NETAPP TECHNICAL REPORT

Verification Test with MetroCluster in a Xen Environment

Jim Lanson, NetApp
March, 2009 TR-3755

ABSTRACT

This document discusses the results of functional testing of MetroCluster in a Citrix XenServer environment. Proper operation is verified along with expected behavior during each of the test cases. Specific equipment, software, and functional failover tests are included along with results.

TABLE OF CONTENTS

1 INTRODUCTION

1.1 DOCUMENT PURPOSE

This paper documents the results of interoperability testing performed by NetApp between the following products:

- MetroCluster
- FAS deduplication technology
- Citrix XenServer 5.0

This document contains detailed descriptions of the tests performed, the test environment, and the results of those tests. It does not include performance-related information, and it is not intended as any kind of formal performance certification.

1.2 ASSUMPTIONS

Throughout this document, the examples assume two physical sites, SITEA and SITEB. SITEA represents the main data center on campus. SITEB is the campus DR location that provides protection in the event of a complete data center outage. Naming of all components clearly shows where they are physically located.

It is also assumed that the reader has basic familiarity with both NetApp and Citrix Xen products.

2 PRODUCT OVERVIEW

2.1 CITRIX XENSERVER

Citrix XenServer is a server virtualization system that makes data centers more agile and efficient through faster application deployments, higher levels of availability, and improved use of IT resources. It delivers the advanced features required by mission-critical workloads without sacrificing the ease of use necessary for wide-scale deployments. The unique streaming technology of XenServer can rapidly deliver workloads across virtual or physical servers, making it the ideal virtualization platform for every server in the enterprise.

2.2 NETAPP METROCLUSTER

NetApp® **MetroCluster** is a unique, cost-effective, synchronous replication solution for combining high availability and disaster recovery in a campus or metropolitan area, to protect against both site disasters and hardware outages. MetroCluster provides automatic recovery for any single storage component failure, and single-command recovery in the case of major site disasters, It also helps provide zero data loss and recovery within minutes rather than hours.

- Protects data against human error, system failures, and natural disasters
- Minimizes downtime during these events, with no data loss for business-critical applications
- Meets increased service-level agreements (SLAs) by reducing planned downtime
- Keeps IT costs under control without compromising data protection and high availability

2.3 FAS DEDUPLICATION TECHNOLOGY

NetApp, a leader in data storage efficiency since 1992, has established the first deduplication product to be used broadly across many applications, including data backup, data archiving, and primary data. NetApp deduplication combines the benefits of granularity, performance, and resiliency to provide users with a significant data deduplication advantage.

- NetApp deduplication operates with a high degree of granularity. Newly stored data is divided into small blocks. Each block of data has a digital “signature,” which is compared to all other signatures in the volume. If an exact block match exists on the disk volume, the duplicate block is discarded and its disk space is reclaimed.
- NetApp deduplication is tightly integrated with Data ONTAP® software and the WAFL® file system. Because of this, deduplication is performed with extreme efficiency. Complex hashing algorithms and lookup tables are not required. Instead, NetApp deduplication leverages existing Data ONTAP internal characteristics to create and search digital fingerprints, redirect data pointers, and free up redundant data areas—all with minimal impact on user performance.

- Another key advantage of NetApp deduplication integration with Data ONTAP is the ability to use the error-checking and recovery procedures that are inherent in Data ONTAP, including recovery from power failures, file inconsistencies, and file-system corruption.

3 HIGH-LEVEL TOPOLOGY

The overall solution uses NetApp MetroCluster on the back end for storage availability. On the front end are two IBM 3550 servers, running XenServer 5.0. Figure 1 shows the general layout of components used in this sample configuration.

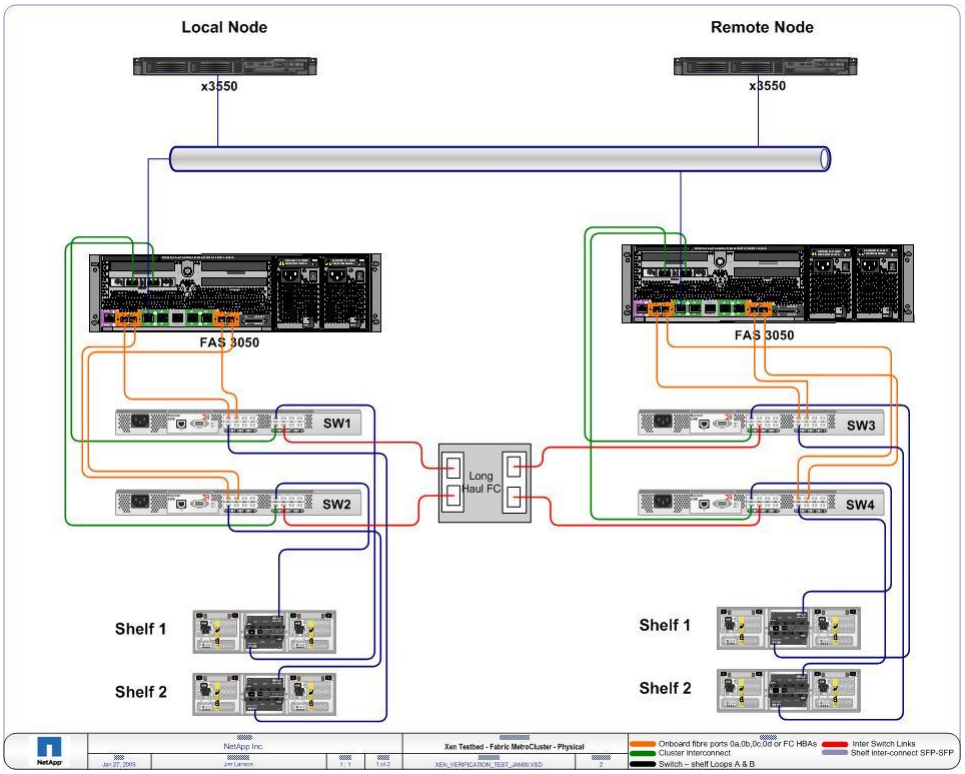


Figure 1) Test environment.

4 PRODUCTION SITE SETUP AND CONFIGURATION (SITEA)

This section covers the steps required to set up the test environment from both a NetApp and a Xen perspective. It does not imply best practices since this is only a functional test.

4.1 NETAPP

The NetApp FAS controller and back-end Fibre Channel switches are installed and configured following the instructions in the *Data ONTAP 7.2.6.1 Active/Active Configuration Guide* and the *Brocade Switch Configuration Guide*. The current software levels are:

- Data ONTAP 7.2.6.1
- Brocade firmware 6.0.0b

The production-site storage controller (FAS3050-SITEA) is a NetApp FAS3050 with two DS14mk2-HA shelves fully populated with 66GB 10k rpm drives. It is the primary node for the fabric MetroCluster and uses an FC/VI interconnect connected through back-end Fibre Channel switch fabrics to another FAS3050 controller (FAS3050-SITEB) at the secondary site.

The switch fabric is actually a dual-fabric configuration using four Brocade 200E switches, two at each site.

The following features are licensed on this controller:

- `cluster`: Required for MetroCluster
- `cluster_remote`: Required for MetroCluster
- `a_sis` for deduplication
- `nearstore_option` for deduplication
- `iscsi`: Used for Xen storage repository
- `nfs`: Used for Xen storage repository
- `syncmirror_local`: Required for MetroCluster

SWITCH CONFIGURATION

The back-end FC switches in a MetroCluster environment must be set up in a specific manner for the solution to function properly. For detailed information, see Appendix B.

VOLUME LAYOUT

The hardware in this configuration is limited to 14 mirrored disks on each controller head. Three of these are for the root volume and one is reserved for a spare. The remaining 10 disks have been used to create an aggregate that will host the volumes. The controller at SITEA has one volume (XenA) to house the iSCSI LUN-based active storage repository. The CIFS share used for the software distributions resides on volume 0 of the root aggregate. The controller at SITEB contains one volume (XenB) to house an iSCSI LUN-based storage repository and another volume (XENB_NFS) for the NFS export for the Xen NFS storage repository.

Note: This layout does not imply best practices or optimum layout. It was chosen because of the constraints of the test environment. Choice of layout did not impede functional test verification.

Table 1) Volume layout.

	Volume Name	Deduplication?	Type	Size
FAS3050-SITEA	XenA	Yes	Flex	100GB
FAS3050-SITEB	XenB	Yes	Flex	56GB
FAS3050-SITEB	XenB_NFS	Yes	Flex	40GB

ISCSI

A single iSCSI LUN was created, as shown in Figure 2. Sizes were arbitrarily chosen for these tests.

```
FAS3050-SITEA> lun show
        /vol/XenA/lun0                50g (53687091200)    (r/w, online, mapped)
```

Figure 2a) FAS3050-SITEA controller LUN configuration.

```
FAS3050-SITEB> lun show
        /vol/XenB/lunone              25g (26843545600)    (r/w, online, mapped)
```

Figure 2b) FAS3050-SITEB controller LUN configuration.

The iSCSI LUN created was then assigned to an igroup called Xen (Figure 3) containing the iSCSI IQN numbers for all XenServers (XenServer-sitea and XenServer-siteb).

```
FAS3050-SITEA> igroup show
    Xen (iSCSI) (ostype: vmware):
        iqn.2009-01.com.example:6020a60d (logged in on: e0a)
```

Figure 3a) FAS3050-SITEA controller igroup configuration.

```
FAS3050-SITEB> igroup show Xen
    Xen (iSCSI) (ostype: vmware):
        iqn.2009-01.com.example:cd4fd1c2 (logged in on: e0a)
```

Figure 3b) FAS3050-SITEB controller igroup configuration.

NFS

A single NFS Export was created, as shown in Figure 4.

```
FAS3050-SITEB> exportfs
/vol/XenB_NFS    -sec=sys,rw,root=10.61.132.18
```

Figure 4) FAS3050-SITEB NFS export.

CIFS

A single CIFS share was used to contain the software distributions for VMs, as shown in Figure 5.

Figure 5) FAS3050-SITEB controller CIFS share.

FAS DEDUPLICATION SETUP AND CONFIGURATION

The following steps were performed to set up and configure FAS deduplication for the Xen storage repositories.

Table 2) Commands to set up and configure FAS deduplication.

Step	Command	Description
1	FAS3050-SITEA*> **license add SAXCNCG OSFYRVH**	Add "nearstore_option site" license first, followed by the "a_sis site" license.
2	FAS3050-SITEA *> **sis on /vol/XenA** SIS for "/vol/VM_VOL_DEDUP" is enabled.	Enable DEDUP on volumes: sis on /vol/<volname>.
3	FAS3050-SITEA *> **sis on /vol/XenA_NFS** SIS for "/vol/VM_NFS_DEDUP" is enabled.	
4	FAS3050-SITEA *> **sis start -s /vol/XenA** The file system will be scanned to process existing data in /vol/XenA. This operation may initialize related existing metafiles. Are you sure you want to proceed with scan (y/n)? **y** Thu Feb 14 15:15:44 GMT [mc-u31: wafl.scan.start:info]: Starting SIS volume scan on volume XenA. The SIS operation for "/vol/XenA" is started.	To start DEDUP on volumes manually.
5	FAS3050-SITEA *> **sis config -s - /vol/XenA**	Disable automatic running of the deduplication process.

		Process can be a better controller for test purposes. For more information on scheduling DEDUPE operations, see the man page for na_sis.
6	FAS3050-SITEA*> **sis status -l** Path: /vol/XenA State: Enabled Status: Active Progress: 2341192 KB (25%) Done Type: Regular Schedule: Last Operation Begin: Thu Feb 14 15:09:07 GMT 2008 Last Operation End: Thu Feb 14 15:09:07 GMT 2008 Last Operation Size: 0 KB Last Operation Error: - FAS3050-SITEA *> **sis status -l** Path: /vol/XenA State: Enabled Status: Idle Progress: Idle for 11:47:23 Type: Regular Schedule: Last Operation Begin: Wed Feb 20 00:00:00 GMT 2008 Last Operation End: Wed Feb 20 00:09:54 GMT 2008 Last Operation Size: 259 GB Last Operation Error: - FAS3050-SITEA *> **sis status** Path State Status Progress /vol/VM_VOL_DEDUP Enabled Active 2248 MB (24%) Done FAS3050-SITEA *> **sis status** Path State Status Progress /vol/VM_VOL_DEDUP Enabled Idle Idle for 11:49:02 FAS3050-SITEA *> **sis status -l** Path: /vol/XenA State: Enabled Status: Idle Progress: Idle for 23:48:41 Type: Regular Schedule: Last Operation Begin: Tue Feb 19 13:00:00 GMT 2008 Last Operation End: Tue Feb 19 13:06:25 GMT 2008 Last Operation Size: 225 GB Last Operation Error: -	Check DEDUPE progress:
7	FAS3050-SITEA *> **df -s** Filesystem used saved %saved /vol/vol0/ 51273444 0 0% /vol/XenA/ 97270580 599304 1% /vol/XenA_NFS/ 220 1 0%	Check for space savings on a DEDUPE volume.
8	FAS3050-SITEA *> **sis check /vol/XenA** Fingerprints of "/vol/XenA" are being verified.	Verify and update fingerprint database for the specified flexible volume and include purging stale fingerprints.
10	Repeat the steps above for the /vol/XenA_NFS volume.	
11	Repeat steps 1-8 for FAS3050-SITEB volumes, /vol/XenB and /vol/XenB_NFS	

4.2 XEN

Two Xen 5.0 XenServers were installed according to the vendor-supplied procedures in the XenServer 5.0 Installation guide.

XENSERVER ENVIRONMENT

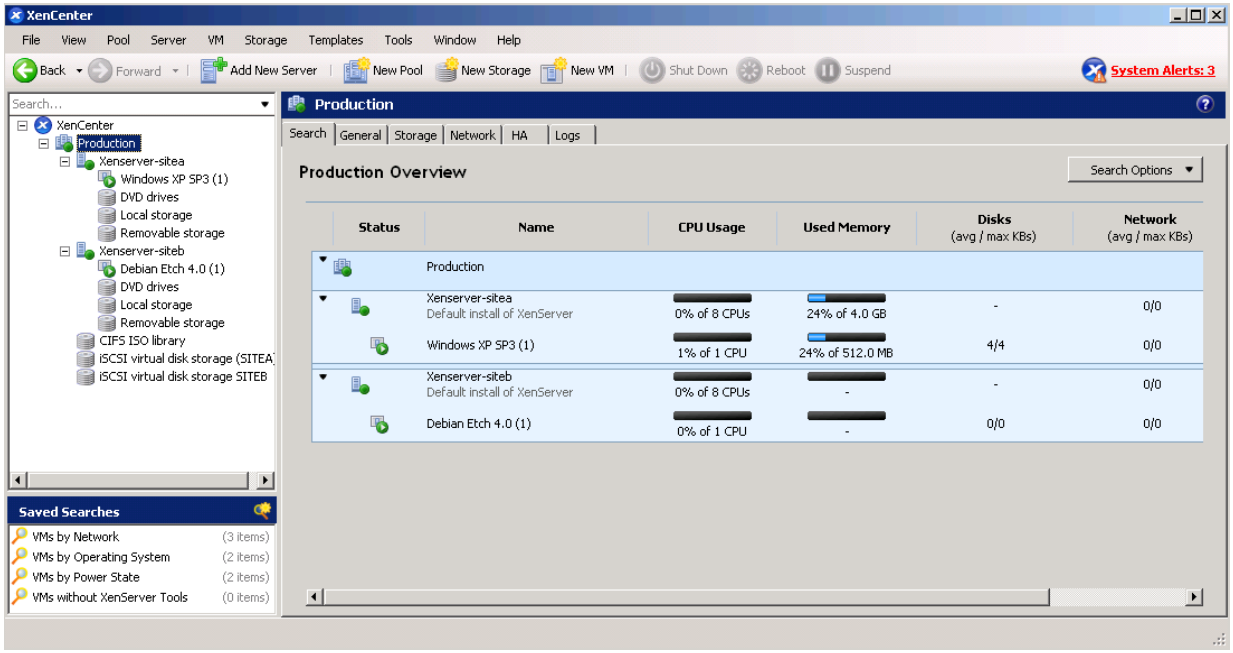


Figure 6) Xen environment.

STORAGE REPOSITORIES

Three storage repositories were created for the following purposes, as shown in Table 3.

Table 3) Summary of storage repositories

Name	Use
FAS3030-SITEA:/vol/xen/lun0	Primary storage for VMs (XenServer-sitea)
FAS3030-SITEB:/vol/xenB/lunOne	Primary storage for VMs (XenServer-siteb)
FAS3050-SITEB:/vol/XenB_NFS	Storage for testing NFS

Once the storage repositories were created, visibility was verified in the XenCenter server, as shown in Figure 7.

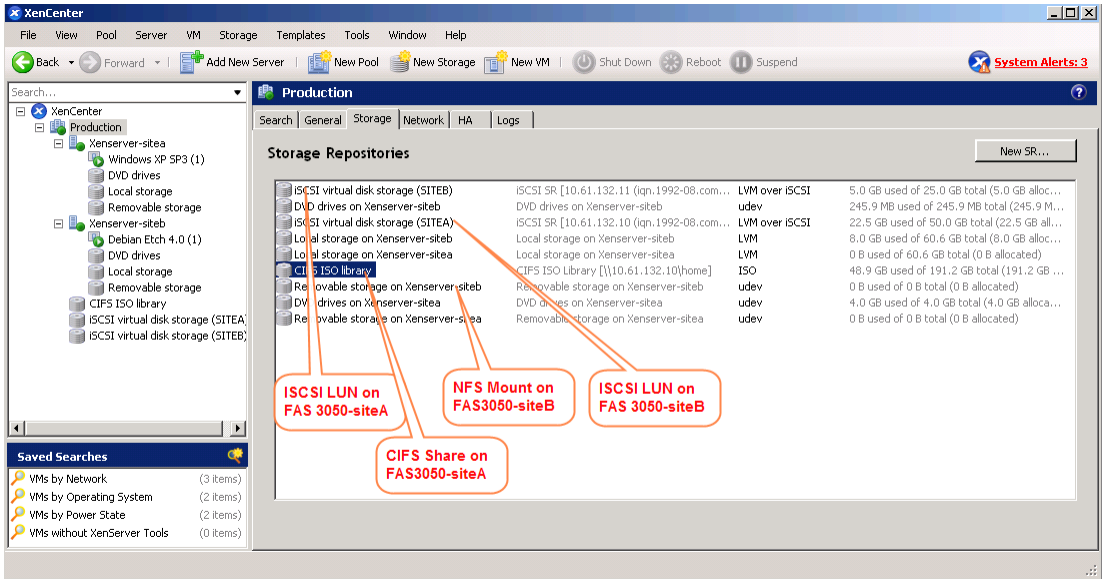


Figure 7) Storage repository in production Xen resource pool.

VIRTUAL MACHINES

For purposes of testing, a resource pool called Production was set up containing two XenServers, Xenserver-sitea and XenServer-siteb. A single Windows® XP virtual machine was created on Xenserver-sitea, with data drives set as shown in Figure 8.

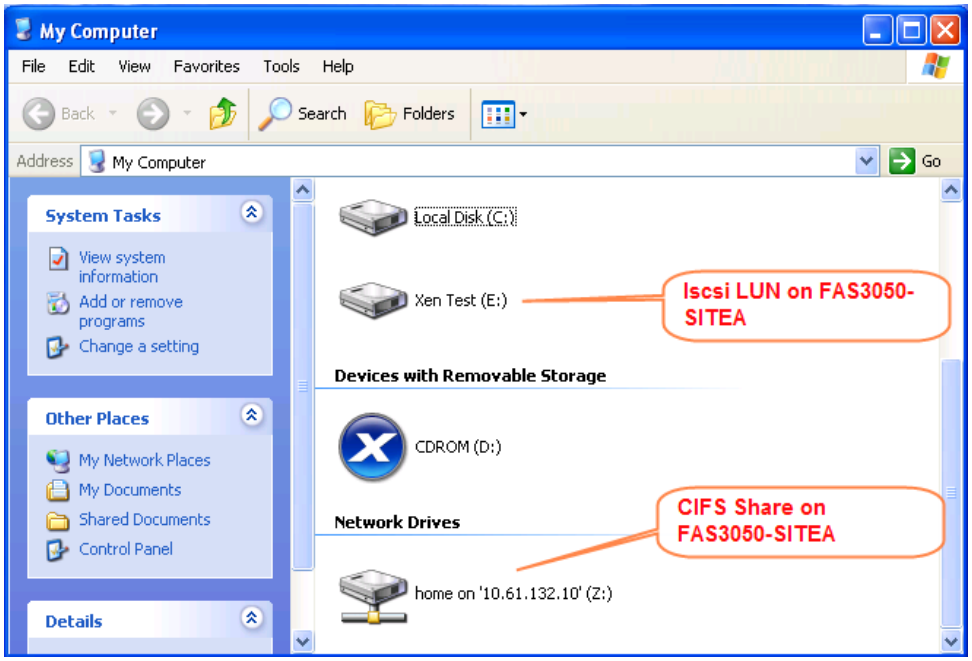


Figure 8) Windows XP virtual machine setup.

Another virtual machine was created on Xenserver-siteb running Debian Linux® with data drives, as shown in Figure 9.

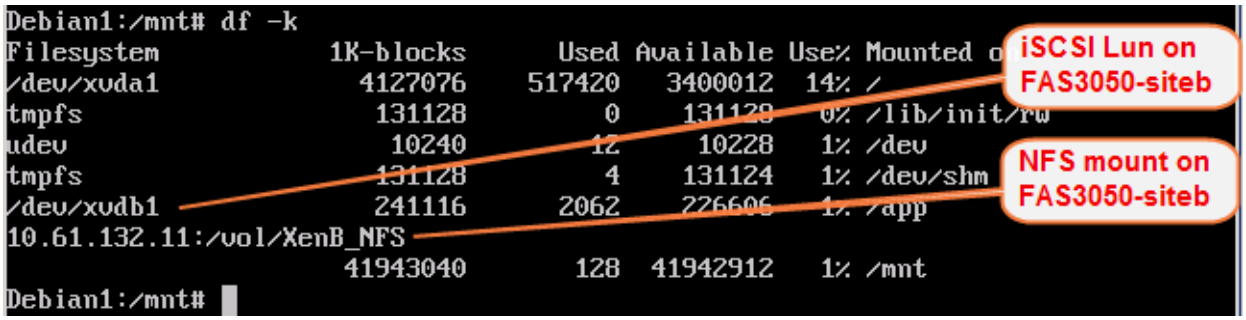


Figure 9) Debian Linux virtual machine setup.

5 FUNCTIONAL TESTS

To test storage repository availability in the failure scenarios described in this section, IOmeter was installed in the Windows XP virtual machine and set up to create activity (50% reads) on the storage repositories. Perfmon was used to monitor the activity and record any interruptions due to failures. Access from the Debian Linux virtual machine was also verified.

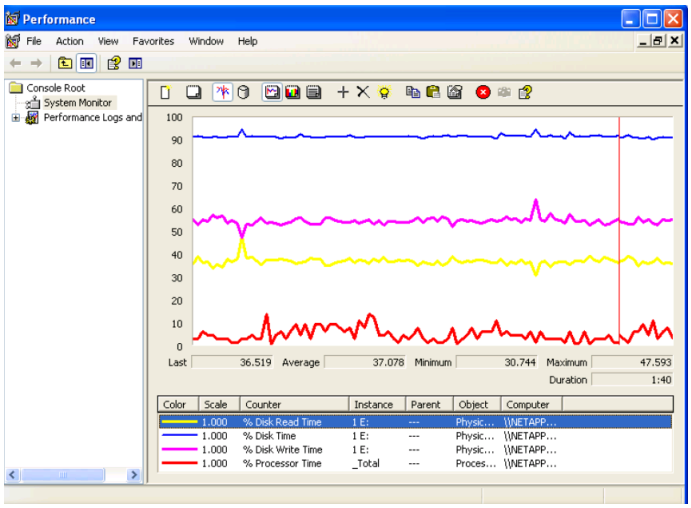


Figure 10) Perfmon recording activity.

The following subsections describe test scenarios that are executed upon successful installation of the test environment described in this document. Test scenarios include various component failure scenarios. Unless stated otherwise, before the execution of each test, the environment is reset to the “normal” running state with virtual machines running on the XenServers generating disk activity

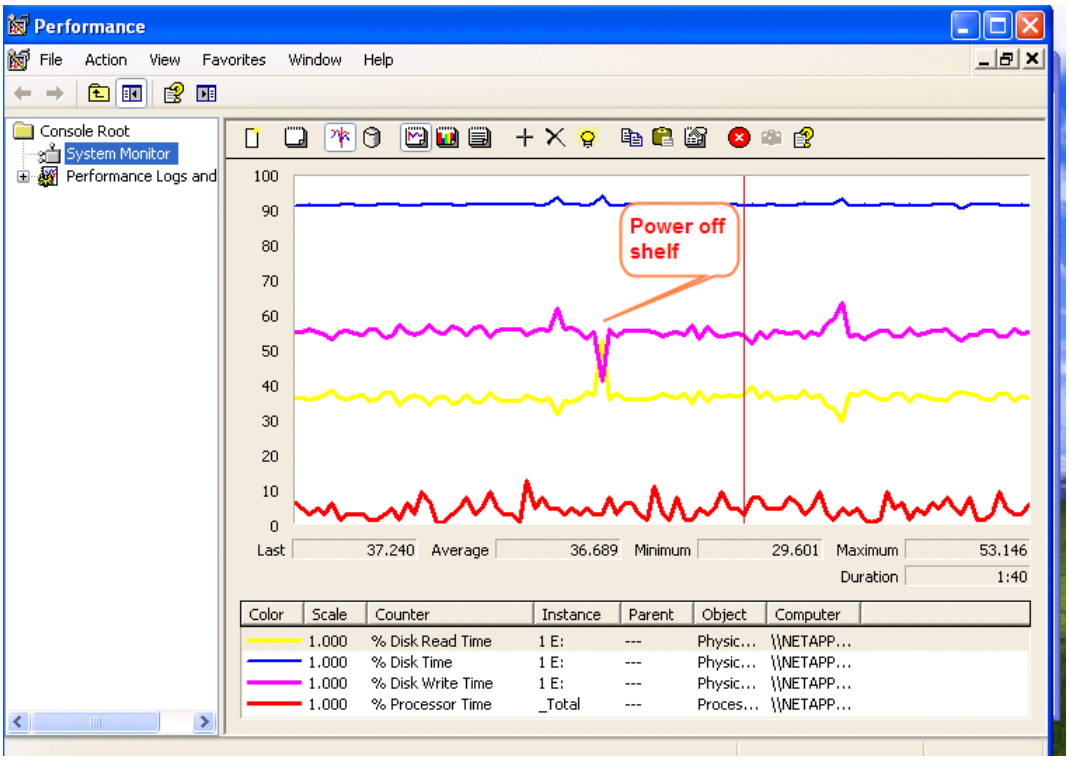
For each of the following test scenarios, both the volumes (LUN and NFS) were verified.

Test Case #	Description
1	Complete Loss of Power to Disk Shelf
2	Loss of One Link on One Disk Loop
3	Loss of Fibre Channel Switch
4	Loss of One ISL
5	Failure of Controller
6	Failback of Controller
7	Loss of Entire Site/Disaster Declared

6 TEST CASES

6.1 TC 1: COMPLETE LOSS OF POWER TO DISK SHELF

Description	No single point of failure should exist in the solution. Therefore, the loss of an entire shelf is tested. This test is accomplished by simply turning off both power supplies while the deduplication process is running.
Task	Power off the FAS3050-SITEA Pool0 shelf, observe the results, and then power it back on.
Expected Results	Relevant disks go offline, plex is broken, but service to clients (availability and performance) is unaffected. When power is returned to the shelf, the disks are detected and a resync of the plexes occurs without any manual action.
Results	Results were as expected and are shown in the following graph.



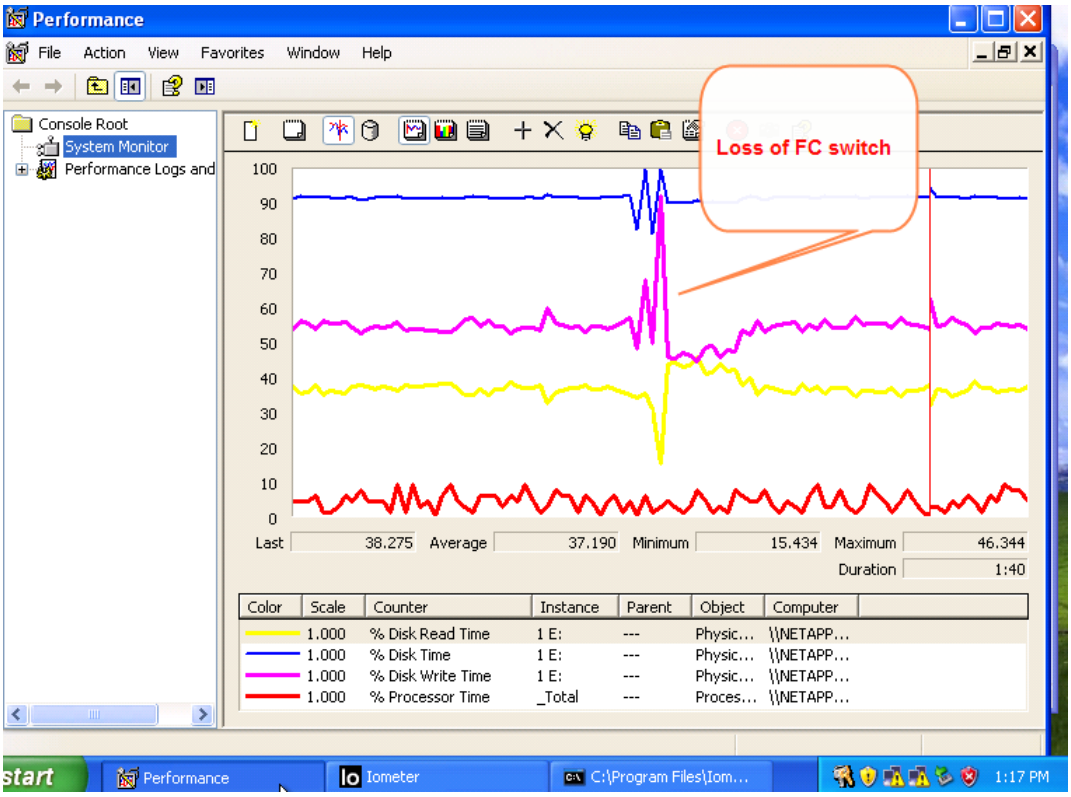

```
FAS3050-SITEA> aggr status
               Aggr State      Status
               vol0 online     raid4, trad
                               resyncing
               aggr1 online    raid_dp, aggr
FAS3050-SITEA>
```

After shelf powered back on

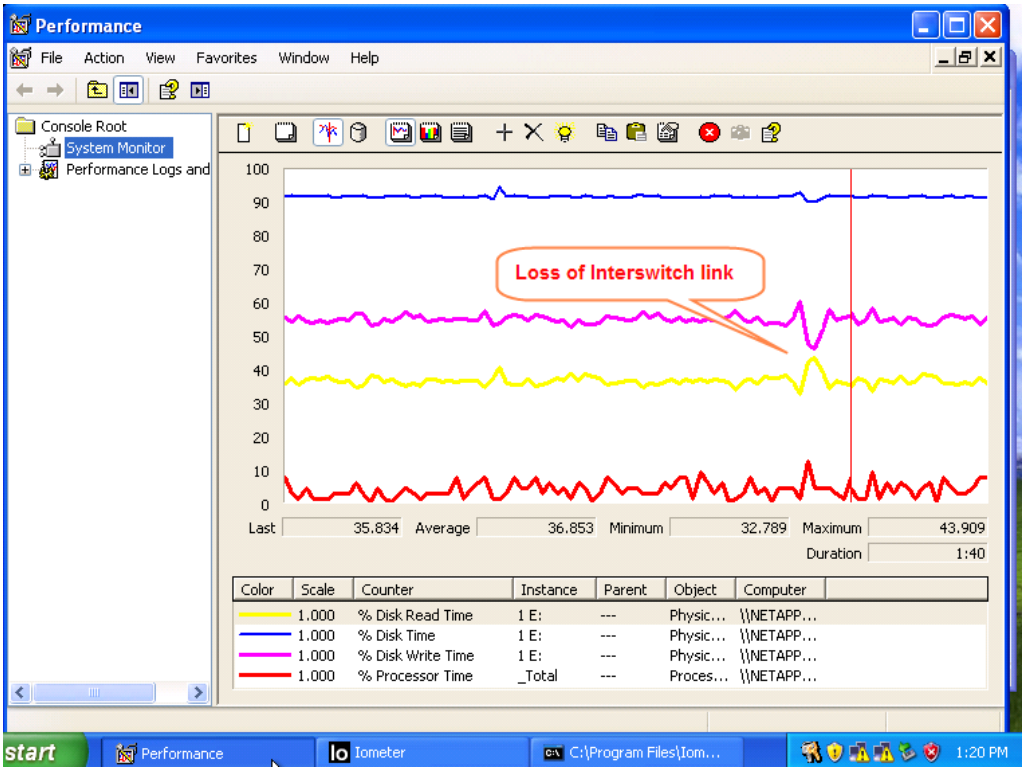
6.2 TC 2: LOSS OF ONE LINK ON ONE DISK LOOP

Description	No single point of failure should exist in the solution. Therefore, the loss of one disk loop is tested. This test is accomplished by removing a fiber patch lead from one of the shelves. The deduplication process is running during this test.
Task	Remove the fiber entering FAS3050-SITEA Pool0, ESH A, observe the results, and then reconnect the fiber.
Expected Results	Controller reports that some disks are connected to only one switch, but service to clients (availability and performance) is unaffected. When the fiber is reconnected, the controller displays the message that disks are now connected to two switches.
Results	Results were as expected and are shown in the following graph. 

6.3 TC 3: LOSS OF FIBRE CHANNEL SWITCH

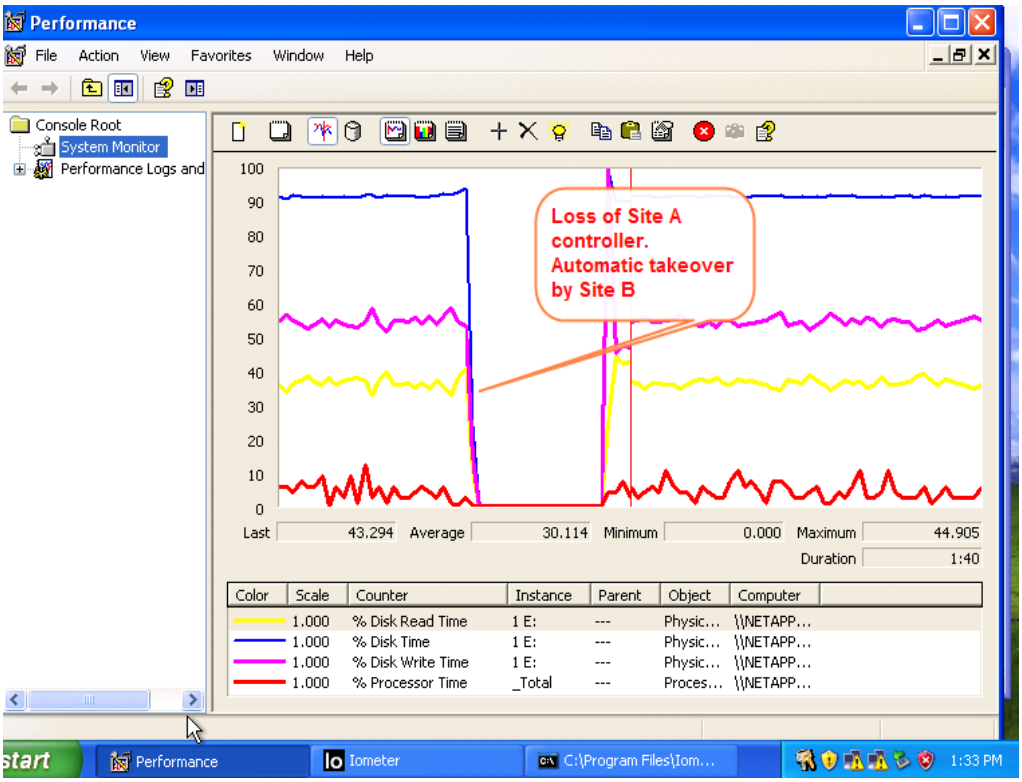
Description	No single point of failure should exist in the solution. Therefore, the loss of an entire Brocade switch is tested. This test is accomplished by simply removing the power cord from the switch while a load is applied.
Task	Power off the Fibre Channel switch SITEA-SW2, observe the results, and then power it back on.
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 power is restored and the switch completes its boot process, the controller displays messages to indicate that the disks are now connected to two switches and that the second cluster interconnect is again active.
Results	Results were as expected and are shown in the following graph. 

6.4 TC 4: LOSS OF ONE ISL

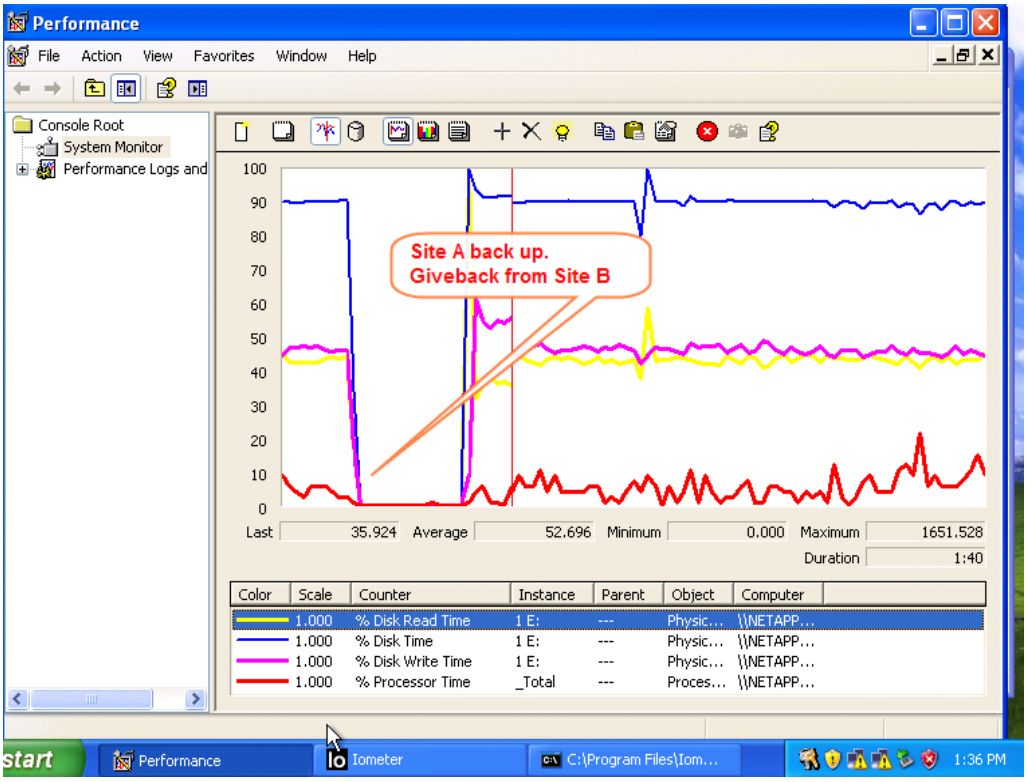
Description	No single point of failure should exist in the solution. Therefore, the loss of one of the interswitch links (ISLs) is tested. This test is accomplished by simply removing the fiber between two of the switches while a load is applied.
Task	Remove the fiber between SITEA-SW1 and SITEB-SW3.
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 messages to indicate that the disks are now connected to two switches and that the second cluster interconnect is again active.
Results	Results were as expected and are shown in the following graph. 

6.5 TC 5 - Failure of Controller

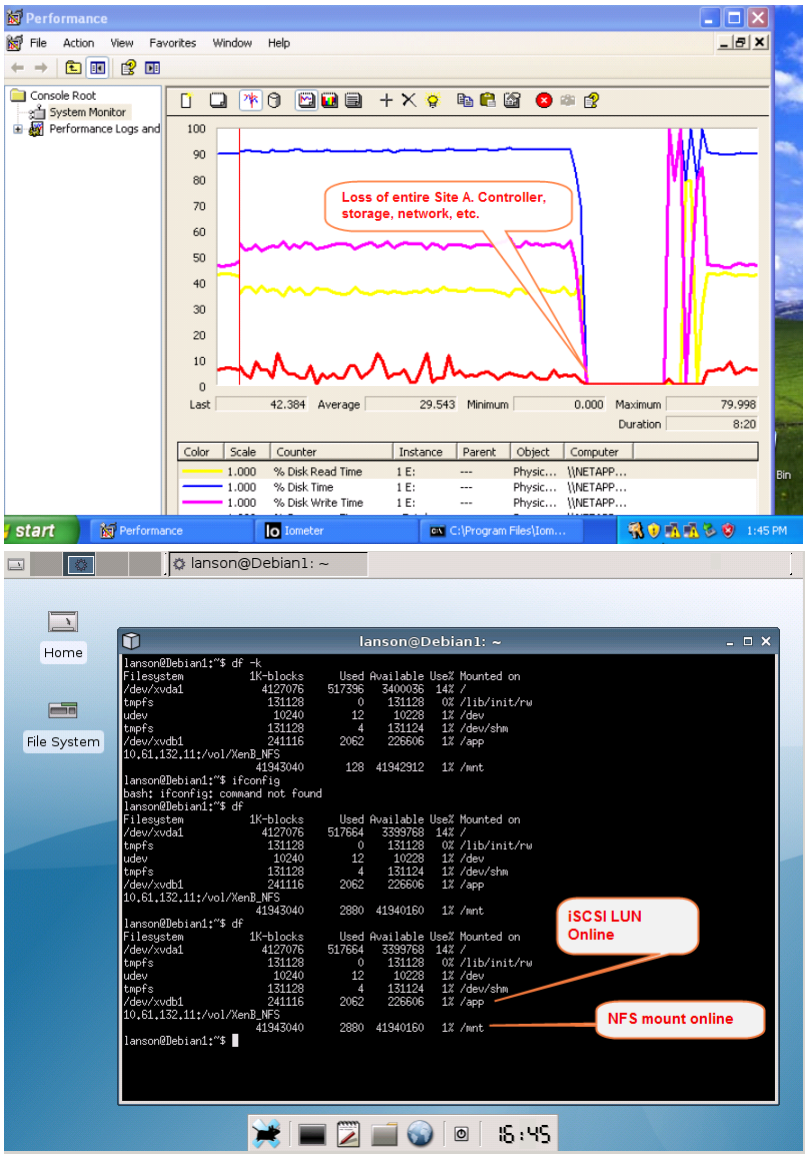
Description	No single point of failure should exist in the solution. Therefore, the loss of one of the controllers is tested.
Task	Power off the FAS3050-SITEA controller by simply turning off both power supplies.
Expected Results	A slight delay from a host perspective occurs while iSCSI tears down and rebuilds the connection because of the change of processing from one controller to the other. The deduplication process should terminate. It should not be possible to run deduplication on the volumes until giveback.
Results	. Results were as expected and are shown in the following graph.



6.6 TC 6: FAILBACK OF CONTROLLER

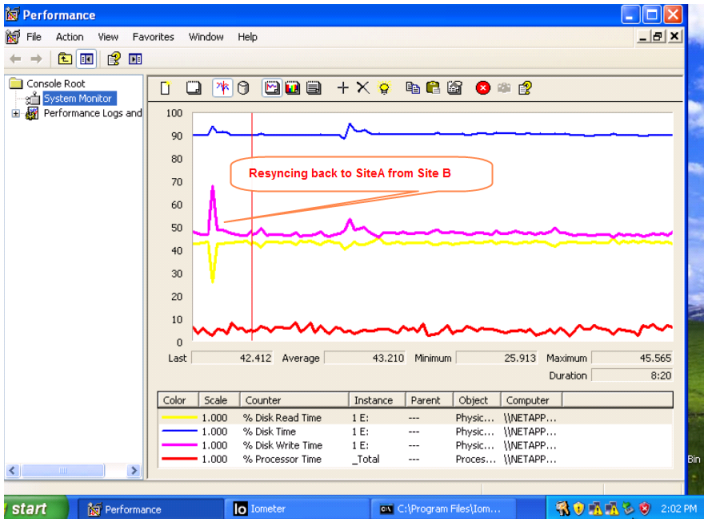
Description	As a follow-up to the previous test, the data serving must be failed back to the previously failed controller to return to the normal operating state. This test is accomplished by issuing a command on the surviving controller to request that processing be returned to the previously failed controller.
Task	Power on SITEA. Issue a `cf giveback` command on FAS3050-SITEB to cause the failback to occur.
Expected Results	A slight delay occurs from a host perspective while iSCSI tears down and rebuilds the connection because of the change of processing from one controller to the other. No errors should be displayed at the application level.
Results	Results were as expected and are shown in the following graph: 

6.7 TC 7: LOSS OF ENTIRE SITE/DISASTER DECLARED

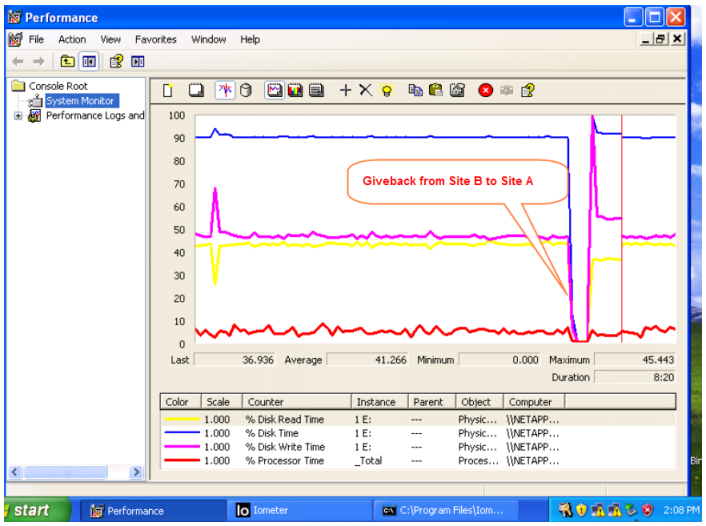
Description	To test the availability of the overall solution, the loss of an entire site is simulated.
Task	Test the failure of SITEA by interrupting the following components in this order, in rapid succession: 1. Disconnect both ISLs. 2. Remove power from the FAS3050-SITEA disk shelves. 3. Remove power from ESX-PROD1. 4. Remove power from FAS3050-SITEA. 5. Declare a site disaster and perform a takeover at the surviving site (SITEB); issue a `cf forcetakeover -d` command on FAS3050-SITEB. 
Expected Results	Takeover is successful. The deduplication process is no longer running and cannot be restarted until giveback is complete. Volumes continue to have SIS enabled, so changed and new blocks can be logged.
Results	Results were as expected. After a brief period, operation resumed after recovery procedures (bringing LUNs online) were performed.

6.8 TC 8: RESTORE OF ENTIRE SITE/RECOVER FROM DISASTER

Description	To test the availability of the overall solution, recovery after the loss of an entire site is simulated.
Task	Power on the disk shelves only on FAS3050-SITEA. Reconnect the ISL between sites so that FAS3050-SITEB can see the disk shelves from FAS3050-SITEA. After connection, the SITEB Pool1 volumes automatically begin to resync.




```
FAS3050-SiteB> cf status
Cluster enabled, FAS3050-SiteA is up.
```

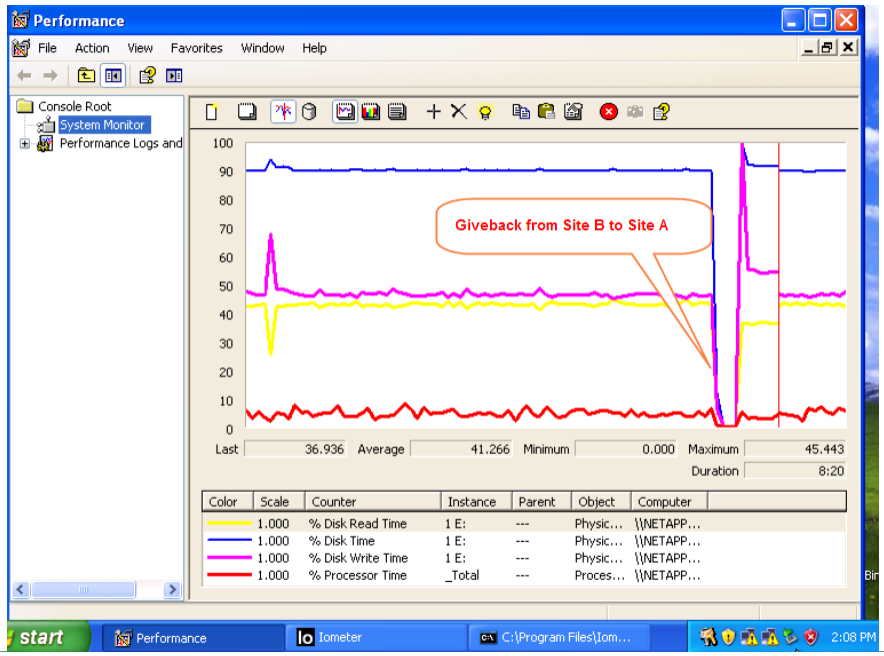


In partner mode on FAS3050-SITEB, reestablish the mirrors in accordance with the *Active-Active Installation Guide* located on NOW™ (NetApp on the Web).

Using the command `aggr status`, make sure that all mirror resynchronization is complete before proceeding.

Power on FAS3050-SITEA. 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 SITEA 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 to the VMs or to the XenServer. After giveback was completed, the deduplication process was restarted and successfully completed.



7 CONCLUSION

The tests and results described in this document complement the far more exhaustive tests performed by NetApp QA relative to the FAS deduplication technology and MetroCluster. It was determined that a basic functional reverification was necessary for these products in a Xen environment. Performance tests were beyond the scope of these scenarios.

In each of the tests performed, the results were as expected, both in terms of NetApp product operation and the resulting impact (or lack of impact) on the Xen 5.0 servers and virtual machines. No anomalies or unexpected behavior was exhibited at either the Xen or the NetApp layer. It is clear from the combination of QA and these tests that the value provided by FAS deduplication in a Xen environment can be realized when combined with MetroCluster.

8 APPENDIX A: MATERIALS LIST

Table 4) Hardware materials list.

STORAGE	Vendor	Name	Version	Description
	NetApp	FAS 3050C		
HOSTS	IBM	1 x IBM eServer xSeries 3550 (2.5 GHz/4GB RAM) (Xenserver-sitea) 1 x IBM eServer xSeries 3550 (2.5 Ghz/4GB RAM) (Xenserver-siteb)		
BACK-END SAN (METROCLUSTER)	Brocade	200E (4)	6.0.0b	16 Port FC Switch
SOFTWARE				
STORAGE	NetApp	SyncMirror®	7.2.6.1	Replication
	NetApp	Data ONTAP	7.2.6.1	Operating System
	NetApp	Cluster_Remote	7.2.6.1	Failover
HOSTS	Citrix	Citrix XenServer	5.0	Operating System

9 APPENDIX B: BROCADE SWITCH CONFIGURATION

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

Table 5) SITEA Switch 1.

Switchname		SITEA-SW1	
PORT	**BANK/POOL**	**CONNECTED WITH**	**PURPOSE**
0	1/0	SITEA, 5a	Site A FC HBA
1	1/0	SITEA, 8a	Site A FC HBA
2	1/0		
3	1/0		
4	1/1		
5	1/1	SITEB pool 1, Shelf 3B	
6	1/1		
7	1/1		
8	2/0		
9	2/0	SITEA pool 0, Shelf 1B	
10	2/0		
11	2/0		
12	2/1	SITEA FCVI, 6a	Cluster interconnect
13	2/1	STB-SW3, port 5	ISL
14	2/1		
15	2/1		

Table 6) SITEA Switch 2.

Switchname		SITEA-SW2	
PORT	**BANK/POOL**	**CONNECTED WITH**	**PURPOSE**
0	1/0	SITEA, 5b	Disk HBA for bank 2 shelves
1	1/0	SITEA, 8b	Disk HBA for bank 2 shelves
2	1/0		
3	1/0		
4	1/1		
5	1/1	SITEB pool 1, Shelf 3A	
6	1/1		
7	1/1	SITEA FCVI, 6b	Cluster interconnect
8	2/0		
9	2/0	SITEA pool 0, Shelf 1A	
10	2/0		
11	2/0		
12	2/1		
13	2/1	STB-SW3, port 4	ISL
14	2/1		
15	2/1		

Table 7) SITEB Switch 3.

Switchname	SITEA-SW3		
Port	Bank/Pool	Connected with	Purpose
0	1/0	SITEA pool 1, Shelf 3B	
1	1/0		
2	1/0		
3	1/0	SITEB FCVI, 6a	Cluster interconnect
4	1/1		
5	1/1	STB-SW1, port 13	ISL
6	1/1		
7	1/1		
8	2/0	SITEB, 5a	Disk HBA for bank 2 shelves
9	2/0	SITEB, 8a	Disk HBA for bank 2 shelves
10	2/0		
11	2/0		
12	2/1	SITEB pool 0, Shelf 1B	
13	2/1		
14	2/1		
15	2/1		

Table 8) SITEB Switch 4.

Switchname	SITEB-SW4		
PORT	BANK/POOL	CONNECTED WITH	PURPOSE
0	1/0	SITEA pool 1, Shelf 3A	
1	1/0		
2	1/0		
3	1/0		
4	1/1	STB-SW1, port 13	ISL
5	1/1		
6	1/1		
7	1/1		
8	2/0	SITEB, 5b	Disk HBA for bank 2 shelves
9	2/0	SITEB, 8b	Disk HBA for bank 2 shelves
10	2/0		
11	2/0		
12	2/1	SITEB pool 0, Shelf 1A	
13	2/1	SITEB FCVI, 6b	Cluster interconnect
14	2/1		
15	2/1		