# Technical report:

# VMware and IBM System Storage N series with SnapMirror and MetroCluster

*High availability and disaster recovery*

*Document NS3606-0*

February 3, 2008

## Table of contents

# Abstract

*This document discusses disaster recovery for a VMware environment that uses IBM System Storage N series for its primary storage. Data protection levels are based on customer needs and can be considered independently. Protection can be at a data center, campus, regional level or all three. There also might be more than one solution at each level. This document provides one example of how IBM System Storage N series with MetroCluster, SnapMirror, and FlexClone volumes can be combined. Specific equipment, software, and functional failover tests are included along with results. Combining VMware and IBM N series products creates a simple, cost-effective, robust, and scalable disaster recovery solution.*

# Executive summary

As companies have become increasingly dependent on data and access to that data, disaster recovery (DR) has become a major concern. The ability to provide a highly available, 24X7 operation is very desirable. Unfortunately, the high cost typically associated with a DR strategy has forced some companies to delay their DR plans. However, with the advent of new technologies, many companies are turning to virtualization as a method of solving their DR challenges in a cost-effective manner. With VMware as a leader in the virtual infrastructure arena, this paper discusses why and how so many companies combine VMware Infrastructure 3 (VI3) software with IBM® System Storage® N series products to provide a cost-effective and reliable DR solution.

A recent survey of VMware end users showed that over 60% are using VI3 as part of their DR solution. This is a high percentage for customers who started out with a plan to use virtualization for consolidation and ended up using it as part of their DR solution as well. In short, many of them found that by trying to solve one problem, they had put the pieces in place to solve a second business problem as well. Combining VI3 and IBM N series products creates a wonderfully simple, cost-effective, and robust DR solution that scales and evolves as the organization’s needs evolve.

As more customers look to virtualization as a means to cost effectively simplify their server environment, they realize the importance of a highly available storage environment with DR capabilities. Levels of protection may be needed in the data center, at a campus level, at a regional level, or a combination of all three. This document provides a detailed plan for setting up and testing both IBM N series and VMware products, providing all three levels of protection, in a virtualized environment. This document outlines the hardware and software used, the installation and configuration steps performed, and operational scenarios. Options for testing DR capabilities are also discussed. Although this example uses iSCSI protocol, the steps are identical for an implementation using Fibre Channel (FC) Protocol (FCP) for host storage connectivity. This technical report is intended as an architecture reference for such specific customer scenarios.

## Document purpose

The intent of this document is to provide an example of how IBM N series products provide a highly available storage environment for VI3, including in disaster recovery scenarios.

The purpose of the reference configuration herein is to show the following:

Reliable and predictable solution behavior
How IBM N series products and VMware host virtualization can work together for a mission-critical application
Continued availability upon loss of any component (that is, no single point of failure)

- Rapid business continuance and DR in case of a full site disaster.

This document does not include performance-related information, and it is not intended as any kind of formal certification.

## Assumptions

Throughout this document, the examples assume three physical sites, named SITEA, SITEB, and DR. 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. DR is the regional DR location that provides geographic protection. Naming of all components clearly shows where they are physically located.

It is also assumed that the reader has basic familiarity with both IBM N series and VMware products.

# Product overview

## VMware

VMware products provide enterprise-class virtualization that increases server and other resource utilization, improves performance, increases security, and minimizes system downtime, reducing the cost and complexity of delivering enterprise services. By leveraging existing technology, VMware enables the roll-out of new applications with less risk and lower platform costs.

### VI3

VI3 (VMware Infrastructure 3.0) and generally referred to as VMware Infrastructure, is a feature-rich suite that delivers the production-proven efficiency, availability, and dynamic management needed to create a responsive data center. The suite includes:

VMware ESX Server: Platform for virtualizing servers, storage and networking
VMware Virtual Machine File System (VMFS): High-performance cluster file system for storage virtualization
VMware Virtual SMP: Multiprocessor support for virtual machines (VMs)
VMware VirtualCenter: Centralized management, automation, and optimization for IT infrastructure
VMware High Availability (HA): Cost-effective high availability for VMs
VMware Distributed Resource Scheduler (DRS): Dynamic balancing and allocation of resources for VMs

VMware VMotion: Live migration of VMs without service interruption
VMware Consolidated Backup: Centralized backup enabler for VMs.

## IBM N series

IBM System Storage N series with MetroCluster is a unique, cost-effective, synchronous replication solution for combining high availability and DR 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 case of major site disasters, ensuring zero data loss and making recovery possible within minutes rather than hours. Metrocluster:

Ensures data protection against human error, system failures, and natural disasters
Ensures minimal 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

IBM System Storage N series with SnapMirror® software is the value leader in the industry when it comes to DR. Its simplicity and flexibility make it affordable for customers to deploy a DR solution for more of their application infrastructures than would be possible with competitive alternatives. SnapMirror supports synchronous replication limited to metro distances, ensuring zero data loss; semisynchronous replication that supports recovery point objectives (RPOs) in seconds with minimal impact on the host application; and asynchronous replication, which is the most cost-effective solution that can meet RPOs ranging from 1 minute to 1 day. Its functionality and configuration flexibility enable SnapMirror to support multiple uses, including DR, data distribution, remote access, data migration, data replication, and load balancing. SnapMirror assists with the:

Need to protect against component and system failures, site failures, and natural disasters
Desire to lower the cost of secondary site and network infrastructure
Ability to reduce the complexity of deployment, and failover and recovery processes
Capability to meet RPOs and recovery time objectives (RTOs).

IBM System Storage N series with FlexClone™ and IBM System Storage N series with FlexVol™ technologies enable entirely new opportunities and ways of working for organizations that are grappling with the challenges of increased overhead, management costs, and data risk.

FlexVol technology delivers true storage virtualization solutions that can lower overhead and capital expenses, reduce disruption and risk, and provide the flexibility to adapt quickly and easily to the dynamic needs of the enterprise. FlexVol technology pools storage resources automatically and enables you to create multiple flexible volumes on a large pool of disks.

FlexClone technology enables true cloning—instant replication of data volumes and data sets without requiring additional storage space at the time of creation. Each cloned volume is a transparent, virtual copy that you can use for essential enterprise operations, such as testing and bug fixing, platform and upgrade checks, multiple simulations against large data sets, remote office testing and staging, and market-specific product variations.

## Tiers of protection

Combining VMware and IBM N series technologies offers a unique value proposition. The combination resolves a number of customer challenges from both a server and a storage perspective. Additionally, having both technologies offers the ability to have a tiered DR environment. While VMware offers DR capabilities in a data center from a host perspective through such features as HA, IBM N series offers storage DR in a data center, across campus, and at a regional level (Figure 1).

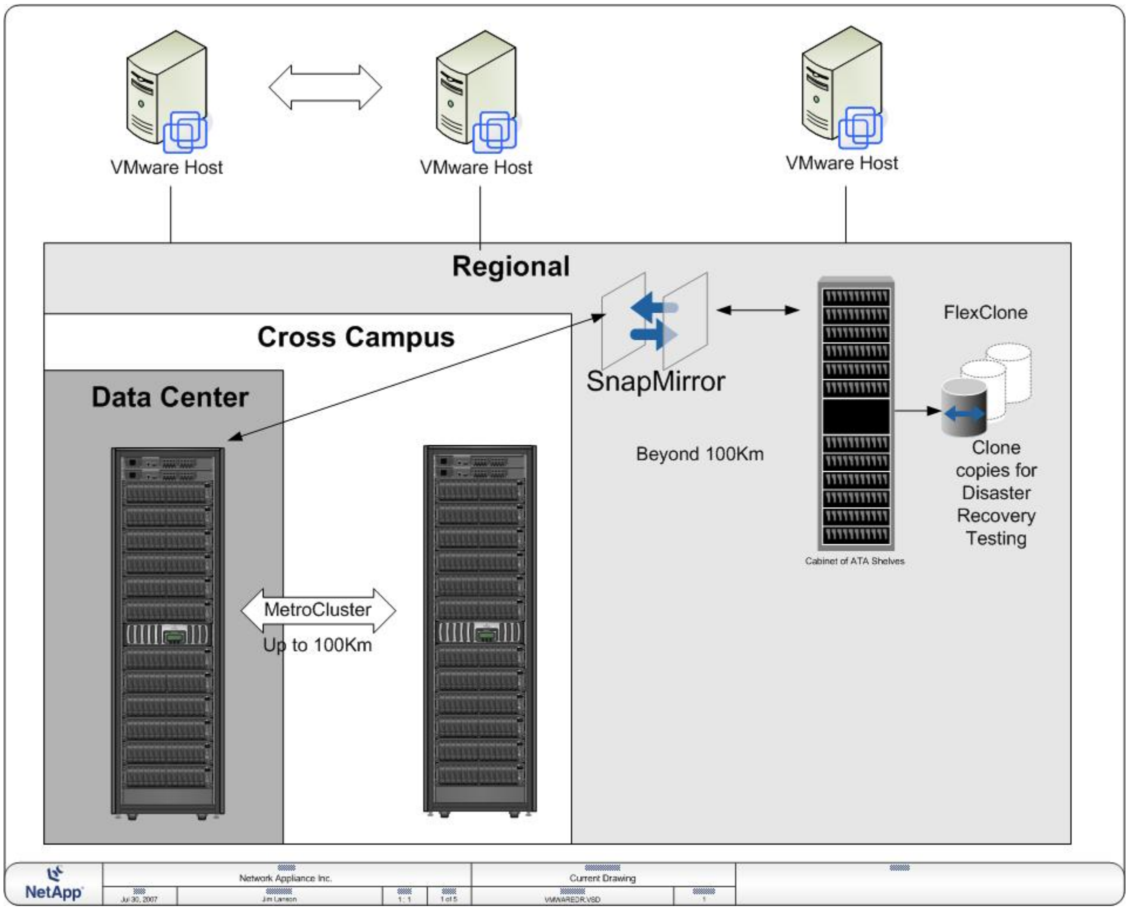


*Figure 1) IBM N series storage at data-center, campus, and regional levels.*

Tiered storage can also increase return on investment (ROI), because this architecture utilizes hardware from both sites. While usage of hardware at the primary site is a high percentage, in a typical DR architecture, hardware at the secondary site sits idle. The secondary site is typically used as a standby site, and hardware is rarely used for anything else (in a non-VMware/IBM N series infrastructure).

With VMware/IBM N series DR architecture, you can create tiered storage architecture so that the primary site continues to be used as it currently is; however, the secondary site’s hardware can also be used for tiered applications such as test /development or critical actions such as testing the DR capabilities of the architecture. VI3 allows such utilization at the alternate sites due to the ability of the administrator to start a VM copy on any server. The encapsulation of the VM’s environment into files allows this to happen.

## Data center

VMware Infrastructure provides inherent high availability at several levels. By their nature, VMs leverage high-availability features in a physical server across all the VMs on that server. Fault-tolerant hardware features such as teamed network interfaces, multiple SAN storage adapters, and redundant power supplies and memory may be prohibitively expensive for a server running a single application, but they become economical when their costs are divided among many VMs.

VMware Infrastructure changes the way that information systems are designed. Featuring such advanced capabilities as migration of VMs between any virtualization platforms, IBM System Storage N series with Snapshot™ copies, automated restarts on alternate hosts in a resource pool, and VMotion, VMware Infrastructure creates environments where outages are limited to brief restarts at most. For a continuous availability solution to guard against application or hardware failure, VMware HA provides easy-to-use, cost-effective protection for applications running on VMs. In the event of server failure, affected VMs are automatically restarted on other physical servers in a VMware Infrastructure resource pool that have spare capacity.

VMware HA minimizes downtime and IT service disruption while eliminating the need for dedicated standby hardware and installation of additional software. VMware HA provides uniform high availability across the entire virtualized IT environment without the cost and complexity of failover solutions tied to either operating systems or specific applications.

When 100% uptime is imperative, IT managers can create a cluster between a physical machine running mission-critical workloads and a similarly configured VM. The VMs do not consume computing resources in standby mode and can be consolidated to one or a few physical platforms at a very high consolidation ratio. As a result, the enterprise can realize high-availability benefits without having to invest in twice the amount of hardware or having to manage and patch sprawling servers. Redundancy is reduced from 2N to N+1.

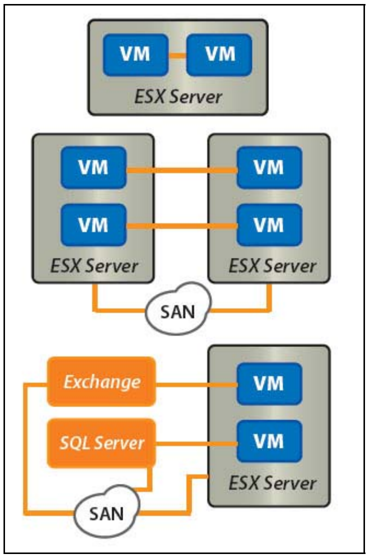


*Figure 2) VI3 layout.*

Physical-to-virtual machine clustering supports the same clustering software as physical-to-physical machine clustering. In fact, the same clustering software is supported for VMs as for their physical equivalent, including Microsoft® clustering, Veritas™ clustering, Legato AAM, and NSI Double-Take, so no IT ramp up is required. At the same time, reduced cost allows implementation of high availability and SLAs for more workloads.

Many IBM N series standard product features can provide protection and quick recovery in the event of a data center disaster such as a power outage, environmental failures such as air conditioning, hardware failures, and human error. Human error includes the unintentional erasure of data or mistakes in following data protection procedures. A single IBM N series storage node can protect against the types of failure shown in the following table.

| **Failure** | **Protection** |
|---|---|
| Failure of power supply, fan, or disk controller | Built-in redundancy |
| Single or dual disk failure | IBM System Storage N series with RAID-DP™ |
| Single disk path or port failure | Multipath |
| Accidental erasure or destruction of data | Snapshot copies |

## Cross-campus

Virtual Infrastructure deployed in conjunction with a storage-area network (SAN) has an additional built-in level of robustness. Any VM that resides on a SAN can survive a crash of the server hardware that runs this VM, and can be restarted on another ESX Server at an alternate campus location. Utilizing a SAN's replication technology, a VM can be replicated and restored anywhere in the world, whether it's cross-campus or cross-country, with little IT staff intervention.

From a storage perspective, MetroCluster provides protection in the event of a complete campus data center failure. From the loss of a disk shelf or controller to the loss of the building itself, MetroCluster offers quick recovery, minimizing resource outages and data loss. In addition to the protections provided in the data center, MetroCluster protects against the failures shown in the following table.

| **Failure** | **Protection** |
|---|---|
| Triple disk failure | IBM System Storage N series with SyncMirror® |
| Complete disk shelf failure | SyncMirror |
| HBA or port failure | Redundancy |
| Storage controller failure | Active-active controller configuration |
| Data center power or environmental outage | MetroCluster |

## Regional

One of the main benefits of virtualization for DR is independence of the recovery process from the recovery hardware. Because VMs encapsulate the complete environment, including data, application, operating system, BIOS, and virtualized hardware, applications can be restored to any hardware with a virtualization platform without concern for the differences in underlying hardware. The physical world limitation of having to restore to an identical platform does not apply. Not only does hardware independence allow IT managers to eliminate manual processes associated with adjusting drivers and BIOS versions to reflect the change in platform, it also eliminates Microsoft Windows® registry issues and plug-and-play issues. By leveraging the hardware independence of VMware VMs,,customers no longer need to worry about the need for identical hardware at their DR sites, which can significantly reduce the cost and complexity of regional DR. VMware enterprise customers actively take advantage of VMware consolidation benefits for their production and staging servers. These consolidation benefits are even greater for the failover hardware, because customers can consolidate servers at the primary data center to fewer physical servers at their DR centers.

Another benefit of VMs that helps ease the complexity of DR is the VMware flexible networking features. Because VMware handles virtual local area networks (VLANs) on its virtual switches, entire complex network environments can be isolated, contained, or migrated very easily with little setup at the DR site.

For the most critical applications, many enterprises turn to storage-array-based data replication to a failover site. This approach provides the most up-to-date copy of the data and applications at a remote location, thereby protecting data from a regional disaster as well as from hardware failure. Virtual Infrastructure combined with array-based replication allows enterprises to replicate the encapsulated VMs to the secondary site and to bring it up at the secondary site in a programmatic way, without human intervention, on any available ESX Server. The hardware independence of VMs means that the ESX Server hardware at the secondary data center does not need to match the ESX Server hardware configuration at the primary data center. Furthermore, a higher ratio of server consolidation can be maintained at the secondary site.

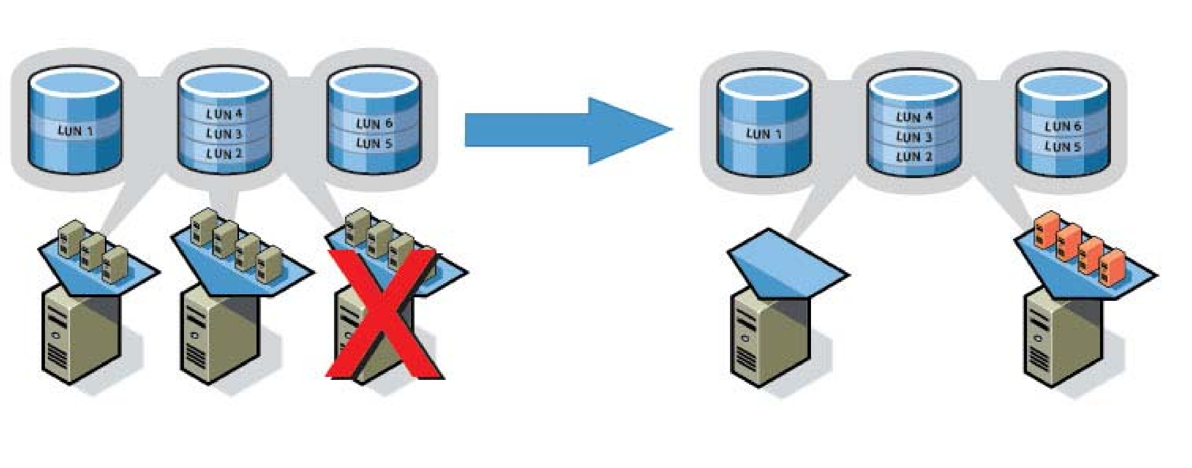


*Figure 3) VMware hardware independence.*

If an incident occurs that makes the entire campus unavailable, SnapMirror provides long-distance replication to protect against such incidents. Operating either asynchronously or synchronously, SnapMirror utilizes snapshot copies to make replication both easy and efficient.

Many customers do not test their remote DR capability because they simply cannot have any downtime. Consequently, they may assume that everything is ready to go at the DR site. One of the tests performed in this report is the use of FlexClone technology to create a copy of the data for testing purposes. This can be done without any disruption to the active system or the replication of data.

The key to this architecture is that the entire primary site can be copied to the secondary or tertiary site (using MetroCluster or SnapMirror) and brought up in order to test or develop against data created by the primary site. Alternate sites are always created but are seldom tested. This is another key benefit to using VI3 and IBM N series storage.

The tested storage infrastructure includes MetroCluster on the main campus for data center protection. One MetroCluster node is inside the data center and the other is located in a building across campus. A VMware ESX server at each of these locations provides host-level protection (VMware HA). The servers are running ESX 3.0.1 with six Windows 2003 VMs. Five of these VMs use an iSCSI logical unit number (LUN) for their storage. The other uses a network file system (NFS) storage device. For remote DR there is an ESX Server at the DR site, along with a third storage system. Also at the data center, SnapMirror is installed to provide storage protection against a complete main campus disaster. SnapMirror can replicate asynchronously or synchronously so that data can be replicated to the DR site according to distance and data currency needs. FlexClone technology is used at the DR site for nondisruptive testing of remote DR. Figure 1 (on page 7 of the report) shows the general layout of components used in this sample configuration. For detailed lists of materials used, see Appendix A.

# Production site setup and configuration (SITEA)

## IBM N series

### Configuration

The IBM N series controller and back-end FC switches are installed and configured using the instructions in the IBM System Storage N series with Data ONTAP® configuration guide. As of this report writing, the most current software version levels were:

Data ONTAP 7.2.3
Brocade firmware 5.1.0.

The production site storage controller (referred to hereafter as METRO3050-SITEA) is an IBM N series N5500 with two EXP4000 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 FC switch fabrics to another N5500 controller at the campus DR 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:

cifs
cluster: required for MetroCluster
cluster_remote: required for MetroCluster
flex_clone: required for DR testing
iscsi: used for VMware datastores
nfs: used for VMware datastores
syncmirror_local: required for MetroCluster
snapmirror.

### 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 C.

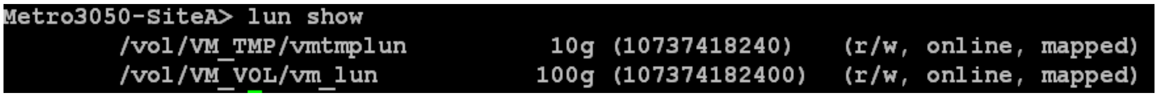
### Volume layout

The hardware in this configuration is limited to 14 mirrored disks on the controller head. Three of these are for the root volume and one is reserved for a spare. The remaining 24 disks have been used to create an aggregate to host the volumes. The controller at SITEA has one volume (VM_VOL) to house the iSCSI LUN-based active datastores. Another volume (VMNFS) contains the NFS export for the VMware NFS datastore.

A third volume (VM_TMP) contains another iSCSI LUN to be used as a datastore for the VM's temporary files. Both VM_VOL and VM_TMP are replicated to the off-campus DR site using SnapMirror. For more details on the LUNs, aggregates, and volumes, see Appendix B.

### iSCSI

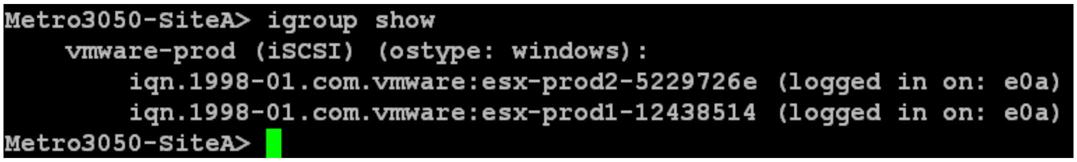
Two iSCSI LUNs were created as shown in Figure 4. Sizes were arbitrarily chosen for these tests.



```
Metro3050-SiteA> lun show
        /vol/VM_TMP/vmtmplun           10g (10737418240)   (r/w, online, mapped)
        /vol/VM_VOL/vm_lun            100g (107374182400)  (r/w, online, mapped)
```

*Figure 4) IBM N series controller LUN configuration.*

The two iSCSI LUNs created were then assigned to an igroup called vmware-prod containing the iSCSI IQN numbers for all servers in the VMware cluster (ESX-PROD1 and ESX-PROD2).



```
Metro3050-SiteA> igroup show
    vmware-prod (iSCSI) (ostype: windows):
        iqn.1998-01.com.vmware:esx-prod2-5229726e (logged in on: e0a)
        iqn.1998-01.com.vmware:esx-prod1-12438514 (logged in on: e0a)
Metro3050-SiteA>
```

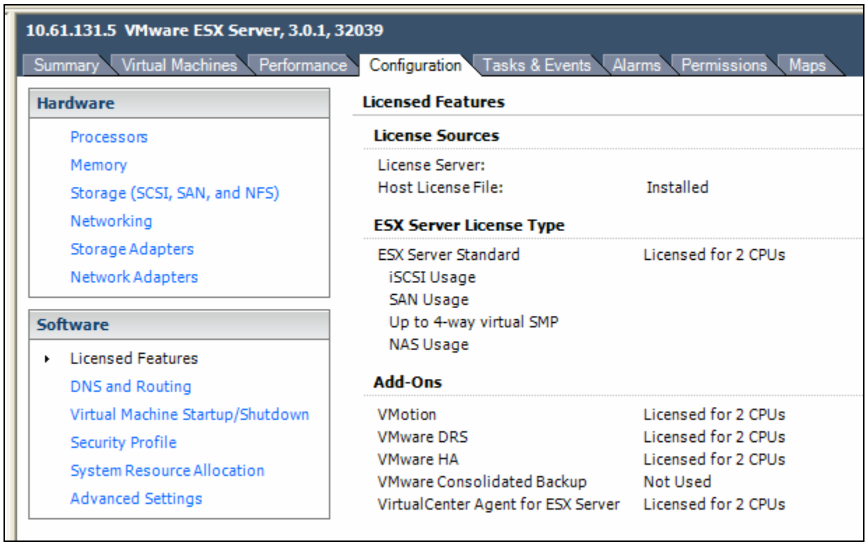
*Figure 5) IBM N series controller Igroup configuration.*

## VMware

The two ESX Servers in the cluster are installed according to the vendor-supplied procedures in the VMware ESX 3.0.1 installation guide.

### Server configuration

Features that are licensed/enabled…



**Licensed Features**

**License Sources**

| | |
|---|---|
| License Server: | |
| Host License File: | Installed |

**ESX Server License Type**

| | |
|---|---|
| ESX Server Standard | Licensed for 2 CPUs |
| iSCSI Usage | |
| SAN Usage | |
| Up to 4-way virtual SMP | |
| NAS Usage | |

**Add-Ons**

| | |
|---|---|
| VMotion | Licensed for 2 CPUs |
| VMware DRS | Licensed for 2 CPUs |
| VMware HA | Licensed for 2 CPUs |
| VMware Consolidated Backup | Not Used |
| VirtualCenter Agent for ESX Server | Licensed for 2 CPUs |

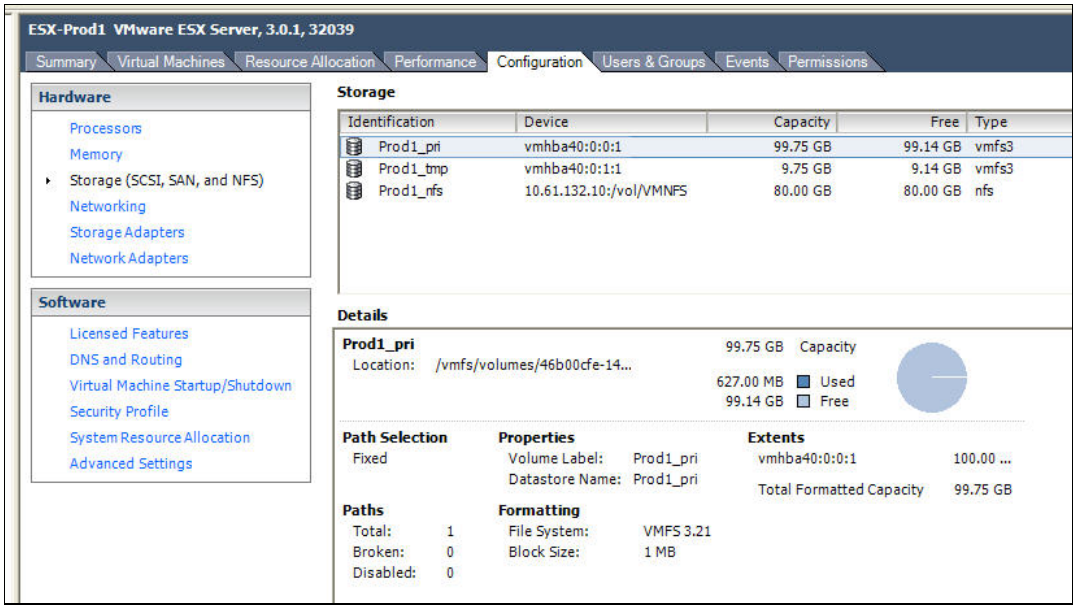
*Figure 6) VMware ESX Server licenses.*

## Datastores

Three datastores were created for the following purposes, as shown in the following table.

| Name | Use | SnapMirror used? |
|---|---|---|
| Prod1_pri | Primary storage for VMs (iSCSI LUNS) | Y |
| Prod1_tmp | Temporary or swap storage for VMs | N |
| Prod1_nfs | Primary storage for VMs (NFS) | Y |

Once the datastores were created, visibility was verified in the ESX Server, as shown in Figure 7.



Storage

| Identification | Device | Capacity | Free | Type |
|---|---|---|---|---|
| Prod1_pri | vmhba40:0:0:1 | 99.75 GB | 99.14 GB | vmfs3 |
| Prod1_tmp | vmhba40:0:1:1 | 9.75 GB | 9.14 GB | vmfs3 |
| Prod1_nfs | 10.61.132.10:/vol/VMNFS | 80.00 GB | 80.00 GB | nfs |

Details

Prod1_pri
Location: /vmfs/volumes/46b00cfe-14...

99.75 GB Capacity
627.00 MB ■ Used
99.14 GB □ Free

Path Selection
Fixed

Properties
Volume Label: Prod1_pri
Datastore Name: Prod1_pri

Extents
vmhba40:0:0:1 100.00 ...
Total Formatted Capacity 99.75 GB

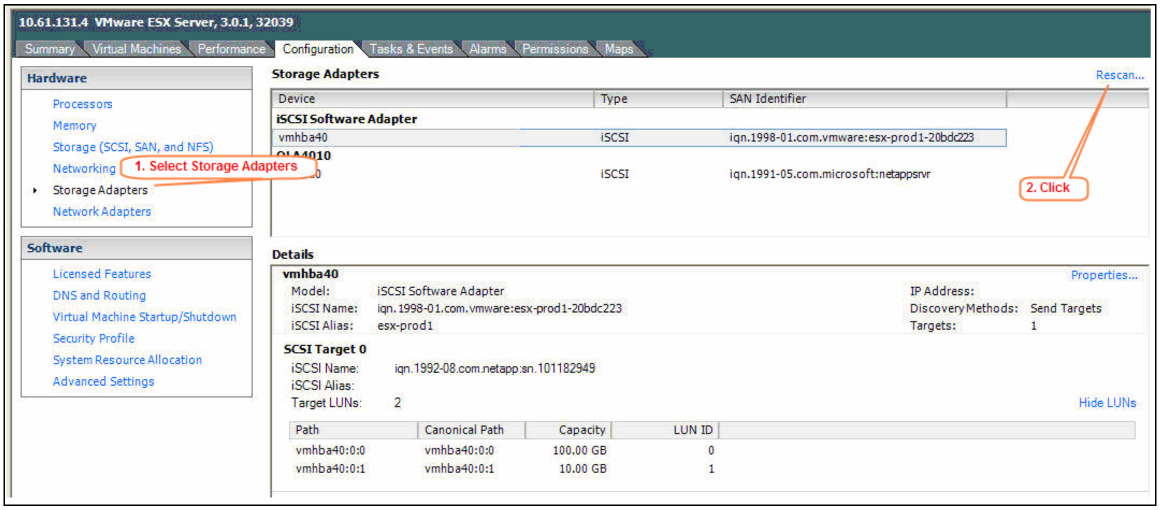
Paths
Total: 1
Broken: 0
Disabled: 0

Formatting
File System: VMFS 3.21
Block Size: 1 MB

*Figure 7) Datastore setup on primary (ESX-Prod1).*

## iSCSI and LUN setup

For the ESX Server to see the LUNs created, one or more steps are necessary. If the iSCSI storage adapter is already running, then all that may be necessary is to tell the iSCSI storage adapter to rescan for devices. After a few moments, the new LUNs are displayed.



Storage Adapters

| Device | Type | SAN Identifier |
|---|---|---|
| **iSCSI Software Adapter** | | |
| vmhba40 | iSCSI | iqn.1998-01.com.vmware:esx-prod1-20bdc223 |
| **QLA4010** | | |
| 0 | iSCSI | iqn.1991-05.com.microsoft:netappsrvr |

Details

**vmhba40**

Model: iSCSI Software Adapter
iSCSI Name: iqn.1998-01.com.vmware:esx-prod1-20bdc223
iSCSI Alias: esx-prod1

IP Address:
Discovery Methods: Send Targets
Targets: 1

**SCSI Target 0**

iSCSI Name: iqn.1992-08.com.netapp:sn.101182949
iSCSI Alias:
Target LUNs: 2

| Path | Canonical Path | Capacity | LUN ID |
|---|---|---|---|
| vmhba40:0:0 | vmhba40:0:0 | 100.00 GB | 0 |
| vmhba40:0:1 | vmhba40:0:1 | 10.00 GB | 1 |

*Figure 8) iSCSI adapter and LUNs.*

If no SAN identifier is displayed for the iSCSI storage adapter, go back to the storage controller and execute the igroup show command. If it shows that the ESX Server iSCSI initiator is not logging into the controller, there may be a communication problem. Make sure that you can ping the VM Kernel IP address from the IBM N series system (Figures 9 and 10).

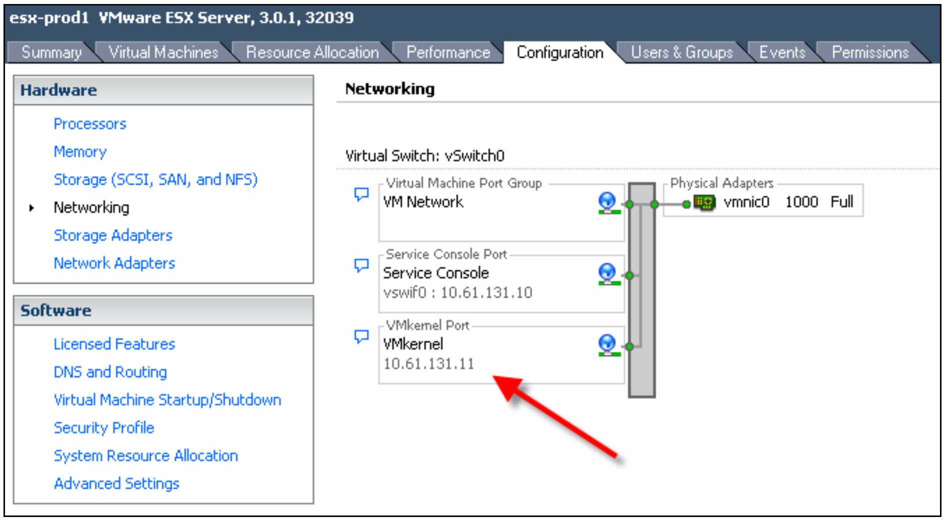
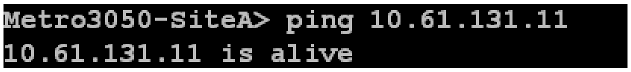


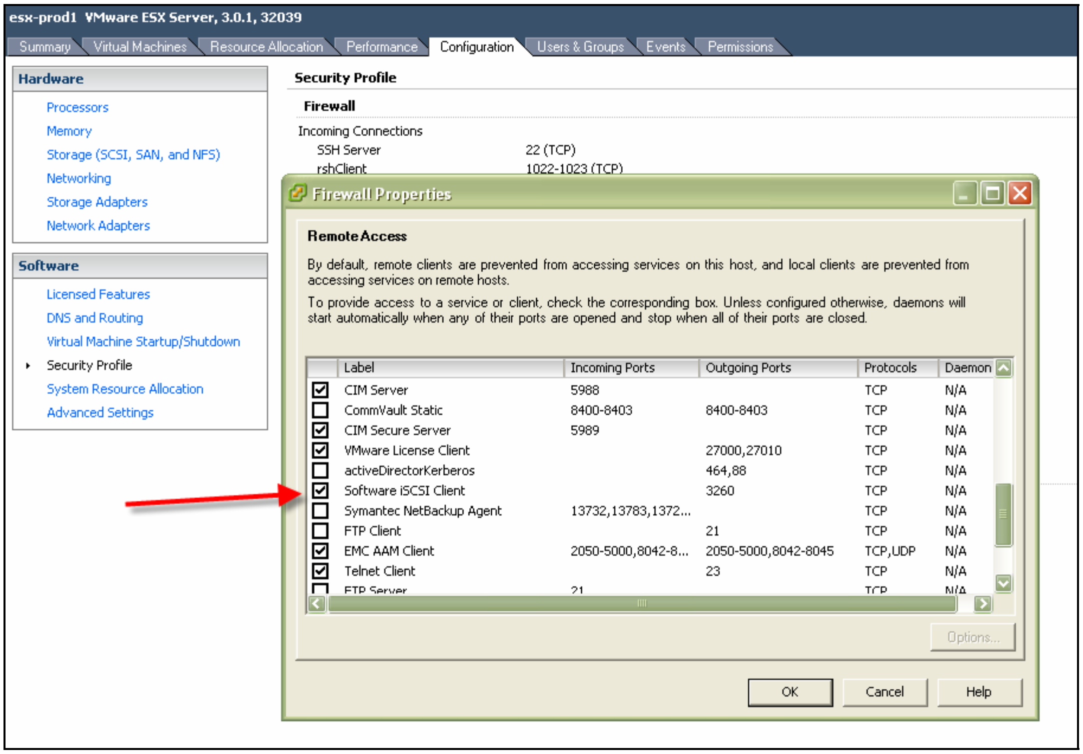
*Figure 9) Obtaining a VMkernel IP address.*



```
Metro3050-SiteA> ping 10.61.131.11
10.61.131.11 is alive
```

*Figure 10) Verifying iSCSI communication.*

Also verify the ESX Firewall configuration in Configuration > Security Profile > Properties to make sure that the iSCSI initiator port is open (Figure 11). This allows the ESX Server iSCSI initiator to log into the storage controller.



Security Profile

Firewall

Incoming Connections

| | |
|---|---|
| SSH Server | 22 (TCP) |
| rshClient | 1022-1023 (TCP) |

**Remote Access**

By default, remote clients are prevented from accessing services on this host, and local clients are prevented from accessing services on remote hosts.

To provide access to a service or client, check the corresponding box. Unless configured otherwise, daemons will start automatically when any of their ports are opened and stop when all of their ports are closed.

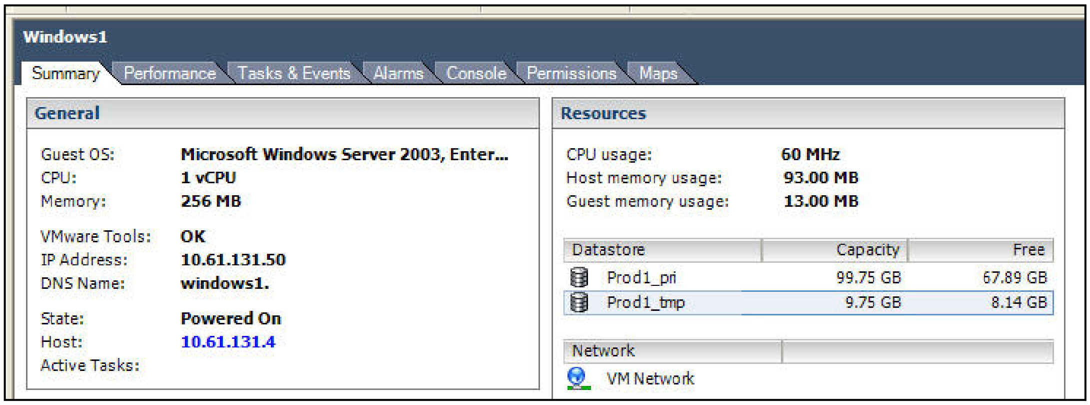
| | Label | Incoming Ports | Outgoing Ports | Protocols | Daemon |
|---|---|---|---|---|---|
| ☑ | CIM Server | 5988 | | TCP | N/A |
| □ | CommVault Static | 8400-8403 | 8400-8403 | TCP | N/A |
| ☑ | CIM Secure Server | 5989 | | TCP | N/A |
| ☑ | VMware License Client | | 27000,27010 | TCP | N/A |
| □ | activeDirectorKerberos | | 464,88 | TCP | N/A |
| ☑ | Software iSCSI Client | | 3260 | TCP | N/A |
| □ | Symantec NetBackup Agent | 13732,13783,1372... | | TCP | N/A |
| □ | FTP Client | | 21 | TCP | N/A |
| ☑ | EMC AAM Client | 2050-5000,8042-8... | 2050-5000,8042-8045 | TCP,UDP | N/A |
| ☑ | Telnet Client | | 23 | TCP | N/A |
| □ | FTP Server | 21 | | TCP | N/A |

*Figure 11) Enabling iSCSI firewall port.*

## VMs

Six VMs were then created. All were Windows 2003 and used a separate datastore for temporary storage (Prod1_tmp). The purpose of this, as a best practice, was to avoid SnapMirror replication of temporary data.

Five of the VMs use an iSCSI LUN (Prod1_pri) for primary storage. The sixth uses an NFS drive (Prod1_nfs) for primary storage (Figures 12a and 12b).



General

| | |
|---|---|
| Guest OS: | Microsoft Windows Server 2003, Enter... |
| CPU: | 1 vCPU |
| Memory: | 256 MB |
| VMware Tools: | OK |
| IP Address: | 10.61.131.50 |
| DNS Name: | windows1. |
| State: | Powered On |
| Host: | 10.61.131.4 |
| Active Tasks: | |

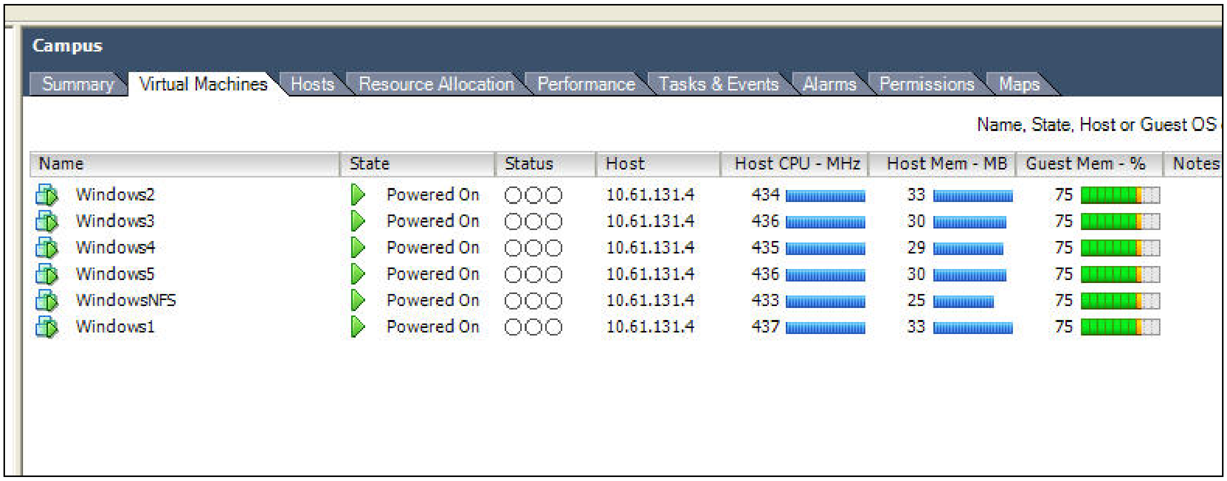
Resources

| | |
|---|---|
| CPU usage: | 60 MHz |
| Host memory usage: | 93.00 MB |
| Guest memory usage: | 13.00 MB |

| Datastore | Capacity | Free |
|---|---|---|
| Prod1_pri | 99.75 GB | 67.89 GB |
| Prod1_tmp | 9.75 GB | 8.14 GB |

| Network |
|---|
| VM Network |

*Figure 12a) Windows VM setup.*



Campus

Name, State, Host or Guest OS

| Name | State | Status | Host | Host CPU - MHz | Host Mem - MB | Guest Mem - % | Notes |
|---|---|---|---|---|---|---|---|
| Windows2 | Powered On | ○○○ | 10.61.131.4 | 434 | 33 | 75 | |
| Windows3 | Powered On | ○○○ | 10.61.131.4 | 436 | 30 | 75 | |
| Windows4 | Powered On | ○○○ | 10.61.131.4 | 435 | 29 | 75 | |
| Windows5 | Powered On | ○○○ | 10.61.131.4 | 436 | 30 | 75 | |
| WindowsNFS | Powered On | ○○○ | 10.61.131.4 | 433 | 25 | 75 | |
| Windows1 | Powered On | ○○○ | 10.61.131.4 | 437 | 33 | 75 | |

*Figure 12b) All completed VMs.*

Figure 12c shows the completed production site setup. This site includes the datastores and the VMs.

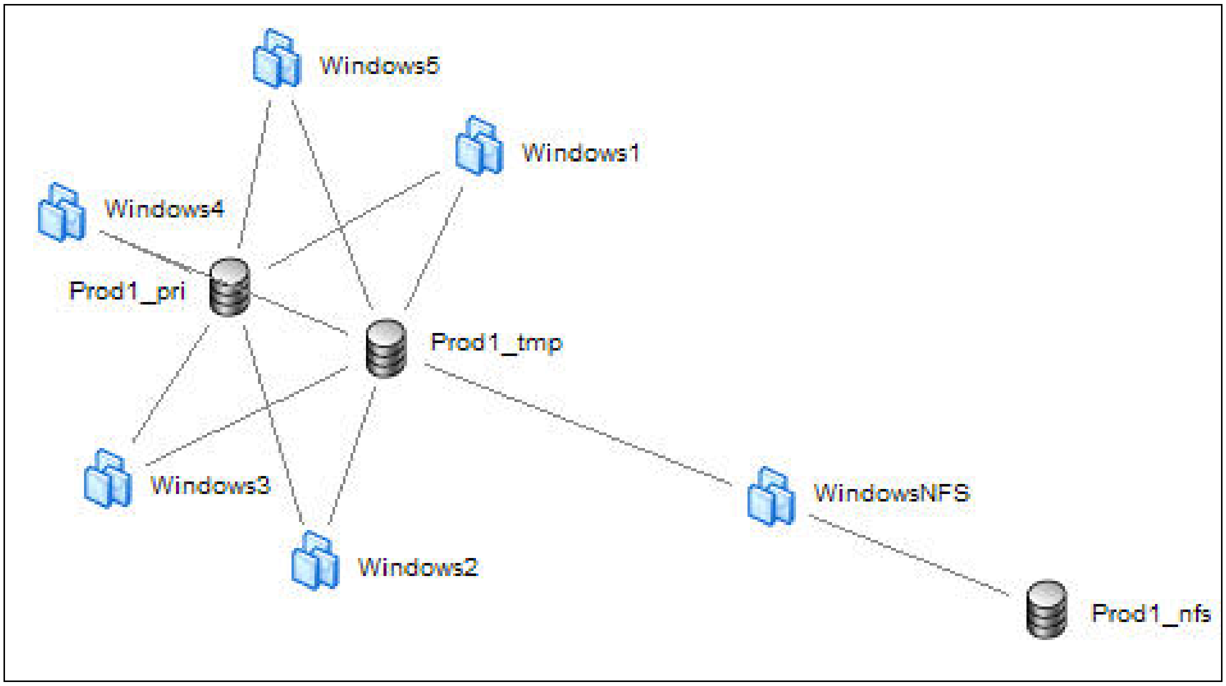


*Figure 12c) Production site setup.*

## On-campus DR site (SITEB)

To simplify this functional testing, the MetroCluster IBM N series storage systems were used in an active/passive configuration. Consequently there was little configuration and setup to perform other than to set up the proper licenses and verify connectivity for failover purposes. Also, VMware VMotion was configured to verify nondisruptive migration of VMs between the two sites.

### IBM N series

The IBM N series controller at the on-campus DR location is also installed and configured using the instructions in the Data ONTAP configuration guide. At the time of this report writing, the most current software version levels are:

Data ONTAP 7.2.3
Brocade firmware 5.1.0.

The SITEB storage controller (referred hereinafter as METRO3050-SITEB) is also an IBM N series 5500 with two EXP4000 shelves fully populated with 66GB 10k rpm drives. It is the passive node for the fabric MetroCluster that communicates with the SITEA controller using the FC/VI interconnect by way of the switch fabrics described earlier.

The controller at SITEB has just the root volume mirrored to SITEA, a requirement of MetroCluster. In this case it is a passive member of the MetroCluster, so no other aggregates or volumes were created.

### VMware

#### Configuration

The ESX Server at this site is configured for an active/passive role. It is a member of a VMware cluster (VMware HA). Because it is the passive partner of an active/passive configuration, it has no active datastores but has full access to the production datastores in the event of a data center disaster. Other than that. it is configured identically to the production ESX Server (ESX-PROD1).

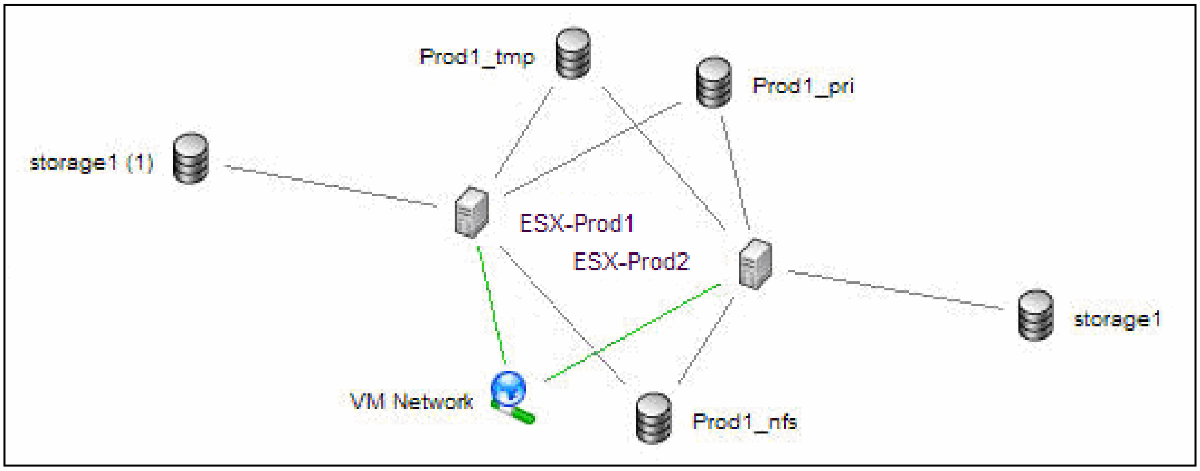


*Figure 13) Primary and campus DR complete.*

To verify proper operation of the now complete VMware cluster, VMotion, VMware DRS, and VMware HA were enabled. VMotion was used to verify nondisruptive migration between the two nodes of the cluster. When tested, it was discovered that although the central processing units (CPUs) at each site were the same manufacturer and type (Intel® Xeon™), they had different clock speeds and different CPU IDs. For this reason, migration involved powering down the VM, moving it to the other server, then powering it on. This was disruptive. When the CPU IDs were masked out (information from www.vmware.com), the process became nondisruptive, allowing the VMs to be migrated while powered on.

# Regional DR site

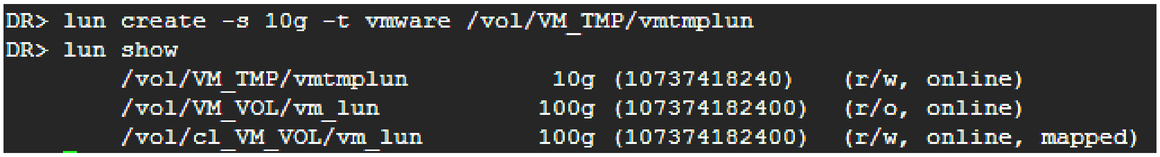
## IBM N series

A standalone IBM N series n7800 storage controller was used as DR site storage (thus the SnapMirror destination) to provide protection in the case of a complete campus disaster. It was not part of the MetroCluster but was configured in a similar fashion (see Appendix B). Its purpose was to enable a complete VMware environment to be brought up in case of loss of the entire main campus.

All data was replicated using SnapMirror, except for temporary file data. Because the LUN for temporary data was not replicated, a LUN had to be created on DR (Figure 14) in order to facilitate operation during this failover situation.

The following features are licensed on this controller:

flex_clone: required for DR testing
iscsi: used for VMware datastores
nfs: used for VMware datastores
snapmirror.



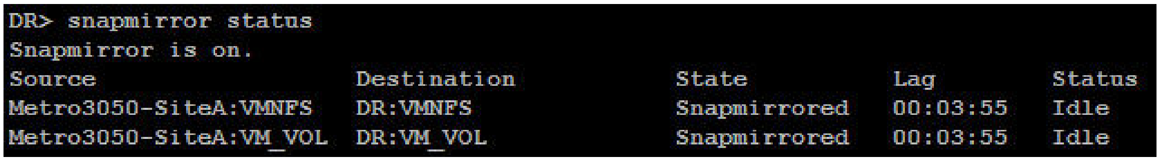
```
DR> lun create -s 10g -t vmware /vol/VM_TMP/vmtmplun
DR> lun show
        /vol/VM_TMP/vmtmplun          10g (10737418240)    (r/w, online)
        /vol/VM_VOL/vm_lun           100g (107374182400)   (r/o, online)
        /vol/cl_VM_VOL/vm_lun        100g (107374182400)   (r/w, online, mapped)
```

*Figure 14) Creation of LUN on DR site for temporary data.*

FlexClone was licensed to provide the ability to perform nondisruptive DR testing. Steps performed are covered in a later report section, “Operational Scenarios.”

### SnapMirror

A SnapMirror relationship was set up for the two volumes (`/vol/VM_VOL` and `/vol/VMNFS`) containing the primary datastores. METRO3050-SITEA is defined as the source with DR as the destination and an update interval of every 15 minutes.



```
DR> snapmirror status
Snapmirror is on.
Source                      Destination            State           Lag         Status
Metro3050-SiteA:VMNFS       DR:VMNFS               Snapmirrored    00:03:55    Idle
Metro3050-SiteA:VM_VOL      DR:VM_VOL              Snapmirrored    00:03:55    Idle
```

*Figure 15) SnapMirror configuration.*

## VMware

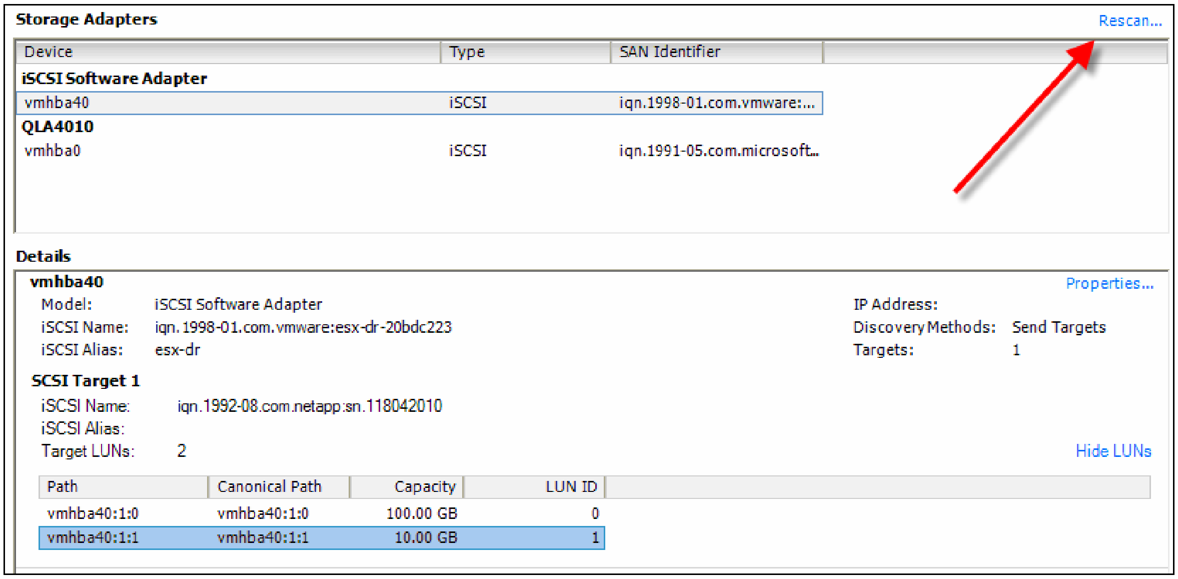
A third ESX Server (ESX-DR) was configured to represent a DR server for the VMware environment. It was configured identically to the other two, except that it was not part of the VMware cluster. It had no access to the datastores and had only one of its own configured (for temporary data).

### Temporary LUN configuration

For the ESX Server to see the LUNs created, several steps are necessary.

In Advanced Settings, set LVM.EnableResignature to 1. Under Hardware and Storage, click Refresh.

In Configuration > Storage Adapters, click Rescan at the right (Figure 16).



**Storage Adapters**

| Device | Type | SAN Identifier |
|---|---|---|
| **iSCSI Software Adapter** | | |
| vmhba40 | iSCSI | iqn.1998-01.com.vmware:... |
| **QLA4010** | | |
| vmhba0 | iSCSI | iqn.1991-05.com.microsoft... |

**Details**

**vmhba40**

Model: iSCSI Software Adapter
iSCSI Name: iqn.1998-01.com.vmware:esx-dr-20bdc223
iSCSI Alias: esx-dr

IP Address:
Discovery Methods: Send Targets
Targets: 1

**SCSI Target 1**

iSCSI Name: iqn.1992-08.com.netapp:sn.118042010
iSCSI Alias:
Target LUNs: 2

| Path | Canonical Path | Capacity | LUN ID |
|---|---|---|---|
| vmhba40:1:0 | vmhba40:1:0 | 100.00 GB | 0 |
| vmhba40:1:1 | vmhba40:1:1 | 10.00 GB | 1 |

*Figure 16) Primary and campus DR complete.*

If the LUN is not detected, follow the troubleshooting tips in the iSCSI/LUN portion of the report.

# Complete DR network

At this point the test IBM N series/VMware Infrastructure environment should be up and running (Figure 17). Operational scenarios can now begin.

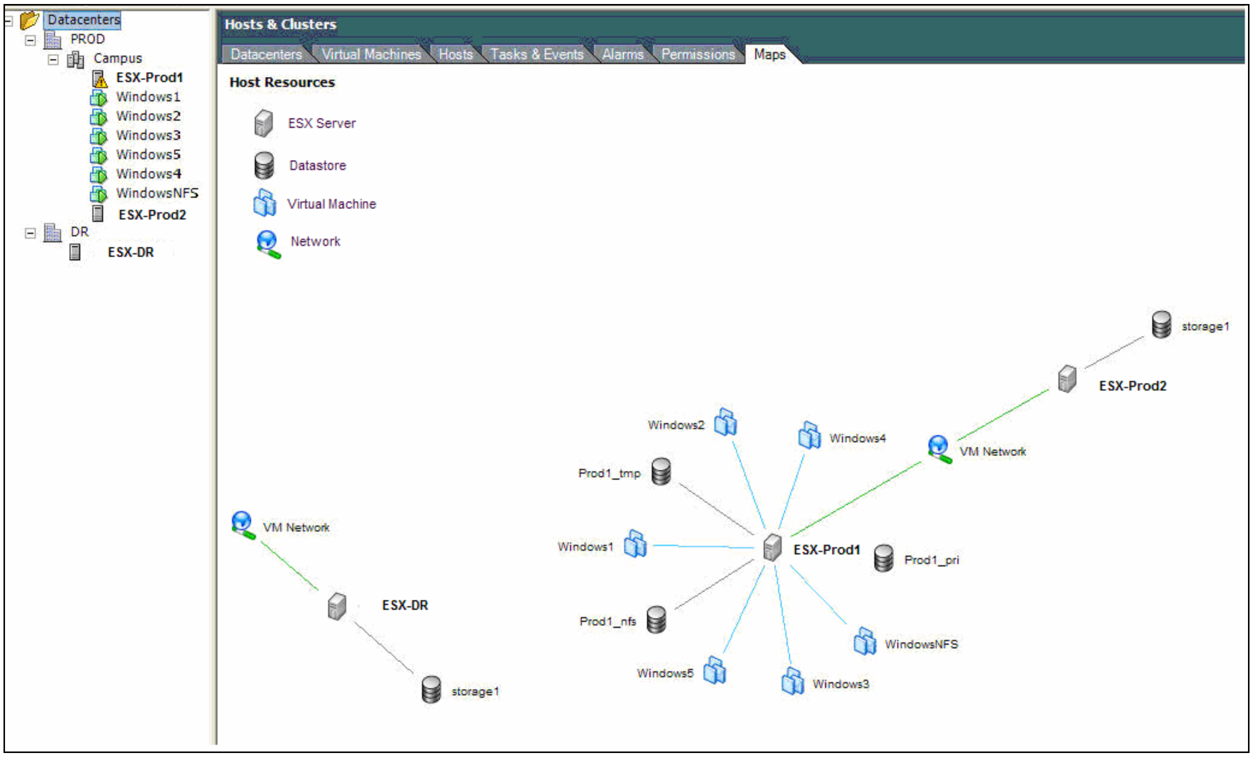


*Figure 17) Primary and campus DR complete.*

# Operational scenarios

The following sections detail various scenarios that were executed after successful installation and configuration of the solution previously described. The purpose of these scenarios was to examine and document, from a DR perspective, the reaction of a VMware/IBM N series environment to various resource losses.

These scenarios include various component, host, and site failure scenarios. Unless stated otherwise, prior to the execution of each test the environment is reset to the “normal” running state. To generate storage activity, the IBM N series spatial input-output (SIO) utility was used.

## Complete loss of power to disk shelf

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

| | |
|---|---|
| **Task** | Power off the METRO3050-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. |
| **Actual Results** | Results were as expected. The storage controllers detected and reported the problem. When the shelf was reconnected, resynchronization began automatically. There was no interruption of disk activity or VM operation. |

## Loss of one link on one disk loop

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

| | |
|---|---|
| **Task** | Remove the fiber entering METRO3050-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. |
| **Actual Results** | Results were as expected. The storage controllers detected and reported the problem. When the loop was reconnected, the controllers responded accordingly. There was no interruption of disk activity or VM operation. |

## Loss of brocade switch

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 was applied.

| | |
|---|---|
| **Task** | Power off the FC 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. |
| **Actual Results** | Results were as expected. The storage controllers detected and reported the problem. When the switch was reconnected, the controllers responded accordingly. There was no interruption of disk activity or VM operation. |

## Loss of one interswitch link (ISL)

No single point of failure should exist in the solution. Therefore the loss of one of the interswitch links was tested. This test was accomplished by simply removing the fiber between two of the switches while a load was 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 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. |
| **Actual Results** | Results were as expected. The storage controllers detected and reported the problem. When the ISL was reconnected, the controllers responded accordingly. There was no interruption of disk activity or VM operation. |

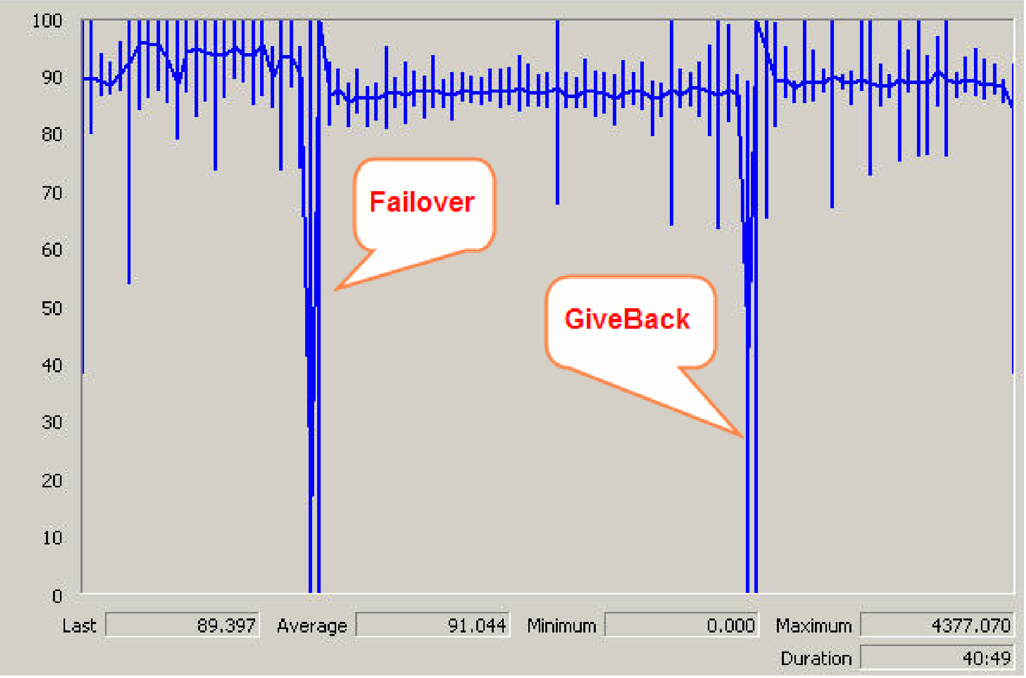
## Failure of controller

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

| | |
|---|---|
| **Task** | Power off the METRO3050-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. |
| **Actual Results** | The partner controller reported the outage and began automatic takeover. There was a momentary pause in disk activity. When takeover was complete, activity returned to normal. Neither the VMs nor the ESX Server detected any problem. |

## Failback of controller

As a follow-up to the previous test, the production data service was failed back to the previously failed controller (METRO3050-SITEA) to return to the normal operating state. This test was accomplished by issuing a command on the surviving controller (METRO3050-SITEB) to request that processing be returned to the previously failed controller.

| | |
|---|---|
| **Task** | Power on SITEA. Issue a cf giveback command on METRO3050-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. |
| **Actual Results** | The partner controller reported the outage and began automatic takeover. There was a momentary pause in disk activity. When takeover was complete, activity returned to normal. Neither the VMs nor the ESX Server detected any problem.<br><br> <br>Last 89.397 Average 91.044 Minimum 0.000 Maximum 4377.070<br>Duration 40:49<br>*Figure 19) Failback of controller test—power on and takeover of site.* |

## Loss of primary data center, disaster declared

To test the availability of the overall solution, recovery after loss of the primary campus site was simulated.

**Task**

Test the failure of SITEA by interrupting the following components in this order, in rapid succ

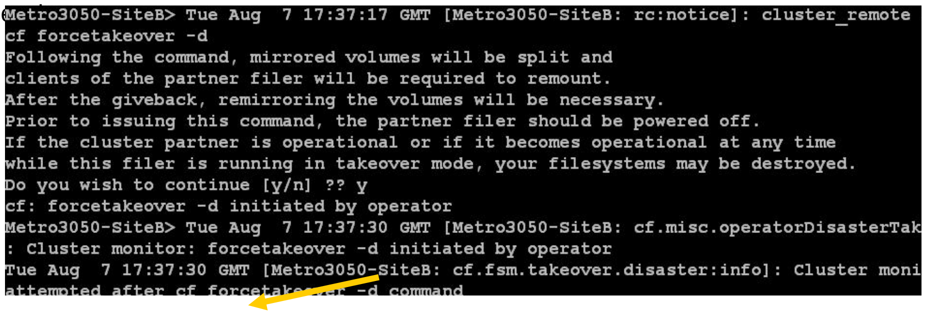
1.

2.

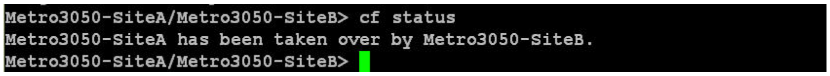
3.

4.

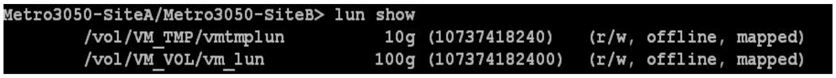
5.



```
Metro3050-SiteB> Tue Aug  7 17:37:17 GMT [Metro3050-SiteB: rc:notice]: cluster_remote
cf forcetakeover -d
Following the command, mirrored volumes will be split and
clients of the partner filer will be required to remount.
After the giveback, remirroring the volumes will be necessary.
Prior to issuing this command, the partner filer should be powered off.
If the cluster partner is operational or if it becomes operational at any time
while this filer is running in takeover mode, your filesystems may be destroyed.
Do you wish to continue [y/n] ?? y
cf: forcetakeover -d initiated by operator
Metro3050-SiteB> Tue Aug  7 17:37:30 GMT [Metro3050-SiteB: cf.misc.operatorDisasterTak
: Cluster monitor: forcetakeover -d initiated by operator
Tue Aug  7 17:37:30 GMT [Metro3050-SiteB: cf.fsm.takeover.disaster:info]: Cluster moni
attempted after cf forcetakeover -d command
```



```
Metro3050-SiteA/Metro3050-SiteB> cf status
Metro3050-SiteA has been taken over by Metro3050-SiteB.
Metro3050-SiteA/Metro3050-SiteB>
```

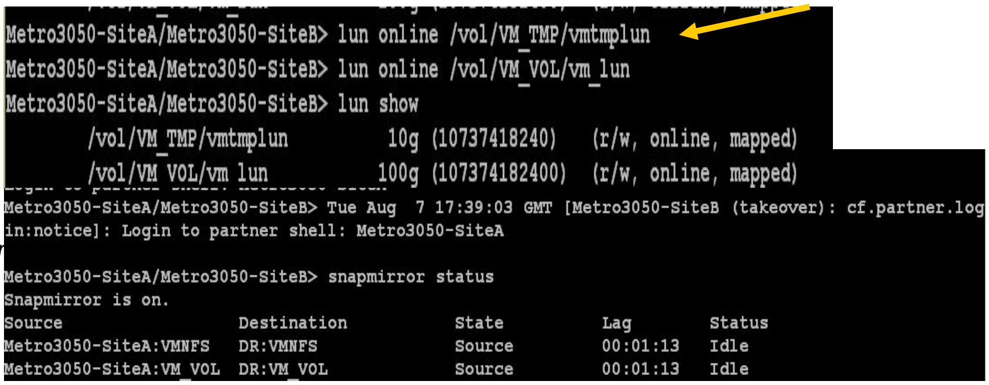


```
Metro3050-SiteA/Metro3050-SiteB> lun show
        /vol/VM_TMP/vmtmplun          10g (10737418240)    (r/w, offline, mapped)
        /vol/VM_VOL/vm_lun           100g (107374182400)   (r/w, offline, mapped)
```

*Figures 20a-c) Primary data center loss test.*

The LUNs from the dead controller (N5500-SITEA) must be brought online, because cf forcetakeover –d sets the LUNs from the SITEA controller offline.

Using IBM System Storage N series FilerView® capabilities or the command line interface (CLI), bring online all LUNs that were brought offline by the cf forcetakeover –d command on SITEA (now running on the same controller as SITEB).



```
Metro3050-SiteA/Metro3050-SiteB> lun online /vol/VM_TMP/vmtmplun
Metro3050-SiteA/Metro3050-SiteB> lun online /vol/VM_VOL/vm_lun
Metro3050-SiteA/Metro3050-SiteB> lun show
        /vol/VM_TMP/vmtmplun          10g (10737418240)    (r/w, online, mapped)
        /vol/VM_VOL/vm lun           100g (107374182400)   (r/w, online, mapped)
Metro3050-SiteA/Metro3050-SiteB> Tue Aug  7 17:39:03 GMT [Metro3050-SiteB (takeover): cf.partner.log
in:notice]: Login to partner shell: Metro3050-SiteA

Metro3050-SiteA/Metro3050-SiteB> snapmirror status
Snapmirror is on.
Source                 Destination          State          Lag        Status
Metro3050-SiteA:VMNFS  DR:VMNFS             Source         00:01:13   Idle
Metro3050-SiteA:VM_VOL  DR:VM_VOL           Source         00:01:13   Idle
```

*Figur*

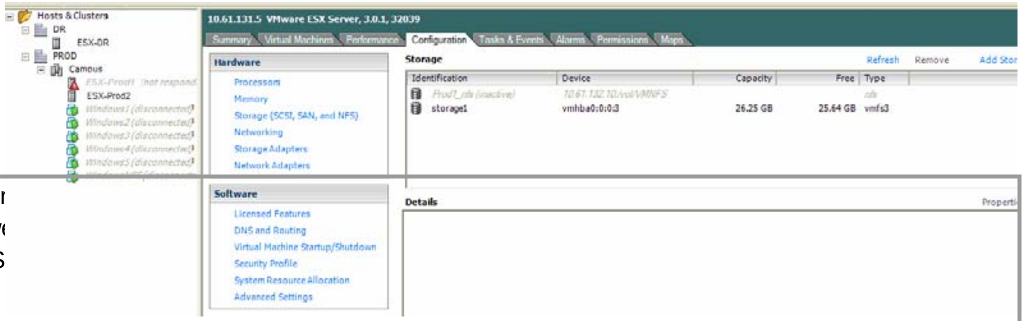
6.

*Figure 20e) Primary data center loss test.*

Takeover was successful.

LUNs are now online and mapped.

SnapMirror relationships have been transparently moved to the SITEB controller.

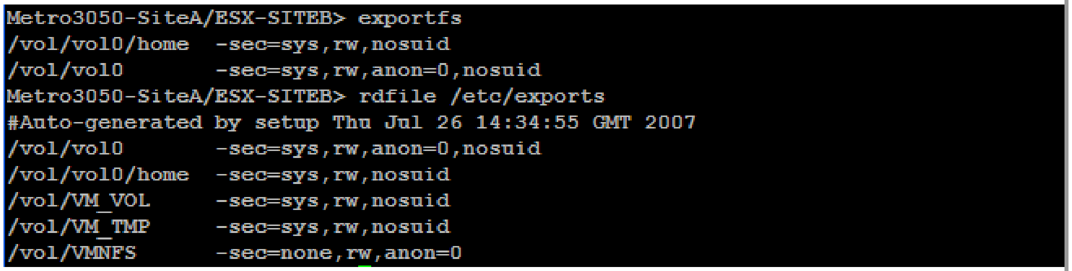


| Identification | Device | Capacity | Free | Type |
|---|---|---|---|---|
| Prod1_nfs (inactive) | 10.61.132.10:/vol/VMNFS | | | nfs |
| storage1 | vmhba0:0:0:3 | 26.25 GB | 25.64 GB | vmfs3 |

After
pow
NFS

*Figure 20f) Primary data center loss test.*

At the METRO3050-SITEB controller command prompt, execute exportfs –a.



```
Metro3050-SiteA/ESX-SITEB> exportfs
/vol/vol0/home  -sec=sys,rw,nosuid
/vol/vol0       -sec=sys,rw,anon=0,nosuid
Metro3050-SiteA/ESX-SITEB> rdfile /etc/exports
#Auto-generated by setup Thu Jul 26 14:34:55 GMT 2007
/vol/vol0       -sec=sys,rw,anon=0,nosuid
/vol/vol0/home  -sec=sys,rw,nosuid
/vol/VM_VOL     -sec=sys,rw,nosuid
/vol/VM_TMP     -sec=sys,rw,nosuid
/vol/VMNFS      -sec=none,rw,anon=0
```

*Figure 20g) Primary data center loss test.*

On the surviving ESX Server (ESX-PROD2), the previous NFS datastore is not available.

Remove the stale datastore and add it back in, being sure to use the same name.

VMFS3 metadata identifies the volumes by several properties, including the LUN number and the LUN ID (UUID or serial number).

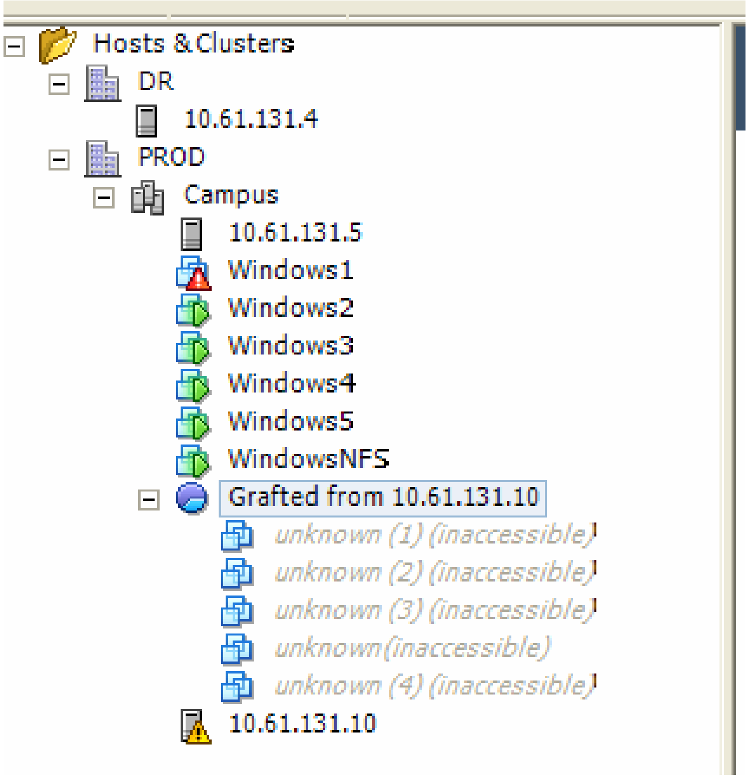
The process of breaking the SyncMirror during this forced takeover results in the volumes being assigned new file system ID number (fsid). Because of this, the LUNs now have new UUIDs, resulting in a mismatch with the metadata, forcing the LVM to identify the LUNs as snapshot copies.

At this writing, the Data ONTAP version necessitated the following process. Future versions of Data ONTAP allow the FSID to be preserved, making the following process unnecessary.

The following procedure was run to make all of the VMFS3 volumes visible again.

1. Enable LVM Resignature on the first ESX Server host (set LVM.EnableResignature to 1):
   a. Log on to the ESX Server host with VI Client.
   b. Click the Configuration tab.
   c. Select the Advanced setting option.
   d. Select the LVM section.
   e. Set the value of LVM.EnableResignature to 1.
   f. Save the change.
   g. Click the storage adapter.
   h. Click Rescan Adapter.
   i. Leave the default option and proceed.

You should now be able to see the VMFS volumes with labels prefixed with snap.

| | |
|---|---|
| | 2. Relabel the volume:<br>a. Log on to the ESX Server with VI Client.<br>b. Disconnect the failed server.<br>c. Remove the server from inventory.<br>d. In Inventory view, select Datastores view.<br>e. Select Hosts & Clusters view.<br>f. In the Summary tab, you should see the list of datastores.<br>g. Click in the Name field for the volume in question and change it to the original name.<br>You now have the correct original label associated with the resignatured volume. |
| | 3. Rescan from all ESX Server hosts. |
| | 4. Reregister all the VMs. Because the VMs are registered against the old UUID, you must reregister them in VirtualCenter:<br>a. Log on to the ESX Server host with VI client.<br>b. Click the Configuration tab.<br>c. Select Storage (SCSI, SAN & NFS).<br>d. Double-click any of the datastores to open the Datastore browser.<br>e. Navigate to the .vmx file of each of the VMs by clicking the folders.<br>f. Right-click and select Add to Inventory. |
| | 5. Power up all VMs and verify proper operation. Power on the VMs. If prompted about a new UUID, click Yes.<br><br> Hosts & Clusters<br>DR<br>10.61.131.4<br>PROD<br>Campus<br>10.61.131.5<br>Windows1<br>Windows2<br>Windows3<br>Windows4<br>Windows5<br>WindowsNFS<br>Grafted from 10.61.131.10<br>unknown (1) (inaccessible)<br>unknown (2) (inaccessible)<br>unknown (3) (inaccessible)<br>unknown(inaccessible)<br>unknown (4) (inaccessible)<br>10.61.131.10<br>*Figure 20i) Primary data center loss test.* |

| | |
|---|---|
| | 6. For each VM, edit the settings to remove the old tmp drive, which has an erroneous name, and add it back in using the appropriate temp datastore. |
| | If any of the VMs refer to missing disks when they power up, check the.vmx file and make sure that the SCSI disk references are not made against the old UUID instead of against the label (or the new label, if you changed it). |
| **Expected Results** | After the SITEB takeover command is issued, the steps of connecting any disk resources on the SIT controller should be completed. Obviously, there should be no loss of data or corruption. |
| **Actual Results** | Operation was as expected. |

## Restore primary data center, recover from disaster

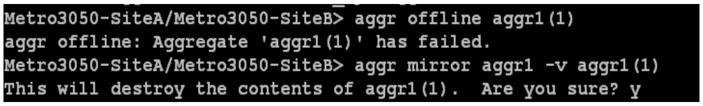
To test the availability of the overall solution, recovery after loss of an entire site was simulated.

**Task**

Power on the disk shelves only on N5500-SITEA.

Reconnect the ISL between sites so that N5500-SITEB can see the disk shelves from N5500-SITEA. After connection, the SITEB Pool1 volumes automatically begin to resync.

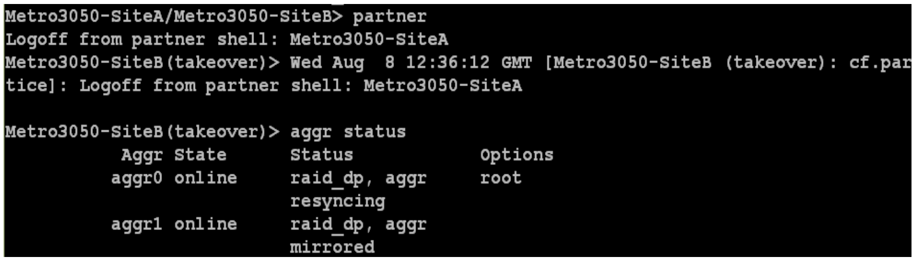
In partner mode on N5500-SITEB, reestablish the mirrors in accordance with the installation guide.



```
Metro3050-SiteA/Metro3050-SiteB> aggr offline aggr1(1)
aggr offline: Aggregate 'aggr1(1)' has failed.
Metro3050-SiteA/Metro3050-SiteB> aggr mirror aggr1 -v aggr1(1)
This will destroy the contents of aggr1(1).  Are you sure? y
```

*Figure 21a) Primary data center restoration—mirror reestablishment.*

Make sure that all mirror resynchronization is complete before proceeding.



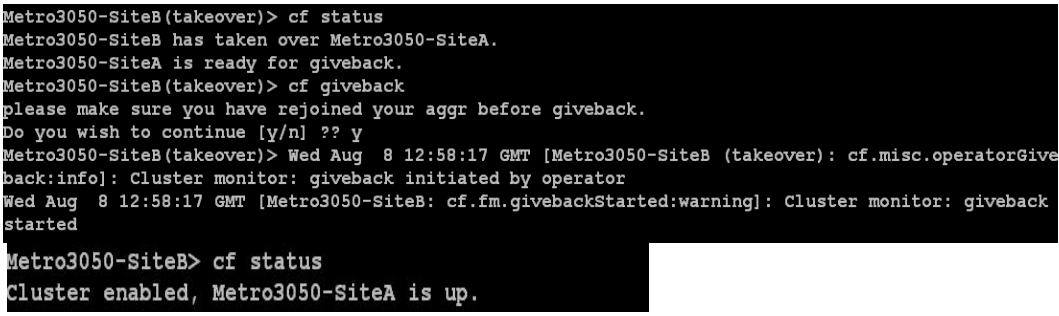
```
Metro3050-SiteA/Metro3050-SiteB> partner
Logoff from partner shell: Metro3050-SiteA
Metro3050-SiteB(takeover)> Wed Aug  8 12:36:12 GMT [Metro3050-SiteB (takeover): cf.par
tice]: Logoff from partner shell: Metro3050-SiteA

Metro3050-SiteB(takeover)> aggr status
           Aggr State      Status            Options
          aggr0 online     raid_dp, aggr     root
                           resyncing
          aggr1 online     raid_dp, aggr
                           mirrored
```

*Figure 21b) Primary data center restoration—mirror resynchronization completion.*

Power on the ESX Server on SITEA (ESX-PROD1).

When the ESX-PROD1 server is online properly, power on N5500-SITEA. Use the cf status command to verify that a giveback is possible and use cf giveback to failback.



```
Metro3050-SiteB(takeover)> cf status
Metro3050-SiteB has taken over Metro3050-SiteA.
Metro3050-SiteA is ready for giveback.
Metro3050-SiteB(takeover)> cf giveback
please make sure you have rejoined your aggr before giveback.
Do you wish to continue [y/n] ?? y
Metro3050-SiteB(takeover)> Wed Aug  8 12:58:17 GMT [Metro3050-SiteB (takeover): cf.misc.operatorGive
back:info]: Cluster monitor: giveback initiated by operator
Wed Aug  8 12:58:17 GMT [Metro3050-SiteB: cf.fm.givebackStarted:warning]: Cluster monitor: giveback
started
Metro3050-SiteB> cf status
Cluster enabled, Metro3050-SiteA is up.
```

*Figure 21c) Primary data center restoration.*

The VMs are migrated back to ESX-PROD1. With VMotion, the process is nondisruptive.

**Expected Results**

The resync of volumes should be completed successfully. On cluster giveback to the SITEA controller, the results should be similar to a normal giveback, as tested previously. This is a maintenance operation involving some amount of downtime.

| | |
|---|---|
| **Actual 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 ESX Server. |

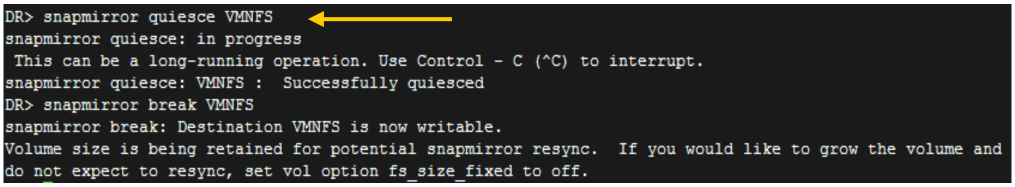
## Loss of entire campus, disaster declared

To test the availability of the overall solution, the loss of an entire site was simulated. This involved bringing up the VMware environment at the DR site using the SnapMirror replica.

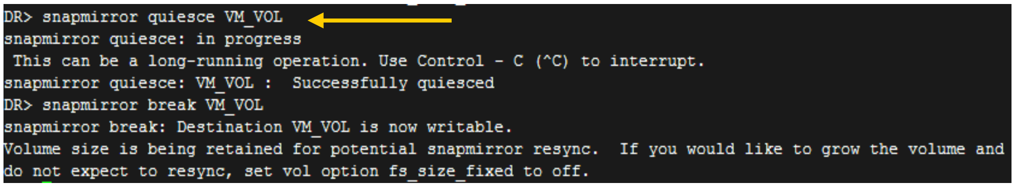
**Task**

Test the failure of the entire main campus (both SITE A and B):

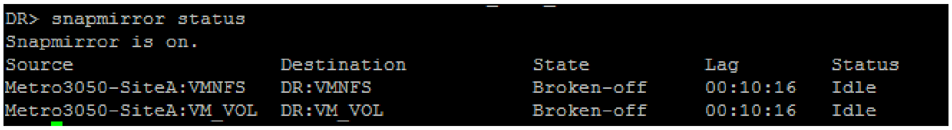
1. At the DR storage controller, quiesce and break SnapMirror volumes to make the DR volumes writable.



```
DR> snapmirror quiesce VMNFS
snapmirror quiesce: in progress
 This can be a long-running operation. Use Control - C (^C) to interrupt.
snapmirror quiesce: VMNFS :  Successfully quiesced
DR> snapmirror break VMNFS
snapmirror break: Destination VMNFS is now writable.
Volume size is being retained for potential snapmirror resync.  If you would like to grow the volume and
do not expect to resync, set vol option fs_size_fixed to off.
```



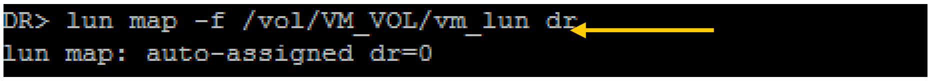
```
DR> snapmirror quiesce VM_VOL
snapmirror quiesce: in progress
 This can be a long-running operation. Use Control - C (^C) to interrupt.
snapmirror quiesce: VM_VOL :  Successfully quiesced
DR> snapmirror break VM_VOL
snapmirror break: Destination VM_VOL is now writable.
Volume size is being retained for potential snapmirror resync.  If you would like to grow the volume and
do not expect to resync, set vol option fs_size_fixed to off.
```



```
DR> snapmirror status
Snapmirror is on.
Source                    Destination           State          Lag        Status
Metro3050-SiteA:VMNFS     DR:VMNFS              Broken-off     00:10:16   Idle
Metro3050-SiteA:VM_VOL    DR:VM_VOL             Broken-off     00:10:16   Idle
```

*Figure 22a) Campus loss test.*

2. Map the LUN at the DR site.

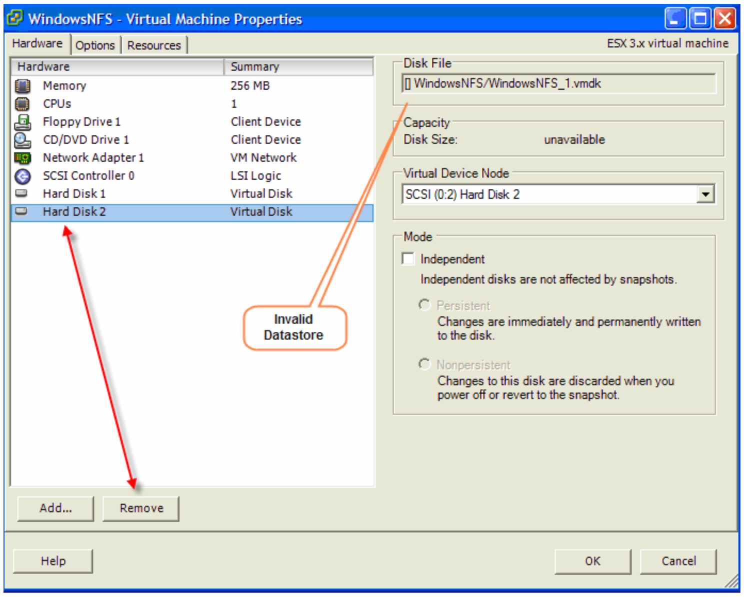


```
DR> lun map -f /vol/VM_VOL/vm_lun dr
lun map: auto-assigned dr=0
```

*Figure 22b) Campus lost test.*

3. NFS: Browse to the datastore and it add to the inventory.

Disk2 shows an invalid disk name for the TEMP datastore.



| Hardware | Summary |
|---|---|
| Memory | 256 MB |
| CPUs | 1 |
| Floppy Drive 1 | Client Device |
| CD/DVD Drive 1 | Client Device |
| Network Adapter 1 | VM Network |
| SCSI Controller 0 | LSI Logic |
| Hard Disk 1 | Virtual Disk |
| Hard Disk 2 | Virtual Disk |

Disk File
[] WindowsNFS/WindowsNFS_1.vmdk

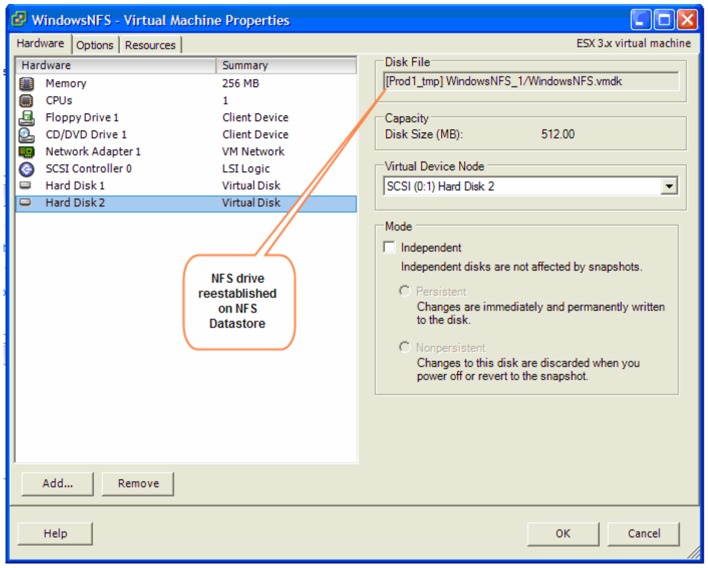
Capacity
Disk Size: unavailable

Virtual Device Node
SCSI (0:2) Hard Disk 2

Mode
□ Independent
Independent disks are not affected by snapshots.
○ Persistent
Changes are immediately and permanently written to the disk.
○ Nonpersistent
Changes to this disk are discarded when you power off or revert to the snapshot.

*Figure 22c) Campus loss test.*

4. Delete Hard Disk2.

| | |
|---|---|
| | 5. Add Hard Disk2, select the new virtual disk option, and select the existing datastore (Prod1_tmp).<br><br> WindowsNFS - Virtual Machine Properties<br>Hardware<br>Options<br>Resources<br>ESX 3.x virtual machine<br>Add...<br>Remove<br>Help<br>OK<br>Cancel<br>Hardware — Summary<br>Memory — 256 MB<br>CPUs — 1<br>Floppy Drive 1 — Client Device<br>CD/DVD Drive 1 — Client Device<br>Network Adapter 1 — VM Network<br>SCSI Controller 0 — LSI Logic<br>Hard Disk 1 — Virtual Disk<br>Hard Disk 2 — Virtual Disk<br>Disk File: [Prod1_tmp] WindowsNFS_1/WindowsNFS.vmdk<br>Capacity: Disk Size (MB): 512.00<br>Virtual Device Node: SCSI (0:1) Hard Disk 2<br>Mode<br>□ Independent<br>Independent disks are not affected by snapshots.<br>○ Persistent<br>Changes are immediately and permanently written to the disk.<br>○ Nonpersistent<br>Changes to this disk are discarded when you power off or revert to the snapshot.<br>NFS drive reestablished on NFS Datastore<br>*Figure 22c) Campus loss test.*<br>6. Repeat steps 3 through 6 for each of the VMs.<br>7. Power on the VMs.<br>8. When prompted, select to create a new identifier. |
| **Expected Results** | The VMware infrastructure should come online at the DR site with no data corruption. Operations should resume as normal. |
| **Actual Results** | Results were as expected. |

## Restore entire campus, recover from disaster

To test the availability of the overall solution, recovery after loss of an entire site was simulated.

| | |
|---|---|
| **Task** | Bring the main campus back online and establish normal operating relationships (SiteA to SiteB and SiteA to DR site).<br>**Scenario 1: All data lost, production site rebuilt.**<br>In this scenario, all the data must be reinitialized before SnapMirror can do any updates.<br>Refer to setting up in section 6, *Regional DR Site*. Repeat the process for the production site.<br>When all volumes are set up, initialize the SnapMirror relationships:<br>On the METRO3050-SITEA system:<br>Metro3050-SiteA> snapmirror initialize –S DR:<vol-name> Metro3050-SiteA:<vol-name><br>When initialization is complete, run a couple of SnapMirror updates to catch the system up with the most recent data:<br>On the METRO3050-SITEA system:<br>Metro3050-SiteA> snapmirror update –S DR:<vol-name> Metro3050-SiteA:<vol-name><br>When you are ready to fail back to SITEA:<br>1. Shut down the applications at the DR site.<br>2. Do another update and break the relationships at SITEA:<br>Metro3050-SiteA> snapmirror update –S DR:<vol-name> Metro3050-SiteA:<vol-name><br>Metro3050-SiteA> snapmirror break Metro3050-SiteA:<vol-name><br>Now the SITEA volumes are writable. Bring up the VMs using the new writable volumes at SITEA. When you have started modifying data at SITEA, you can start replicating data to the DR site.<br>At the DR site, perform a resync to get only the modified data from SITEA to DR:<br>DR> snapmirror resync –S Metro3050-SiteA:<vol-name> DR:<vol-name><br>From this point on, SnapMirror updates resume as per the SnapMirror configuration file ( /etc/snapmirror.conf).<br>**Scenario 2: Data is present but is out of date.**<br>In this scenario, the SITEA data is out of date because DR is being used for production purposes. Next the original SITEA systems are updated with the new data at the DR site.<br>At this point, SnapMirror relationships are in a “broken-off” state.<br>On the Metro3050-SiteA system:<br>Metro3050-SiteA> snapmirror resync –S DR:<vol-name> Metro3050-SiteA:<vol-name><br>When you are ready to fail back to SITEA:<br>1. Shut down the applications at the DR site.<br>2. Do another update and break the relationships at SITEA.<br>Metro3050-SiteA> snapmirror update –S DR:<vol-name> Metro3050-SiteA:<vol-name><br>Metro3050-SiteA> snapmirror break Metro3050-SiteA:<vol-name><br>Now the SITEA volumes are writable. Bring up the VMs, using the new writable volumes at SITEA. When you have started modifying data at SITEA, you can start replicating data to the DR site.<br>At the DR site, perform a resync to get only the modified data from SITEA to DR: |

| | |
|---|---|
| | DR> snapmirror resync –S Metro3050-SiteA:<vol-name> DR:<vol-name><br>From this point on, SnapMirror updates resume as per the SnapMirror configuration file (/etc/snapmirror.conf).<br>**Cleaning up old relationships.**<br>To see the old relationships, run snapmirror destinations:<br>DR> snapmirror destinations<br>Path                    Destination<br><vol-name>    Metro3050-SiteA:<vol-name><br>Delete the old destinations:<br>DR> snapmirror release <vol-name> Metro3050-SiteA:<vol-name> |
| **Expected Results** | After a SnapMirror resynchronization back to the primary server (METRO3050-SITEA), the VMs should be brought up and normal operation should resume, including resetting SnapMirror from primary to DR. |
| **Actual Results** | To ensure consistency, it was important to quiesce applications at the DR site before resynchonizing back to the primary. Results were as expected. |

# Testing the DR environment

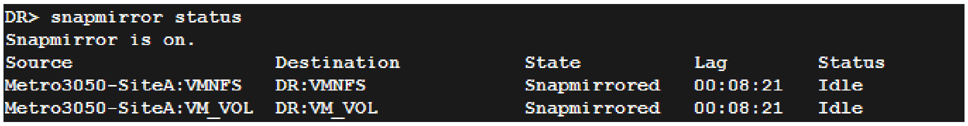
Testing the DR site is a process that sometimes gets neglected. The reason is that there can be no disruption to the active mission-critical applications. It is unfortunately assumed that everything is ready at the DR site.

Utilizing FlexClone technology in the VMware environment, DR tests can now be performed on a regular basis without any disruption to the active applications and the replication operations from primary to DR sites. The volumes that are mirrored by SnapMirror at the DR site are read-only. However, using FlexClone technology, the read-only volumes can be cloned and made writable for testing purposes.

The following steps can be performed to ensure that the DR site is ready when needed.

Note: For simplicity, only the cloning of the LUN volume (vm_vol) is shown. The process would be the same for the NFS volume.

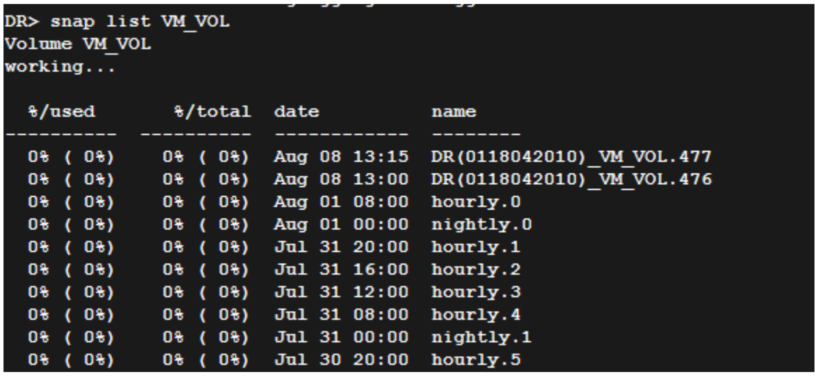
1. At the DR storage controller, verify the SnapMirror relationships.



```
DR> snapmirror status
Snapmirror is on.
Source                      Destination             State          Lag        Status
Metro3050-SiteA:VMNFS       DR:VMNFS                Snapmirrored   00:08:21   Idle
Metro3050-SiteA:VM_VOL      DR:VM_VOL               Snapmirrored   00:08:21   Idle
```

*Figure 23a) DR test.*

2 Check snapshot copies for the LUNs volume.

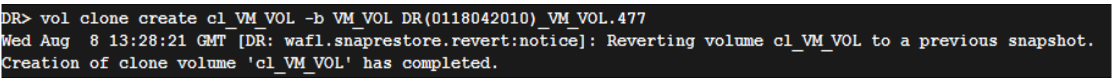


```
DR> snap list VM_VOL
Volume VM_VOL
working...

  %/used       %/total  date          name
----------  ----------  ------------  --------
  0% ( 0%)    0% ( 0%)  Aug 08 13:15  DR(0118042010)_VM_VOL.477
  0% ( 0%)    0% ( 0%)  Aug 08 13:00  DR(0118042010)_VM_VOL.476
  0% ( 0%)    0% ( 0%)  Aug 01 08:00  hourly.0
  0% ( 0%)    0% ( 0%)  Aug 01 00:00  nightly.0
  0% ( 0%)    0% ( 0%)  Jul 31 20:00  hourly.1
  0% ( 0%)    0% ( 0%)  Jul 31 16:00  hourly.2
  0% ( 0%)    0% ( 0%)  Jul 31 12:00  hourly.3
  0% ( 0%)    0% ( 0%)  Jul 31 08:00  hourly.4
  0% ( 0%)    0% ( 0%)  Jul 31 00:00  nightly.1
  0% ( 0%)    0% ( 0%)  Jul 30 20:00  hourly.5
```

*Figure 23b) DR test.*

3 Create a clone for the iSCSI LUN volume using the latest snapshot copy

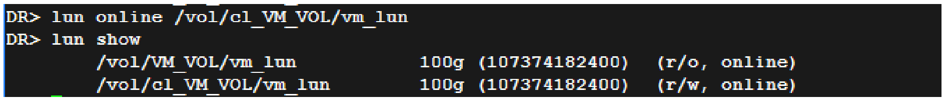


```
DR> vol clone create cl_VM_VOL -b VM_VOL DR(0118042010)_VM_VOL.477
Wed Aug  8 13:28:21 GMT [DR: wafl.snaprestore.revert:notice]: Reverting volume cl_VM_VOL to a previous snapshot.
Creation of clone volume 'cl_VM_VOL' has completed.
```

*Figure 23c) DR test.*

Vol Status shows that the clone volume is now writable. Notice that the first LUN is read-only,  whereas its clone is writable.

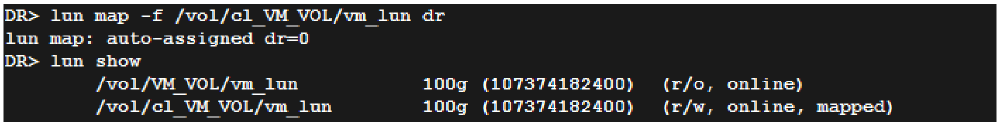
4 Bring LUNs online and check status



```
DR> lun online /vol/cl_VM_VOL/vm_lun
DR> lun show
        /vol/VM_VOL/vm_lun           100g (107374182400)   (r/o, online)
        /vol/cl_VM_VOL/vm_lun        100g (107374182400)   (r/w, online)
```

*Figure 23d) DR test.*

6 The clone LUN must now be mapped to the ESX-DR Server.

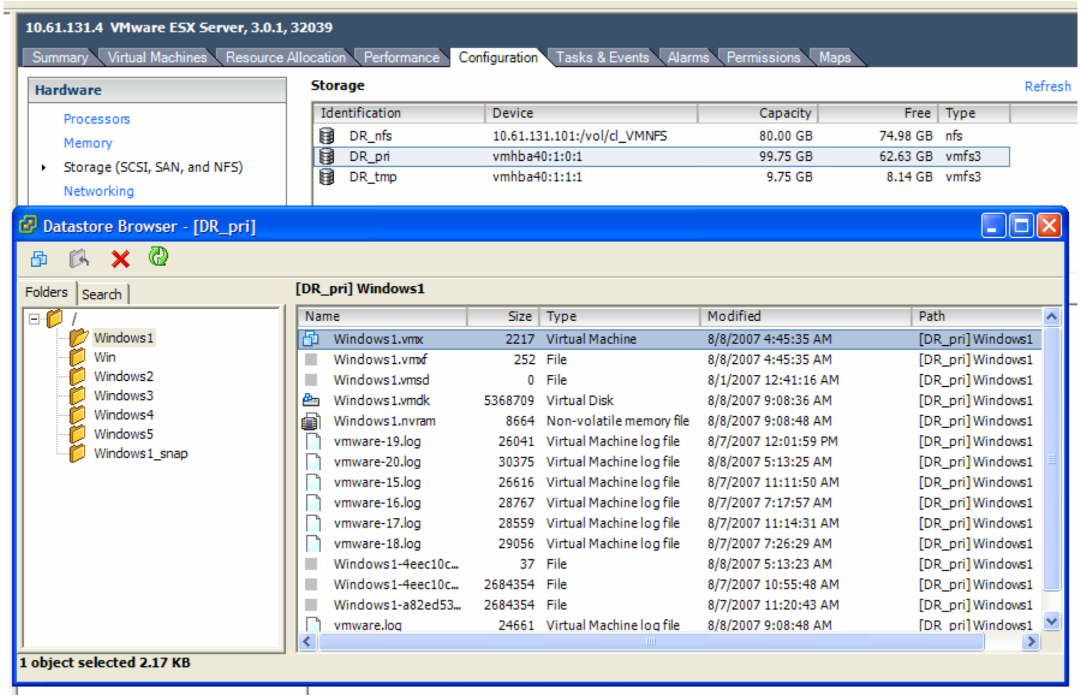


```
DR> lun map -f /vol/cl_VM_VOL/vm_lun dr
lun map: auto-assigned dr=0
DR> lun show
        /vol/VM_VOL/vm_lun           100g (107374182400)   (r/o, online)
        /vol/cl_VM_VOL/vm_lun        100g (107374182400)   (r/w, online, mapped)
```

*Figure 23e) DR test.*

7 A rescan may be necessary for the DR ESX Server to pick up the cloned LUN.

8



| Identification | Device | Capacity | Free | Type |
|---|---|---|---|---|
| DR_nfs | 10.61.131.101:/vol/cl_VMNFS | 80.00 GB | 74.98 GB | nfs |
| DR_pri | vmhba40:1:0:1 | 99.75 GB | 62.63 GB | vmfs3 |
| DR_tmp | vmhba40:1:1:1 | 9.75 GB | 8.14 GB | vmfs3 |

[DR_pri] Windows1

| Name | Size | Type | Modified | Path |
|---|---|---|---|---|
| Windows1.vmx | 2217 | Virtual Machine | 8/8/2007 4:45:35 AM | [DR_pri] Windows1 |
| Windows1.vmxf | 252 | File | 8/8/2007 4:45:35 AM | [DR_pri] Windows1 |
| Windows1.vmsd | 0 | File | 8/1/2007 12:41:16 AM | [DR_pri] Windows1 |
| Windows1.vmdk | 5368709 | Virtual Disk | 8/8/2007 9:08:36 AM | [DR_pri] Windows1 |
| Windows1.nvram | 8664 | Non-volatile memory file | 8/8/2007 9:08:48 AM | [DR_pri] Windows1 |
| vmware-19.log | 26041 | Virtual Machine log file | 8/7/2007 12:01:59 PM | [DR_pri] Windows1 |
| vmware-20.log | 30375 | Virtual Machine log file | 8/8/2007 5:13:25 AM | [DR_pri] Windows1 |
| vmware-15.log | 26616 | Virtual Machine log file | 8/7/2007 11:11:50 AM | [DR_pri] Windows1 |
| vmware-16.log | 28767 | Virtual Machine log file | 8/7/2007 7:17:57 AM | [DR_pri] Windows1 |
| vmware-17.log | 28559 | Virtual Machine log file | 8/7/2007 11:14:31 AM | [DR_pri] Windows1 |
| vmware-18.log | 29056 | Virtual Machine log file | 8/7/2007 7:26:29 AM | [DR_pri] Windows1 |
| Windows1-4eec10c... | 37 | File | 8/8/2007 5:13:23 AM | [DR_pri] Windows1 |
| Windows1-4eec10c... | 2684354 | File | 8/7/2007 10:55:48 AM | [DR_pri] Windows1 |
| Windows1-a82ed53... | 2684354 | File | 8/7/2007 11:20:43 AM | [DR_pri] Windows1 |
| vmware.log | 24661 | Virtual Machine log file | 8/8/2007 9:08:48 AM | [DR_pri] Windows1 |

1 object selected 2.17 KB

*Figure 23f) DR test.*

Add the VMs to the data center.

The following figure shows that there are now two independent networks (except for the SnapMirror replication that continues). Applications can be brought up and tested without disruption.

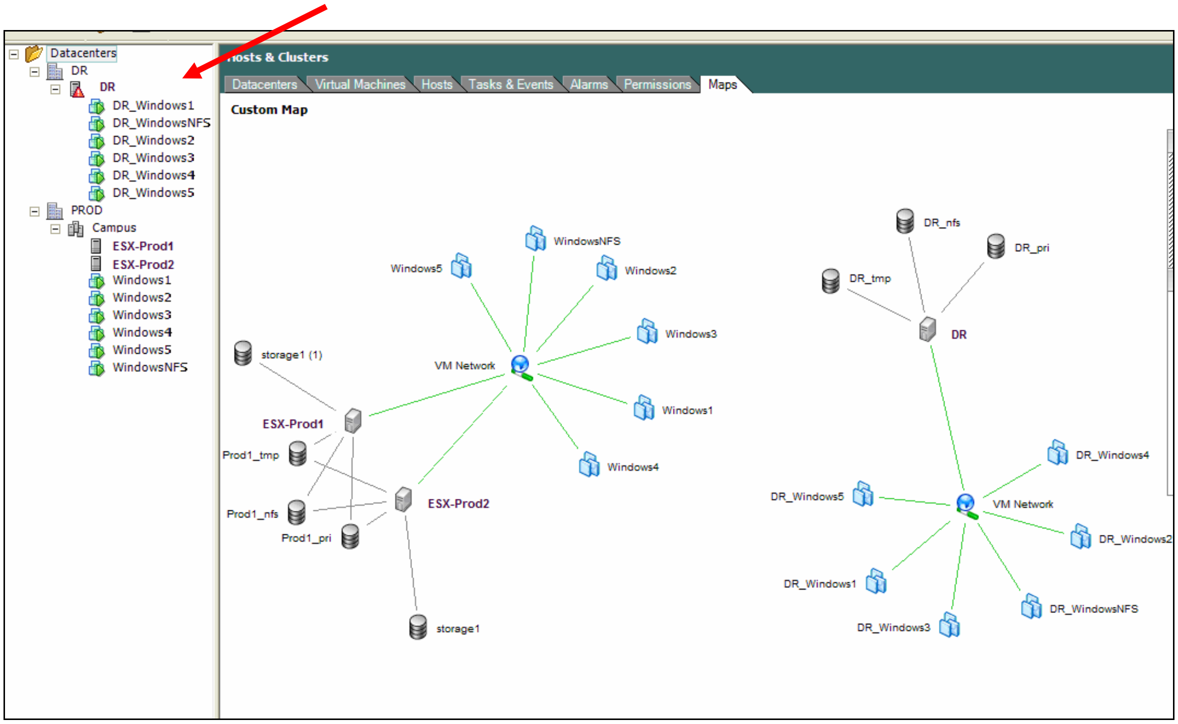


*Figure 24) Independent DR test network*

After testing at the DR site is complete, perform the following steps to clean up:

Power off the VMs.  
Remove them from the inventory.  
Destroy the clone volumes with the `vol offline` and `vol destroy` commands.

# Conclusion

The combination of VMware Infrastructure 3 and MetroCluster, SnapMirror, and FlexClone provides a solid server consolidation solution with DR protection at different levels. In all of the operational scenarios, it was demonstrated that protection was maximized while application disruptions were minimized.

# Appendix A: Materials List

**Hardware**

| Storage | Vendor | Name | Version | Description |
|---|---|---|---|---|
| | IBM N series | N5500C | | |
| | IBM N series | N7800 Gateway | | |
| Hosts | IBM | IBMX335 | | 3.06Ghz, 4Gb RAM, 40Gb disk |
| | IBM | IBMX306 | | |
| MetroCluster | Brocade | 200E (4) | 5.1.0 | 16-port FC switch |
| | | **Software** | | |
| Storage | IBM N series | SyncMirror | 7.2.3 | Replication |
| | IBM N series | Data ONTAP | 7.2.3 | Operating system |
| | IBM N series | Cluster_Remote | 7.2.3 | Failover |
| | IBM N series | SnapMirror | 7.2.3 | DR replication |
| Hosts | VMware | Infrastructure 3 | 3.0.1 | |
| | Microsoft | Windows Server 2003 Enterprise Edition-SP1 (x86) | 2003 | Operating system |

# Appendix B: Platform specifications

## IBM N series storage controller

### Configuration

For purposes of this paper, the controller and back-end FC switches are configured using the instructions in the Data ONTAP configuration guide and the current firmware levels and other notes found on the IBM N series support site. The versions used are:

Data ONTAP 7.2.3
Brocade firmware 5.1.0.

Two N5500 storage controllers (each with two EXP4000shelves full of 66GB 10k drives) connected with the VI-MC interconnect and four Brocade 200E switches were used in this test, representing the main data center and the on-campus DR site. The controllers were named METRO3050-SITEA and METRO3050-SITEB, and the switches were named SITEA-SW1, SITEA-SW2, SITEB-SW3, and SITEB-SW4. For these tests, the controllers were used in an active/passive configuration. Functionally, it would work the same for an active/active configuration other than performing on-campus recovery in either direction.

An N7800 storage controller was used as DR site storage (thus the SnapMirror destination) to provide protection in case of a complete campus disaster.

### Storage controllers

| Name | Description | IPAddress |
|---|---|---|
| METRO3050-SITEA | N5500 | 10.61.132.10 |
| METRO3050-SITEB | N5500 | 10.61.132.11 |
| DR | N7800 Gateway | 10.61.131.101 |

### Aggregate layout

| Controller | Aggregate name | Options | # Disks | Purpose |
|---|---|---|---|---|
| METRO3050-SITEA | Aggr0 | RAID_DP, aggr mirrored | 3 | Root volume |
| METRO3050-SITEA | Aggr1 | RAID_DP, aggr mirrored | 10 | Datastores |
| METRO3050-SITEB | Aggr0 | RAID_DP, aggr mirrored | 3 | Root |
| DR | Aggr0 | RAID_DP | 3 | Root |
| DR | Aggr1 | RAID_DP | 39 | DR datastores |

### Volume layout

The hardware in this configuration is limited to 14 mirrored disks on the controller head. Three of these are for the root volume and one is reserved as a spare. The remaining 24 disks have been used to create volumes. The controller at SITEA has one volume to house the iSCSI LUN-based datastores. The controller at SITEB has just the root volume. All volumes are mirrored using SyncMirror with pool0 and pool1 disks in SITEA and SITEB respectively. The IBM N series controller at the DR site is configured as a standalone and is the destination for SnapMirror relationships.

The following table shows the volume layout in detail.

| Controller | Volume Name | Options | Total Volume Size | Purpose |
|---|---|---|---|---|
| METRO3050-SITEA | vol0 | RAID_DP, flex mirrored | 191GB | Root volume |
| METRO3050-SITEA | VM_VOL | RAID_DP, flex mirrored | 200GB | VMDK datastore |
| METRO3050-SITEA | VM_TMP | RAID_DP, flex mirrored | 20G | VMDK temporary (pagefile, etc.) |
| METRO3050-SITEA | VMNFS | RAID_DP, flex mirrored | 50G | NFS mount for NFS datastore |
| METRO3050-SITEB | vol0 | RAID_DP, flex mirrored | 172GB | Root volume |
| DR | vol0 | RAID_DP | 268GB | Root volume |
| DR | VM_VOL | | | iSCSI LUNs for datastores |
| DR | VM_TMP | | | iSCSI LUN for temp |
| DR | VMNFS | | | NFS mount for NFS datastore |

## MetroCluster setup

### 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 sections, the switch and port connections are detailed and should be implemented exactly as documented.

# Host servers

## Software configuration

The hosts in the cluster are installed according to the vendor-supplied procedures documented in the VMWare installation/upgrade guide. For this paper, ESX version 3.0.1 was used.

## Network settings

The following tables provide the network settings for the ESX Servers.

| Hostname | Purpose |
|---|---|
| ESX-PROD1 | SITEA (primary) |
| ESX-PROD2 | SITEB (secondary) |
| ESX-DR | DR SITE |

## iSCSI/LUN setup

A couple of volumes have been created for the LUN files. These LUNs have the following attributes.

| Purpose | Drive | LUN Size | LUN File |
|---|---|---|---|
| VMDK | C: | 200G | /vol/vm_vol/vm_lun |
| VMDK | D: | 20G | /vol/vm_tmp/tmp_lun |

# Appendix C: Switch configuration

## SITEA, Switch 1

| Port | Bank/<br>Pool | Connected with | Purpose |
|---|---|---|---|
| 0 | 1/0 | METRO3050-SITEA, 0a | Site A FC HBA |
| 1 | 1/0 | METRO3050-SITEA, 0c | Site A FC HBA |
| 2 | 1/0 | | |
| 3 | 1/0 | | |
| 4 | 1/1 | | |
| 5 | 1/1 | METRO3050-SITEB pool 1, Shelf 3B | |
| 6 | 1/1 | | |
| 7 | 1/1 | | |
| 8 | 2/0 | | |
| 9 | 2/0 | METRO3050-SITEA pool 0, Shelf 1B | |
| 10 | 2/0 | | |
| 11 | 2/0 | | |
| 12 | 2/1 | METRO3050-SITEA FCVI, 2a | Cluster interconnect |
| 13 | 2/1 | SITEB-SW3, port 5 | ISL |
| 14 | 2/1 | | |
| 15 | 2/1 | | |

## SITEA, Switch 2

| Port | Bank /Pool | Connected with | Purpose |
|---|---|---|---|
| 0 | 1/0 | METRO3050-SITEA, 0b | Disk HBA for bank 2 shelves |
| 1 | 1/0 | METRO3050-SITEA, 0d | Disk HBA for bank 2 shelves |
| 2 | 1/0 | | |
| 3 | 1/0 | | |
| 4 | 1/1 | | |
| 5 | 1/1 | METRO3050-SITEB pool 1, Shelf 3A | |
| 6 | 1/1 | | |
| 7 | 1/1 | METRO3050-SITEA FCVI, 2b | Cluster interconnect |
| 8 | 2/0 | | |
| 9 | 2/0 | METRO3050-SITEA pool 0, Shelf 1A | |
| 10 | 2/0 | | |
| 11 | 2/0 | | |
| 12 | 2/1 | | |
| 13 | 2/1 | SiTEB-SW3, port 4 | ISL |
| 14 | 2/1 | | |
| 15 | 2/1 | | |

## SITEB, Switch 2

| Port | Bank/Pool | Connected with | Purpose |
|---|---|---|---|
| 0 | 1/0 | METRO3050-SITEA pool 1, Shelf 3B | |
| 1 | 1/0 | | |
| 2 | 1/0 | | |
| 3 | 1/0 | METRO3050-SITEB FCVI, 2a | Cluster interconnect |
| 4 | 1/1 | | |
| 5 | 1/1 | SiTEB-SW1, port 13 | ISL |
| 6 | 1/1 | | |
| 7 | 1/1 | | |
| 8 | 2/0 | METRO3050-SITEB, 0a | Disk HBA for bank 2 shelves |
| 9 | 2/0 | METRO3050-SITEB, 0c | Disk HBA for bank 2 shelves |
| 10 | 2/0 | | |
| 11 | 2/0 | | |
| 12 | 2/1 | METRO3050-SITEB pool 0, Shelf 1B | |
| 13 | 2/1 | | |
| 14 | 2/1 | | |
| 15 | 2/1 | | |

## SITEB, Switch 4

| Port | Bank/ Pool | Connected with | Purpose |
|---|---|---|---|
| 0 | 1/0 | METRO3050-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 | METRO3050-SITEB, 0b | Disk HBA for bank 2 shelves |
| 9 | 2/0 | METRO3050-SITEB, 0d | Disk HBA for bank 2 shelves |
| 10 | 2/0 | | |
| 11 | 2/0 | | |
| 12 | 2/1 | METRO3050-SITEB pool 0, Shelf 1A | |
| 13 | 2/1 | METRO3050-SITEB FCVI, 2b | Cluster interconnect |
| 14 | 2/1 | | |
| 15 | 2/1 | | |

# Trademarks and special notices

References in this document to IBM products or services do not imply that IBM intends to make them available in every country.

Data ONTAP, FlexClone, FlexVol, Network Appliance, the Network Appliance logo, Data ONTAP, RAID-DP, ReplicatorX, SnapMirror, Snapshot and SyncMirror are trademarks or registered trademarks of Network Appliance, Inc., in the U.S. and other countries.

Microsoft, Windows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

Intel, Intel Inside (logos), MMX, and Pentium are trademarks of Intel Corporation in the United States, other countries, or both.

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

Information is provided "AS IS" without warranty of any kind.

Information concerning non-IBM products was obtained from a supplier of these products, published announcement material, or other publicly available sources and does not constitute an endorsement of such products by IBM. Sources for non-IBM list prices and performance numbers are taken from publicly available information, including vendor announcements and vendor worldwide homepages. IBM has not tested these products and cannot confirm the accuracy of performance, capability, or any other claims related to non-IBM products. Questions on the capability of non-IBM products should be addressed to the supplier of those products.

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