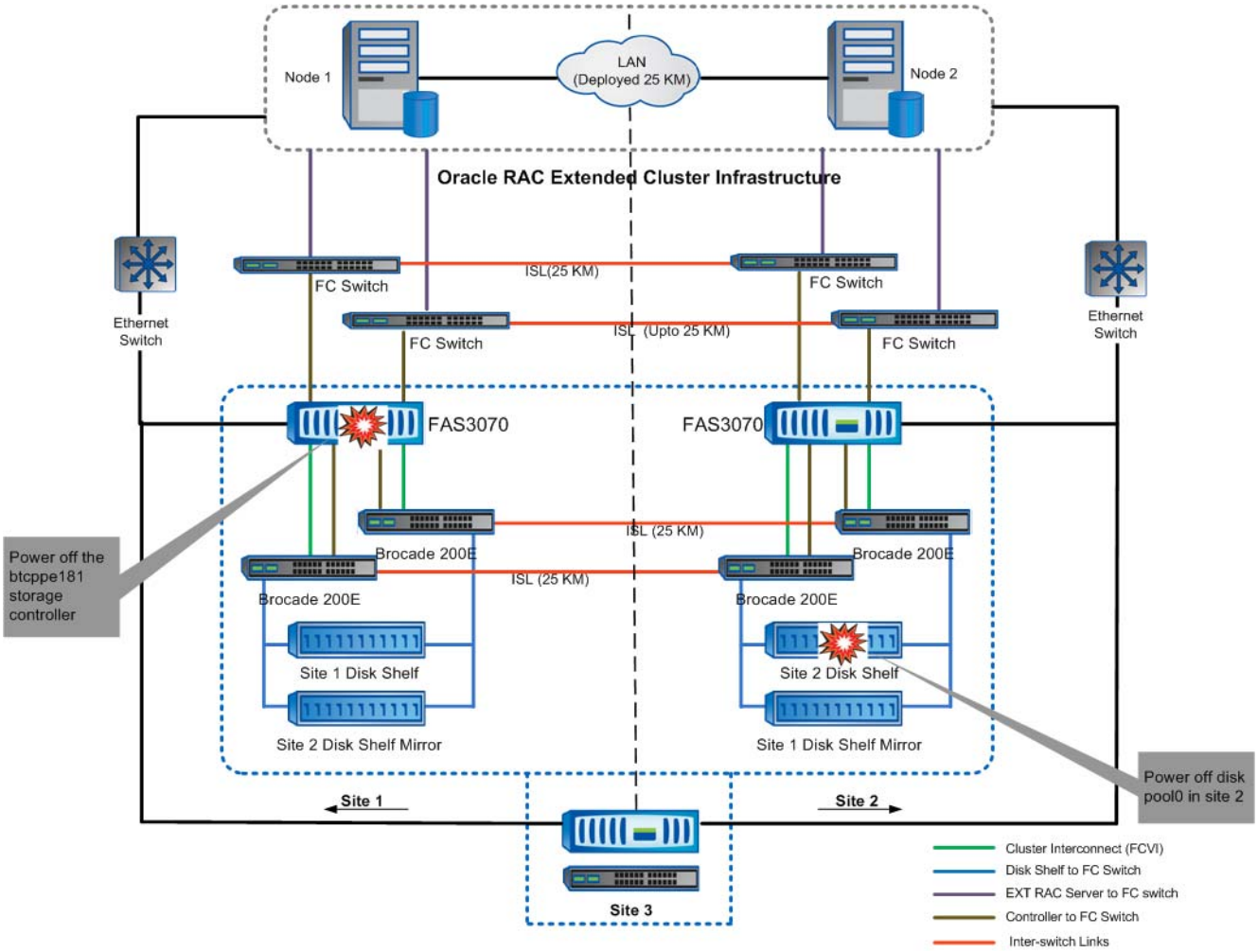

Figure 16) Failures in both sites (3).

Table 19) Combination tests (simultaneous failures in both sites).

Tests Performed	1. Power off storage controller in site 1. 2. Power off disk pool 0 in site 2.
Expected Results	Oracle client should not see any change and continue operation normally.
Actual Results	Actual results were in line with the expected behavior, and the tests passed as expected.
MetroCluster Behavior	Surviving storage controller performs automatic takeover; there is no disruption of data access to either site.
Oracle RAC HA behavior	No event
Disruptive?	No

7 TIEBREAKER SOLUTION

When not using ASM normal redundancy, manual intervention is required in the event of primary site loss due to a disaster. However, there are some scenarios where automated recovery from a site disaster is not only desirable, but required. One such scenario is that of “twin-site” architecture. “Twin site” or “colocation” architecture refers to the concept of a single virtual data center traversing two geographically separated sites. The two physically separate data centers might even share FC SAN and network infrastructure with intentionally little or no differentiation between the two sites. Service level demands for such deployments are such that even a catastrophic loss of one of the two data centers must not result in any service interruption. All solutions deployed into such architecture must comply with a simple requirement: if either site drops, the other site must keep all services running.

The demands of this kind of cross-site fault tolerant solution are perfectly suited to MetroCluster.

This section describes a simple extension to the MetroCluster solution to remove the manual element of the disaster recovery decision and thus provide a true “hands-off” failover solution that complies with “twin-site” availability requirements. We refer to this solution as “MetroCluster tiebreaker.”

THE SPLIT-BRAIN PROBLEM AND THE IMPORTANCE OF A THIRD SITE

The very notion of a twin site is in itself somewhat flawed. To provide the required service levels, a third site or “triple-site” architecture is required. The reason for this is the possibility of a “split brain” or “partition” occurring. If the sites or subsets of the sites lose contact with each other (loss of networks or fiber rings between the twin-site locations), none of the individual solutions deployed in the twin site can be relied upon to behave “correctly” on its own, and it’s possible that data integrity might become compromised. Many host/server clustering software solutions have relatively robust methods solutions to address this issue. However, it generally comes down to the following:

- A third site is required with full, reliable network connectivity to each of the primary sites (especially for the third Oracle RAC Voting disk).
- A daemon process on the third site arbitrates takeover. This process is often referred to as a “tiebreaker” or “steward.” In some cases this is a mature, packaged product; in many cases it is a simple script.
- In some cases Fibre Channel access might even be required in order to provide a SCSI-reserved quorum volume to use as an alternative arbitration mechanism.
- Both primary nodes can be shut down either using out-of-band communication or committing suicide (usually by panicking), in the event that they are not visible to either the steward process or the partner node in the cluster.

Such solutions are rarely deployed due to the difficulties in locating and funding a suitable third site. It is also highly unlikely the third site will have separate networks to each of the twin-site locations. It is far more likely that the connection to one twin site will run through the other. However, in the case of a true twin-site deployment, for at least the most critical services, tiebreaker solutions must be deployed along with the infrastructure required to support them.

TIEBREAKER SOLUTION OVERVIEW

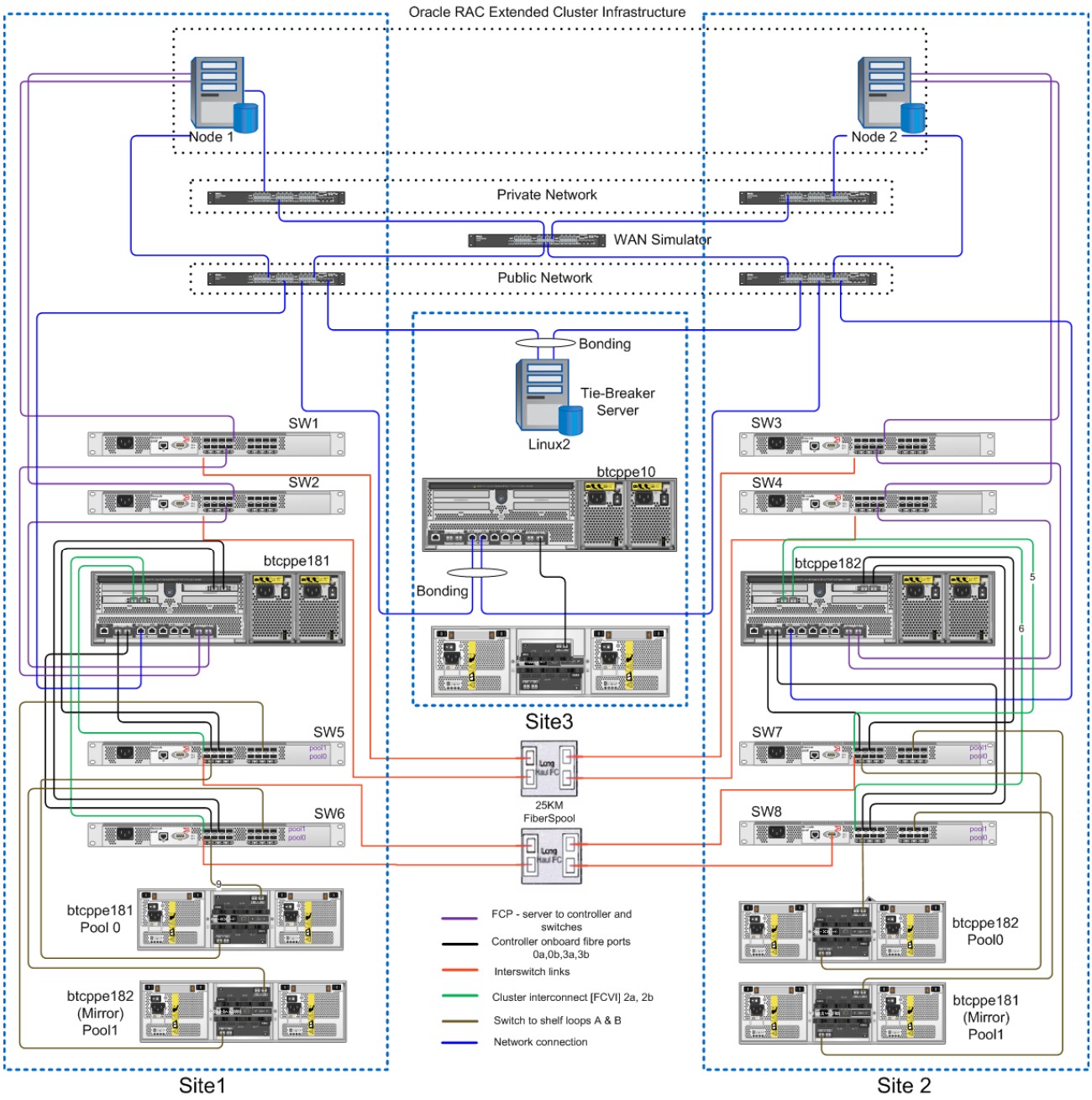

Figure 17) Tiebreaker solution overview.

The NetApp MetroCluster tiebreaker solution automates the cluster failover process (to avoid split brain) and will minimize any database downtime.

ASM normal redundancy in conjunction with NetApp Fabric MetroCluster provides zero downtime for the application because Oracle will automatically access the other plex on the ASM mirror without any interruption for the users.

In conclusion, Oracle Extended RAC together with Oracle ASM normal redundancy, NetApp Fabric MetroCluster, and NetApp MetroCluster Tiebreaker enables customers to deploy an environment with zero downtime for Oracle Databases with automatic site failover.

8 IMPLEMENTATION SCENARIOS

Table 20) Implementation scenario 1.

	General High Availability	Loss of 3 disks on the same diskgroup	Loss of Complete Raid Group	Loss of NetApp Storage Controller	Loss of Fibre Channel Switch	Loss of One ISL	Loss of an Entire Site
NetApp MetroCluster	✓	✓	✓	✓	✓	✓	✓*

Note: The majority of these scenarios are addressed for NetApp MetroCluster solution without the need to use ASM normal redundancy.

The “Loss of an Entire Site” scenario requires ASM normal redundancy to provide zero downtime and automatic site failover, with NetApp Tiebreaker to automate site failover.

The scenario “Loss of an Entire Site” is an example of Split Brain where it is necessary for an entity to determine the surviving site. This is the same situation as the implementation of third voting disk in a tertiary site.

The voting file is used by the cluster synchronization service (CSS) component, which is part of Oracle Clusterware, to resolve network splits, commonly referred to as split brain. A “split brain” in the cluster describes the condition where each side of the split cluster cannot see the nodes on the other side.

The voting files are used as the final arbiter on the status of the configured nodes (either up or down) and are used as the medium to deliver eviction notices. This means, once it is decided that a particular node must be evicted, it is marked as such in the voting file. If a node does not have access to the majority of the voting files in the cluster in a way that it can write a disk heartbeat, the node will be evicted from the cluster.

As far as voting files are concerned, a node must be able to access more than the half of the voting files at any time (simple majority). To be able to tolerate a failure of *n* voting files, one must have at least *2n+1* configured. (*n*= number of voting files) for the cluster.

Extended clusters are generally implemented to provide system availability and to protect from site failures. The goal is that each site can run independently of when one site fails. The problem in a stretched cluster configuration is that most installations only use two storage systems (one at each site), which means that the site that hosts the majority of the voting files is a potential single point of failure for the entire cluster. If the storage or the site where the *n+1* voting files are configured fails, the whole cluster will go down, because Oracle Clusterware will lose the majority of voting files.

To prevent a full cluster outage, Oracle supports a third voting file standard NFS mounted installed typically in a third site.

Table 21) Implementation scenario 2.

	General High Availability	Loss of 3 disks on the same disk group	Loss of Complete Raid Group	Loss of NetApp Storage Controller	Loss of Fibre Channel Switch	Loss of One ISL	Loss of an Entire Site
ASM Normal Redundancy	✓	✓*	✓*	✓*	✓*	✓*	✓*

Note: Using only ASM normal redundancy *all* scenarios above will require an ASM rebalance (redistributing file data evenly across all the disks of the disk group). This operation will result in database server CPU utilization as well as LAN/WAN utilization. You can minimize this system resource impact by using `alter diskgroup <diskgroupname> rebalance power <number>` or `ASM_POWER_LIMIT`. Where, the valid values range from 1 to 11, with 1 being the default. The higher the limit, the more resources are allocated, resulting in faster rebalancing operations with higher CPU and network impact along with higher disk latencies.

Table 22) Implementation scenario 3.

	General High Availability	Loss of 3 disks on the same Disk Group	Loss of Complete Raid Group	Loss of NetApp Storage Controller	Loss of Fibre Channel Switch	Loss of One ISL	Loss of an Entire Site
NetApp MetroCluster + ASM Normal Redundancy + NetApp Tiebreaker	✓	✓	✓	✓	✓	✓	✓

Note: A combination of NetApp MetroCluster, ASM Normal Redundancy, and NetApp Tiebreaker is optimal because the majority of the failures (that are, loss of a shelf, loss of ISL, loss of a storage controller and others) are handled by NetApp Fabric MetroCluster without any CPU or network impact to the Oracle RAC nodes. The only scenario that is be handled by ASM normal redundancy is the loss of an entire site. The probability of this occurrence is much lower than the other scenarios.

9 SUMMARY

NetApp Fabric MetroCluster (FMC) combined with Oracle Real Applications Clusters (RAC) on Extended Distance Clusters, and Oracle Automated Storage Management (ASM) Normal Redundancy provides a robust zero downtime solution for disaster recovery. Oracle ASM and FMC provide zero downtime, maximizing availability, and avoiding both planned and unplanned downtime. Planned and unplanned site failovers can be triggered without any impact to the environment.

The combination of NetApp Fabric MetroCluster and ASM Normal Redundancy is optimal, because the majority of the failures (that is, loss of a shelf, loss of ISL, loss of a storage controller) will be handled by NetApp Fabric MetroCluster without any CPU impact to the Oracle RAC Nodes. The only failure scenario that will be handled by ASM Normal Redundancy is the loss of entire site.

The operation to recombine failed disks to ASM disk group (rebalance operation) will cause an increase of CPU and network utilization for Oracle RAC Nodes.

Table 23 summarizes the failure scenarios discussed in this document, showcasing the combination of Oracle RAC and NetApp MetroCluster which come together to deliver complete protection from server and storage failures.

Table 23) Summary of failure scenarios and their effect on data availability.

#	Failure Scenario	Data Availability Effect
1	Complete loss of power to disk shelf	None
2	Loss of one link in one disk loop	None
3	Loss of brocade switch	None
4	Loss of one interswitch link (ISL)	None
5	Failure and failback of storage controller	None
6	Loss of an entire site	None. ASM normal redundancy will be responsible for data availability for the complete failed site disks
7	Loss of an Oracle node	None
8	Loss of an Oracle host HBA	None
9	Loss of disk(s)	None
10	Loss of complete disk shelf	None
11	Loss of NetApp storage controller	None

This paper is not intended to be a definitive implementation or solutions guide for high-availability solutions in Oracle RAC with NetApp storage. Many factors related to specific customer environments are not addressed in this document. Contact NetApp support to speak with one of our Oracle solutions experts for any deployment requirement. Please forward any errors, omissions, differences, new discoveries, or comments about this paper to the authors.

10 APPENDIX A: BROCADE SWITCH CONNECTION DETAILS FOR FABRIC METROCLUSTER

Table 24 lists the Brocade switch connection details for fabric MetroCluster (software-based disk ownership).

Site 1 / btcppe181

Site 2 / btcppe182

Table 24) Brocade switch connection details for fabric MetroCluster.

Switch Name	SW5		
Port	**Bank/Pool**	**Connected To**	**Purpose**
0	1/0	Site 1 FCVI	Cluster interconnect
1	1/0	Onboard HBA -0a	
2	1/0	Onboard HBA – 3a	
4	1/0	ISL	Interswitch link between SW5-SW7
5	1/0	Site 1 Shelf 1 B port	To access the site 1 controller “Root – aggr0” and “Oracle RAC - racaggr” aggregate pool0
10	2/1	Site 2 Shelf 1 Mirror Pool 1 Bport	To access the site 2 controller “Root – aggr0” and “Oracle RAC – extrac_B” aggregate pool1
Switch Name	**SW6**		
Port	**Bank/Pool**	**Connected To**	**Purpose**
0	1/0	Site 1 FCVI	Cluster interconnect
1	1/0	Onboard HBA -0b	
2	1/0	Onboard HBA -3b	
4	1/1	ISL	Interswitch link between SW6-SW8
5	1/0	Site 1 Shelf 1 A Port	To access the site 1 controller “Root – aggr0” and “Oracle RAC - racaggr” aggregate pool0 - Multipath
10	2/1	Site 2 Shelf 1 Mirror Pool 1 A Port	To access the site 2 controller “Root – aggr0” and “Oracle RAC – extrac_B” aggregate pool1 - Multipath
Switch Name	**SW7**		
Port	**Bank/Pool**	**Connected To**	**Purpose**
0	1/0	Site 2 FCVI	Cluster interconnect
1	1/0	Onboard HBA 0a	
2	1/0	Onboard HBA 3a	
4	1/1	ISL	Interswitch link
5	1/0	Site 2 Shelf 1 B port	To access the site 2 controller “Root – aggr0” and “Oracle RAC – extrac_B” aggregate pool0
10	2/1	Site 1 Shelf 1 Mirror Pool 1 B port	To access the site 2 controller “Root – aggr0” and “Oracle RAC – racaggr” aggregate pool1
Switch Name	**SW8**		
Port	**Bank/Pool**	**Connected To**	**Purpose**
0	1/0	Site 2 FCVI	Cluster interconnect
1	1/0	Onboard HBA -0b	

2	1/0	Onboard HBA -3b	
4	1/1	ISL	Interswitch link between SW6-SW8
5	1/0	Site 2 Shelf 1 A Port	To access the site 2 controller “Root – aggr0” and “Oracle RAC – extrac_B” aggregate pool0 – multipath
10	2/1	Site 1 Shelf 1 Mirror Pool 1 A Port	To access the site 2 controller “Root – aggr0” and “Oracle RAC – racaggr” aggregate pool1 - multipath
Switch Name	**SW1**		
Port	**Bank/Pool**	**Connected To**	**Purpose**
0	1/0	ISL	Interswitch link between SW1-SW3
3	1/0	Node1 HBA port1	To access the FC LUN from site 1 and site 2 storage controllers.
7	1/1	Onboard HBA -0c	
Switch Name	**SW2**		
Port	**Bank/Pool**	**Connected To**	**Purpose**
0	1/0	ISL	Interswitch link between SW2-SW4
3	1/0	Node1 HBA port2	To access the FC LUN from site 1 and site 2 storage controllers – multipath.
7	1/1	Onboard HBA -0d	
Switch Name	**SW3**		
Port	**Bank/Pool**	**Connected To**	**Purpose**
0	1/0	ISL	Interswitch link between SW3-SW1
3	1/0	Node2 HBA port1	To access the FC LUN from site 1 and site 2 storage controllers.
7	1/1	Onboard HBA -0c	
Switch Name	**SW4**		
Port	**Bank/Pool**	**Connected To**	**Purpose**
0	1/0	ISL	Interswitch link between SW2-SW4
3	1/0	Node2 HBA port2	To access the FC LUN from site 1 and site 2 storage controllers – multipath.
7	1/1	Onboard HBA -0d	

11 APPENDIX B: AVOIDING THE “DEVICE/FILE NEEDS TO BE SYNCHRONIZED WITH THE OTHER DEVICE” ERROR

```
[root@node1 ~]# ocrcheck
Status of Oracle Cluster Registry is as follows :
         Version                  :          2
         Total space (kbytes)     :    3139148
         Used space (kbytes)      :       4920
         Available space (kbytes) :    3134228
         ID                       : 1633336445
         Device/File Name         : /dev/raw/raw1
                                    Device/File integrity check succeeded
         Device/File Name         : /dev/raw/raw2
                                    Device/File needs to be synchronized with the other
device

         Cluster registry integrity check succeeded

[root@node1 ~]# ocrconfig -replace ocrmirror /dev/raw/raw2
[root@node1 ~]# ocrcheck
Status of Oracle Cluster Registry is as follows :
         Version                  :          2
         Total space (kbytes)     :    3139148
         Used space (kbytes)      :       4920
         Available space (kbytes) :    3134228
         ID                       : 1633336445
         Device/File Name         : /dev/raw/raw1
                                    Device/File integrity check succeeded
         Device/File Name         : /dev/raw/raw2
                                    Device/File integrity check succeeded

         Cluster registry integrity check succeeded
```

12 APPENDIX C: AVOIDING THE "SELECT" QUERY FAILURE FROM CLIENT "SQLPLUS" DURING SITE FAILURE

1. In the `tnsnames.ora` file, enable the following parameters:
   FAILOVER
   LOAD_BALANCE
   FAILURE_MODE TYPE=select
   METHOD=basic

   The following is an example of `tnsnames.ora` entry for the `ext` service:

```
EXT =
  (DESCRIPTION =
    (ADDRESS = (PROTOCOL = TCP)(HOST = node1-vip.btcppe.netapp.com)(PORT =
        1521))
    (ADDRESS = (PROTOCOL = TCP)(HOST = node2-vip.btcppe.netapp.com)(PORT =
        1521))
    (LOAD_BALANCE = yes)
    (FAILOVER = true)
    (CONNECT_DATA =
      (SERVER = DEDICATED)
      (SERVICE_NAME = ext)
        (FAILOVER_MODE=
                (TYPE=select)
                (METHOD=basic)
                (RETRIES=20)
                (DELAY=15)
          )
    )
  )
```

2. Verify the instance name before executing the "select * from all_objects" such as "ext2."
3. Simulate the site failure by powering off the node, switches, controller, and disk shelves all at once. For example, for the architecture described above, shut down:
   - Server - node2
   - Storage controller - btcppe182
   - Front end switches – SW3, SW4
   - Back end switches – SW7,SW8
   - Disk shelves
     - btcppe182 - pool0
     - btcppe181 – pool1
   - Private and public switch

   The "select" query freezes for 15-30 seconds and then resumes the select query without errors.
4. Check the instance name by substituting with another instance name such as "ext1."
   The query continues to run.

13 APPENDIX D: COMMANDS OUTPUT

File name: `multipath.conf`

Output:

```
blacklist {
        wwid SServeRA_Drive_1_C3442A45
        devnode "^(ram|raw|loop|fd|md|dm-|sr|scd|st)[0-9]*"
        devnode "^hd[a-z]"
        devnode "^cciss!c[0-9]d[0-9]*[p[0-9]*]"
}


defaults {
        user_friendly_names           yes
        max_fds                               max
        queue_without_daemon          no
}
multipaths {
        multipath {
                wwid                    360a98000486e5851636f544658793478
                alias                   oradata1.lun
        }
        multipath {
                wwid                    360a98000486e5851636f544658793535
                alias                   oraarch1.lun
        }
        multipath {
                wwid                    360a98000486e5851636f544658793476
                alias                   oralog1.lun
        }
        multipath {
                wwid                    360a98000486e5851636f544658793530
                alias                   oraocr1.lun
        }
        multipath {
                wwid                    360a98000486e5851636f54465879347a
                alias                   votedisk1.lun
        }
        multipath {
                wwid                    360a98000486e58514c34544b55465033
                alias                   oradata2.lun
        }
        multipath {
                wwid                    360a98000486e58514c34544b55465038
                alias                   oraarch2.lun
        }
        multipath {
                wwid                    360a98000486e58514c34544b55465034
                alias                   oralog2.lun
        }
        multipath {
                wwid                    360a98000486e58514c34544b55465041
                alias                   oraocr2.lun
        }
        multipath {
                wwid                    360a98000486e58514c34544b5546502d
                alias                   votedisk2.lun
        }
}

devices {
        device {
                vendor                  "NETAPP"
                product                 "LUN"
                getuid_callout          "/sbin/scsi_id -g -u -s /block/%n"
```

```
                prio_callout            "/sbin/mpath_prio_alua /dev/%n"
                features                "0"
                hardware_handler        "0"
                path_grouping_policy    group_by_prio
                failback                immediate
                rr_weight               uniform
                rr_min_io               128
                path_checker            directio
                flush_on_last_del       yes
                no_path_retry           fail
        }
}
```

ALUA setting on both storage controllers

```
igroup set node1-port2 alua yes
igroup set node1-port1 alua yes
igroup set node2-port1 alua yes
igroup set node2-port2 alua yes
igroup show -v

node1-port2 (FCP):
        OS Type: linux
        Member: 21:01:00:1b:32:30:94:c2 (logged in on: 0d, vtic)
        ALUA: Yes
node2-port1 (FCP):
        OS Type: linux
        Member: 21:00:00:1b:32:10:34:c4 (logged in on: 0c, vtic)
        ALUA: Yes
node1-port1 (FCP):
        OS Type: linux
        Member: 21:00:00:1b:32:10:94:c2 (logged in on: 0c, vtic)
        ALUA: Yes
node2-port2 (FCP):
        OS Type: linux
        Member: 21:01:00:1b:32:30:34:c4 (logged in on: 0d, vtic)
               ALUA: Yes
```

Command: `sanlun lun show all`

Output:

```
controller:  lun-pathname          device filename   adapter  protocol        lun size   lun state
btcppe181:  /vol/oradata1/oradata1.lun    /dev/sda         host0    FCP        380.0g (408063836160)   GOOD
btcppe181:  /vol/oraarch1/oraarch1.lun    /dev/sdb         host0    FCP        155.0g (166435225600)   GOOD
btcppe181:  /vol/oralog1/oralog1.lun      /dev/sdc         host0    FCP           75g (80530636800)    GOOD
btcppe182:  /vol/oradata2/oradata2.lun    /dev/sdd         host0    FCP        380.0g (408063836160)   GOOD
btcppe182:  /vol/oraarch2/oraarch2.lun    /dev/sde         host0    FCP        155.0g (166435225600)   GOOD
btcppe182:  /vol/oralog2/oralog2.lun      /dev/sdf         host0    FCP           75g (80530636800)    GOOD
btcppe181:  /vol/ocr1/ocr1.lun            /dev/sdg         host0    FCP            3g (3221225472)     GOOD
btcppe182:  /vol/ocr2/ocr2.lun            /dev/sdh         host0    FCP            3g (3221225472)     GOOD
btcppe181:  /vol/votedisk1/votedisk1.lun  /dev/sdi         host0    FCP            3g (3221225472)     GOOD
btcppe182:  /vol/votedisk2/votedisk2.lun  /dev/sdj         host0    FCP            3g (3221225472)     GOOD
btcppe181:  /vol/oradata1/oradata1.lun    /dev/sdn         host0    FCP        380.0g (408063836160)   GOOD
btcppe181:  /vol/oraarch1/oraarch1.lun    /dev/sdo         host0    FCP        155.0g (166435225600)   GOOD
btcppe181:  /vol/oralog1/oralog1.lun      /dev/sdp         host0    FCP           75g (80530636800)    GOOD
btcppe182:  /vol/oradata2/oradata2.lun    /dev/sdq         host0    FCP        380.0g (408063836160)   GOOD
btcppe182:  /vol/oraarch2/oraarch2.lun    /dev/sdr         host0    FCP        155.0g (166435225600)   GOOD
btcppe182:  /vol/oralog2/oralog2.lun      /dev/sds         host0    FCP           75g (80530636800)    GOOD
btcppe181:  /vol/ocr1/ocr1.lun            /dev/sdt         host0    FCP            3g (3221225472)     GOOD
btcppe182:  /vol/ocr2/ocr2.lun            /dev/sdu         host0    FCP            3g (3221225472)     GOOD
btcppe181:  /vol/votedisk1/votedisk1.lun  /dev/sdv         host0    FCP            3g (3221225472)     GOOD
btcppe182:  /vol/votedisk2/votedisk2.lun  /dev/sdw         host0    FCP            3g (3221225472)     GOOD
btcppe181:  /vol/tiedata/tiedata.lun1     /dev/sdx         host0    FCP         40.0g (42953867264)    GOOD
btcppe181:  /vol/tiearch/tiearch.lun1     /dev/sdy         host0    FCP         17.0g (18254659584)    GOOD
btcppe181:  /vol/tielog/tielog.lun1       /dev/sdz         host0    FCP         17.0g (18254659584)    GOOD
btcppe181:  /vol/oradata1/oradata1.lun    /dev/sddf        host1    FCP        380.0g (408063836160)   GOOD
btcppe181:  /vol/oraarch1/oraarch1.lun    /dev/sddg        host1    FCP        155.0g (166435225600)   GOOD
btcppe181:  /vol/oralog1/oralog1.lun      /dev/sddh        host1    FCP           75g (80530636800)    GOOD
btcppe182:  /vol/oradata2/oradata2.lun    /dev/sddi        host1    FCP        380.0g (408063836160)   GOOD
btcppe182:  /vol/oraarch2/oraarch2.lun    /dev/sddj        host1    FCP        155.0g (166435225600)   GOOD
btcppe182:  /vol/oralog2/oralog2.lun      /dev/sddk        host1    FCP           75g (80530636800)    GOOD
btcppe181:  /vol/ocr1/ocr1.lun            /dev/sddl        host1    FCP            3g (3221225472)     GOOD
btcppe182:  /vol/ocr2/ocr2.lun            /dev/sddm        host1    FCP            3g (3221225472)     GOOD
btcppe181:  /vol/votedisk1/votedisk1.lun  /dev/sddn        host1    FCP            3g (3221225472)     GOOD
btcppe182:  /vol/votedisk2/votedisk2.lun  /dev/sddo        host1    FCP            3g (3221225472)     GOOD
btcppe181:  /vol/oradata1/oradata1.lun    /dev/sdds        host1    FCP        380.0g (408063836160)   GOOD
btcppe181:  /vol/oraarch1/oraarch1.lun    /dev/sddt        host1    FCP        155.0g (166435225600)   GOOD
btcppe181:  /vol/oralog1/oralog1.lun      /dev/sddu        host1    FCP           75g (80530636800)    GOOD
btcppe182:  /vol/oradata2/oradata2.lun    /dev/sddv        host1    FCP        380.0g (408063836160)   GOOD
btcppe182:  /vol/oraarch2/oraarch2.lun    /dev/sddw        host1    FCP        155.0g (166435225600)   GOOD
btcppe182:  /vol/oralog2/oralog2.lun      /dev/sddx        host1    FCP           75g (80530636800)    GOOD
btcppe181:  /vol/ocr1/ocr1.lun            /dev/sddy        host1    FCP            3g (3221225472)     GOOD
btcppe182:  /vol/ocr2/ocr2.lun            /dev/sddz        host1    FCP            3g (3221225472)     GOOD
btcppe181:  /vol/votedisk1/votedisk1.lun  /dev/sdea        host1    FCP            3g (3221225472)     GOOD
btcppe182:  /vol/votedisk2/votedisk2.lun  /dev/sdeb        host1    FCP            3g (3221225472)     GOOD
```

Command: `sanlun lun show -p`

Output:

```
btcppe182:/vol/oraarch2/oraarch2.lun (LUN 4)            Lun state: GOOD

Lun Size:  155.0g (166435225600) Controller_CF_State: Cluster Enabled

Protocol: FCP           Controller Partner: btcppe181

DM-MP DevName: oraarch2.lun     (360a98000486e58514c34544b55465038)     dm-8

Multipath-provider: NATIVE

--------- ---------- ------- ------------ ------------------------------------------- ---------------

   sanlun Controller                                                        Primary         Partner

     path       Path   /dev/         Host                                Controller      Controller

    state       type    node          HBA                                      port            port

--------- ---------- ------- ------------ ------------------------------------------- ---------------

     GOOD  primary      sddw        host1                                       0d              --

     GOOD  primary       sdr        host0                                       0c              --

     GOOD  secondary    sddj        host1                                       --              0d
```

```
     GOOD  secondary     sde       host0                                          --            0c

btcppe182:/vol/ocr2/ocr2.lun (LUN 7)           Lun state: GOOD
Lun Size:      3g (3221225472)   Controller_CF_State: Cluster Enabled
Protocol: FCP           Controller Partner: btcppe181
DM-MP DevName: oraocr2.lun      (360a98000486e58514c34544b55465041)     dm-13
Multipath-provider: NATIVE
--------- ---------- ------- ------------ ------------------------------------------ --------------
   sanlun Controller                                                      Primary        Partner
     path       Path   /dev/        Host                                Controller     Controller
    state       type    node         HBA                                     port           port
--------- ---------- ------- ------------ ------------------------------------------ --------------
     GOOD  primary      sddz       host1                                      0d             --
     GOOD  primary       sdu       host0                                      0c             --
     GOOD  secondary    sddm       host1                                      --             0d
     GOOD  secondary     sdh       host0                                      --             0c

btcppe181:/vol/oraarch1/oraarch1.lun (LUN 1)           Lun state: GOOD
Lun Size:  155.0g (166435225600) Controller_CF_State: Cluster Enabled
Protocol: FCP           Controller Partner: btcppe182
DM-MP DevName: oraarch1.lun     (360a98000486e5851636f544658793535)     dm-1
Multipath-provider: NATIVE
--------- ---------- ------- ------------ ------------------------------------------ --------------
   sanlun Controller                                                      Primary        Partner
     path       Path   /dev/        Host                                Controller     Controller
    state       type    node         HBA                                     port           port
--------- ---------- ------- ------------ ------------------------------------------ --------------
     GOOD  primary       sdb       host0                                      0c             --
     GOOD  primary      sddg       host1                                      0d             --
     GOOD  secondary    sddt       host1                                      --             0d
     GOOD  secondary     sdo       host0                                      --             0c

btcppe182:/vol/votedisk2/votedisk2.lun (LUN 9)         Lun state: GOOD
Lun Size:      3g (3221225472)   Controller_CF_State: Cluster Enabled
Protocol: FCP           Controller Partner: btcppe181
DM-MP DevName: votedisk2.lun    (360a98000486e58514c34544b5546502d)     dm-18
Multipath-provider: NATIVE
--------- ---------- ------- ------------ ------------------------------------------ --------------
   sanlun Controller                                                      Primary        Partner
     path       Path   /dev/        Host                                Controller     Controller
    state       type    node         HBA                                     port           port
--------- ---------- ------- ------------ ------------------------------------------ --------------
     GOOD  primary      sdeb       host1                                      0d             --
```

```
     GOOD  primary       sdw       host0                                          0c             --
     GOOD  secondary    sddo       host1                                          --             0d
     GOOD  secondary     sdj       host0                                          --             0c

btcppe181:/vol/ocr1/ocr1.lun (LUN 6)            Lun state: GOOD
Lun Size:      3g (3221225472)   Controller_CF_State: Cluster Enabled
Protocol: FCP           Controller Partner: btcppe182
DM-MP DevName: oraocr1.lun      (360a98000486e5851636f544658793530)     dm-11
Multipath-provider: NATIVE
--------- ---------- ------- ------------ ------------------------------------------ ---------------
   sanlun Controller                                                      Primary         Partner
     path       Path   /dev/        Host                                Controller      Controller
    state       type    node         HBA                                     port            port
--------- ---------- ------- ------------ ------------------------------------------ ---------------
     GOOD  primary      sddl       host1                                          0d             --
     GOOD  primary       sdg       host0                                          0c             --
     GOOD  secondary    sddy       host1                                          --             0d
     GOOD  secondary     sdt       host0                                          --             0c


btcppe182:/vol/oradata2/oradata2.lun (LUN 3)            Lun state: GOOD
Lun Size:  380.0g (408063836160) Controller_CF_State: Cluster Enabled
Protocol: FCP           Controller Partner: btcppe181
DM-MP DevName: oradata2.lun     (360a98000486e58514c34544b55465033)     dm-6
Multipath-provider: NATIVE
--------- ---------- ------- ------------ ------------------------------------------ ---------------
   sanlun Controller                                                      Primary         Partner
     path       Path   /dev/        Host                                Controller      Controller
    state       type    node         HBA                                     port            port
--------- ---------- ------- ------------ ------------------------------------------ ---------------
     GOOD  primary      sddv       host1                                          0d             --
     GOOD  primary       sdq       host0                                          0c             --
     GOOD  secondary    sddi       host1                                          --             0d
     GOOD  secondary     sdd       host0                                          --             0c

btcppe181:/vol/votedisk1/votedisk1.lun (LUN 8)          Lun state: GOOD
Lun Size:      3g (3221225472)   Controller_CF_State: Cluster Enabled
Protocol: FCP           Controller Partner: btcppe182
DM-MP DevName: votedisk1.lun    (360a98000486e5851636f54465879347a)     dm-15
Multipath-provider: NATIVE
--------- ---------- ------- ------------ ------------------------------------------ ---------------
   sanlun Controller                                                      Primary         Partner
     path       Path   /dev/        Host                                Controller      Controller
```

```
    state       type    node          HBA                                          port            port
--------- ---------- ------- ------------ ---------------------------------------- ---------------
     GOOD  primary      sddn        host1                                          0d              --
     GOOD  primary       sdi        host0                                          0c              --
     GOOD  secondary    sdea        host1                                          --              0d
     GOOD  secondary     sdv        host0                                          --              0c

btcppe181:/vol/oradata1/oradata1.lun (LUN 0)            Lun state: GOOD
Lun Size:  380.0g (408063836160) Controller_CF_State: Cluster Enabled
Protocol: FCP           Controller Partner: btcppe182
DM-MP DevName: oradata1.lun     (360a98000486e5851636f544658793478)     dm-0
Multipath-provider: NATIVE
--------- ---------- ------- ------------ ---------------------------------------- ---------------
   sanlun Controller                                                      Primary         Partner
     path       Path   /dev/        Host                                Controller      Controller
    state       type    node          HBA                                     port            port
--------- ---------- ------- ------------ ---------------------------------------- ---------------
     GOOD  primary       sda        host0                                          0c              --
     GOOD  primary      sddf        host1                                          0d              --
     GOOD  secondary    sdds        host1                                          --              0d
     GOOD  secondary     sdn        host0                                          --              0c


btcppe182:/vol/oralog2/oralog2.lun (LUN 5)              Lun state: GOOD
Lun Size:     75g (80530636800)  Controller_CF_State: Cluster Enabled
Protocol: FCP           Controller Partner: btcppe181
DM-MP DevName: oralog2.lun      (360a98000486e58514c34544b55465034)     dm-9
Multipath-provider: NATIVE
--------- ---------- ------- ------------ ---------------------------------------- ---------------
   sanlun Controller                                                      Primary         Partner
     path       Path   /dev/        Host                                Controller      Controller
    state       type    node          HBA                                     port            port
--------- ---------- ------- ------------ ---------------------------------------- ---------------
     GOOD  primary      sddx        host1                                          0d              --
     GOOD  primary       sds        host0                                          0c              --
     GOOD  secondary    sddk        host1                                          --              0d
     GOOD  secondary     sdf        host0                                          --              0c

btcppe181:/vol/oralog1/oralog1.lun (LUN 2)              Lun state: GOOD
Lun Size:     75g (80530636800)  Controller_CF_State: Cluster Enabled
Protocol: FCP           Controller Partner: btcppe182
DM-MP DevName: oralog1.lun      (360a98000486e5851636f544658793476)     dm-2
Multipath-provider: NATIVE
```

```
--------- ---------- ------- ------------ ------------------------------------------- --------------
   sanlun Controller                                                      Primary         Partner
     path       Path   /dev/         Host                                Controller      Controller
    state       type    node          HBA                                     port            port
--------- ---------- ------- ------------ ------------------------------------------- --------------
     GOOD  primary       sdc        host0                                       0c              --
     GOOD  primary      sddh        host1                                       0d              --
     GOOD  secondary    sddu        host1                                       --              0d
     GOOD  secondary     sdp        host0                                       --              0c
```

Command: `sanlun fcp show adapter -v`

Output:

```
adapter name:      host0
WWPN:              2100001b321094c2
WWNN:              2000001b321094c2
driver name:       qla2xxx
model:             QLE2462
model description: QLogic QLE2462
serial number:     RFC0808K98388
hardware version:  PX2510401-23  C
driver version:    v.8.02.23
firmware version:  v. 4.06.03
Number of ports:   1
port type:         Fabric
port state:        Operational
supported speed:   1 GBit/sec, 2 GBit/sec, 4 GBit/sec
negotiated speed:  2 GBit/sec
OS device name:    /proc/scsi/qla2xxx/0

adapter name:      host1
WWPN:              2101001b323094c2
WWNN:              2001001b323094c2
driver name:       qla2xxx
model:             QLE2462
model description: QLogic QLE2462
serial number:     RFC0808K98388
hardware version:  PX2510401-23  C
driver version:    v.8.02.23
firmware version:  v. 4.06.03
Number of ports:   1
port type:         Fabric
port state:        Operational
supported speed:   1 GBit/sec, 2 GBit/sec, 4 GBit/sec
negotiated speed:  2 GBit/sec
OS device name:    /proc/scsi/qla2xxx/1
```

Command: multipath –ll

Output:

```
oralog1.lun (360a98000486e5851636f544658793476) dm-2 NETAPP,LUN
[size=75G][features=0][hwhandler=0][rw]
\_ round-robin 0 [prio=100][active]
 \_ 0:0:0:2   sdc  8:32    [active][ready]
 \_ 1:0:0:2   sddh 70:240  [active][ready]
\_ round-robin 0 [prio=20][enabled]
 \_ 1:0:57:2  sddu 71:192  [active][ready]
 \_ 0:0:57:2  sdp  8:240   [active][ready]
oralog2.lun (360a98000486e58514c34544b55465034) dm-9 NETAPP,LUN
[size=75G][features=0][hwhandler=0][rw]
\_ round-robin 0 [prio=100][active]
 \_ 1:0:57:5  sddx 71:240  [active][ready]
 \_ 0:0:57:5  sds  65:32   [active][ready]
\_ round-robin 0 [prio=20][enabled]
 \_ 1:0:0:5   sddk 71:32   [active][ready]
 \_ 0:0:0:5   sdf  8:80    [active][ready]
oradata1.lun (360a98000486e5851636f544658793478) dm-0 NETAPP,LUN
[size=380G][features=0][hwhandler=0][rw]
\_ round-robin 0 [prio=100][active]
 \_ 0:0:0:0   sda  8:0     [active][ready]
 \_ 1:0:0:0   sddf 70:208  [active][ready]
\_ round-robin 0 [prio=20][enabled]
 \_ 1:0:57:0  sdds 71:160  [active][ready]
 \_ 0:0:57:0  sdn  8:208   [active][ready]
votedisk1.lun (360a98000486e5851636f54465879347a) dm-15 NETAPP,LUN
[size=3.0G][features=0][hwhandler=0][rw]
\_ round-robin 0 [prio=100][active]
 \_ 1:0:0:8   sddn 71:80   [active][ready]
 \_ 0:0:0:8   sdi  8:128   [active][ready]
\_ round-robin 0 [prio=20][enabled]
 \_ 1:0:57:8  sdea 128:32  [active][ready]
 \_ 0:0:57:8  sdv  65:80   [active][ready]
oradata2.lun (360a98000486e58514c34544b55465033) dm-6 NETAPP,LUN
[size=380G][features=0][hwhandler=0][rw]
\_ round-robin 0 [prio=100][active]
 \_ 1:0:57:3  sddv 71:208  [active][ready]
 \_ 0:0:57:3  sdq  65:0    [active][ready]
\_ round-robin 0 [prio=20][enabled]
 \_ 1:0:0:3   sddi 71:0    [active][ready]
 \_ 0:0:0:3   sdd  8:48    [active][ready]
oraocr1.lun (360a98000486e5851636f544658793530) dm-11 NETAPP,LUN
[size=3.0G][features=0][hwhandler=0][rw]
\_ round-robin 0 [prio=100][active]
 \_ 1:0:0:6   sddl 71:48   [active][ready]
 \_ 0:0:0:6   sdg  8:96    [active][ready]
\_ round-robin 0 [prio=20][enabled]
 \_ 1:0:57:6  sddy 128:0   [active][ready]
 \_ 0:0:57:6  sdt  65:48   [active][ready]
```

```
votedisk2.lun (360a98000486e58514c34544b5546502d) dm-18 NETAPP,LUN
[size=3.0G][features=0][hwhandler=0][rw]
\_ round-robin 0 [prio=100][active]
 \_ 1:0:57:9  sdeb 128:48  [active][ready]
 \_ 0:0:57:9  sdw  65:96   [active][ready]
\_ round-robin 0 [prio=20][enabled]
 \_ 1:0:0:9   sddo 71:96   [active][ready]
 \_ 0:0:0:9   sdj  8:144   [active][ready]
oraarch1.lun (360a98000486e5851636f544658793535) dm-1 NETAPP,LUN
[size=155G][features=0][hwhandler=0][rw]
\_ round-robin 0 [prio=100][active]
 \_ 0:0:0:1   sdb  8:16    [active][ready]
 \_ 1:0:0:1   sddg 70:224  [active][ready]
\_ round-robin 0 [prio=20][enabled]
 \_ 1:0:57:1  sddt 71:176  [active][ready]
 \_ 0:0:57:1  sdo  8:224   [active][ready]
oraocr2.lun (360a98000486e58514c34544b55465041) dm-13 NETAPP,LUN
[size=3.0G][features=0][hwhandler=0][rw]
\_ round-robin 0 [prio=100][active]
 \_ 1:0:57:7  sddz 128:16  [active][ready]
 \_ 0:0:57:7  sdu  65:64   [active][ready]
\_ round-robin 0 [prio=20][enabled]
 \_ 1:0:0:7   sddm 71:64   [active][ready]
 \_ 0:0:0:7   sdh  8:112   [active][ready]
oraarch2.lun (360a98000486e58514c34544b55465038) dm-8 NETAPP,LUN
[size=155G][features=0][hwhandler=0][rw]
\_ round-robin 0 [prio=100][active]
 \_ 1:0:57:4  sddw 71:224  [active][ready]
 \_ 0:0:57:4  sdr  65:16   [active][ready]
\_ round-robin 0 [prio=20][enabled]
 \_ 1:0:0:4   sddj 71:16   [active][ready]
 \_ 0:0:0:4   sde  8:64    [active][ready]
```

Command: `crs_stat –t`

Output:

```
Name           Type           Target    State     Host
------------------------------------------------------------
ora.ext.db     application    ONLINE    ONLINE    node2
ora....t1.inst application    ONLINE    ONLINE    node1
ora....t2.inst application    ONLINE    ONLINE    node2
ora....SM1.asm application    ONLINE    ONLINE    node1
ora....ER.lsnr application    ONLINE    ONLINE    node1
ora....E1.lsnr application    ONLINE    ONLINE    node1
ora.node1.gsd  application    ONLINE    ONLINE    node1
ora.node1.ons  application    ONLINE    ONLINE    node1
ora.node1.vip  application    ONLINE    ONLINE    node1
ora....SM2.asm application    ONLINE    ONLINE    node2
ora....E2.lsnr application    ONLINE    ONLINE    node2
ora.node2.gsd  application    ONLINE    ONLINE    node2
ora.node2.ons  application    ONLINE    ONLINE    node2
ora.node2.vip  application    ONLINE    ONLINE    node2
```

Command: `ocrcheck`

Output:

```
Status of Oracle Cluster Registry is as follows:
         Version                  :          2
         Total space (kbytes)     :    3139148
         Used space (kbytes)      :       4920
         Available space (kbytes) :    3134228
         ID                       : 1633336445
         Device/File Name         : /dev/raw/raw1
                                    Device/File integrity check succeeded
         Device/File Name         : /dev/raw/raw2
                                    Device/File integrity check succeeded

         Cluster registry integrity check succeeded
```

Command: `crsctl query css votedisk`

Output:

```
 0.     0    /dev/raw/raw4
 1.     0    /dev/raw/raw5
 2.     0    /votedisk3/votedisk3.crs

located 3 votedisk(s).
```

14 AUTHORS

Antonio Jose Rodrigues Neto, Consulting Systems Engineer, NetApp

Jeffrey Steiner, PS Consultant, NetApp

Jim Lanson, Sr. Technical Marketing Engineer, Data Protection Solutions, NetApp

Karthikeyan Nagalingam, Product and Partner Engineer, NetApp

Lou Lydiksen, Consulting Systems Engineer, NetApp

Neil Gerren, Consulting Systems Engineer, NetApp

15 ACKNOWLEDGEMENTS

Various teams from NetApp and Oracle have helped to a great extent to develop this paper. Their invaluable guidance and participation in all phases made sure of a technical report that relied on their real-world experience and expertise. This paper would like to acknowledge the help received particularly from the experts:

Bill Heffelfinger, Cloud, Database, and Business Apps Global Field Technology Lead, NetApp

Jorge Costa, EMEA Solutions Specialists, NetApp

Lee Dorrier, Director, VEABU, NetApp

Lynne Thieme, Senior Manager, NetApp

Matt Mercer, Staff Engineer, NetApp

Michael Kiernan, Consulting Systems Engineer, NetApp

Rajashekhar A, SAN Linux Integration Engineer, NetApp

Steve Daniel, Director Database Platforms and Performance Technology, NetApp

Steven Schuettinger, Technical Alliance Manager, NetApp

Uday Shet, Senior Manager, NetApp

NetApp provides no representations or warranties regarding the accuracy, reliability or serviceability of any information or recommendations provided in this publication, or with respect to any results that may be obtained by the use of the information or observance of any recommendations provided herein. The information in this document is distributed AS IS, and the use of this information or the implementation of any recommendations or techniques herein is a customer's responsibility and depends on the customer's ability to evaluate and integrate them into the customer's operational environment. This document and the information contained herein may be used solely in connection with the NetApp products discussed in this document.

www.netapp.com