EMC²
TECHNICAL NOTES

VMware ESX 5.x iSCSI NMP Connectivity Guide for VNX

Technical Notes

P/N 302-001-019
REV 01
May 2014

Contents

Introduction

VMware ESXi version 3.0 introduced support for host connectivity to iSCSI storage systems. In order to accommodate the iSCSI target implementation, ESXi offers multiple configuration options to satisfy the varied connectivity requirements of different iSCSI target implementations. This document clarifies the connectivity options and recommendations for iSCSI with the VNX iSCSI target.

There are two basic options for iSCSI connectivity:

- Hardware Initiator: Uses a dedicated iSCSI adapter card to access the target. This adapter is functionally similar to other SAN adapters, using Ethernet networks in place of Fiber Channel. Many adapters are configured at the BIOS level and appear as another storage adapter within the VMware management interface.
- Software Initiator: Implemented as a VMkernel software driver. It uses one or more standard 1Gb or 10Gb network interfaces on the ESXi host to carry iSCSI traffic.

This Technical Note focuses on the proper configuration of the iSCSI software initiator when using EMC VNX storage systems. It also describes the supported configuration options and presents the considerations of each.

In addition, this Technical Note resolves confusion about the configuration of iSCSI for VNX and provides the information necessary for determining the best supported iSCSI topologies for using ESX iSCSI with VNX.

ESXi Networking Brief

Since the software initiator is built upon shared ESXi host network resources, this section gives a brief overview of the network configuration to support subsequent discussion points.

Each host should have multiple physical network interfaces (VMNICs) for redundancy and scaling. For 1Gb networks, each interface should support a specific service like iSCSI or Virtual Machine networking. iSCSI NICs should not be shared. Each interface is configured as either:

- Virtual machine network ports for VM connectivity, or
- VMkernel ports for Storage VMotion, iSCSI, NFS, and host management.

You implement the iSCSI software initiator as a VMkernel device driver. It uses the hosts' Ethernet adapters to transport the iSCSI commands and data. The software initiator is dependent on one or more VMkernel ports (VMKNICs) and each VMkernel port depends on one or more physical network interfaces (VMNICs).

Figure 1 shows a host with four physical network interfaces identified as vmnic1 through vmnic4.

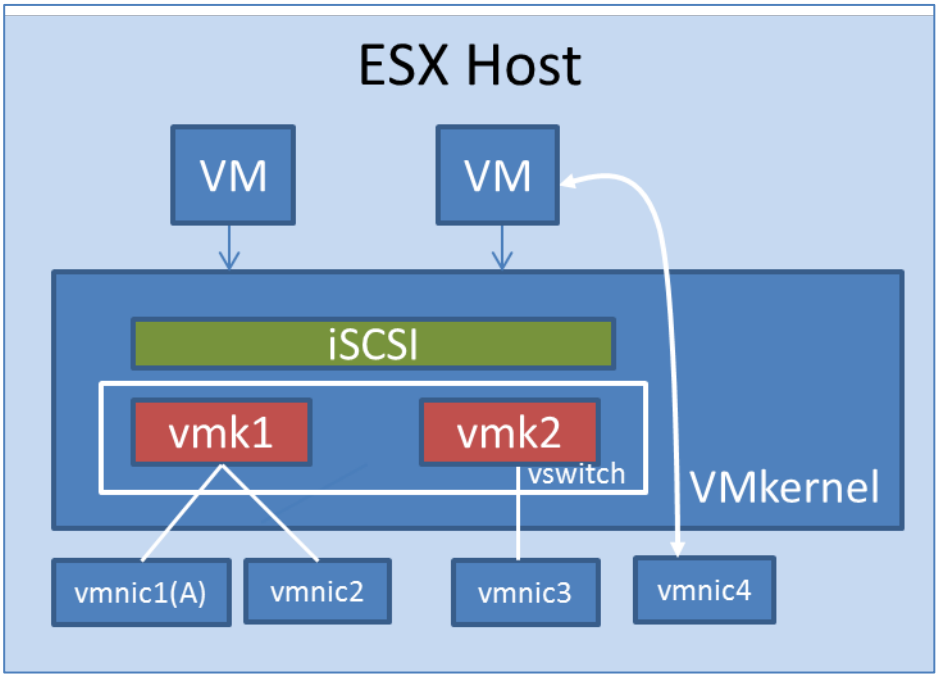

Figure 1. iSCSI Network block

VMKernel port vmk1 is configured using a NIC team consisting of one active NIC (vmnic1) and a standby device (vmnic2), which will be used if vmnic1 fails. However, it should be noted that NIC teaming for VMkernel interfaces does not increase the throughput of the VMkernel interface, and therefore iSCSI initiator. Multi-pathing is provided by adding multiple vmk ports to the iSCSI initiator. In this example, vmk2 (consisting of a single vmnic) provides a second path that can be used for multipath.

VMware Software Initiator

When configuring the iSCSI software initiator, VMkernel ports are associated with the initiator in one of two ways:

1. Port binding: One or more VMkernel ports are explicitly assigned to software initiator through vCenter or the ESXcli commands.

   The following rules apply to port binding:
   - Port cannot be used for anything else.
   - All VMkernel ports must use the same subnet address.

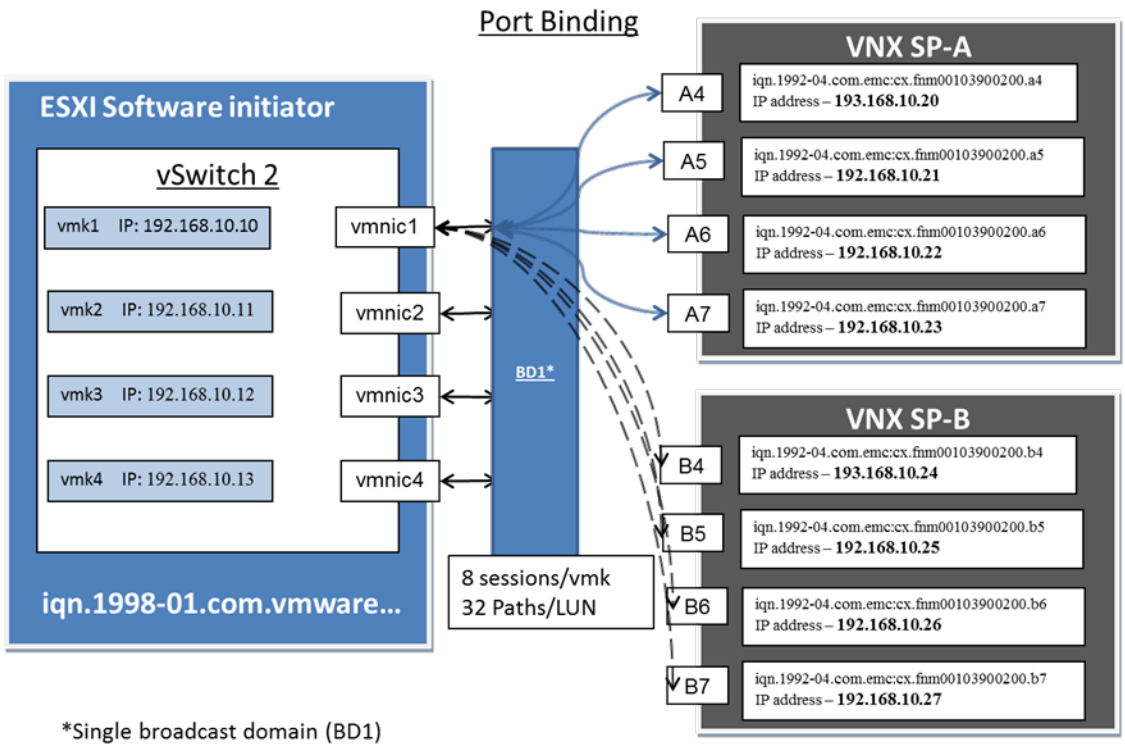

Figure 2: Port Binding topology diagram

2. Non-port binding (Implicit VMkernel assignment): All configured VMkernel ports are available for use by the software initiator.

During the iSCSI target discovery phase, the ESXi VMkernel uses the dynamic discovery address to connect to the target and issues a SCSI send-targets inquiry command. The iSCSI target responds with all of the configured addresses. The addresses are visible in the static address tab of the vCenter software initiator properties window.

Prior to performing the login phase of the iSCSI session, the VMkernel performs a network connectivity test between each VMkernel network port and VNX target port using the network addresses returned from the send targets SCSI inquiry (via dynamic discovery). Validated network paths are used to establish iSCSI sessions with the VNX. Path or login failure is not retried. Figure 3 shows that if all VMkernel network ports have a route to all VNX target ports, each LUN will have 32 potential IO paths.

- Network ports can be shared with iSCSI and NFS

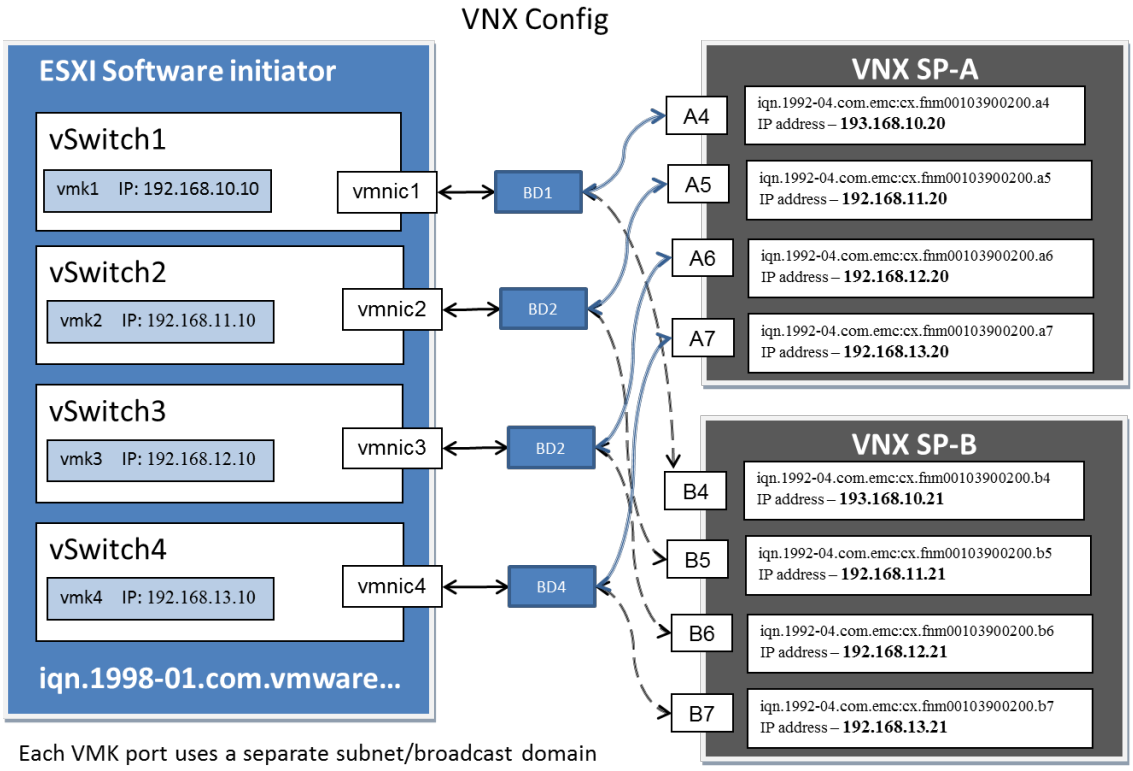

Figure 3: Topology Diagram for multi-subnet iSCSI configuration

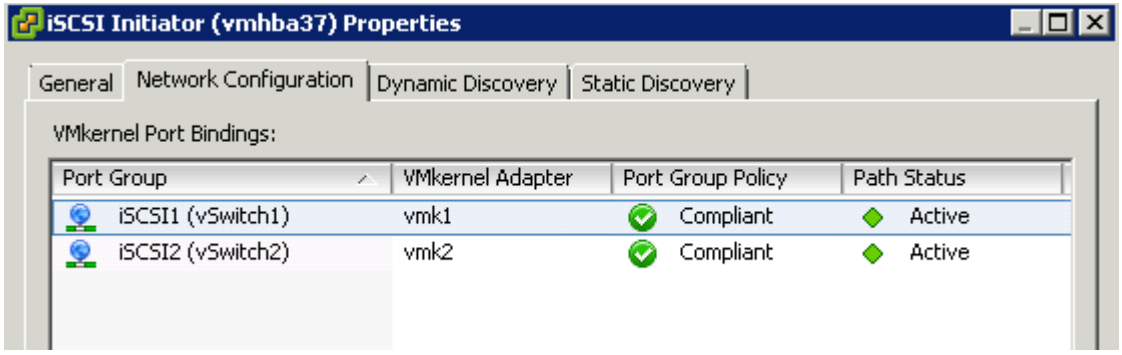

Figure 4: Explicit Port Binding (Not Recommended)

Implicit iSCSI binding is used when the iSCSI target uses multiple subnets. This configuration is established by:

- Configuring VMkernel ports on the same subnets as the VNX iSCSI network portals.
- Defining one of the VNX target addresses to the initiator through the Dynamic Discovery interface of the iSCSI configuration UI.

You can think of the different binding configurations as being associated with the network topologies they support. Port Binding can only be used when the initiator and all targets are part of a single subnet (i.e., broadcast domain); this is the first example shown in Figure 1.

Implicit port assignment is accomplished by not assigning any adaptors to the iSCSI initiator. When the software initiator is enabled, all hosts will attempt to use all VMkernel paths.

Configuration

Configuring the ESXi software initiator can seem confusing at first. But if you've configured the iSCSI software initiator, these steps will be familiar. The key point between port binding for a single subnet and the multi-subnet configuration is that you do not perform an explicit port binding step. You simply configure the VMkernel portal groups and define the VNX discovery address.

> Note: The iSCSI software initiator is not enabled by default. To enable it from vSphere: Select the ESXi host > Configuration tab > Storage Adapters > Add > Add Software iSCSI Adapter. If the iSCSI adapter was previously added, verify that it is enabled by viewing the iSCSI adapter's properties.

The VMkernel will:

1. Contact the VNX iSCSI target.
2. Issue a send targets command.
3. Validate the network path between each VMkernel port and VNX target address.
4. Attempt to establish a session between the initiators and targets using the paths that have been validated. (Note: an unreachable target would not be included as a valid path).
5. If routing is configured on the VMkernel port and all target ports on the VNX are accessible, the number of iSCSI sessions will be doubled. For example, if the 192.168.10 network contains a route to 192.168.11, both VMkernel interfaces (192.168.10.10 and 192.168.11.10) will attempt to log into the port resulting in a total of 8 target paths. However, the recommendation for iSCSI is not to configure routing, so the result is that 4 paths are used.

   If the network is not routable, each VMkernel interface will log into the target using the local subnet addresses, which will create (for this example) 4 iSCSI sessions.

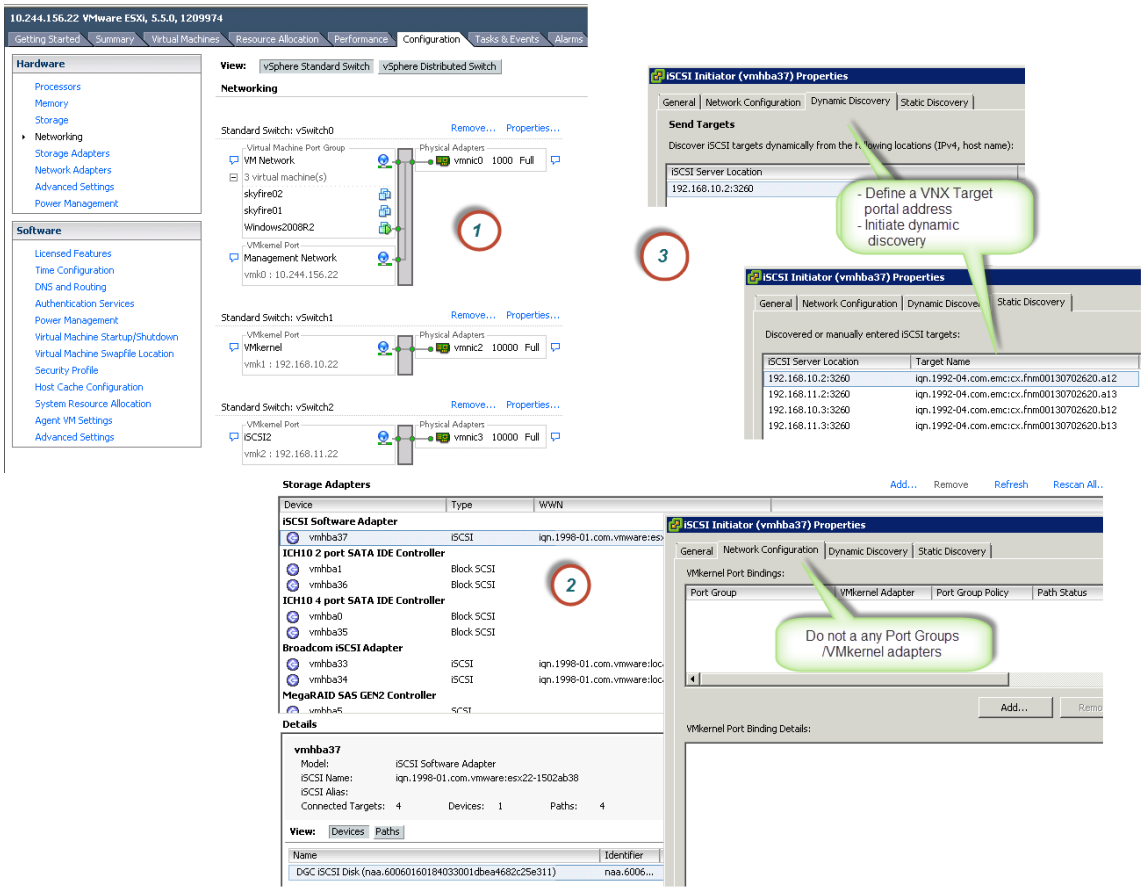

Figure 5: iSCSI configuration

Delayed ACK

During periods of network congestion, the TCP/IP implementation between the VNX and ESXi host's iSCSI initiator can potentially cause slow read performance.

Rather than implementing either a slow start algorithm or congestion avoidance algorithm (or both), the VNX takes the very conservative approach of retransmitting only one lost data segment at a time and waiting for the host's ACK before retransmitting the next one. This process continues until all lost data segments have been recovered.

Coupled with the delayed ACK implemented on the ESXi host, this approach slows read performance to a halt in a congested network. Consequently, frequent timeouts are reported in the kernel log on hosts that use this type of array. Most notably, the VMFS heartbeat experiences a large volume of timeouts because VMFS uses a short timeout value. This configuration also

experiences excessively large maximum read-response times (on the order of several tens of seconds) reported by the guest. This problem is exacerbated when reading data in large block sizes. In this case, the higher bandwidth contributes to network congestion, and each I/O is comprised of many more data segments, requiring longer recovery times.

When in the planning stages, consider designing your IP storage network with enough capacity to account for the peak usage and lower the risk of congestion. If you experience slow read performance and are unable to alter your network configuration or find ways to guarantee a congestion-free environment, you can experiment with the following workaround.

This workaround involves disabling delayed ACK on your ESXi host through a configuration option.

Configuring Delayed ACK in ESX/ESXi 4.x and ESXi 5.x

To implement this workaround in ESX/ESXi 4.x and ESXi 5.x, use the vSphere Client to disable delayed ACK.

Disabling Delayed ACK in ESX/ESXi 4.x and ESXi 5.x

1. Log in to the vSphere Client and select the host.
2. Navigate to the Configuration tab.
3. Click Storage Adapters.
4. Click the iSCSI vmhba that you want to modify.
5. Click Properties.
6. Modify the delayed ACK setting, using the option that best matches your site's needs:

   Modify the delayed ACK setting on a discovery address (recommended):
   a) On a discovery address, click the Dynamic Discovery tab.
   b) Click the Server Address tab.
   c) Click Settings > Advanced.

   Modify the delayed ACK setting on a specific target:
   a) Static Discovery tab.
   b) Select the target.
   c) Click Settings > Advanced.

   Modify the delayed ACK setting globally:
   a) Select the General tab.
   b) Click Advanced.

7. In the Advanced Settings dialog box, scroll down to the delayed ACK setting.
8. Deselect Inherit From parent.
9. Deselect DelayedAck.
10. Reboot the host.

Re-enabling Delayed ACK in ESX/ESXi 4.x and ESXi 5.x:

1. Log in to the vSphere Client and select the host.
2. Navigate to the Advanced Settings page, as described in the preceding task Disabling Delayed ACK in ESX/ESXi 4.x and ESXi 5.x.
3. Click Inherit From parent > DelayedAck.
4. Reboot the host.

Checking the Current Setting of Delayed ACK in ESX/ESXi 4.x and ESXi 5.x:

1. Log in to the vSphere Client and select the host.
2. Navigate to the Advanced Settings page, as described in the preceding task Disabling Delayed ACK in ESX/ESXi 4.x and ESXi 5.x.
3. Observe the setting for DelayedAck.

   If the DelayedAck setting is checked, this option is enabled. If you perform this check after changing the delayed ACK setting but before you reboot the host, the result shows the new setting rather than the setting currently in effect.

   Notes:
   To disable delayed_ack, run this command from the command line:

```
vmkiscsi-tool -W -a delayed_ack=0 -j vmhbaXX
```

   To enable delayed_ack, run this command:

```
vmkiscsi-tool -W -a delayed_ack=1 -j vmhbaXX
```

   To check this parameter, run this command:

```
vmkiscsi-tool -W vmhbaXX
```

For more information on this workaround, please see VMware Knowledge Base article ID 1002598.

VNX

VNX Pool LUNs are accessible from all iSCSI ports configured on the system. When configured in ALUA mode, the optimal paths to each Pool LUN are provided through paths on the SP that owns that LUN.

> Note: VNX2 Classic LUNs provide active/active access with all SP ports offering an optimized IO path to the LUN.

Network Segmentation (i.e., broadcast domain) is the primary method of controlling initiator access to the VNX iSCSI network portals. This is the preferred method of connectivity.

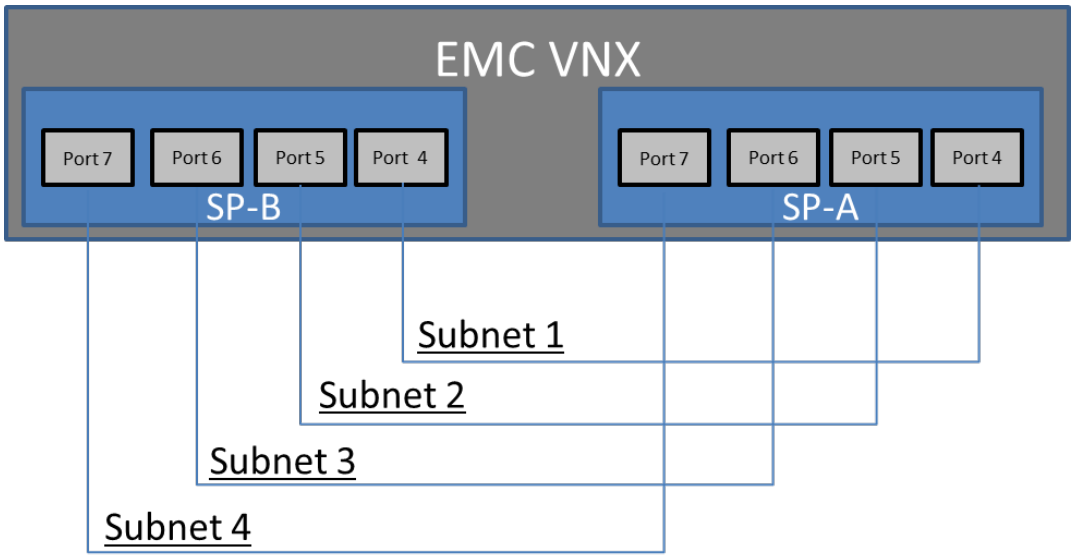

Figure 6: VNX multiple subnet iSCSI target port addressing

Configure separate broadcast domains for each iSCSI initiator and VNX target pair (i.e. SPA4 and SPB4). A broadcast domain can be created in two ways:

1. Configure a network subnet and place all of the nodes in that subnet.
2. Configure a Virtual Local Area Network (VLAN) that is port-based or one that uses packet tagging.

   With port-based VLANs, all switch ports are assigned a particular VLAN ID. Network traffic is limited to ports assigned to that VLAN. Packet tagging uses a field in the Ethernet frame to carry the VLAN ID. Only systems that have tagging configured for that VLAN ID will see those packets.

VNX iSCSI Target Addressing

VNX provides flexible iSCSI portal configuration. Each port can support multiple networks or VLANs, and each port (physical or logical) can be on the same or different subnets.

The traditional approach to VNX iSCSI configuration has been to configure each pair of VNX ports (i.e., SPA4/SPB4) on a different subnet or VLAN. This configuration was required in VNX releases prior to Flare 30 because the VNX iSCSI target did not support multiple logins from the same host port IQN to the same SP port. So if the software initiator established a session using SPA port 0 and then tried to create a second session on SPA port 1, Flare terminated the session on port 0 in order to create the new session on A1.

As of Flare 30, each FE port represents a unique iSCSI target that can be used to support independent iSCSI sessions between hosts and the SP. Flare 30 provides support such that an initiator may establish multiple sessions into a single SP port without terminating an existing session. However, best practice is to use a single initiator-single target configuration.

The following section describes the behavior when running Flare 30 or later with multiple iSCSI sessions per port supported on the VNX.

iSCSI Multipath

Multipath is a connectivity option that can be provided natively with VMware Native Multipathing (NMP) or by third party software like EMC PowerPath. NMP requires multiple paths to the iSCSI storage device.

ESXi offers several Path Selection Policies (PSPs) when configuring the host with a VNX storage system. As of vSphere 5.1 the default storage array type (SATP) configuration for VNX is:

Name	Default PSP	Description
VMW_SATP_ALUA_CX	VMW_PSP_RR	Supports EMC VNX/CX that use the ALUA protocol

Two basic options support this configuration. The appropriate selection for your environment is determined by the architecture of the iSCSI target.

iSCSI targets that support a single-network portal and IP subnet require that all VMkernel ports be configured to use the same subnet. Architectures like VNX Targets that support multiple network portals and subnets can be configured to use different broad cast domains defined by

network addresses. The examples below show one host NIC and multiple SP ports. You can add ports to increase the number of connectivity points for scalability or availability.

1. Single network address space (This configuration is not recommended and should only be used in lab environments that do not have the switch configuration to comply with best practices).

   In Figure 7, all of the ports are on the same subnet (192.168.10.0/24). Based on the Flare 30 changes, this configuration will work, and allow you to configure multiple VMkernel adapters and all VNX iSCSI FE ports using the same subnet.

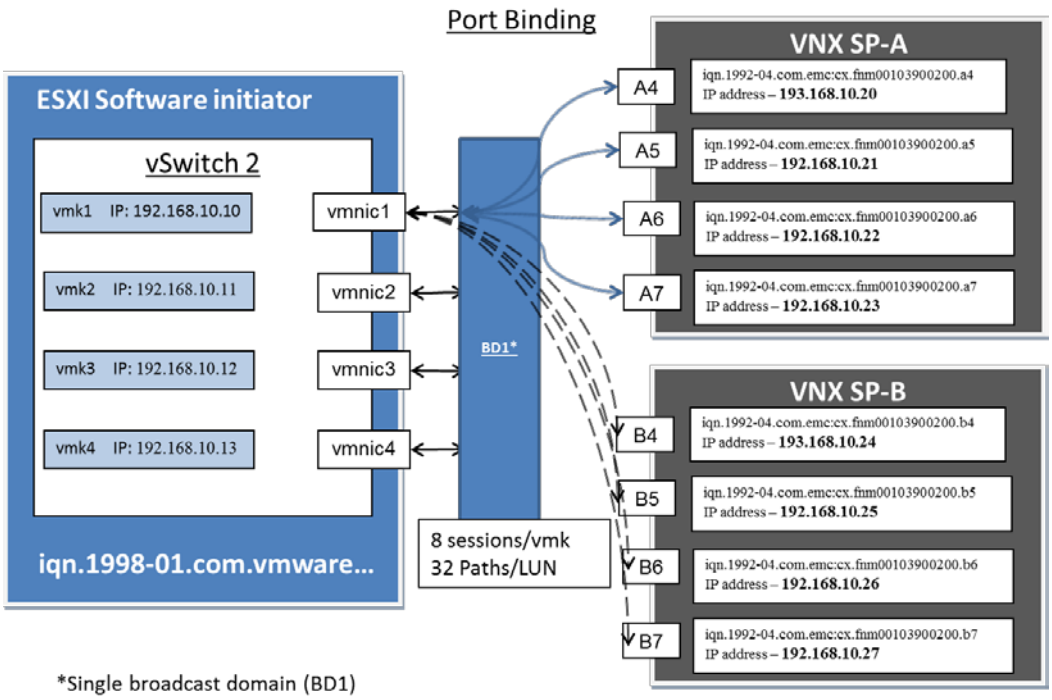

Figure 7: Single network address space Port Binding

Single Subnet	
Pros:	Simple to configure, scalable.
Cons:	Each Initiator establishes a session with all VNX target ports. • 8 iSCSI sessions in this example • 16 sessions when using 4 VMkernel and target ports. • Requires NIC teaming or VLANs to take advantage of all paths. • Has potential to overload the host NIC – four SP ports going to a single host NIC. With one host NIC you should limit the SP ports to one per SP, for a total of 2 paths.

2. The second option is to create multiple subnets or broadcast domains, which is achieved by defining separate network addresses for each VMkernel NIC and VNX SP port pair. Figure 8 shows a modified version of the previous graphic using two subnets (192.168.10 and 192.168.11). The node address is the same for all interfaces.

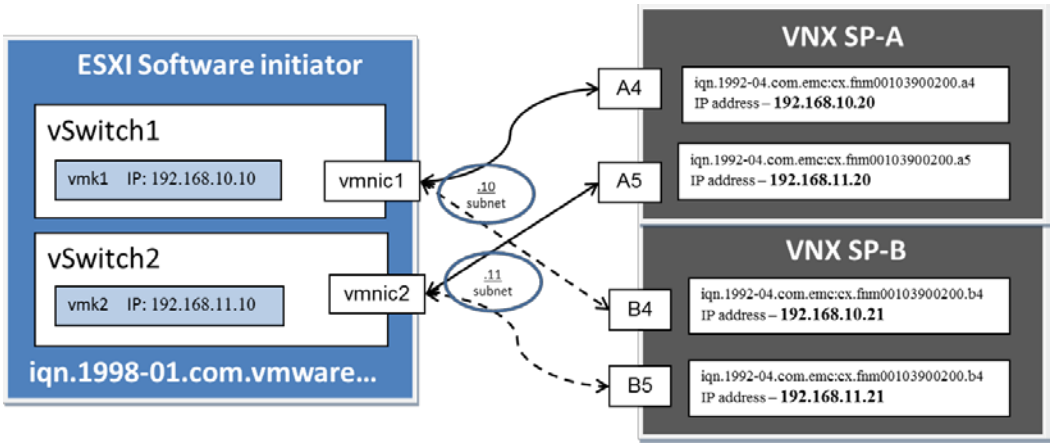

Figure 8: Multi-Subnet Port Configuration (Recommended)

Multi-Subnet Config (VNX Config)	
Pros:	1:1 source to target/LUN session subscription. Performs IO distribution across all paths when configured with multipath or PowerPath.
Cons:	Requires additional network setup. The host can still establish 8 iSCSI sessions if subnets are routable.

ESXi Software Initiator Configuration

The ESXi software initiator must be configured with one or more physical NICs to access the VNX iSCSI port. The network ports are either explicitly assigned when using Port Binding, or implicitly associated with the initiator through the iSCSI target discovery.

Explicit configuration is performed through the UI by selecting the Port Groups associated with the iSCSI VMkernel interfaces.

> Note: This configuration is used when the storage system has a single target that does not support multiple network addresses. It requires that all VMkernel adapters are configured to use the same subnet. It is not a recommended configuration.

MPIO Load Balancing

EMC recommends using the Round Robin multi-pathing policy on the ESXi host to increase performance with the VNX array.

The ESXi default Round Robin policy has an IO operation limit parameter set to 1000, which determines the number of IOs that go down each path before switching to the next path. EMC recommends leaving the parameter at the default value of 1000.

However, changing the Round Robin IO operation limit parameter from the default value of 1000 to 16 or 8 can improve the performance of sequential workloads, in some cases.

For information on changing the multi-pathing policy and Round Robin IOPS setting, please see the following VMware Knowledge Base articles:

- 1011340
- 1017760
- 2000552