Redbooks
ibm.com/redbooks

High Availability for Oracle Database with IBM PowerHA SystemMirror and IBM Spectrum Virtualize HyperSwap

Ian MacQuarrie

Storage

IBM

Redpaper

High Availability for Oracle Database with IBM PowerHA SystemMirror and IBM Spectrum Virtualize HyperSwap

This IBM® Redpaper™ publication describes the use of the IBM Spectrum™ Virtualize HyperSwap® function to provide a high availability (HA) storage infrastructure for Oracle databases across metro distances, using the IBM SAN Volume Controller. The HyperSwap function is available on all IBM storage technologies that use IBM Spectrum Virtualize™ software, which include the IBM SAN Volume Controller, IBM Storwize® V5000, IBM Storwize V7000, IBM FlashSystem® V9000, and IBM Spectrum Virtualize as software.

The following reference architectures are covered:

- Active-Passive Oracle on IBM PowerHA® with HyperSwap

  Single instance Oracle database on IBM PowerHA SystemMirror® with SAN Volume Controller HyperSwap.

- Active-Active Oracle RAC with HyperSwap

  Oracle RAC Extended Distance Cluster with SAN Volume Controller HyperSwap.

This paper focuses on the functional behavior of HyperSwap when subjected to various failure conditions and provides detailed timings and error recovery sequences that occur in response to these failure conditions.

This paper does not provide the details necessary to implement the reference architectures (although some implementation detail is provided).

© Copyright IBM Corp. 2017. All rights reserved.

Introduction to HyperSwap

The IBM HyperSwap function is a high availability feature that provides dual-site, active-active access to volumes. HyperSwap functions are available on all systems running IBM Spectrum Virtualize that are supporting more than one I/O group.

HyperSwap volumes have a copy at one site and a copy at another site. Data that is written to the volume is automatically sent to both copies; if one site is no longer available, either site can provide access to the volume.

To construct HyperSwap volumes, active-active relationships are created between copies at each site. These relationships automatically run and switch direction according to which copy or copies are online and up to date. The relationships provide access to whichever copy is up to date through a single volume. Relationships can be grouped into consistency groups in the same way as Metro Mirror and Global Mirror relationships.

When the system topology is set for HyperSwap, each node, controller, and host in the system configuration must have a site attribute set to 1 or 2. Both nodes of an I/O group must be at the same site along with the controller that is providing the managed disks to that I/O group. When managed disks are added to storage pools, their site attributes must match. This requirement ensures that each copy in a HyperSwap volume is fully independent and is at a distinct site.

The Small Computer System Interface (SCSI) protocol allows storage devices to indicate the preferred ports for hosts to use when they submit I/O requests. Using the Asymmetric Logical Unit Access (ALUA) state for a volume, a storage controller can inform the host of which paths are active and which ones are preferred. In a HyperSwap system topology, the system suggests that the host use “local” nodes over remote nodes. A *local node* is a node that is configured at the same site as the host.

Figure 1 shows an overview of a generic HyperSwap topology.

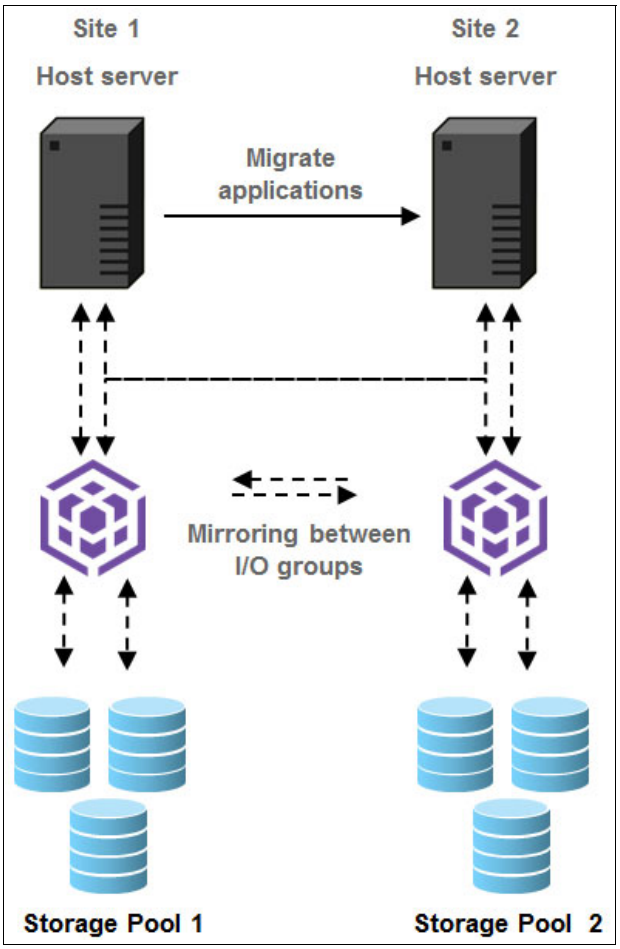

Figure 1 HyperSwap system topology

Implications of inter-site latency on performance

The HyperSwap cluster maintains a primary and secondary volume relationship as determined by the replication direction. When the replication direction is from Site1 to Site2 (Site1 → Site2), the primary volumes will be on Site1; when the replication direction is from Site2 to Site1 (Site2 → Site1), the primary volumes will be on Site2.

Read operations that occur on primary volumes will be satisfied directly from the respective site, whereas read operations that occur on secondary volumes will be forwarded to the remote site for execution and will therefore incur additional latency attributed to the inter-site latency. HyperSwap minimizes the impact of inter-site latency by automatically switching the replication direction to favor the site with greater than 75% of the write operations.

With active-active workloads, where I/O is active to both I/O groups simultaneously, a performance difference will occur between the two I/O groups with an elevation in read response time on the secondary I/O group of a magnitude relative to the inter-site latency.

Write operations that occur on either primary or secondary volumes will incur additional latency attributed to the inter-site latency. This is because of the requirement for all writes to be mirrored between sites.

With shorter distances between sites, where inter-site latency is low, the impact that read-forwarding and write-mirroring have on volume performance is negligible. With longer distances and associated higher inter-site latency, the impact on volume performance becomes significant.

For this paper, the distance was simulated using 10 km fiber cables between I/O groups, which introduced approximately 100 microseconds between sites. Although the impact this latency has on volume response times is small, it does show in the test results and demonstrates the influence that inter-site latency will have on HyperSwap in its various nominal and degraded states.

Criteria for replication reversal

HyperSwap will reverse the replication direction when 75% of the total write I/O is processed by the secondary volumes of the replication relationship for a sustained period. A *sustained* period can be as long as 20 minutes or as short as 2.5 minutes depending on the I/O distribution across sites. For cases where the primary site consistently performs 75% of the write activity, such as with an active-passive configuration the reverse will occur 20 minutes following a workload shift to the secondary volumes.

In cases where the workload is more evenly balanced between sites, such as with active/active configurations that have instances where 75% of the write activity shifts to the secondary volumes for short periods, the reverse can occur in under 20 minutes, and as low as 2.5 minutes following a workload shift to the secondary volumes. When a reverse occurs, the volumes swap their primary and secondary roles and maintain that relationship until the write I/O criteria is met again.

Note regarding performance evaluation

A slight difference exists in nominal response times between Site1 and Site2 that were considered in the performance sections of the test case summaries. To remove these differences from the evaluation of HyperSwap influence on response time, we use the phrase “return to nominal response time for SiteX.”

HyperSwap configuration

Figure 2 shows the reference HyperSwap configuration used for this paper.

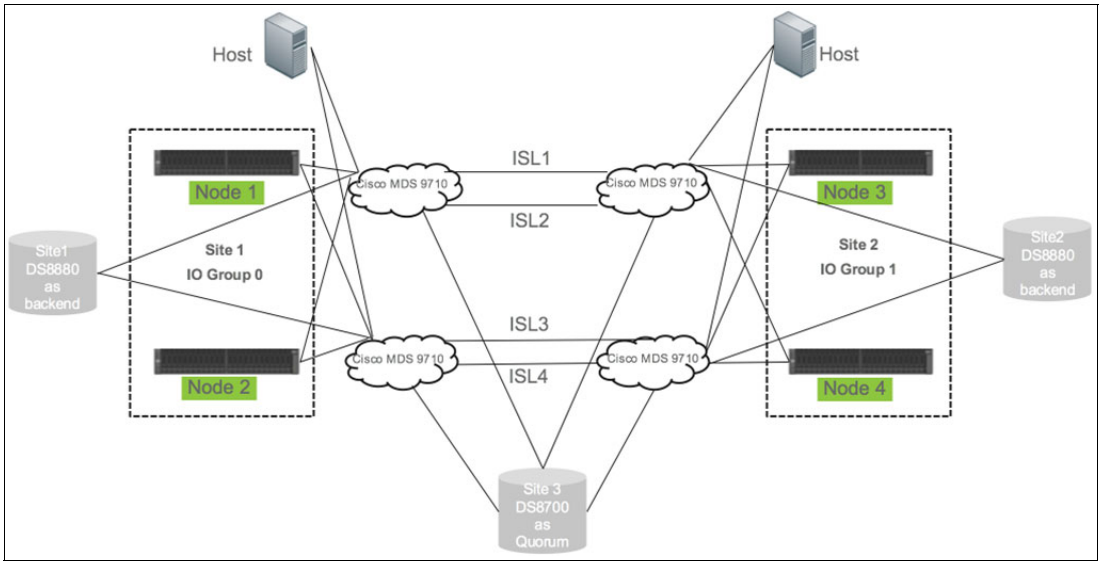

Figure 2 HyperSwap configuration

Hardware

- LPARs: 2 logical partitions (LPARs) on IBM Power Server P770 (memory: 16 GB; CPU: 2 cores; host bus adapter (HBA): 8 GB Fibre Channel (FC))
- Nodes: 4 SAN Volume Controller DH8 nodes (2 I/O groups); HBA: 8 GB FC; cache: 32 GB per node
- Pools: 2 DS8880 SSD pools for back end at Site1 and Site2, 1 DS8700 at Site3 for quorum
- Switches: 4 8-GB FC Cisco MDS switches
- Cables: 10 km fiber cables for inter-switch links (ISLs) used to add metro distance latency between I/O groups

Software

- Operating system: AIX7.1 TL4 SP3 (SAN boot)
- Multipath: Multipath I/O (MPIO) using default AIXPCM
- Cluster: PowerHA SystemMirror v7.2 standard edition (two-node cluster)
- Database: Oracle 12c R1
- SAN Volume Controller/Storwize microcode level 7.8.1.0
- Disk based quorum
- IBM Spectrum Control™ v5.2.13.0

Host specifications

- Host type: IBM 9117-MMC
- RAM: 16 GB
- Processor type: PowerPC_POWER7
- Processor speed: 3304 MHz
- Number of processors: 2
- Threads per processor: 4
- System firmware: AM770_112

AIX hdisk parameters

- algorithm: shortest_queue
- queue_depth: 64
- reserve_policy: no_reserve

HBA specifications and parameters

- HBA make and model: IBM FC 5735
- HBAs: 1
- Ports per HBA: 2
- HBA speed: 8 Gbps, connected at 8 Gbps
- HBA firmware: US2.03X5
- HBA parameters: `fc_err_recov=fast_fail`, `dyntrk=yes`

SAN topology

Figure 3 shows the SAN topology diagram.

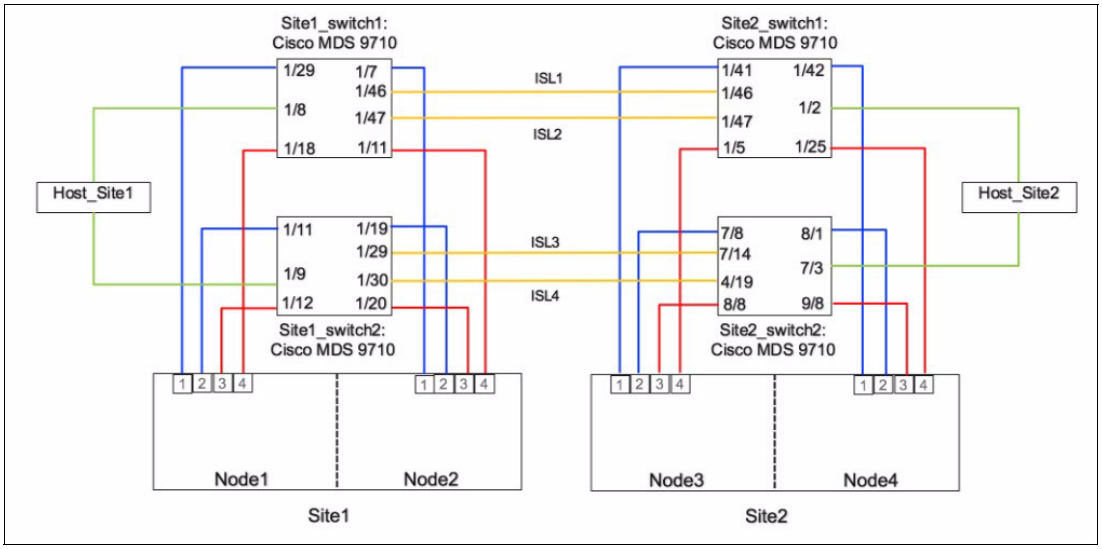

Figure 3 SAN topology

Public and private SAN

The topology has a public and a private SAN:

- Public SAN (VSAN 100), marked in blue on SAN topology diagram

  Site1_Node1_Port1, Site1_Node1_Port2, Site1_Node2_Port1, Site1_Node2_Port2, Site2_Node3_Port1, Site2_Node3_Port2, Site2_Node4_Port1, Site2_Node4_Port2

- Private SAN (VSAN 200), marked as red in SAN topology diagram

  Site1_Node1_Port3, Site1_Node1_Port4, Site1_Node2_Port3, Site1_Node2_Port4, Site2_Node3_Port3, Site2_Node3_Port4, Site2_Node4_Port3, Site2_Node4_Port4

- Local port masking

  Port3 and Port4 on each node are masked (`chsystem -localfcportmask 1100`)

ISL configuration

Configuration is as follows:

- Two ISLs (ISL1 and ISL3) in public SAN
- Two ISLs (ISL2 and ISL4) in private SAN

Host to SAN Volume Controller nodes

Nodes are as follows:

- Site1_host_zoning:
  - Zone_cfg1: Host1_Port1, Site1_Node1_Port1, Site1_Node2_Port1, Site2_Node3_Port1, Site2_Node4_Port1
  - Zone_cfg2: Host1_Port2, Site1_Node1_Port2, Site1_Node2_Port2, Site2_Node3_Port2, Site2_Node4_Port2
- Site2_host_zoning:
  - Zone_cfg1: Host2_Port1, Site1_Node1_Port1, Site1_Node2_Port1, Site2_Node3_Port1, Site2_Node4_Port1
  - Zone_cfg2: Host2_Port2, Site1_Node1_Port2, Site1_Node2_Port2, Site2_Node3_Port2, Site2_Node4_Port2

Active-Passive Oracle on PowerHA with HyperSwap

The following test cases are covered in this section:

- Test Case 1: Loss of access from SAN Volume Controller to back-end storage
- Test Case 2: SAN Volume Controller I/O group failure
- Test Case 3: PowerHA node failover
- Test Case 4: Site failure
- Test Case 5: Quorum site failure

Site configuration

Site1 is the primary production site. All I/O will be performed to Site1 under nominal (fully redundant) conditions. Site2 will be failed over to in response to various failure conditions introduced by the test cases.

Site1

- Server: ARCPMMC37D47P14 (P14)
- I/O group: 0 (IOG0)

Site2

- Server: ARCPMMC37D47P13 (P13)
- I/O group: 1 (IOG1)

Oracle database configuration

- A 50 GB SOE schema data file, single instance.
- ASM disk group consists of four 100 GB physical volumes (PVs) configured for HyperSwap.
- All volumes (hdisk8, hdisk9, hdisk10, hdisk11) reside in a single consistency group.

Test workload

The Swingbench load generator (v2.6) is used to generate transactions against the database. We measure and document any interruption that occurred in transaction processing as a result of any recovery actions that are invoked in response to failure conditions.

Figure 4 shows the benchmark configuration and connection pool settings.

Benchmark Configuration

Parameter	Value
Connect String	//9.11.101.15/orcl
Driver	oracle.jdbc.pool.OracleDataSource
Total run time	1:01:55
Number of users/threads logged on	200
Minimum inter sleep time	0
Maximum inter sleep time	0
Minimum intra sleep time	0
Maximum intra sleep time	0
Wait until users logon	true

Connection Pool Settings

Parameter	Value
Initial connection count	200
Minimum connection count	200
Maximum connection count	4000
Connection Wait Timeout (secs)	45
Abandoned Connection Timeout (secs)	240
Inactivity Connection Timeout (secs)	50
Property Check Time (secs)	10

Figure 4 Benchmark configuration and connection pool settings

Test Case 1: Loss of access from SAN Volume Controller to back-end storage

This test case looks at a loss of access from SAN Volume Controller to back-end storage.

Objective

Simulate the failure of a back-end storage controller.

Failure inject method

Remove access from IOG0 to back-end controller ports by using switch zoning.

Figure 5 shows loss of access from SAN Volume Controller to back-end storage.

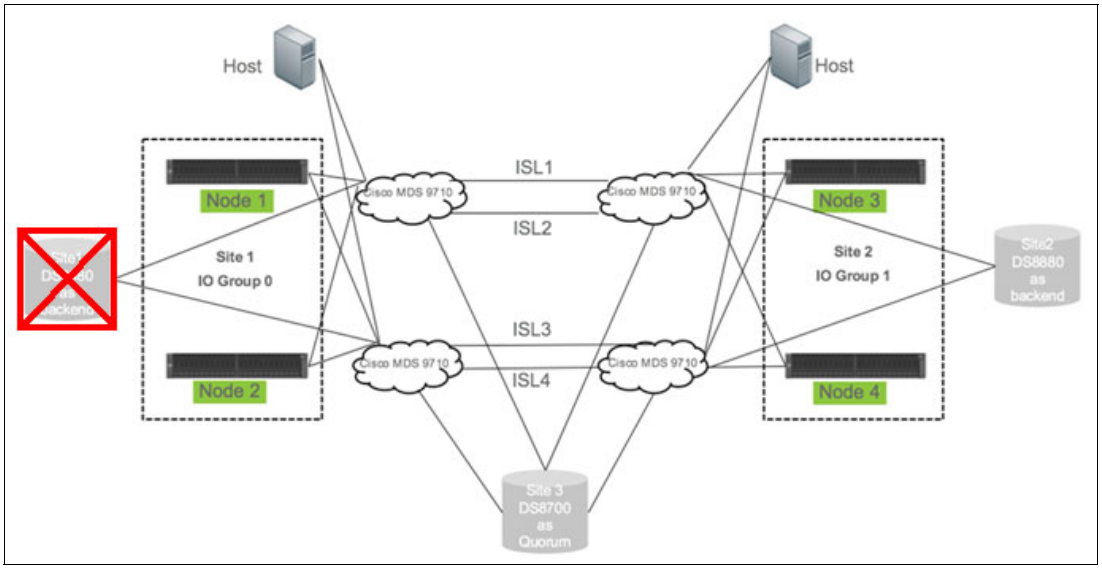

Figure 5 Test Case 1: Loss of access from SAN Volume Controller to back-end storage (part 1 of 3)

HyperSwap state changes

- Replication reverses and suspends.
- Host I/O continues from host P14 at Site1.
- Front-end volume I/O remains on IOG0 at Site1 and I/O is forwarded to IOG1 at Site2.

MPIO path state changes

- Path failure/recovery.
- Active paths remain on preferred paths to IOG0.

Host error recovery

Table 1 shows the host error recovery.

Table 1 Host error recovery

Host	Label	Description
P14	SC_DISK_ERR4	Timeout
P14	SC_DISK_ERR7	Path failure
P14	SC_DISK_ERR9	Path recovery

Interruption to I/O processing

A 32-second pause occurs during replication-suspend.

Figure 6 shows loss of access from SAN Volume Controller to back-end storage.

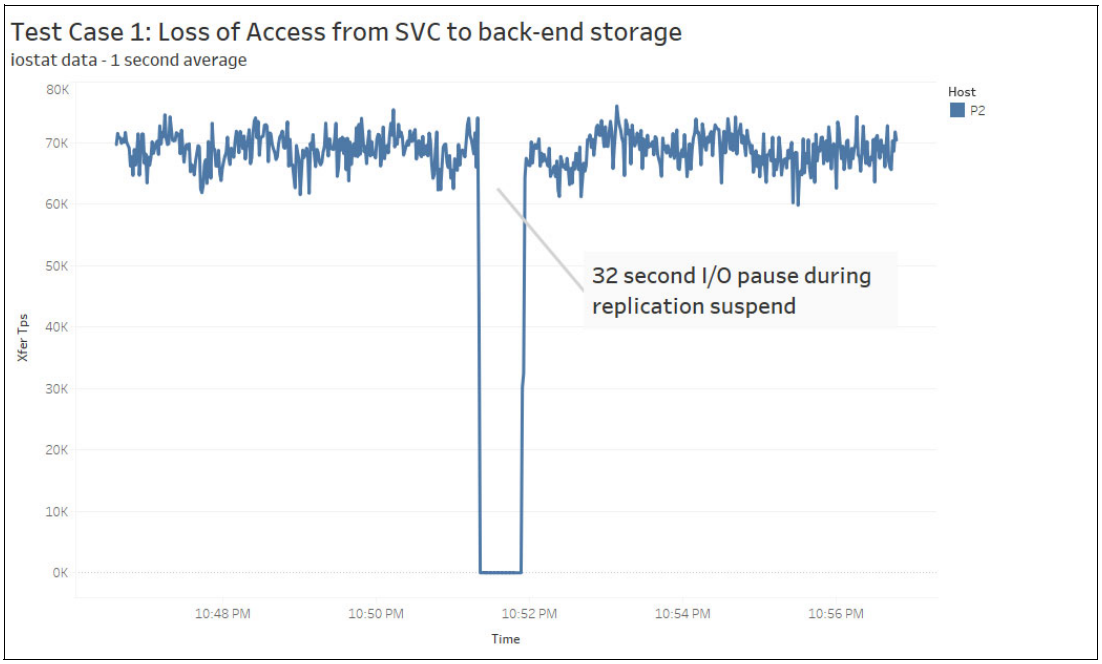

Figure 6 Test Case 1: Loss of access from SAN Volume Controller to back-end storage (part of 2 of 3)

Performance

- Read response time increases due to I/O forwarding.
- Write response time drops due to suspension of mirroring (Figure 7).

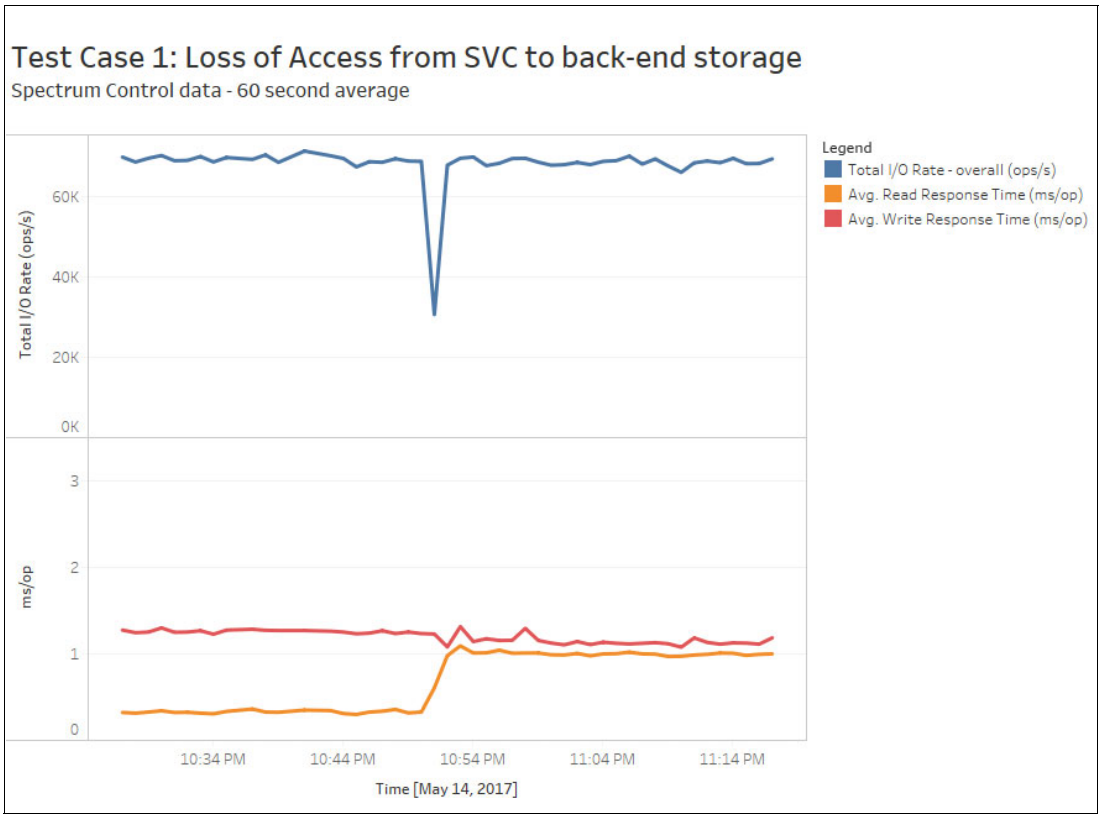

Figure 7 Test Case 1: Loss of access from SAN Volume Controller to back-end storage (part 3 of 3)

Swingbench statistics (Test Case 1)

Table 2 shows the Swingbench summary.

Table 2 Swingbench

Parameters	Value
Total completed transactions	337,372
Average transactions per second	90.81
Maximum transaction rate	8,653
Total failed Transactions	2

Test Case 2: SAN Volume Controller I/O group failure

This test case examines an SAN Volume Controller I/O group failure.

Objective

Simulate the failure of an SAN Volume Controller I/O group failure (both nodes down).

Failure inject method

Remove access to all ports on IOG0 by using switch zoning.

Figure 8 shows the SAN Volume Controller I/O group failure.

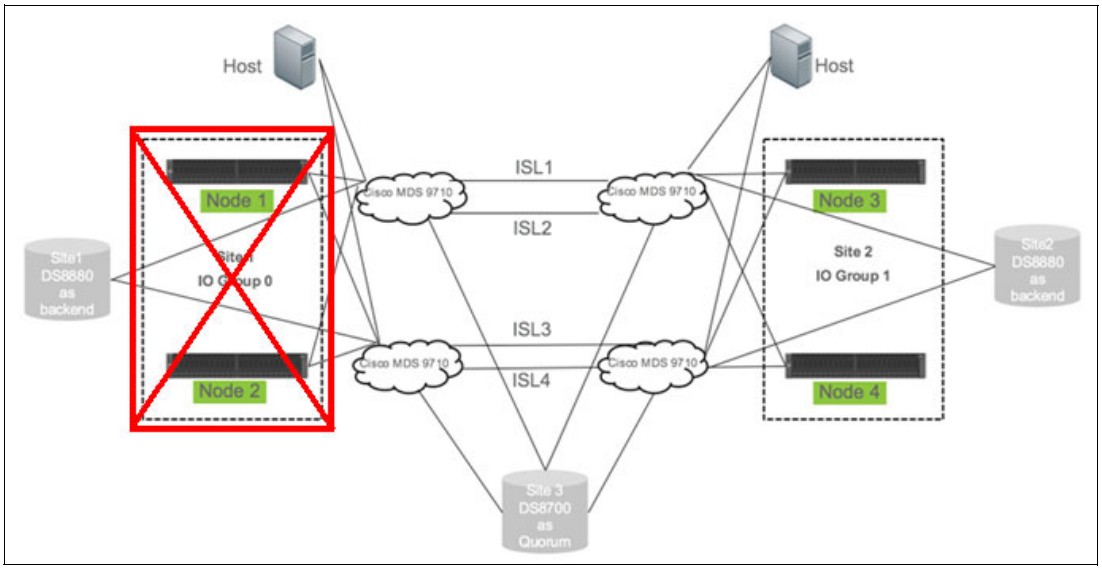

Figure 8 Test Case 2: SAN Volume Controller I/O group failure (part 1 of 3)

HyperSwap state changes

- Replication reverses and suspends.
- Host I/O continues from host P14 at Site1.
- Front-end volume I/O switches from IOG0 at Site1 to IOG1 at Site2.

MPIO path state changes

- Paths fail over from preferred paths on IOG0 to preferred paths on IOG1.

Path states before and after failure inject

Figure 9 shows the path states before/after failure inject.

```
root @ ARCPMMC37D47P14: /
# lsmpio -l hdisk8
name path_id status path_status parent connection
hdisk8 0 Enabled Sel,Opt  fscsi0 500507680140066f,9000000000000
hdisk8 1 Enabled Non      fscsi0 50050768014001ad,9000000000000
hdisk8 2 Enabled Non      fscsi0 500507680140064e,9000000000000
hdisk8 3 Enabled Non      fscsi0 50050768014001ae,9000000000000
hdisk8 4 Enabled Sel,Opt  fscsi1 500507680140066f,9000000000000
hdisk8 5 Enabled Non      fscsi1 50050768014001ad,9000000000000
hdisk8 6 Enabled Non      fscsi1 500507680140064e,9000000000000
hdisk8 7 Enabled Non      fscsi1 50050768014001ae,9000000000000

root @ ARCPMMC37D47P14: /
# lsmpio -l hdisk8
name path_id status path_status parent connection
hdisk8 0 Failed  Non,Deg,Fai fscsi0 500507680140066f,9000000000000
hdisk8 1 Failed  Non,Deg,Fai fscsi0 50050768014001ad,9000000000000
hdisk8 2 Enabled Non         fscsi0 500507680140064e,9000000000000
hdisk8 3 Enabled Sel,Opt     fscsi0 50050768014001ae,9000000000000
hdisk8 4 Failed  Non,Deg,Fai fscsi1 500507680140066f,9000000000000
hdisk8 5 Failed  Non,Deg,Fai fscsi1 50050768014001ad,9000000000000
hdisk8 6 Enabled Non         fscsi1 500507680140064e,9000000000000
hdisk8 7 Enabled Sel,Opt     fscsi1 50050768014001ae,9000000000000
```

Figure 9 Path states before and after failure inject

Host error recovery

Table 3 shows the host error recovery.

Table 3 Host error recovery

Host	Label	Description
P14	FCP_ERR14	Name server reject due to device being dropped from the fabric
P14	SC_DISK_ERR4	Timeout
P14	SC_DISK_ERR4	No device response
P14	SC_DISK_ERR7	Path failure

Interruption to I/O processing

A 60-second lapse in I/O occurs due to host error recovery that is associated with loss of access to IOG0. See Figure 10.

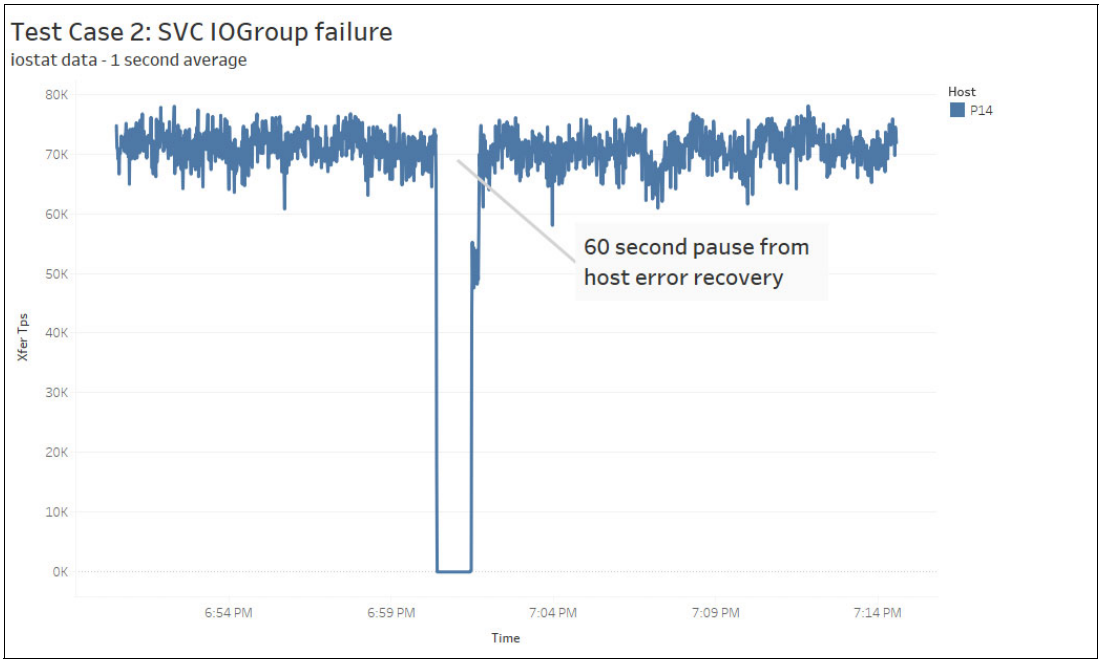

Figure 10 Test Case 2: SAN Volume Controller I/O group failure (part 2 of 3)

Performance

- Read response time shifts to nominal response times for Site2.
- Write response time drops due to suspension of mirroring (Figure 11).

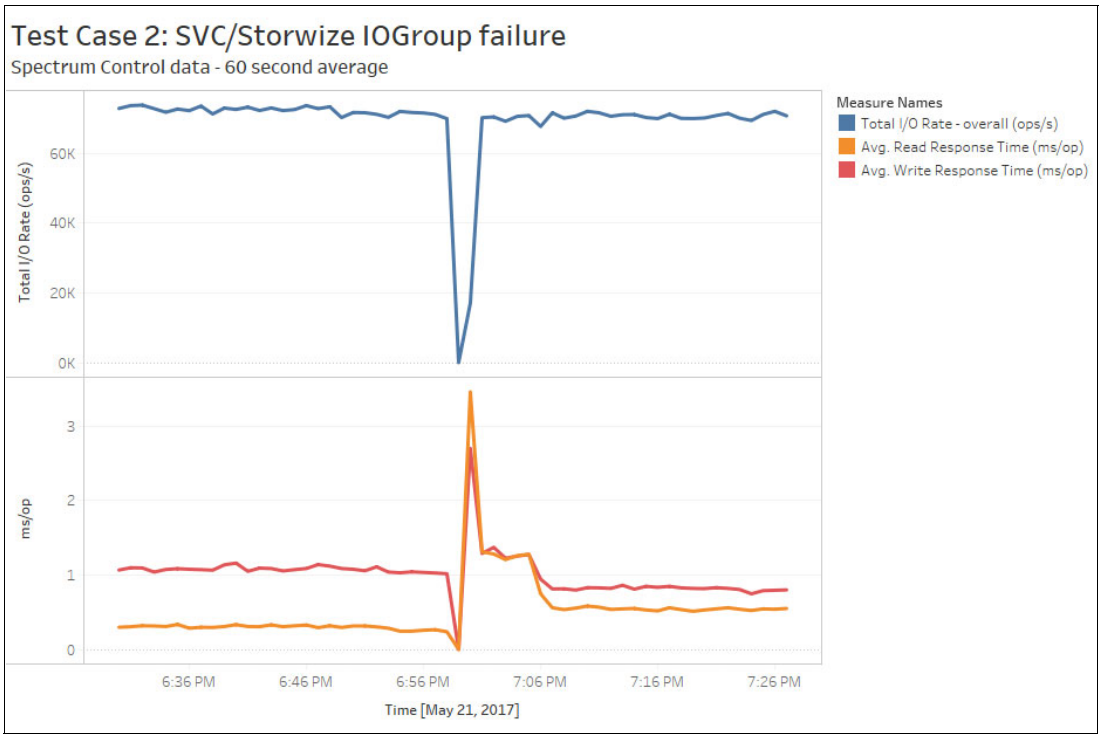

Figure 11 Test Case 2: SAN Volume Controller I/O group failure (part 3 of 3)

Swingbench statistics (Test Case 2)

Table 4 shows the Swingbench summary.

Table 4 Swingbench

Parameters	Value
Total completed transactions	775,956
Average transactions per second	85.34
Maximum transaction rate	8,624
Total failed transactions	3

Test Case 3: PowerHA node failover

This test case examines a PowerHA node failover.

Objective

Simulate the failure of a PowerHA node.

Failure inject method

Halt IBM AIX® kernel on node P14.

Figure 12 shows a PowerHA node failover.

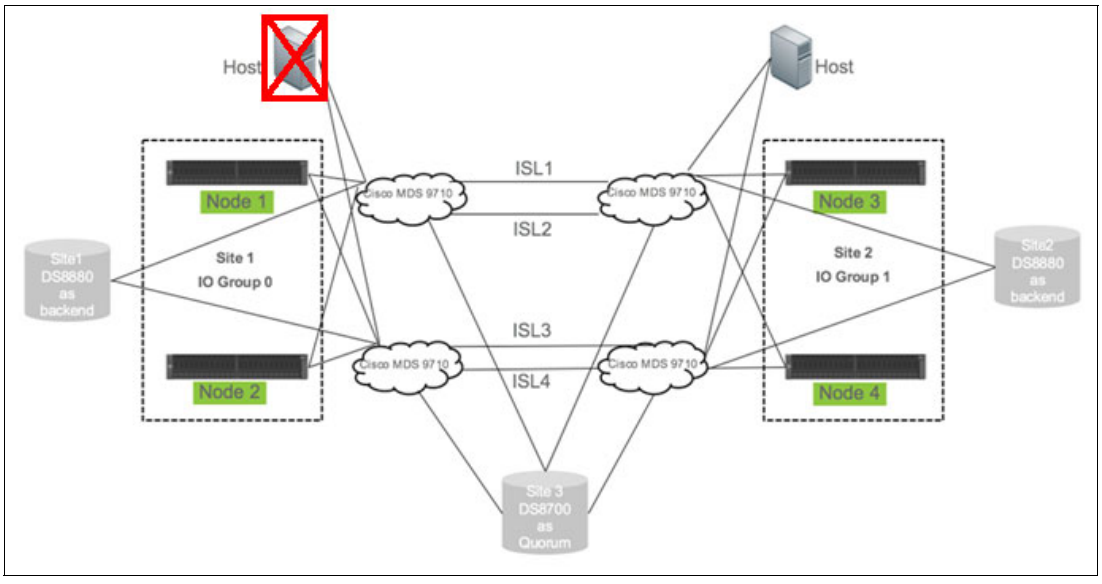

Figure 12 Test Case 3: PowerHA node failover (part 1 of 3)

HyperSwap state changes

- Host I/O switches from host P14 at Site1 to host P13 at Site2.
- Front-end volume I/O switches from IOG0 at Site1 to IOG1 at Site2.
- Replication reverses 20 minutes after PowerHA node failover.

MPIO path state changes

- None

Host error recovery

PowerHA failover occurs due to a node crash.

Interruption to I/O processing

- PowerHA failover time was 2 minutes and 52 seconds.
- A 27-second pause occurs 20 minutes after failover during replication reversal (Figure 13).

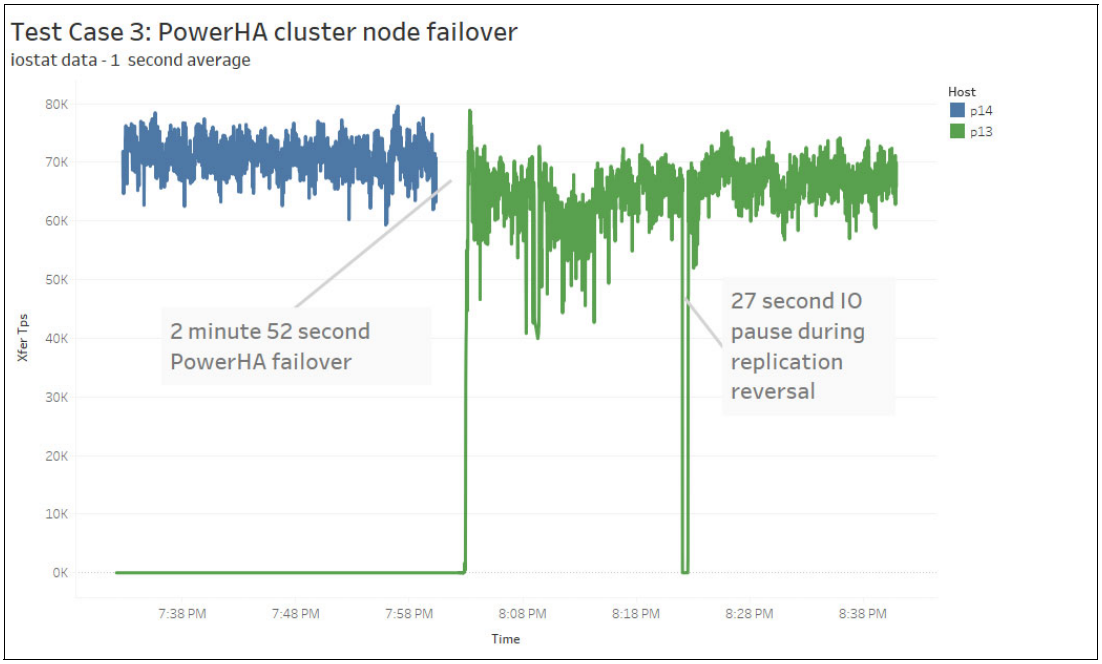

Figure 13 Test Case 3: PowerHA node failover (part 2 of 3)

Performance

- Read and write response times increase due to I/O forwarding.
- Response times return to nominal response times for Site2 after 20 minutes, when the I/O distribution criteria is met, and replication reverses (Figure 14).

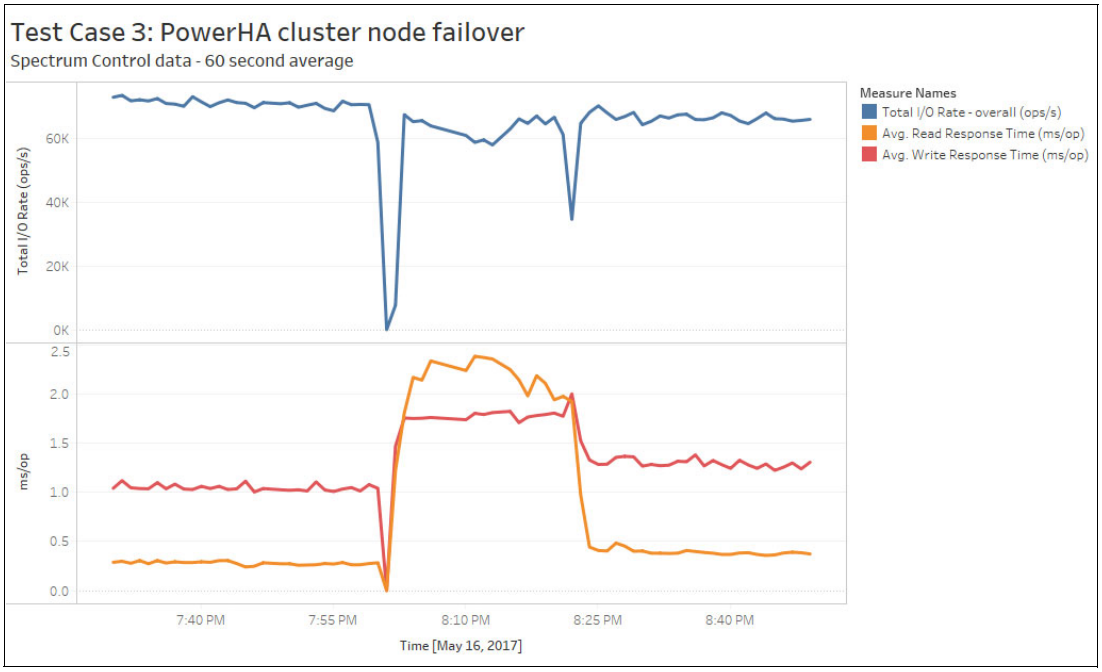

Figure 14 Test Case 3: PowerHA node failover (part 3 of 3)

Swingbench statistics (Test Case 3)

Table 5 shows the Swingbench summary.

Table 5 Swingbench

Parameters	Value
Total completed transactions	391,644
Average transactions per second	79.41
Maximum transaction rate	6,970
Total failed Transactions	12,728

Test Case 4: Site failure

This test case simulates a site failure.

Objective

Simulate the failure of an entire site.

Failure inject method

Simultaneously remove access to all ports on IOG0 using switch zoning and halt AIX kernel on node P14.

Figure 15 shows a site failure.

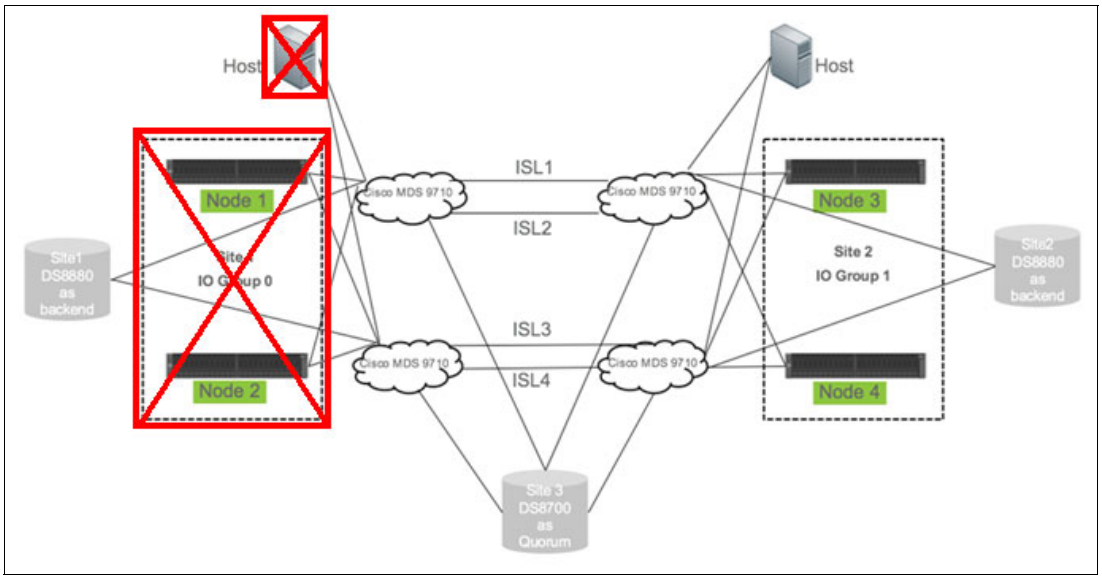

Figure 15 Test Case 4: Site failure (part 1 of 3)

HyperSwap state changes

- Replication reverses and suspends.
- Host I/O switches from host P14 at Site1 to host P13 at Site2.
- Front-end volume I/O switches from IOG0 at Site1 to IOG1 at Site2.

MPIO path state changes

- None

Host error recovery

PowerHA failover is due to a node crash.

Interruption to I/O processing

A 4-minute PowerHA failover time occurs. See Figure 16 on page 19.

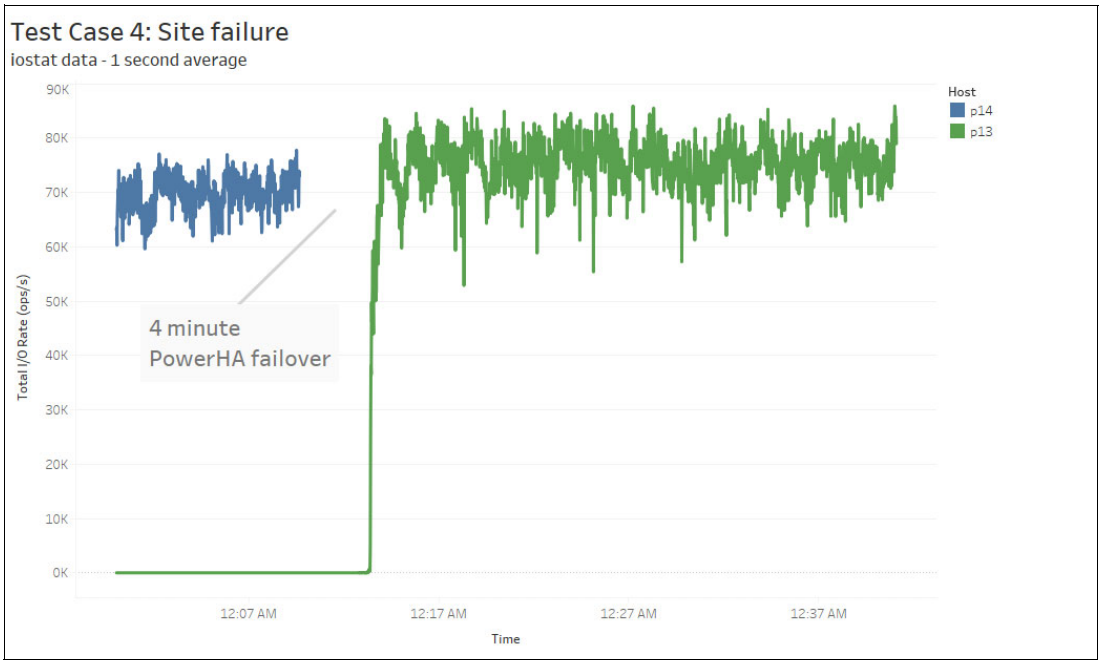

Figure 16 Test Case 4: Site failure (part 2 of 3)

Performance

- Read response time changes to nominal read response time for Site2.
- Write response time drops due to suspension of mirroring (Figure 17).

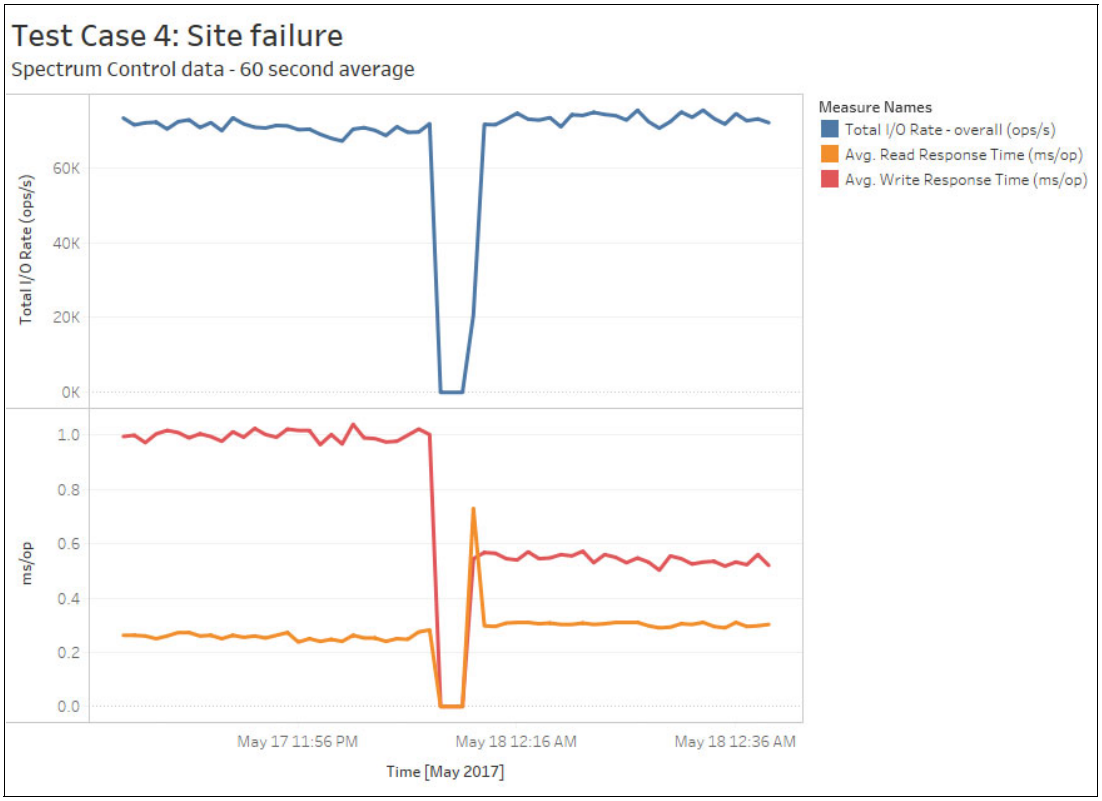

Figure 17 Test Case 4: Site failure (part 3 of 3)

Swingbench statistics (Test Case 4)

Table 6 shows the Swingbench summary.

Table 6 Swingbench

Parameters	Value
Total completed transactions	311,750
Average transactions per second	85.25
Maximum transaction rate	8,314
Total failed transactions	15,259

Test Case 5: Quorum site failure

This test case simulates a quorum site failure.

Objective

Simulate a failure or loss of access to the quorum site.

Failure inject method

Remove access to the quorum disk by using switch zoning (Figure 18).

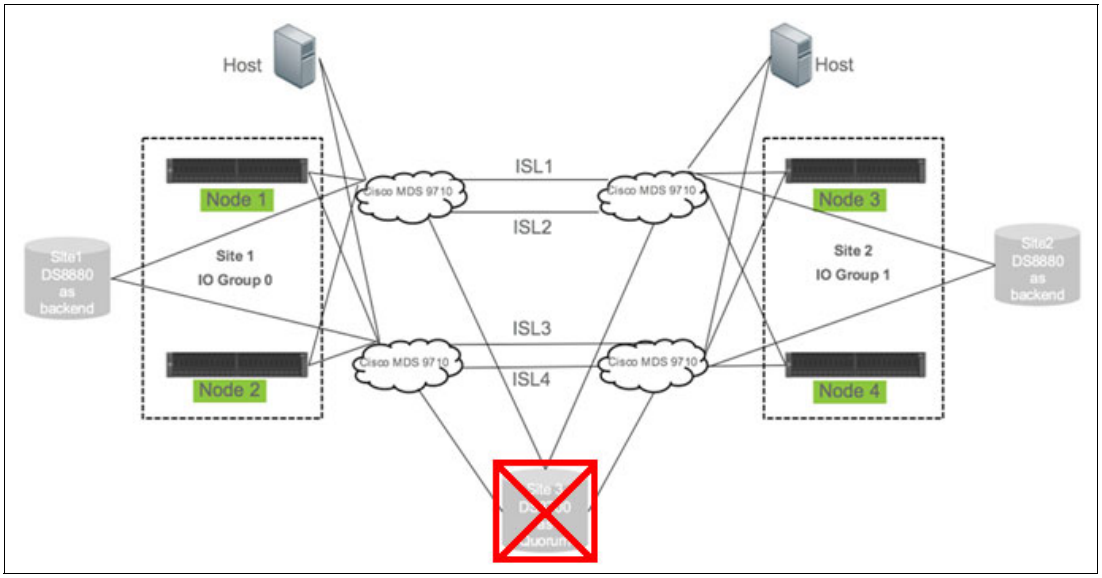

Figure 18 Test Case 5: Quorum site failure

Test results

Loss of access to the quorum disk was completely transparent. No error recovery or interruption in I/O occurred.

Active-Active Oracle RAC with HyperSwap

The following test cases are covered in this section:

- Test Case 6: Loss of access from SAN Volume Controller to back-end storage
- Test Case 7: SAN Volume Controller I/O group failure
- Test Case 8: RAC node failure
- Test Case 9: Site failure

Site configuration

Both Site1 and Site2 are running the production workload. All I/O will be split between sites under nominal (fully redundant) conditions. The Site1 workload will be failed over to Site2 in response to various failure conditions introduced by the test cases.

Site1

- Server: ARCPMMC37D47P2 (P2)
- I/O group: 0 (IOG0)

Site2

- Server: ARCPMMC37D47P3 (P3)
- I/O group: 1 (IOG1)

Oracle database configuration

- A 50 GB SOE schema data file, database instance orcl1 located on server P2 (Site1) and database instance orcl2 located on server P3 (Site2), SCAN listeners located on server P2.
- ASM disk group +DATA consists of four 100 GB PVs configured for HyperSwap (hdisk5, hdisk6, hdisk7, hdisk8).
- A separate ASM disk group configured for the voting disks. Three voting disks with each mapped to a separate storage controller.
- The ASM heartbeat timeout is set to 120 seconds:
  - Parameter| name: `_asm_hbeatiowait`
  - Instance value: 120

Test workload

The Swingbench load generator is used to generate a transaction against the database. Any interruption that occurs in transaction processing as a result of any recovery actions that are invoked in response to failure conditions are measured and documented.

Figure 19 shows the benchmark configuration and connection pool settings.

Benchmark Configuration

Parameter	Value
Connect String	//9.11.101.10/soe_ac
Driver	oracle.jdbc.replay.OracleDataSourceImpl
Total run time	1:08:21
Number of users/threads logged on	60
Minimum inter sleep time	50
Maximum inter sleep time	500
Minimum intra sleep time	0
Maximum intra sleep time	0
Wait until users logon	false

Connection Pool Settings

Parameter	Value
Initial connection count	40
Minimum connection count	20
Maximum connection count	60
Connection Wait Timeout (secs)	45
Abandoned Connection Timeout (secs)	240
Inactivity Connection Timeout (secs)	50
Property Check Time (secs)	10

Figure 19 Benchmark configuration and connection pool settings

Test Case 6: Loss of access from SAN Volume Controller to back-end storage

This test case simulates the loss of access from SAN Volume Controller to back-end storage.

Objective

Simulate the failure of a back-end storage controller.

Failure inject method

Remove access from IOG0 to back-end controller ports using switch zoning (Figure 20).

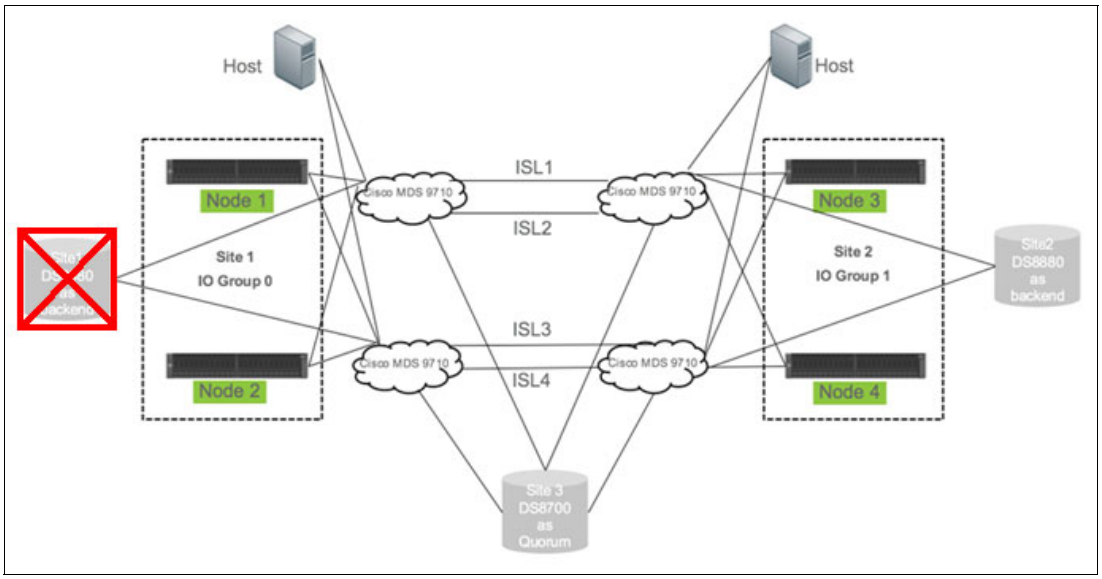

Figure 20 Test Case 6: Loss of access from SAN Volume Controller to back-end storage (part 1 of 3)

HyperSwap state changes

- Replication reverses and suspends.
- Host I/O continues from host P2 at Site1 and host P3 at Site2.
- Front-end volume I/O on IOG0 at Site1 is forwarded to IOG1 at Site2.

MPIO path state changes

- Path failure/recovery
- I/O remains on initial paths

Host error recovery

Table 7 shows the host error recovery.

Table 7 Host error recovery

Host	Label	Description
P2	SC_DISK_ERR4	Timeout
P2	SC_DISK_ERR7	Path failure
P2	SC_DISK_ERR9	Path recovery

Interruption to I/O processing

A 26-second I/O pause occurs during replication reverse and suspend (Figure 21).

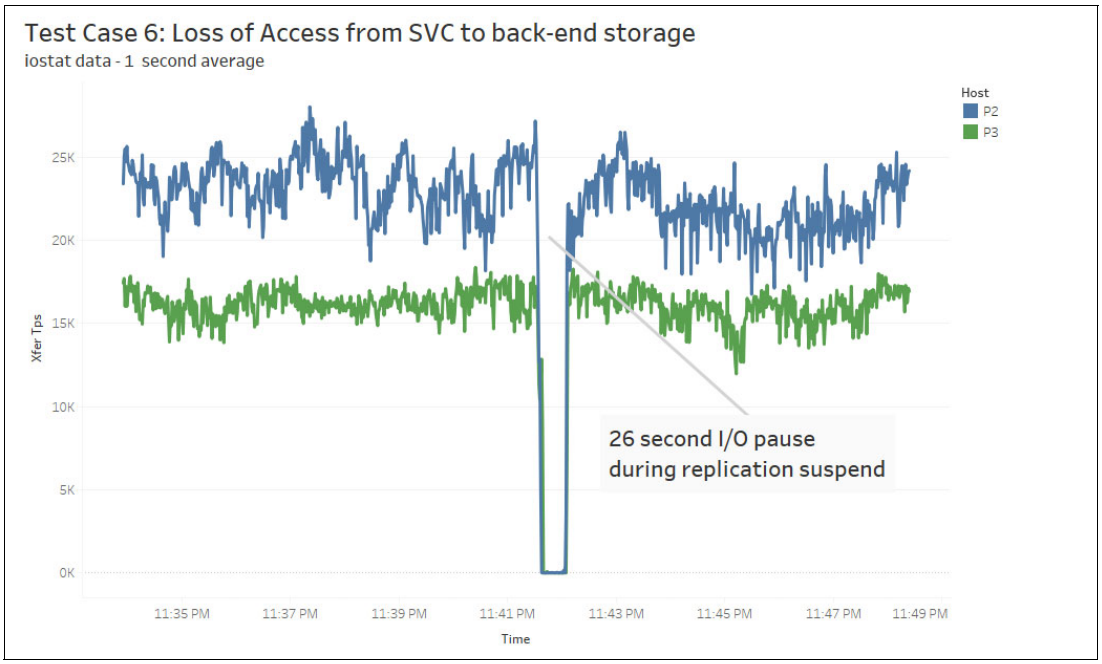

Figure 21 Test Case 6: Loss of access from SAN Volume Controller to back-end storage (part 2 of 3)

Performance

- Read response time increases on Site1 due to I/O forwarding.
- Write response time drops due to suspended mirroring (Figure 22).

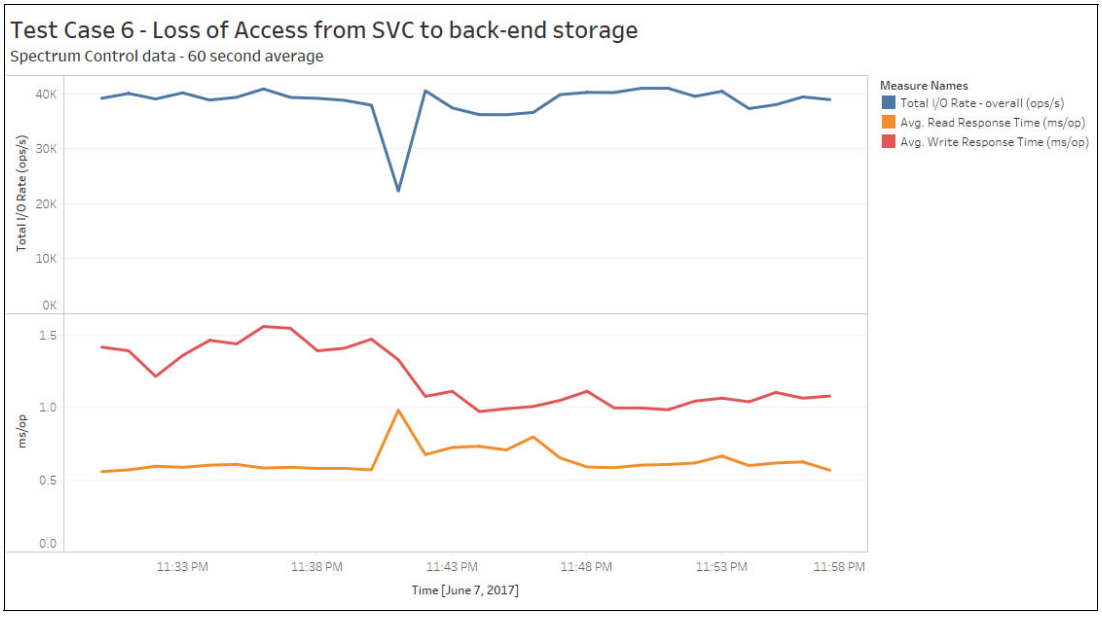

Figure 22 Test Case 6: Loss of access from SAN Volume Controller to back-end storage (part 3 of 3)

Swingbench statistics (Test Case 6)

Table 8 shows the Swingbench summary.

Table 8 Swingbench

Parameters	Value
Total completed transactions	106,576
Average transactions per second	25.99
Maximum transaction rate	3,046
Total failed Transactions	1

Test Case 7: SAN Volume Controller I/O group failure

This test case simulates a SAN Volume Controller I/O group failure.

Objective

Simulate the failure of a SAN Volume Controller I/O group failure (both nodes down).

Failure inject method

Remove access to all ports on IOG0 by using switch zoning (Figure 23).

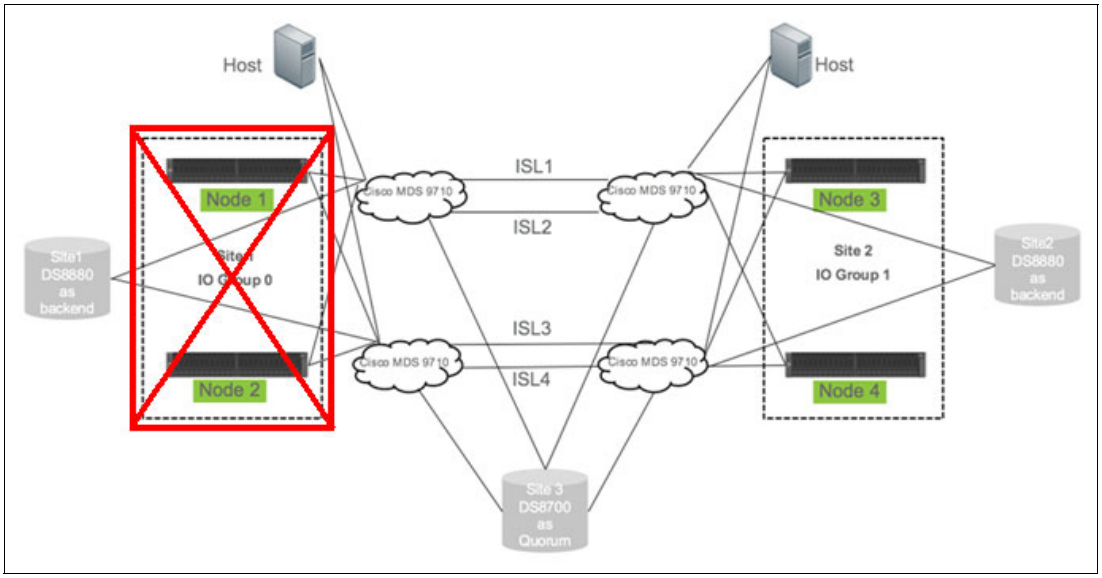

Figure 23 Test Case 7: SAN Volume Controller I/O group failure (part 1 of 3)

HyperSwap state changes

- Replication reverses and suspend.
- Host I/O continues from host P2 at Site1 and host P3 at Site2.
- Front-end volume I/O from host P2 switches from IOG0 at Site1 to IOG1 at Site2.
- Front-end volume I/O from host P3 remains on IOG1.

MPIO path state changes

- Active paths for host P2 paths fail over from preferred paths on IOG0 to preferred paths on IOG1.

Host error recovery

Table 9 shows the host error recovery.

Table 9 Host error recovery

Host	Label	Description
P2	FCP_ERR14	Name server reject due to device dropped from fabric
P2	SC_DISK_ERR4	Timeout
P2	SC_DISK_ERR4	No device response
P2	SC_DISK_ERR7	Path failure

Interruption to I/O processing

A 77-second I/O pause occurs during replication suspend and host failover (Figure 24).

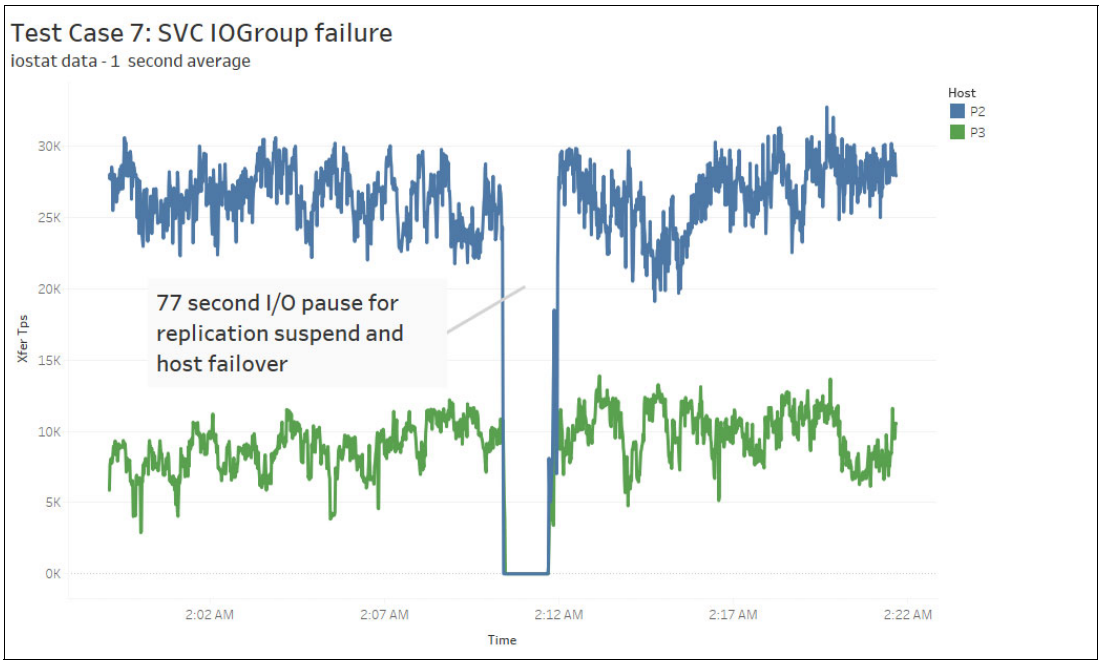

Figure 24 Test Case 7: SAN Volume Controller I/O group failure (part 2 of 3)

Performance

- Read response time returns to nominal response times for Site2.
- Write response time drops due to suspension of mirroring (Figure 25).

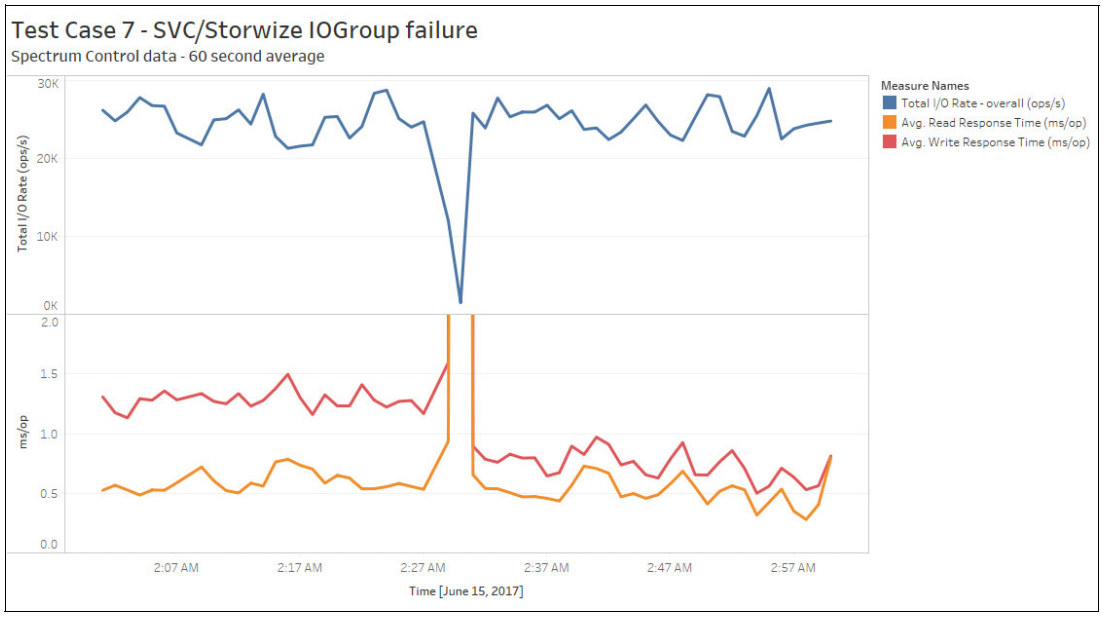

Figure 25 Test Case 7: SAN Volume Controller I/O group failure (part 3 of 3)

Swingbench statistics (Test Case 7)

Table 10 shows the Swingbench summary.

Table 10 Swingbench

Parameters	Value
Total completed transactions	106,991
Average transactions per second	29.00
Maximum transaction rate	3,448
Total failed Transactions	0

Test Case 8: RAC node failure

This test case simulates a RAC node failure.

Objective

Simulate the failure of an Oracle RAC node.

Failure inject method

Halt AIX kernel on node P2 (Figure 26).

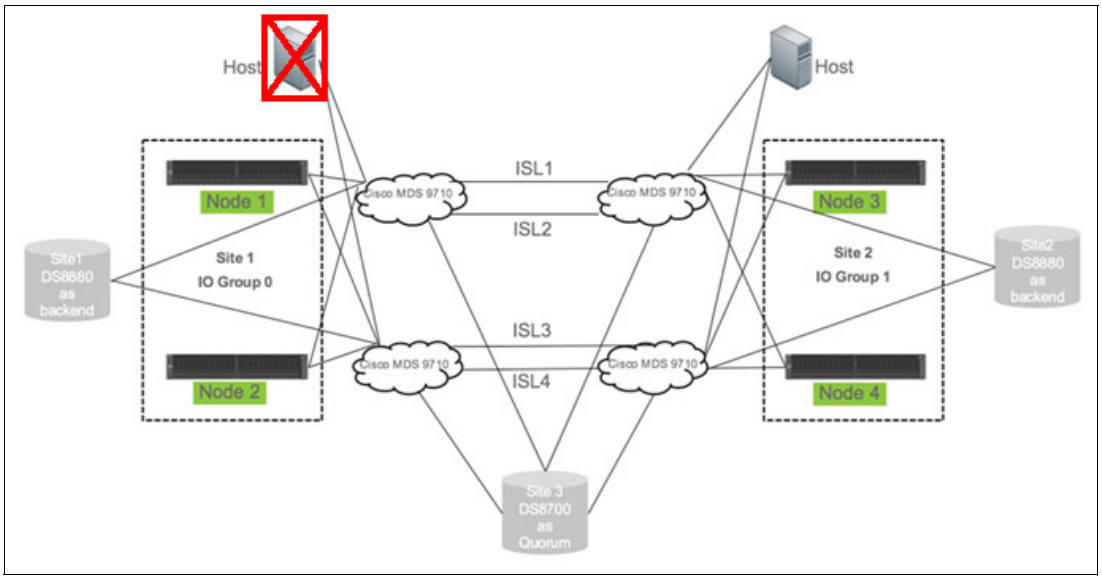

Figure 26 Test Case 8: RAC node failure (part 1 of 3)

HyperSwap state changes

- Replication reverses when IOG1 meets greater than 75% criteria after I/O halts on IOG0.

MPIO path state changes

- None

Host error recovery

- Failed node is evicted from RAC cluster.

Interruption to I/O processing

- A 20-second pause from RAC when node is evicted.
- A 28-second pause when replication reverses (Figure 27).

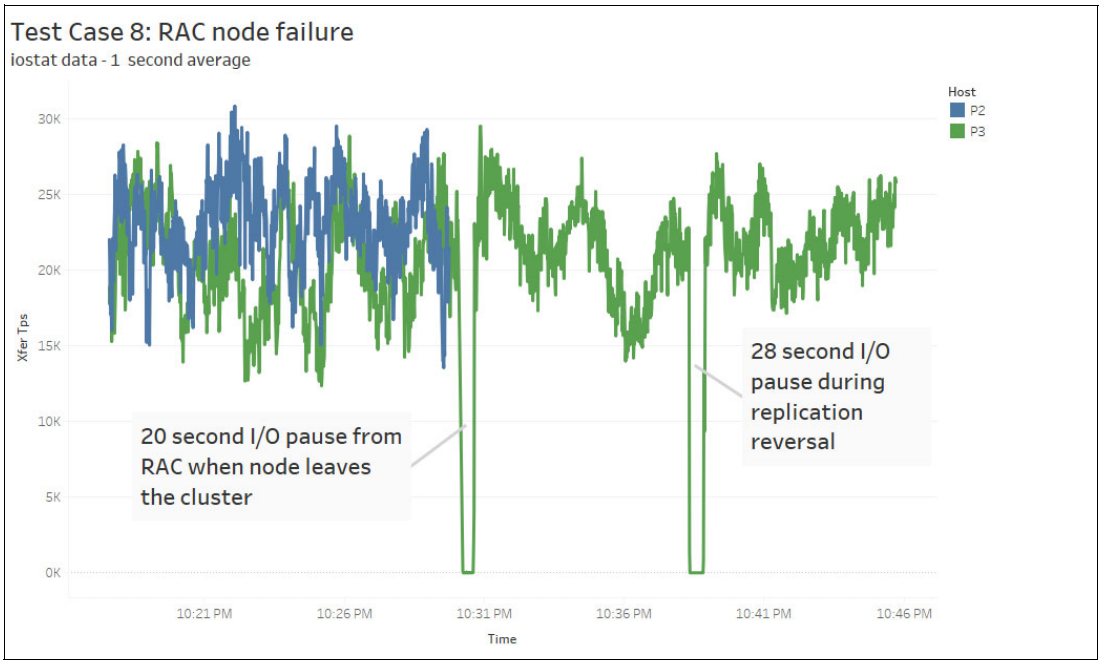

Figure 27 Test Case 8: RAC node failure (part 2 of 3)

Performance

Response times shift to nominal response times for Site2 after 7.5 minutes when the I/O distribution criteria is met and replication reverses. Response times are improved due to 100% directed to optimal site. See Figure 28.

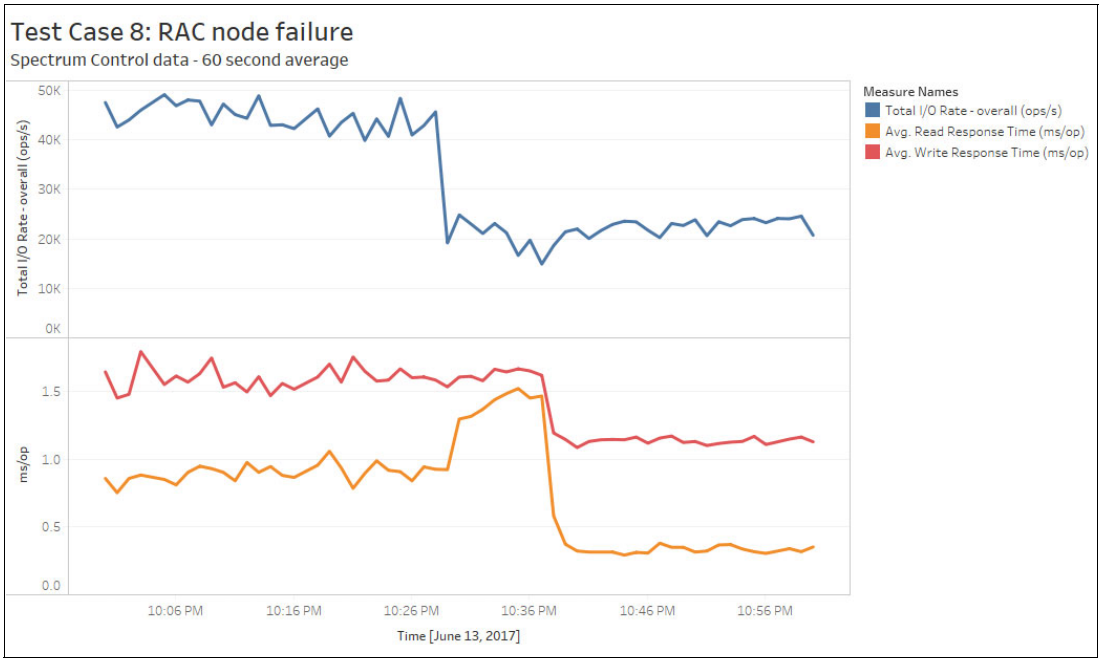

Figure 28 Test Case 8: RAC node failure (part 3 of 3)

Swingbench statistics (Test Case 8)

Table 11 shows the Swingbench summary.

Table 11 Swingbench

Parameters	Value
Total completed transactions	131,641
Average transactions per second	33.7
Maximum transaction rate	3,702
Total failed Transactions	1

Test Case 9: Site failure

This test case simulates the failure of the entire site.

Objective

Simulate the failure of an entire site.

Failure inject method

Simultaneously remove access to all ports on IOG0 by using switch zoning and halt AIX kernel on node P2. See Figure 29.

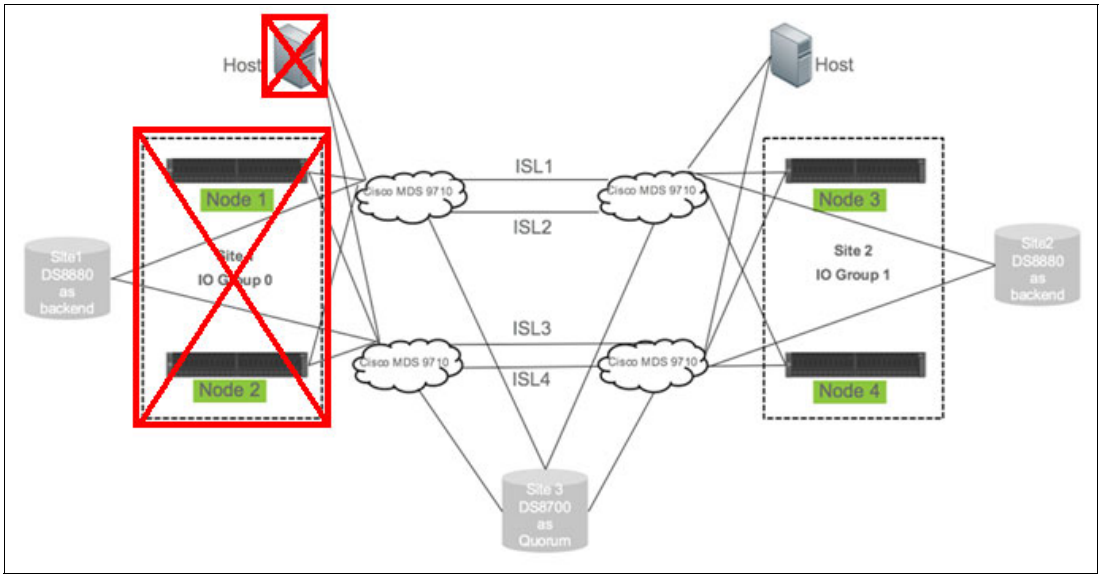

Figure 29 Test Case 9: Site failure (part 1 of 3)

HyperSwap state changes

- Replication reverses and suspends.

MPIO path state changes

- None

Host error recovery

- Failed node is evicted from RAC cluster.

Interruption to I/O processing

- A 57-second pause from RAC node eviction and replication suspend (Figure 30).

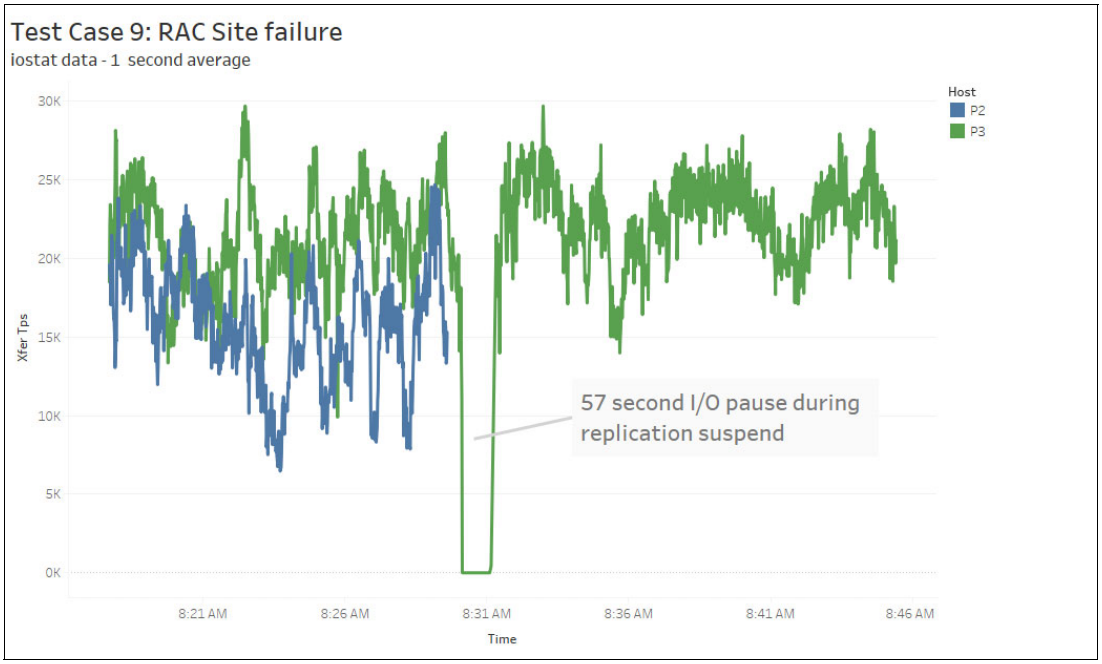

Figure 30 Test Case 9: Site failure (part 2 of 3)

Performance

- Response times drop due to replication suspension (Figure 31).

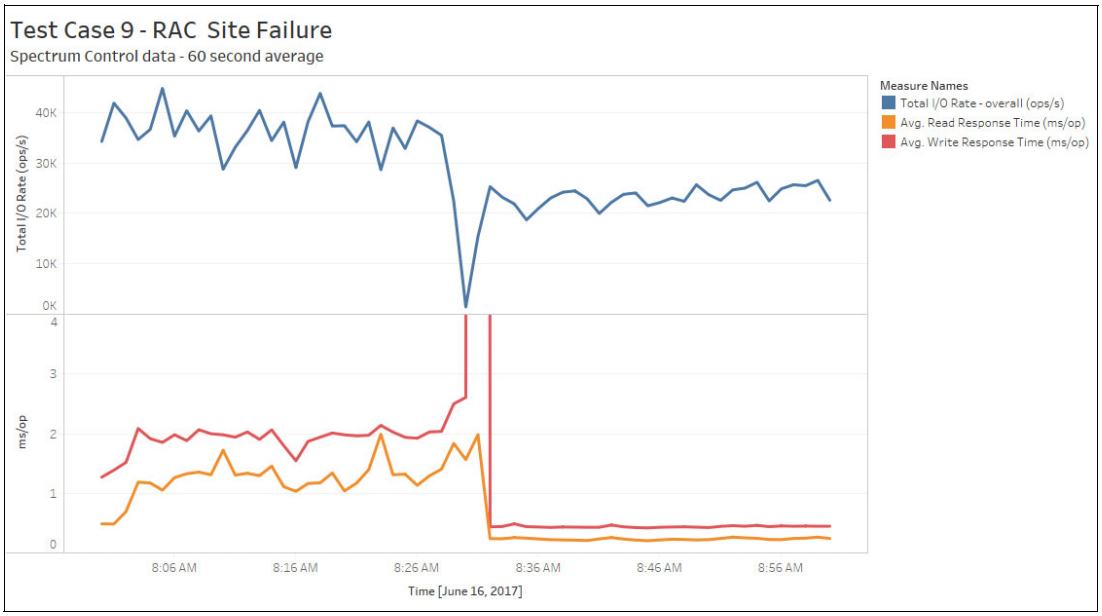

Figure 31 Test Case 9: Site failure (part 3 of 3)

Swingbench statistics (Test Case 9)

Table 12 shows the Swingbench summary.

Table 12 Swingbench

Parameters	Value
Total completed transactions	109,371
Average transactions per second	29.07
Maximum transaction rate	3,251
Total failed Transactions	2

Conclusion

The IBM HyperSwap feature of the SAN Volume Controller is an enterprise-class SAN-based technology that provides dual-site, active-active access to volumes. The sites can be co-located within the same physical data center or geographically dispersed across separate data centers at metro distances. This paper provided detailed examples of how HyperSwap provides a highly available active-active storage platform to provide high availability to Oracle databases in both active-passive and active-active configurations.

Authors

This paper was produced by a team of specialists from around the world working at the International Technical Support Organization, San Jose Center.

Ian MacQuarrie
IBM Systems

Thanks to the following people for their contributions to this project:

Ru Mei Niu, Pan Qun, Dharmesh Kamdar, Chad Collie, Bill Carlson
IBM Systems

This project was managed by:

Jon Tate
IBM International Technical Support Organization

Now you can become a published author, too

Here's an opportunity to spotlight your skills, grow your career, and become a published author—all at the same time! Join an ITSO residency project and help write a book in your area of expertise, while honing your experience using leading-edge technologies. Your efforts will help to increase product acceptance and customer satisfaction, as you expand your network of technical contacts and relationships. Residencies run from two to six weeks in length, and you can participate either in person or as a remote resident working from your home base.

Find out more about the residency program, browse the residency index, and apply online at:

ibm.com/redbooks/residencies.html

Stay connected to IBM Redbooks

- Find us on Facebook:

  http://www.facebook.com/IBMRedbooks
- Follow us on Twitter:

  http://twitter.com/ibmredbooks
- Look for us on LinkedIn:

  http://www.linkedin.com/groups?home=&gid=2130806
- Explore new IBM Redbooks® publications, residencies, and workshops with the IBM Redbooks weekly newsletter:

  https://www.redbooks.ibm.com/Redbooks.nsf/subscribe?OpenForm
- Stay current on recent Redbooks publications with RSS Feeds:

  http://www.redbooks.ibm.com/rss.html

Notices

This information was developed for products and services offered in the US. This material might be available from IBM in other languages. However, you may be required to own a copy of the product or product version in that language in order to access it.

IBM may not offer the products, services, or features discussed in this document in other countries. Consult your local IBM representative for information on the products and services currently available in your area. Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any IBM intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any non-IBM product, program, or service.

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not grant you any license to these patents. You can send license inquiries, in writing, to:
IBM Director of Licensing, IBM Corporation, North Castle Drive, MD-NC119, Armonk, NY 10504-1785, US

INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some jurisdictions do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-IBM websites are provided for convenience only and do not in any manner serve as an endorsement of those websites. The materials at those websites are not part of the materials for this IBM product and use of those websites is at your own risk.

IBM may use or distribute any of the information you provide in any way it believes appropriate without incurring any obligation to you.

The performance data and client examples cited are presented for illustrative purposes only. Actual performance results may vary depending on specific configurations and operating conditions.

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

Statements regarding IBM's future direction or intent are subject to change or withdrawal without notice, and represent goals and objectives only.

This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to actual people or business enterprises is entirely coincidental.

COPYRIGHT LICENSE:

This information contains sample application programs in source language, which illustrate programming techniques on various operating platforms. You may copy, modify, and distribute these sample programs in any form without payment to IBM, for the purposes of developing, using, marketing or distributing application programs conforming to the application programming interface for the operating platform for which the sample programs are written. These examples have not been thoroughly tested under all conditions. IBM, therefore, cannot guarantee or imply reliability, serviceability, or function of these programs. The sample programs are provided "AS IS", without warranty of any kind. IBM shall not be liable for any damages arising out of your use of the sample programs.

© Copyright IBM Corp. 2017. All rights reserved.

Trademarks

IBM, the IBM logo, and ibm.com are trademarks or registered trademarks of International Business Machines Corporation, registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on the web at “Copyright and trademark information” at http://www.ibm.com/legal/copytrade.shtml

The following terms are trademarks or registered trademarks of International Business Machines Corporation, and might also be trademarks or registered trademarks in other countries.

AIX®
HyperSwap®
IBM®
IBM FlashSystem®
IBM Spectrum™
IBM Spectrum Control™
IBM Spectrum Virtualize™
PowerHA®
Redbooks®
Redpaper™
Redbooks (logo)  ®
Storwize®
SystemMirror®

The following terms are trademarks of other companies:

Other company, product, or service names may be trademarks or service marks of others.

REDP-5459-00

ISBN 0738456349

Printed in U.S.A.

Get connected