White Paper

EMC VNXe3200 HIGH AVAILABILITY

A DETAILED REVIEW

Abstract

This white paper discusses the high availability (HA) features in the EMC® VNXe3200 storage system and how you can configure a VNXe system to achieve your goals for data availability.

February, 2015

EMC²

Part Number H12990.1

Table of Contents

Executive Summary

IT administrators have spent countless hours planning and configuring their organization's infrastructure to provide a competitive advantage, increase productivity, and empower users to make faster and more informed decisions. Both small businesses and global enterprises have users around the world who require constant access to their data at all hours of the day. Without this data access, operations can stop and revenue is lost. Because of this, IT administrators are constantly concerned about the accessibility of their organization's business data. EMC® focuses on helping customers maintain the most highly available storage environment possible to help alleviate these concerns. Having a highly available environment involves many factors, including the environment's architecture and design, and how you configure your infrastructure connectivity, storage system, and disk storage. A highly available storage system does not have a single point of failure; that is, if a component fails, the system maintains its basic functionality. For example, having two storage processors (SPs) in your VNXe3200 storage system allows you to configure alternate data paths at the storage system level. This allows the system to remain operational if a component in one path fails.

In many cases, a highly-available system can withstand multiple failures if the failures occur in different component sets. Once detected, failed components in the VNXe3200 can be replaced easily and brought online without affecting users or applications.

This document discusses different factors that affect availability. It also discusses how to design your VNXe system to achieve the level of availability that your environment requires.

Audience

This white paper is intended for EMC customers, partners, and employees who want to understand the features in the EMC VNXe3200 storage system that can maximize data availability in their storage environment.

Terminology

Asymmetric Logic Unit Access (ALUA) – A SCSI standard that allows multiple controllers to route I/O to a given logical unit.

Battery Backup Unit (BBU) – A 3-cell Lithium-Ion battery located within each VNXe3200 Storage Processor enclosure. The BBU is designed to power the system long enough to flush SP cache content to the mSATA device in the case of a power failure.

Fibre Channel (FC) Interface – An interface on the VNXe3200 system that uses the Fibre Channel (FC) protocol to provide access to LUNs and VMFS-based VMware datastores. Fibre Channel interfaces are created automatically on the VNXe3200 system.

iSCSI Interface – An interface on the VNXe3200 system that uses the iSCSI protocol to provide access to LUNs and VMFS-based VMware datastores. These interfaces can be created on any Ethernet network port that is not part of a Link Aggregation group.

Link Aggregation – A high-availability feature based on the IEEE 802.3ad Link Aggregation Control Protocol (LACP) standard that allows Ethernet ports with similar characteristics to connect to the same switch (physical or cross-stack). When aggregated in this way, the ports combine to make a single virtual link with a single MAC address. This virtual link can have multiple IP addresses depending on the number of NAS Servers that are leveraging the aggregated port group. Link Aggregation is not supported for iSCSI interfaces.

MCx – Multicore Everything initiative that delivers high performance and platform efficiency in VNXe3200 storage systems.

Mini-Serial Advanced Technology Attachment (mSATA) – A 32GB MLC Flash device physically found underneath each Storage Processor that contains a partition for the boot image. In the event of a power failure, the memory contents of cache are written to the permanent memory persistence (PMP) within the mSATA device. Even if the mSATA device becomes corrupted, the PMP data can be recovered from the peer SP.

MPIO – A fault-tolerance and performance enhancement technique that allows the use of more than one path to a storage device

Multicore Cache – An MCx component that optimizes storage processor's CPU core and DRAM usage to increase host write and read performance.

Multicore RAID – An MCx component that defines, manages, creates, and maintains RAID protection for storage pools on VNXe3200 storage systems.

NAS Server – A VNXe3200 file server that uses either the CIFS and/or NFS protocol to catalog, organize, and transfer files within designated shares. A NAS Server is required to create File Systems that contain CIFS shares, NFS shares, or VMware NFS datastores. NAS Servers require a Storage Pool to be created first, which is used to store the configuration data such as Anti-Virus configurations, NDMP settings, Network Interfaces and IP addresses, etc.

Storage Processor (SP) – A hardware component that performs storage operations on your VNXe system such as creating, managing, and monitoring storage resources.

High Availability in the Storage System Components

It is important to configure a storage system using a high-availability (HA) design to ensure that business-critical data is always accessible. The VNXe3200 storage system offers N+1 redundant architecture, which provides data protection against any single component failure. With redundant components, including dual SPs and dual-ported disk drives, the VNXe system can overcome many different types of multiple component failure scenarios. System components include all Customer Replaceable Units (CRUs), such as power supplies, battery backup units (BBU), memory, drives, and so on. This is vital when you need to quickly return a storage system back to its HA state.

Storage Processors

In the VNXe system, storage resources are distributed between the two SPs; however, a storage resource is assigned to only one SP at a time. For example, a file system created on SP A will not be associated with SP B unless an SP failover occurs.

An SP failover occurs when an SP reboots, experiences a hardware or software failure, or when a user places it in Service Mode. In these cases, the storage resources fail over to the peer SP. The surviving SP assumes ownership and begins servicing host I/O requests. When one SP services all I/O requests, performance between hosts and the VNXe system can be degraded. Table 1 describes the SP events that can cause a failover.

Table 1 – Events that cause SP failover

Event	Description
SP rebooting	The system or a user rebooted the SP. If the SP is healthy when it comes back online, the storage resources will return to normal. Check the **System Health** page in Unisphere to ensure that the SP is operating normally.
SP in Service Mode	The system or a user placed the SP in Service Mode. An SP automatically enters Service Mode when it is unable to boot due to a hardware or system problem. Use the service actions on the **Service System** page to try to fix the problem. If the SP is healthy, you can reboot it to return it to Normal Mode.
SP powered down	A user powered down the SP.
SP failed	The SP failed and must be replaced.

Failover of the Management Interface

The management services (including the single management IP address used to access the storage system) for the VNXe system run on one SP at a time, and it does not matter on which SP they run. In the event of an SP failure, the management server fails over to the peer SP, and the management stack starts on the peer SP (Figure 1). Assuming both SPs' management ports are cabled and on the same network, this process is not visible to the user, except for a brief time when the management stack restarts on the peer SP. If Unisphere® is open in a web browser at the time of the SP failure, you will see a pop-up message indicating a loss of connection to the failed SP. Another pop-up message appears when the connection is re-established to the peer SP. The management stack remains on this SP even after the failed SP returns to a healthy state.

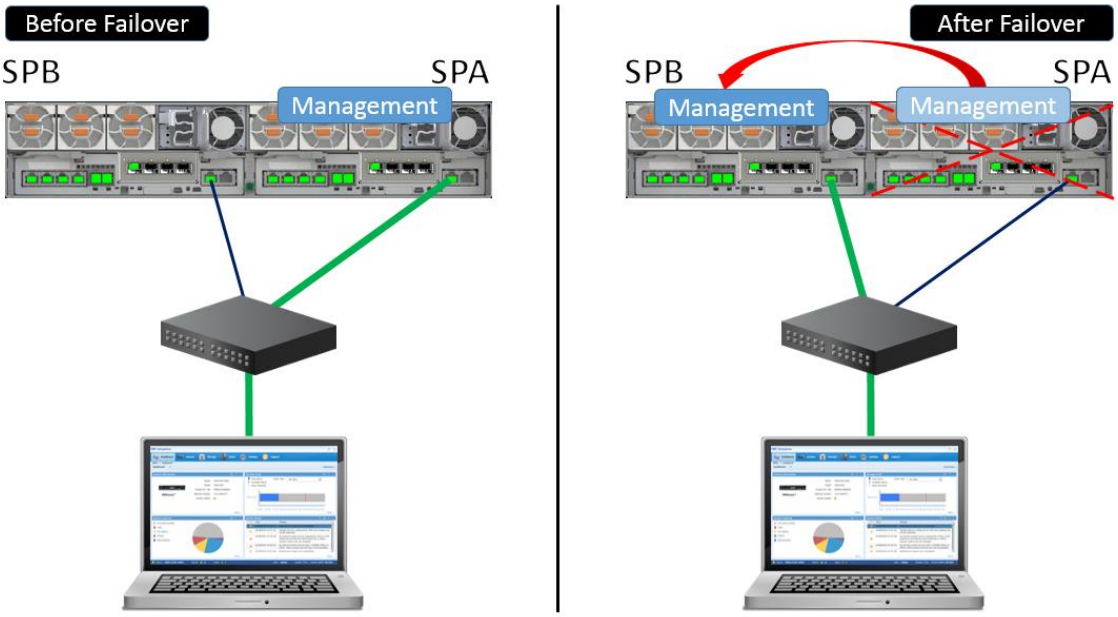

Figure 1 – Failover of the Management services of a VNXe system

Storage Processor Memory

Each Storage Processor has its own dedicated system memory. The VNXe system has 24 GB per SP. This system memory is divided into SP operating system memory and cache memory. The cache memory is used for read and write caching. Read cache is for data that is held in memory in anticipation of it being requested in a future read I/O. Write cache stores write request data waiting to be written to a drive. Write cache memory is mirrored from one SP to its peer SP. Figure 2 shows a conceptual view of the VNXe system memory.

In addition to an SP's own read and write cache, cache memory contains a mirror copy of its peer SP's write cache. This is an important availability feature. Managed by Multicore Cache, memory is dynamically allocated to write cache and read cache depending on the current I/O the system is servicing.

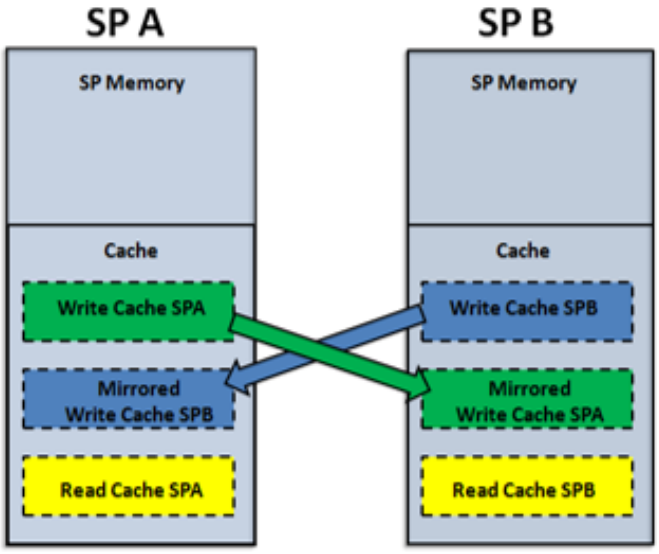

Figure 2 – Mirroring write cache memory

Write Cache

Write Cache is a subcomponent of the adaptive global cache managed by Multicore Cache. The VNXe write cache is a mirrored write-back cache. For every write, the data is stored in cache and copied to the peer SP. Then, the request is acknowledged to the host. In this process, write cache is fully mirrored between the VNXe system SPs to ensure data protection through redundancy.

When the VNXe system is shut down properly, the SP cache is flushed to backend drives and disabled. If an SP fails and then reboots, the cache is kept intact in DRAM through all system resets.

If there is power loss, each SP leverages the power from the internal BBU to write its copy of the write cache to its internal mSATA device, which does not need any power to retain the data. Upon reboot, the SP cache contents are restored on both SPs. The two SPs then determine the validity of the contents. Normally, both copies are valid. In the event that one SP has a newer version of the cache contents (or if one of them is invalid), the SP with the latest valid copy synchronizes its contents with the peer SP before re-enabling the cache and allowing access to storage resources.

Power Modules

VNXe systems have redundant power supplies, power cables, and battery backup units (BBUs) that protect data in the event of internal or external power failures. The VNXe system employs dual-shared power supplies. If one power supply fails, the other one provides power to both Storage Processors.

The BBU is appropriately sized to support the connected SP. During power loss, it will maintain power long enough for write cache to be safely stored to the internal mSATA device. This protects the data in the cache when there is power loss. Unlike an uninterruptable power supply (UPS), the BBU is not designed to keep the storage system up and running for long periods in anticipation of power being restored.

Drive Storage

The VNXe system includes drives in the Disk Processor Enclosure (DPE), which is available in either a 12-drive (3.5" drives) or 25-drive (2.5" drives) configuration. Additional Disk Array Enclosures (DAEs) can be connected to the system for storage expansion, and are available in 12-drive (3.5" drives) or 25-drive (2.5" drives) configurations.

Number of Backend Buses

Each SP in the VNXe system has two dedicated backend buses available (BE Port 0 and BE Port 1) which can be connected to additional Disk Array Enclosures (DAEs) for storage expansion. Each DAE has two ports that are connected to the same backend bus port on both SPs by separate cables. If a cable or port on one of the SPs or DAEs fail, both SPs can still access the DAE through the second connection path.

Keep in mind to balance the backend buses when connecting additional DAEs to your VNXe system in order to eliminate saturation of the backend buses. The DPE is using the BE Port 0 internally and is noted as Bus 0 Enclosure 0. It is recommended to cable the first DAE to BE Port 1 (noted as Bus 1 Enclosure 0) and the second DAE to BE Port 0 (noted as Bus 0 Enclosure 1). Any additional DAEs attached to the VNXe system would then continue this alternating pattern. An example of this recommended solution is shown in Figure 3 below.

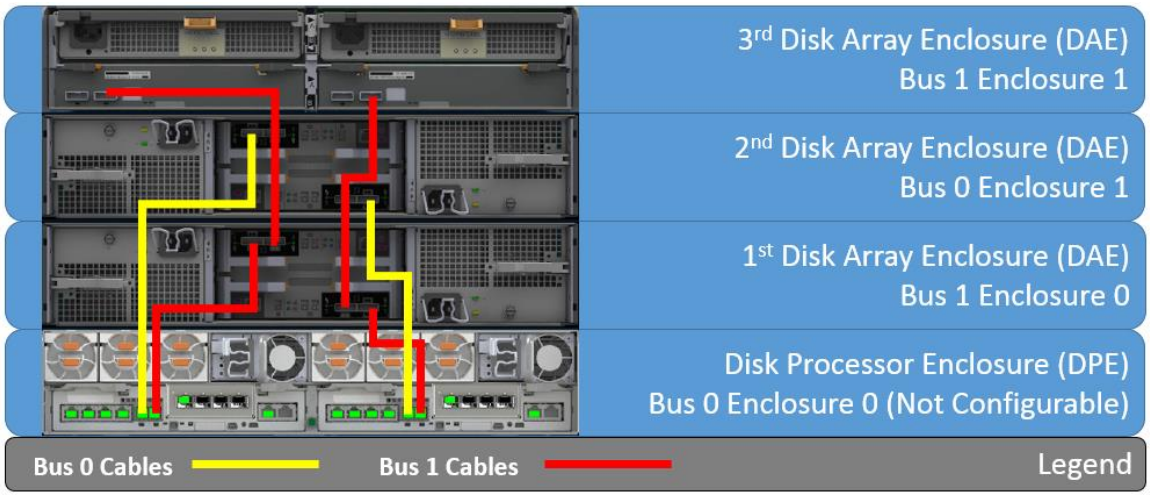

Figure 3 – Recommended solution to balance the DAEs across Backend Buses on the VNXe3200 system

RAID Configurations

To access data on the drives, you must create storage pools. A storage pool consists of drives that are bound into one or more RAID configurations. Storage pool tiers can contain one or more RAID configurations depending on the stripe width and number of drives selected. If you do not have the FAST VP license installed, all drives in a storage pool must be of the same type (SAS, NL-SAS, or Flash). If you have the FAST VP license installed, you can create storage pools leveraging multiple tiers.

NOTE: It is recommended to use the same speed and capacity drives within a tier of a storage pool, since this ensures the highest reliability and consistent performance.

NOTE: The VNXe3200 system will automatically select drives for each storage pool or tier. The system will use the first four drives (DPE drives 0-3) last since those drives are also used for the VNXe Operating Environment (OE). When the system drives are included in a storage pool, storage resources in that pool may experience some reduction in performance. Applications that require high performance should not be provisioned from storage pools that contain these drives.

VNXe systems offer RAID 5, RAID 6, and RAID 1/0 protection. These configurations offer protection against different kinds of drive failures and are suitable for various applications. Table 2 provides a brief summary of the RAID levels supported by VNXe systems. The following sections provide more information about each type of RAID level.

Table 2 – RAID level details and configurations

RAID Type	Protects	Recommended for:	Drive Configuration
RAID 5 Striped, distributed, parity-protected	Against single drive failure.	Transaction processing; is often used for general purpose storage, relational databases, and enterprise resource systems.	4+1 8+1 12+1
RAID 6 Striped, distributed, double-parity-protected	Against double drive failure.	Same uses as RAID 5, only where increased fault tolerance is required.	6+2 8+2 10+2 14+2
RAID 1/0 Mirror-protected	Against multiple drive failures, as long as the drive failures do not occur in the same mirrored pair.	RAID 1/0 may be more appropriate for applications with fast or high processing requirements, such as enterprise servers and moderate-sized database systems.	1+1 2+2 3+3 4+4

RAID 5

RAID 5 stripes data at a block level across several drives and distributes parity among the drives. With RAID 5, no single drive is devoted to parity. This distributed parity protects data if a single drive fails. Failure of a single drive reduces storage performance because some CPU time is needed to rebuild the failed drive data to an available spare drive. The failed drive is replaced by the VNXe3200 storage system automatically from available unused drives and then the data is rebuilt on the spare drive using the RAID parity data. If there are no available unused drives then the failed drive should be replaced immediately.

The failure of two drives in a RAID 5 group causes data loss and renders any storage in the storage pool unavailable. For this reason, it is recommended to have some unused drives to allow the VNXe3200 system to automatically replace failed drives. You should physically replace failed drives as soon as possible.

RAID 6

RAID 6 is similar to RAID 5; however, it uses a double-parity scheme that is distributed across different drives. This offers extremely high fault tolerance and drive-failure tolerance. This configuration provides block-level striping with parity data distributed across all drives. Storage Pools can continue to operate with up to two failed drives per RAID group. Double-parity allows time for the storage pool to rebuild without the data being at risk if a single additional drive fails before the rebuild is complete. RAID 6 is useful for low speed, high capacity drives where the rebuild procedure will take a significant amount of time and the probability of a double drive fault is higher.

The Multicore RAID functionality available in the VNXe3200 storage system includes support for RAID 6 parallel rebuilds. This functionality allows for improved resiliency and protection for RAID 6 drive configurations as faulted drives are rebuilt and restored more quickly.

The failure of three drives in a RAID 6 group causes data loss and renders any storage resources in the storage pool RAID configuration unavailable. As with RAID 5, the failed drives are replaced automatically from an available unused drive and the data then restored based on RAID parity data. RAID 6 provides performance and reliability at medium cost, while providing lower capacity per drive.

RAID 1/0

This configuration requires a minimum of two physical drives to implement in VNXe systems, where one mirrored set in a striped set together provide fault tolerance. In other words, to provide redundancy, one drive out of every two is a duplicate of the other drive. In case of drive failure in a RAID 1/0 set the failed drive is replaced with a hot spare and the data from the remaining drive will be copied to the new drive, re-enabling the mirrored protection. Therefore, it is recommended to have some available unused drives in the storage system to automatically replace faulted drives.

NOTE: When using RAID 1/0 with only two physical drives in a 1+1 configuration, the data is not striped, effectively providing RAID 1 protection. When using a larger configuration with 4, 6, or 8 drives, data will be mirrored and then striped providing RAID 1/0 protection.

Dual Drive Ownership

The VNXe3200 storage systems support dual SP ownership of hard drives. All hard drives are dual-ported and can accept I/O from both SPs at the same time.

Spare Drives

Any appropriate unused spare drive will replace a failed storage pool drive in the event of a drive failure. When a drive fails, the VNXe system rebuilds data to the spare drive, using the remaining drives in the RAID configuration of the storage pool. The spare drive will become a permanent member of the storage pool once the rebuild is complete. When the failed drive is replaced, the VNXe system will initialize the new drive and mark it as an unused spare drive.

The VNXe system enforces a default spare policy to keep one drive unused per every 30 drives of the same type, speed, and capacity. The system will automatically keep the necessary drives unused for each additional group of 30 drives of the same type.

Data Availability in the Connectivity Infrastructure

When designing an HA environment, it is important to carefully plan the connectivity infrastructure. A single point of failure at the host-level, switch-level, or storage system-level can result in data being unavailable for the host applications.

The HA concepts for the VNXe system are different for Block-level and File-level storage resources. The following sections include additional information that explains these differences and provides examples and diagrams to help you design an HA environment.

High Availability for Block-Level Storage

The VNXe system leverages block-level interfaces (iSCSI and Fibre Channel) to allow host access to block-level storage resources, including LUNs and VMware VMFS datastores.

Block-level storage resources have an SP Owner, which can be determined from the details page in Unisphere (Figure 5). However, these resources are not associated with any specific block-level interface. In other words, if a LUN is owned by SPA and there are four block-level interfaces created on SPA, then hosts with proper access and configurations can use any of these available paths (which may require zoning configurations with an FC-SAN or network configurations with an IP-SAN). It is required to leverage multi-pathing software on the hosts in order to manage the multiple connections to the VNXe system.

NOTE: The VNXe3200 does not currently support simultaneous FC and iSCSI access to the same host.

The following sections provide additional information to help you design an HA environment with block-level storage resources provided by the VNXe system.

Multipath Storage Access

Multi Path Input Output (MPIO) is a fault-tolerance and performance enhancement technique that allows the use of more than one path to a storage device from the same host. The VNXe system fully supports MPIO for its block-level storage resources by creating multiple block interfaces and connecting hosts to these available interfaces.

EMC requires that every SAN-attached host is configured with multiple physical connections to the SAN fabric (typically one per pair of fabrics), and multiple SAN zones to every VNXe SP (typically each host HBA port is zoned to a pair of front-end ports, one on each SP). This configuration ensures that in the event of a single connection failure, MPIO technology provides undisrupted access to the data.

The VNXe3200 system utilizes Asymmetric Active/Active LUN access which is based on the Asymmetric Logic Unit Access (ALUA) standard. ALUA uses SCSI 3 primary commands that are part of the standard SCSI SPC-3 specification to determine I/O paths. By leveraging ALUA, I/O can be sent through either SP. For example, if I/O for a LUN is sent to an SP that does not own the LUN, that SP redirects the I/O to the SP

that *does* own the LUN. This redirection is done through internal communication within the VNXe3200 system. It is transparent to the host, and the host is not aware that the other SP processed the I/O. Hence, failover is not required when I/O is sent to the non-owning (or non-optimal) SP. If more than 64,000 I/Os are received on the non-optimal path, the VNXe3200 system will fail over the LUN to the peer SP in order to optimize LUN access.

NOTE: Multi-pathing software is required for hosts to manage the multiple connection paths to the VNXe3200 system. The MPIO software should be configured to first use the optimized paths, then use the non-optimized paths in the event there are no remaining optimized paths.

iSCSI Configuration and Settings

In an HA configuration, it is required to set up host connections to at least one iSCSI interface on each SP. This basic configuration allows your host to continue accessing block-level storage resources in the event that an SP should become inaccessible. It is required that you install multi-pathing software (such as EMC PowerPath) on your hosts to manage the multiple paths to the VNXe system. To create a more robust HA environment, you can setup additional iSCSI interfaces on each SP.

When implementing an HA network for block-level storage resources using iSCSI connectivity, consider the following:

- A block-level storage resource on a VNXe system is only accessible through one SP at a time[1].
- You can configure up to 8 iSCSI interfaces (each with a different VLAN ID) per each physical Ethernet port.
- You can configure one IPv4 or IPv6 address for an iSCSI interface, which is associated with a single physical Ethernet port on a Storage Processor.
- iSCSI interfaces cannot be created on link-aggregated port groups and link-aggregated port groups cannot be created on Ethernet ports with iSCSI interfaces.

When provisioning an iSCSI interface, specify the network information for the network interface, as shown in Figure 4. In the **Advanced** section, you can also enter any VLAN ID information for the iSCSI interfaces you are creating.

[1] The multi-pathing software on hosts should be configured to first use the optimized paths, then use the non-optimized paths in the event there are no remaining optimized paths.

Figure 4 – Provisioning an iSCSI Interface

After the VNXe system creates the iSCSI interfaces on both SPs using the selected Ethernet port, you can view the details in Unisphere (Figure 5). You will notice that Figure 5 only displays two available Ethernet ports. This is because a Link Aggregation port group was created on Ethernet ports eth4 and eth5.

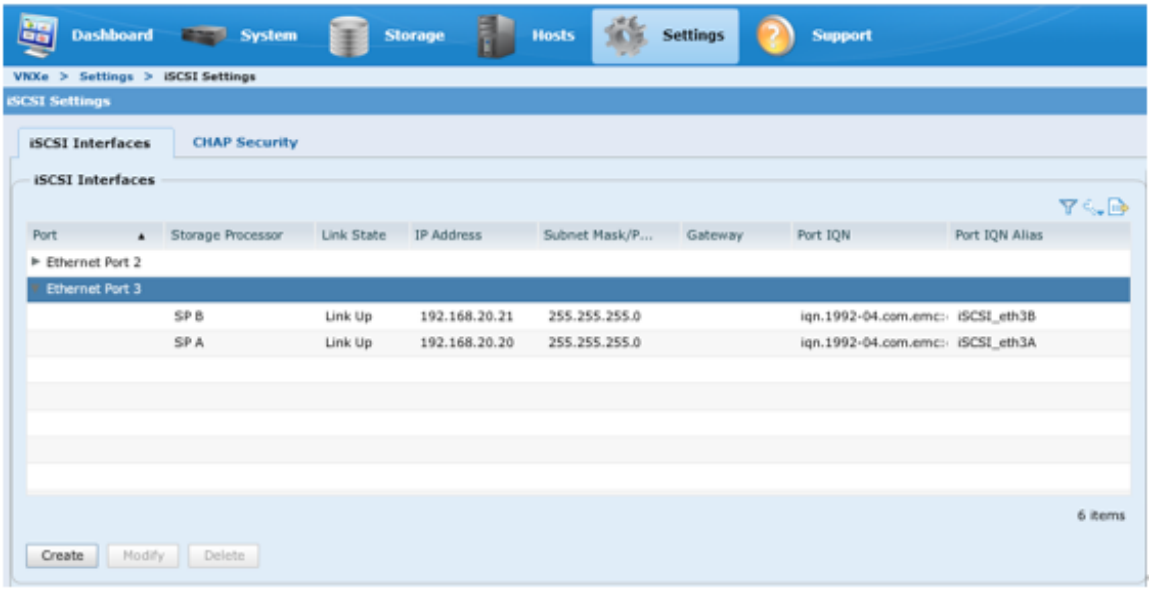

Figure 5 – Managing iSCSI Interfaces

Table 3 shows the storage resource configuration on the VNXe system. There is one IP address associated with each iSCSI interface, and each of these IP addresses is bound to a separate physical port on the given SP. For HA purposes, configure an iSCSI interface on each SP, regardless of whether the block-level storage resources were created on SPA or SPB. In this example, iSCSI interfaces have been configured on SPA and SPB.

Table 3 – Multiple iSCSI Interfaces created on each SP

Component	Interface Name	Port	IP Address
SPA	iSCSI_eth2A	eth2	192.168.10.10
SPA	iSCSI_eth3A	eth3	192.168.20.20
SPB	iSCSI_eth2B	eth2	192.168.10.11
SPB	iSCSI_eth3B	eth3	192.168.20.21

Fibre Channel Configuration and Settings

In an HA configuration, it is required to set up host connections to at least one Fibre Channel (FC) interface on each SP. This basic configuration allows your host to continue accessing block-level storage resources in the event that an SP becomes inaccessible. It is required that you install multi-pathing software (such as EMC PowerPath) on your hosts to manage the multiple paths to the VNXe system. To create a more robust HA environment, you can connect hosts to access the storage system using additional FC Interfaces on each SP.

When implementing an HA network for block-level storage resources using FC connectivity, consider the following:

- A block-level storage resource on a VNXe system is only accessible through one SP at a time[2].
- FC interfaces are system-defined and created automatically by the VNXe system. FC interfaces are not user-configurable.

Fibre Channel (FC) interface details can be found in two places within Unisphere, including the **System Health** page (not shown) and the **Port Settings** page (Figure 6).

[2] The multi-pathing software on hosts should be configured to first use the optimized paths, then use the non-optimized paths in the event there are no remaining optimized paths.

Figure 6 – Port Settings page in Unisphere showing FC Port details

By using this information, you will be able to configure a similar setup for your FC-SAN environment. You'll be able to see all the connected Host FC WWNs on the **Initiators** page in Unisphere (Figure 7).

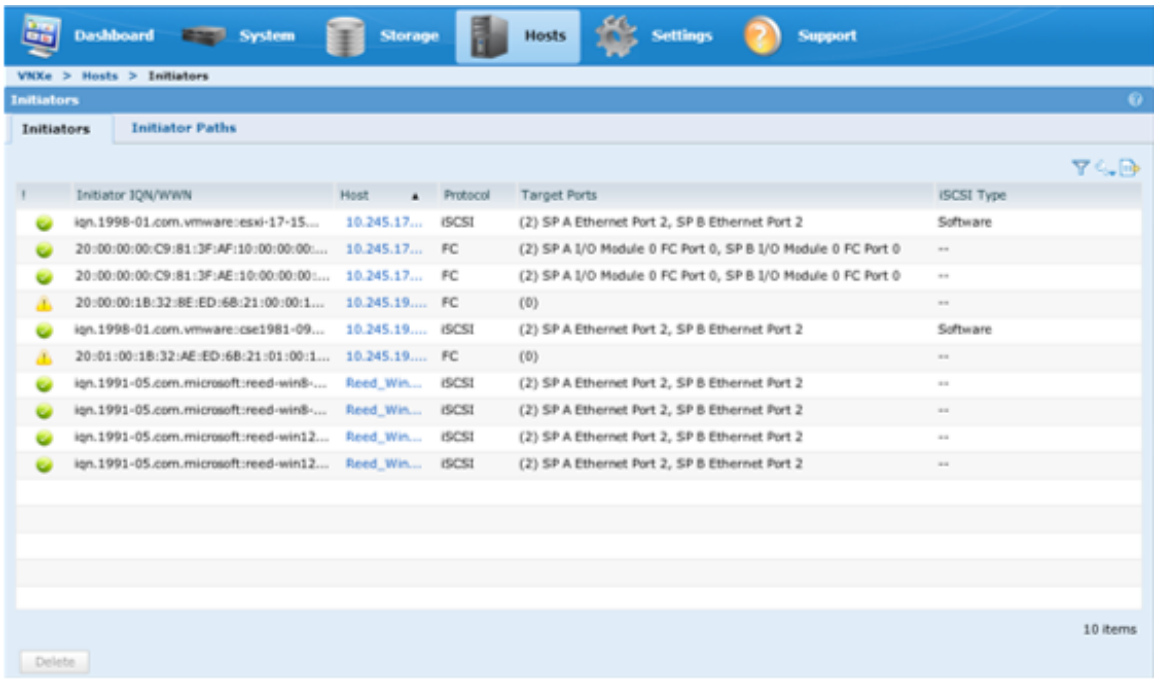

Figure 7 – Initiators page in Unisphere showing connected iSCSI IQNs and FC WWNs of Hosts

High Availability Options for Block-Level Storage

When implementing HA for block-level storage resources, consider the following:

- A block-level storage resource on a VNXe system is associated with only one SP at a given time.
- iSCSI interfaces:
  - Do not failover and cannot leverage the Link Aggregation (LACP) functionality. Because of this, iSCSI interfaces must be created so that there is at least one interface created on each SP. This allows a path failover when an iSCSI connection is lost.
- FC interfaces:
  - Are system-defined and created automatically on the VNXe3200 system.
- Hosts connecting to block-level storage resources on a VNXe system must be configured to connect to at least one interface from each SP. This allows hosts to utilize different paths to the storage resources in the event a connection path becomes unusable. For a more robust HA environment, hosts can be configured to connect to additional interfaces on each SP.
- Using multiple switches enhances the level of protection in an HA environment.

The following examples display the benefits of the HA concepts and options available on the VNXe3200 storage systems for block-level storage resources.

Example

The following is an example HA configuration for a VMware ESXi environment (Figure 8).

- The VMware VMFS datastore created on the VNXe system is owned by SPA.
- The VNXe system has block-level interfaces created on two ports on each SP for a total of four available paths out of the storage system. The cabling is the same for both SPs as well – the first port on each SP is connected to Switch B, and the second port on each SP is connected to Switch A.
- Switch A and Switch B are configured for separate subnets (Switch A–10.10.10.1/24; Switch B–10.20.20.1/24).
- Each ESXi host is configured to access the VMFS datastore through all the available connection paths. Each ESXi Host is configured so that it can connect to the first and second port on each SP.
- All of the ESXi host ports are cabled and configured the same way – the first port on each host is connected to Switch B, and the second port on each host is connected to Switch A.
- In Unisphere, each ESXi host has an entry created including any IQN or WWN information associated with it.

NOTE: This example has been designed in such a way so that it can be applied to SAN environments leveraging iSCSI or FC connectivity options. You cannot use FC and iSCSI at the same time for the same host. The images provided have been created to

show connectivity to the onboard Ethernet ports using iSCSI interfaces. This HA configuration can be implemented for FC-SAN environments through the use of the optional Fibre Channel I/O Module, if installed in your VNXe system.

The links in each of the following diagrams are denoted by integers [1-8] and are referenced when explaining traffic routing throughout the example scenarios.

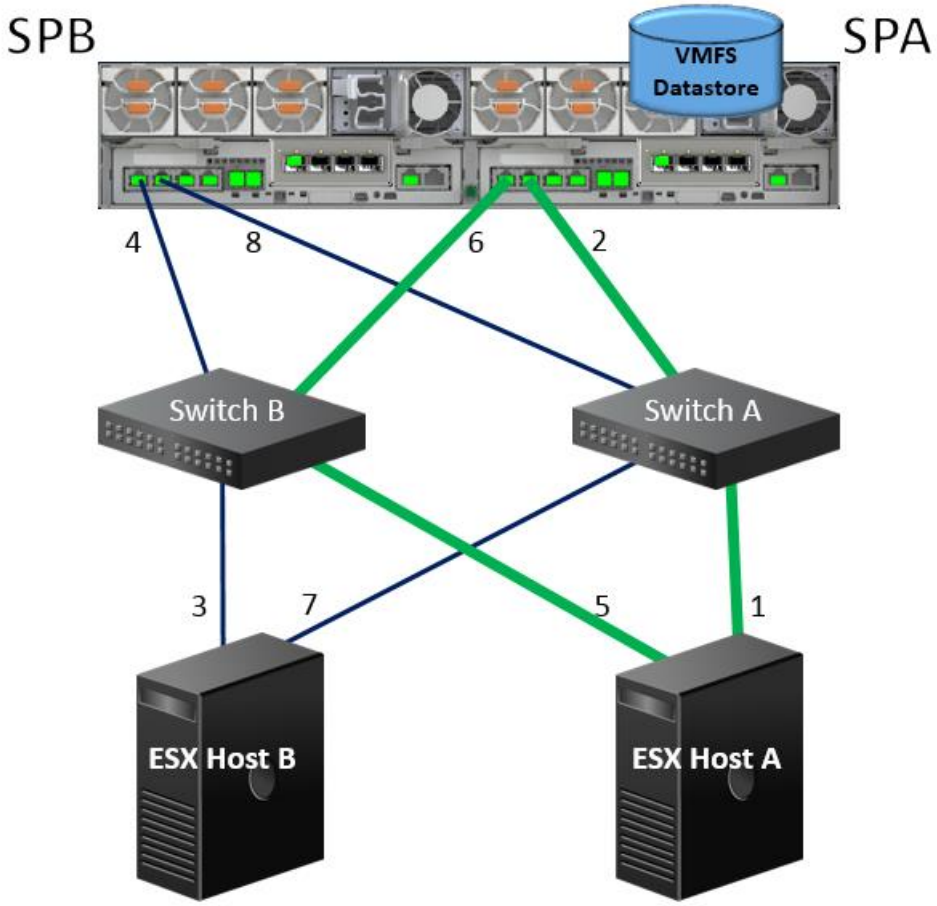

Figure 8 – HA Configuration for iSCSI Storage

In this example environment, both ESXi Host A and ESXi Host B are configured so that they can access all the available paths from the VNXe system. For example, ESXi Host A is configured so that it can connect to the first and second port on each SP. In a fully HA environment, ESXi Host A can access the VMFS storage on SPA through multiple paths - optimally through Switch A (link 1 -› link 2) and Switch B (link 5 -› link 6), and non-optimally[3] through Switch A (link 1 -› link 8) and Switch B (link 5 -› link 4).

Scenario 1: Path Failure

In a fully HA network, ESXi Host A has multiple paths to access the VMFS storage on SPA. In this example, the optimal paths ESXi Host A can use to access the VMFS datastore are through Switch A (link 1 -› link2) and through Switch B (link 5 -› link 6).

[3] The multi-pathing software on hosts should be configured to first use the optimized paths, then use the non-optimized paths in the event there are no remaining optimized paths.

If a port or link fails, such as link 2 as shown in Figure 9, ESXi Host A can still access the VMFS storage using the optimal path through Switch B (link 5 -› link 6).

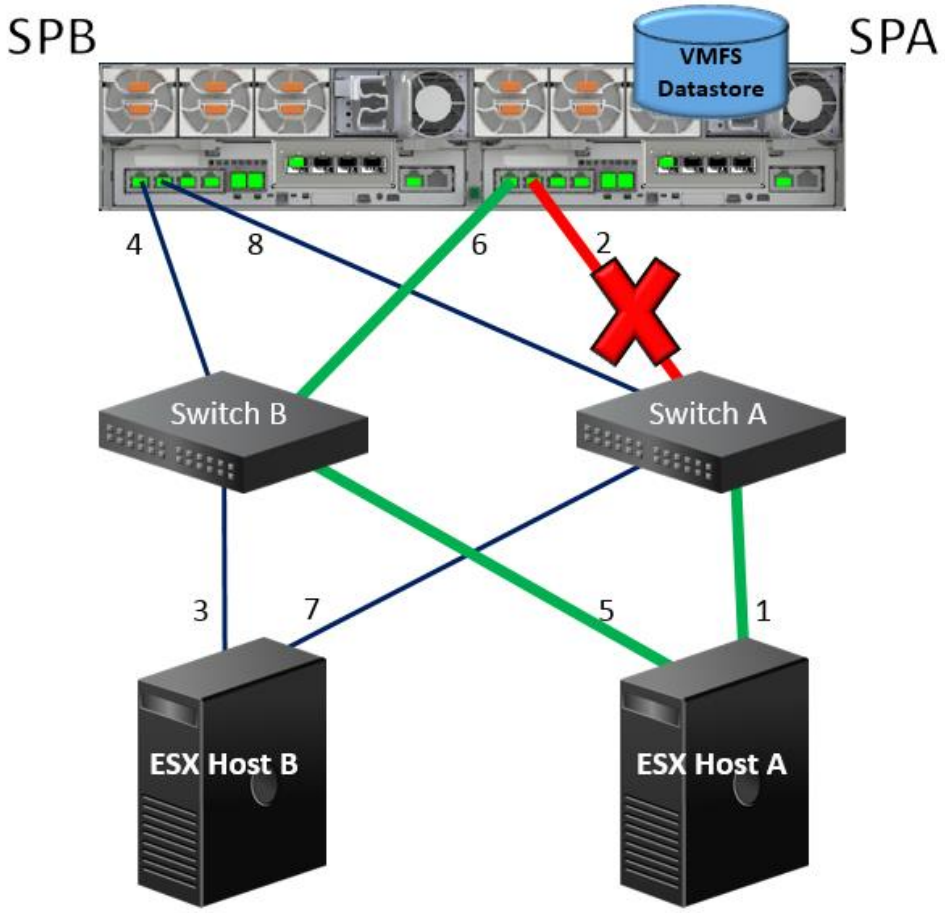

Figure 9 – Alternative path remains after SP port or link failure

If multiple ports or links fail, such as link 2 and link 6 in the example configuration, ESXi Host A can still access the VMFS storage using the non-optimal paths through Switch A (link 1 -› link 8) and through Switch B (link 5 -› link 4).

When using the non-optimal paths to service I/O requests, the VNXe system will route the I/O through the inter-SP communications link to send the request to the correct SP to be processed. I/O responses are then returned through the inter-SP communications link to the peer SP, which then returns the I/O response to the host.

In this instance, the block-level storage resource (LUN or VMFS datastore) will failover to the peer SP once 64,000 I/Os have been made over non-optimal paths or when the host multi-pathing software requests a failover (whichever happens first).

Scenario 2: Switch Failure

If a switch fails, such as Switch A as shown in Figure 10, ESXi Host A can still access the VMFS storage via the optimal path (link 5 -› link 6) and the non-optimal path (link 5 -› link 4).

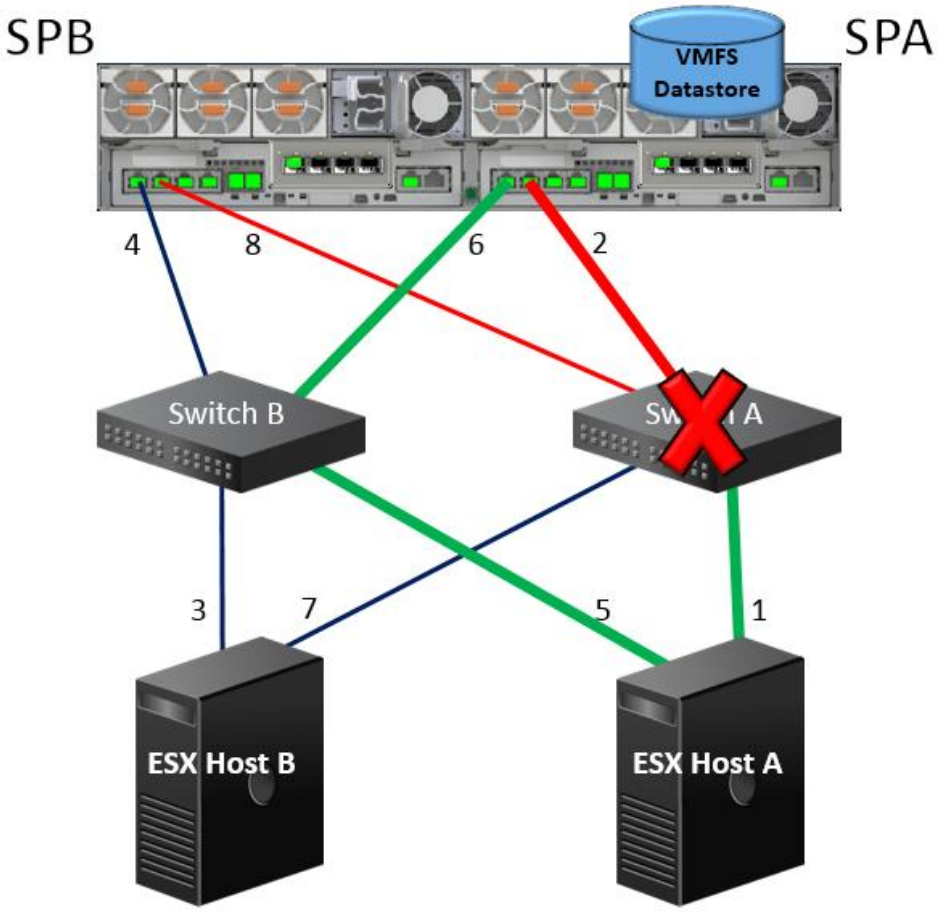

Figure 10 – Alternative path remains after switch failure

Scenario 3: SP Failure

If an SP is faulted and cannot service requests from any hosts, such as SPA in the example configuration, ESXi Host A can still access the VMFS storage using the non-optimal paths through Switch A (link 1 -› link 8) and through Switch B (link 5 -› link 4), as shown in Figure 11.

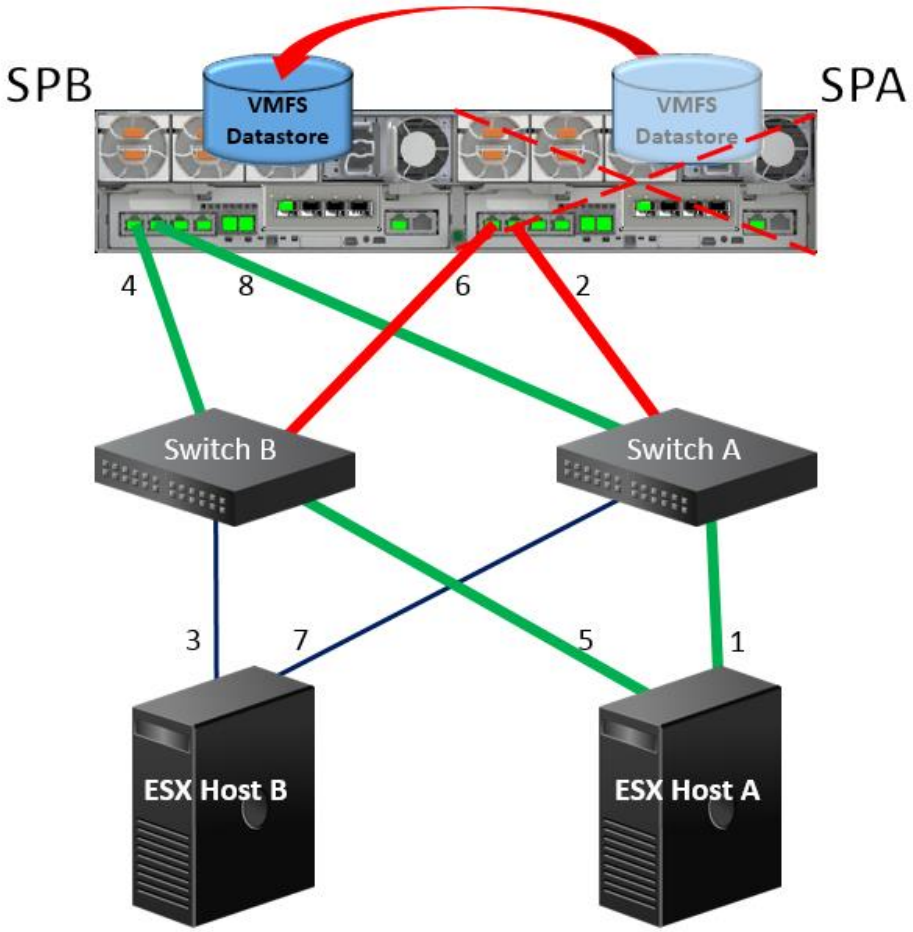

Figure 11 – Alternative paths remain when an SP fails

While SPA remains unavailable, SPB will service and process any I/O requests to the block-level storage resources on the VNXe system. Once SPA returns to Normal Mode, the optimal paths will be available and SPA will begin servicing and processing I/O requests for the block-level storage resources that are associated with it.

When implementing an HA network for block-level storage, you can leverage VMware ESXi's failover software, Native Multipathing Plug-in (NMP). The failover software contains policies for Fixed, Round Robin, and Most Recently Used (MRU) device paths. Once port bindings are added to the software iSCSI adapter, you can configure a single vSwitch with two NICs so that each NIC is bound to one VMkernel port. In addition, you can explicitly associate the VMkernel interfaces with the VMware iSCSI software initiator instance.

High Availability for File-Level Storage

The VNXe system leverages NAS Servers to provide host access to file-level storage resources, including file systems (CIFS or NFS) and VMware NFS datastores.

NAS Servers are associated with an SP and a Storage Pool, which can be determined from the NAS Server details page (Figure 12), and have front-end interfaces that are associated with a specific Ethernet interface. When file-level storage resources are created, they are associated with a NAS Server which provides access to hosts.

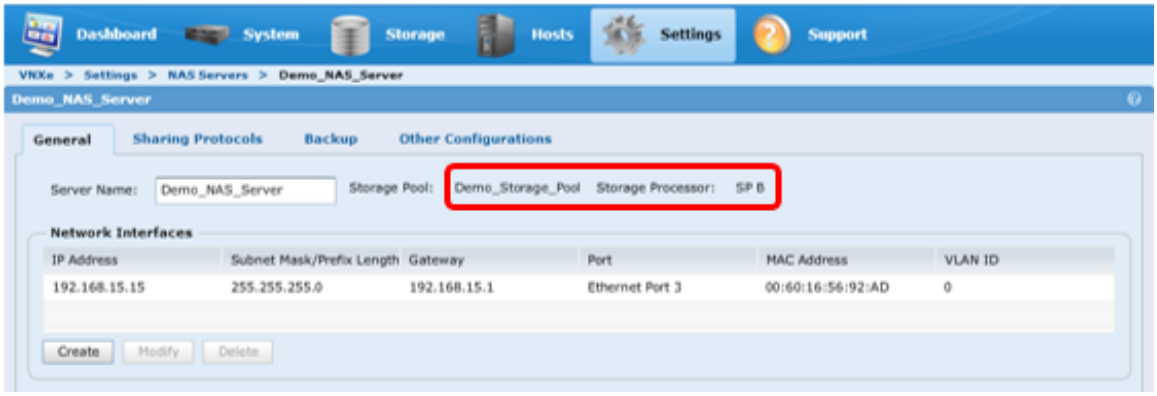

Figure 12 – NAS Server details page showing the assigned Storage Processor

The following sections provide additional information to help design an HA environment with file-level storage resources provided by the VNXe system.

Link Aggregation (LACP)

The VNXe system supports link aggregation that allows up to four Ethernet ports to be connected to the same physical or logical switch to be combined into a single logical link. To configure link aggregation on a VNXe system, each storage processor (SP) must have the same type and number of Ethernet ports. This is because configuring link aggregation actually creates two link aggregations, one on each SP. This provides high availability as follows:

- If one of the ports in the link aggregation fails, the system directs the network traffic to one of the other ports in the group.

VNXe systems use the Link Aggregation Control Protocol (LACP) IEEE 802.3ad standard. A link aggregation appears as a single Ethernet link and has the following advantages:

- High availability of network paths to and from the VNXe system — if one physical port in a link aggregation fails, the system does not lose connectivity.
- Possible increased overall throughput —multiple physical ports are bonded into one logical port with network traffic distributed between the multiple physical ports.

Although link aggregations can provide more overall bandwidth than a single port, the connection to any single client runs through one physical port and is therefore limited by the port's bandwidth. If the connection to one port fails, the switch automatically switches traffic to the remaining ports in the group. When the

connection is restored, the switch automatically resumes using the port as part of the group.

With link aggregation, the cabling on SPA must be identical to the cabling on SPB for failover purposes. Also, the switch must support and be configured for LACP. Figure 13 shows the Unisphere **Port Settings** page, with Ethernet Port 4 and Ethernet Port 5 bound to create one logical port.

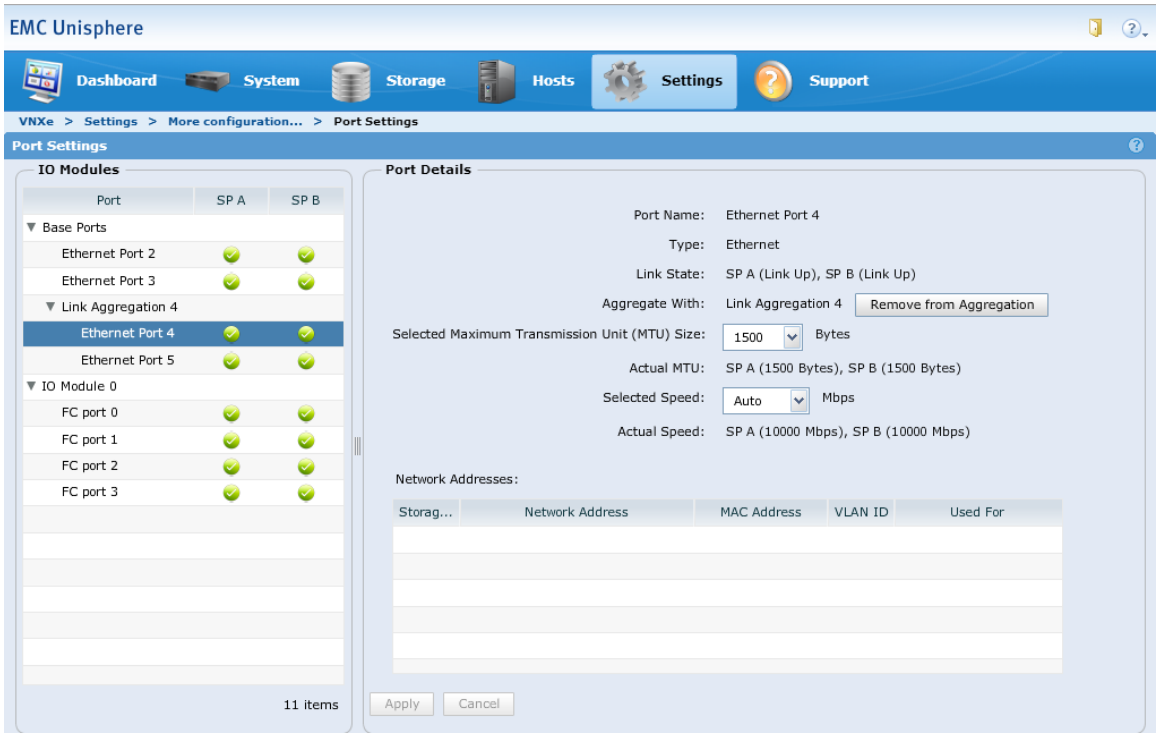

Figure 13 – Port Settings page in Unisphere showing a Link Aggregation port group

NAS Servers Configuration and Settings

NAS Servers are software components used to transfer data and provide the connection ports for hosts to access file-level storage resources. NAS Servers retrieve data from available drives according to specific protocols (NFS or CIFS) in order to make it available to network hosts. The VNXe3200 supports NAS Servers for managing file-level (NFS or CIFS) storage resources, such as VMware NFS datastores or file systems. Before you can provision a file-level storage resource, a NAS Server must be created with the necessary sharing protocols enabled on the VNXe system. When NAS Servers are created, you can specify the storage pool and owning SP – either SPA or SPB. Once NAS Servers are created, you cannot modify its owning SP or its associated storage pool.

In VNXe systems, NAS Servers can leverage the Link Aggregation functionality to create a highly available environment. Once a link aggregated port group is available on the VNXe system, you can create or modify NAS Server network interfaces to leverage the available port group. Because NAS Servers are accessible through only one SP at a time, they will fail over to the other SP when there is an SP failure event,

as noted in Table 1. For this reason, it is important to cable the Ethernet ports the same on both SPs for any ports that used by NAS Servers.

Hosts connect to NAS Servers and file-level storage resources through SPA and SPB. Assume that SPA experienced a hardware failure and the NAS Servers failed over from SPA to SPB. The hosts now access the NAS Servers through SPB. If the NAS Servers and file-level storage resources were hosted on SPB and SPB experienced this hardware failure, the NAS Servers would have failed over to SPA. The **System Health** page in Unisphere shows that SPA is faulted, causing the NAS Server to fail over to SPB. You can view this page to determine if an SP or one of its components has an error.

After the storage resources are failed over, data traffic is re-routed to SPB[4]. Hosts that were accessing the storage resource through SPA may experience a small pause in I/O servicing. Hosts and applications should be configured to wait for the storage resources to become available on SPB. Note that different storage resources may behave differently when there is an SP failover. For example, CIFS users may have to reconnect to their shared storage.

SMB 3.0 Continuous Availability

The SMB 3.0 protocol support is available with Microsoft Windows 8 and Microsoft Windows Server 2012 systems and has significant improvements over the previous SMB versions. The VNXe Operating Environment (OE) version 3.0 has the SMB 3.0 protocol enabled by default. Continuous Availability (CA) is one of the enhancements introduced with this protocol.

CA minimizes the impact on applications in the event of a NAS Server failure or recovery of applications. In the situation where a storage processor is faulted or placed in service mode, storage resources are failed over to the peer storage processor. With CA, application access to all open files that were present prior to the failover is re-established and the failover is transparent to the end users.

For more information on CA, refer to the **Introduction to SMB 3.0 Support** white paper on EMC Online Support.

[4] This is not due to the Link Aggregation (LACP) functionality. This occurs because the route is updated when the original network interface (IP address) is presented through a different interface.

High Availability Options for File-Level Storage

When implementing HA for file-level storage resources, keep in mind that:

- File-level storage resources on a VNXe3200 system are accessed through NAS Servers, which are presented on only one SP at a given time.
- NAS Servers can have network interfaces on a link aggregated port group. The port group combines multiple Ethernet ports to make a single virtual link with a single MAC address. Depending on the number of NAS Servers using the port group, this virtual link may have multiple IP addresses.
- Using stacked switches enhances the level of protection in this HA environment. You can create LACP groups that span physical switches (often referred to as cross-stack LACP) with stacked switches.

The following example displays the benefits of the HA concepts and options available on the VNXe3200 storage systems for file-level storage resources.

Example

The following is an example HA configuration for a VMware ESXi environment (Figure 14).

- The VMware NFS datastore is accessible using a NAS Server which resides on SPA.
- The NAS Server has a network interface that is leveraging a link aggregation port group using Ethernet ports eth4 and eth5.
- The Ethernet ports for both SPA and SPB are cabled the same way – eth4 ports from each SP are cabled to Switch B, and eth5 ports from each SP are cabled to Switch A.
- Switch A and Switch B are set up for a stacked-switch configuration, and have an inter-switch link configured.
- Each ESXi host can access the NFS datastore through separate, individual NICs. In addition, each of the NICs on an ESXi host connects to different switches.

The links in each of the following diagrams are denoted by integers [1-8], and are referenced when explaining traffic routing.

When implementing an HA network for NFS, NIC Teaming on the ESXi hosts provides fault tolerance in case of a NIC port failure. Teaming and Failover policies (found in the Properties page of the vSwitch) help determine how network traffic is distributed between adapters and how to route I/O traffic in the event of an adapter failure. For more information, refer to the **ESXi Configuration Guide** (for v4.X) and **vSphere Networking Guide** (for v5.X) on the VMware technical resource page.

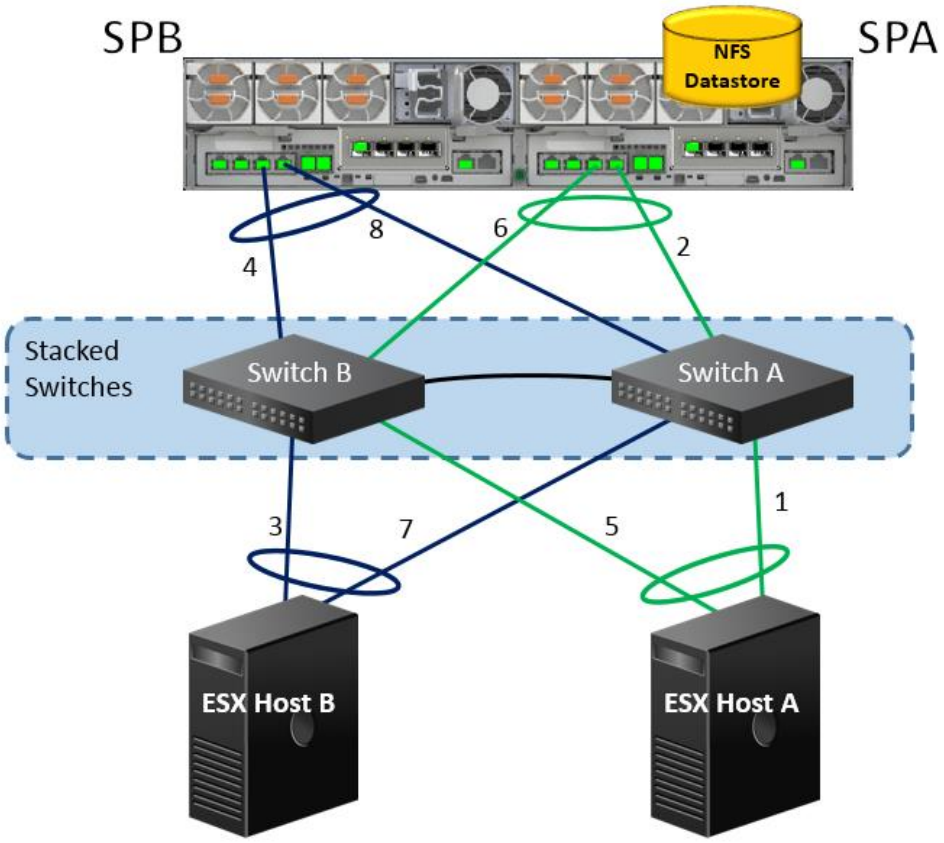

Figure 14 – HA Configuration for NFS Storage

Table 4 shows the storage resource configuration on the VNXe system. A network interface is assigned to each NAS Server leveraging the link aggregation port group (using eth4 and eth5) on the VNXe system. For HA purposes, it is recommended to associate each NAS server with link aggregated port groups, regardless of whether the SP storage resources were created in SPA or SPB.

Table 4 – Aggregating together ports eth4 and eth5

SP	NAS Server	Port Group	IP Address
SPA	NAS_serverA	eth4/eth5	192.168.15.15
SPB	NAS_serverB	eth4/eth5	192.168.25.25

In a fully HA environment, ESXi Host A can access the NFS storage on SPA through Switch A (link 1 -› link 2) and Switch B (link 5 -› link 6).

Scenario 1: Path Failure

If a port or link fails, such as link 2 in the example configuration, ESXi Host A can still access the NFS storage through Switch A by leveraging the Inter-Switch Link (ISL) between the stacked switches (link 1 -› inter-switch link › link 6) or Switch B (link 5 -› link 6) (Figure 15). Throughout this failure, hosts will continue to have access to the NAS Server and NFS storage resource which both still reside on SPA.

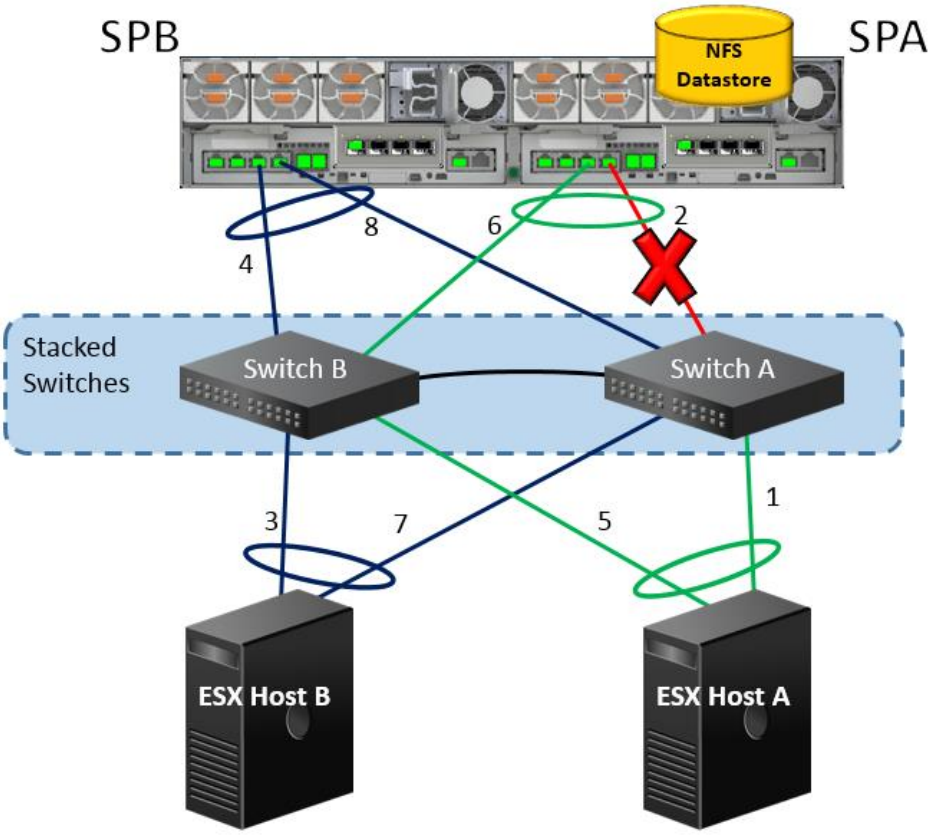

Figure 15 – Link remains active after SP port or link failure

Scenario 2: Switch Failure

If a switch fails, such as Switch A in the example configuration, ESXi Host A can still access the NAS Server and NFS storage through Switch B (link 5 -> link 6, Figure 16). Throughout this failure, hosts will continue to have access to the NAS Server and NFS storage resource which both still reside on SPA.

If stacked switches are used, like in the example configuration, you can create an LACP trunk with links to each switch in the stack. If a switch fails, the links to that switch fail, but traffic fails over to the surviving links in the LACP group to the other switch.

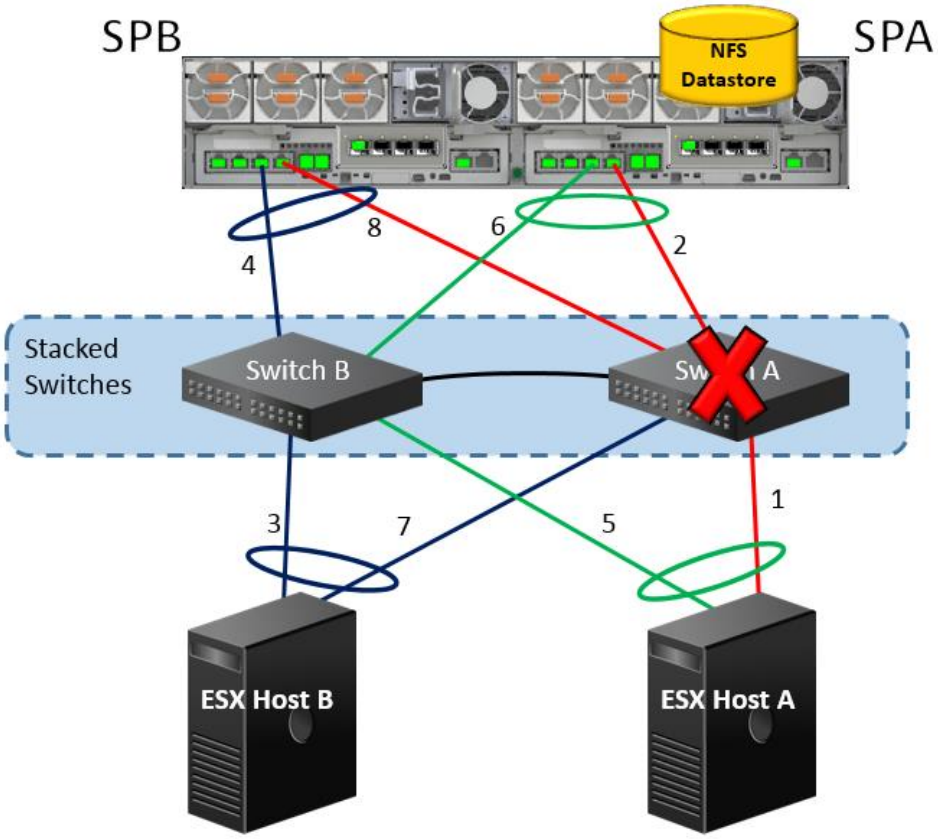

Figure 16 – Links remain active after switch failure

Scenario 3: SP Failure (NAS Server Failover)

If an SP is faulted and cannot service requests from any hosts, such as SPA in the example configuration, the NAS Server and NFS storage resource fails over to the peer, SPB. The I/O traffic is now routed through a different path, leveraging the available ports from SPB. In this scenario, ESXi Host A in our example configuration would access the storage resource through Switch A (link 1 -› link 8) or through Switch B (link 5 -› link 4) as shown in Figure 17.

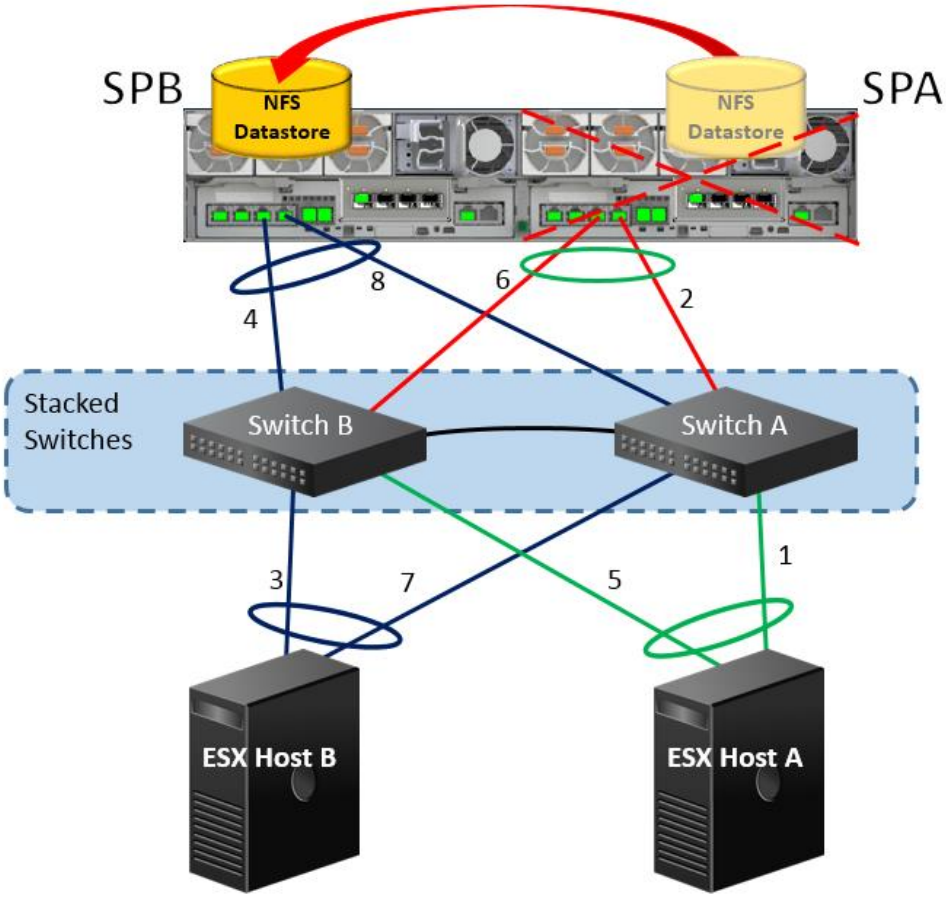

Figure 17 – Alternate path when an SP fails causing a NAS Server failover

Failback Options

Failback is the reverse of failover. It involves moving all storage resources that have failed over back to their original SP. By default, storage resources automatically fail back when the original SP returns to a healthy state.

Administrators can disable automatic failback using the Failback Policy option located in the Management Settings page of Unisphere, shown in Figure 18. If disabled, an administrator must perform a manual failback for the storage resources to return to the original SP. Administrators should disable the automatic Failback Policy only if they need to monitor the failback process or when the failback needs to occur at a specific time to lessen the impact to users.

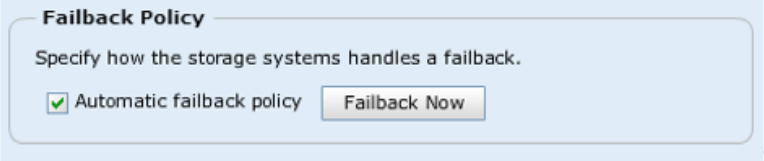

Figure 18 – Failback Policy option

Replication

For additional data availability in your storage environment, you can leverage the replication functionality on your VNXe3200 storage system. Details about the replication functionality is provided in the following sections, but more information can be found in the **EMC VNXe3200 Replication Technologies** white paper available on EMC Online Support.

Native Block Replication

In VNXe Operating Environment 3.1.1 and later, you can replicate block-level storage resources asynchronously from your VNXe3200 system to a destination VNXe3200 system. This applies to LUNs, LUN groups, and VMware VMFS datastores and is entirely managed using the Unisphere interface.

EMC RecoverPoint Integration

In VNXe Operating Environment 3.1.1 and later, you can leverage EMC RecoverPoint to replication block-level storage resources both synchronously and asynchronously from your VNXe3200 system. By leveraging the on-array RecoverPoint splitter, your VNXe3200 system can take advantage of the advanced replication features provided by EMC's RecoverPoint technology, including the ability to replicate to destination VNX, VMAX, and VPLEX systems. For more information regarding RecoverPoint, refer to the **EMC RecoverPoint Administrator's Guide** available on EMC Online Support.

Conclusion

The need to deliver increasing levels of availability continues to accelerate as organizations re-design their infrastructure to gain a competitive advantage. Typically, these new solutions rely on immediate access to business-critical data. When this critical data is not available, the organization's operations fail to function, leading to lost productivity and revenue. A top concern for IT administrators is designing an organization's IT environment with high availability (HA) in mind. Setting solid HA measures assures business continuity and lessens time and cost in recovery efforts, should there be technical difficulties. The EMC VNXe3200 storage systems have these HA measures built in and can ensure that data is readily accessible to the customer.

This white paper described the key HA features that VNXe systems offer at the network and storage levels. Configuring multiple paths to storage resources allows business-critical data to remain accessible in situations of network failure. VNXe systems have an N+1 redundant architecture, which provides data protection against any single component failure. Choosing the appropriate drive configuration alleviates performance impacts and disruptions caused from drive failure.

References

For additional information regarding any of the topics presented in this white paper, please refer to the following documentation available on EMC Online Support:

EMC Unisphere on the VNXe3200: Next-Generation Storage Management white paper

EMC VNXe3200 MCx Multicore Everything white paper

EMC VNXe3200 Replication white paper

Introduction to EMC VNXe3200 SMB 3.0 Support white paper