EMC ISILON SCALE-OUT STORAGE AND VMWARE VSPHERE 5

Best Practices Guide

Abstract

This guide provides technical information to consider when planning and designing a VMware vSphere 5 virtual data center on a foundation of Isilon Scale-out storage from EMC. It also includes EMC’s recommendations for optimizing performance and availability on an Isilon-based vSphere virtual data center.

April 2012

EMC believes the information in this publication is accurate as of its publication date. The information is subject to change without notice.

The information in this publication is provided "as is." EMC Corporation makes no representations or warranties of any kind with respect to the information in this publication, and specifically disclaims implied warranties of merchantability or fitness for a particular purpose.

Use, copying, and distribution of any EMC software described in this publication requires an applicable software license.

For the most up-to-date listing of EMC product names, see EMC Corporation Trademarks on EMC.com.

VMware and DRS, ESX, ESXi, HA, RDM, SRM, Storage VMotion,VADP, VCB, vCenter, Virtual Center, VMDK, VMFS, vMotion, and vSphere are the registered trademarks of VMware, Inc., in the United States and/or other jurisdictions. All other trademarks used herein are the property of their respective owners.

Part Number H10522.1

Table of Contents

About this Guide

Intended Audience

This Guide will provide technical information to consider when designing a VMware® vSphere® infrastructure on a foundation of Isilon® Scale-out storage. The Guide is written for experienced system administrators who are familiar with virtual machine technology and network storage administration.

Assumptions

This Best Practices Guide assumes the reader has:

- An understanding of the iSCSI and NFS storage system protocols
- A working knowledge of Isilon scale-out storage from EMC® and the OneFS® operating system
- A working knowledge of VMware vSphere 5 virtual infrastructure components, including ESXi® and Virtual Center®

Additional Resources

Before reading this Guide, EMC recommends reading the following additional resources.

VMware vSphere documentation:

VMware Performance Best Practices for VMware vSphere 5.0

VMware vSphere Installation and Setup Guide

VMware vSphere 5 Storage Guide

VMware vSphere Networking Guide

EMC Isilon Storage Configuration documentation:

Isilon Scale-out Storage for Virtualization

Solution Brief: EMC Isilon Virtualization at Scale

Insight from Isilon: Understanding the Role Storage Plays in Virtual Environments

EMC Isilon vSphere 5 documentation

EMC Isilon Scale-Out Storage and VMware vSphere: Sizing Guide

EMC Isilon Scale-Out Storage and VMware vSphere 5: Deployment Guide

Revision Summary

Date	Version	Author	Change Summary
July 2010	1.0	Shai Harmelin	Initial Document
December 2010	1.1	Nick Trimbee	Updated for vSphere 4.1
December 2011	2.0	Ian Breitner James Walkenhorst	Updated for vSphere 5.0 Removed deployment-related content for publication in a separate Deployment Guide
April 2012	2.01	James Walkenhorst	Added hyperlinks for relevant documentation Updated formatting for style consistency Added max VMDK file size to comparison table

Introduction

Server virtualization has become a standard in major enterprises, simplifying deployments and reducing management costs for large-scale server environments. While virtualization provides a solution for increasing server utilization while reducing operational expenses, the challenges of using traditional SAN or NAS storage architectures to store large numbers of virtual machines can adversely impact the benefits of server virtualization.

This document outlines the principles and concepts for optimizing availability and performance. Information on specific implementation and configuration steps for deploying a VMware vSphere 5 infrastructure on an EMC Isilon storage cluster is provided in the *EMC Isilon Scale-Out Storage and VMware vSphere 5 Deployment Guide*, which serves as a companion document to this Guide.

Although this Guide is intended to build on, and be consistent with, VMware's *Best Practices Guide for vSphere 5.0*, it is not intended to replace the information contained in that document. Additionally, while the Guide attempts to enumerate general best practices as recommended by EMC and VMware, different conditions with different considerations may apply in every environment.

EMC Isilon Scale-Out Storage for Server Virtualization

The Isilon scale-out storage platform from EMC, powered by the OneFS operating system, was designed from the ground up to reduce storage complexity. In a dynamic, virtualized environment, Isilon eliminates the management complexity inherent in traditional SAN and NAS architectures, while increasing application availability and protection.

Simplify storage management

Consolidate and replace tens, hundreds, or thousands of volumes with a single namespace, single volume, and single file system to host thousands of virtual machines. With Isilon you can also consolidate virtualized and non-virtualized application storage, using standard networking and file sharing protocols.

Quickly adapt to change

In an ever-changing environment where virtual machines are added and expanded daily, Isilon can scale both capacity and performance dynamically to meet growing demand. When a storage node is added to an Isilon cluster, the additional capacity, performance and connectivity are immediately available and shared across all datastores automatically.

Increase efficiency and productivity

Isilon can achieve over 80% storage utilization without degrading performance. The globally-coherent cache in OneFS, combined with the AutoBalance® feature, eliminates complex storage management tasks such as chasing hot-spots (LUN thrashing) and manually rebalancing virtual machines across volumes, allowing administrators to focus on achieving the greatest gains from a virtualized server environment.

Increase protection and disaster recovery in your virtualized environment

The FlexProtect® feature in OneFS, along with a suite of integrated software applications, allows non-disruptive protection-policy changes, unlimited snapshot and replication schedules at the individual VM level, across any number of datastores.

OneFS Operating System

The cornerstone of Isilon's fully distributed scale-out architecture is the OneFS Operating System. OneFS creates a single intelligent, fully-symmetrical file system that spans all nodes within a cluster. In addition to enabling industry-leading scalability of performance and capacity, OneFS combines mission-critical reliability and high availability with state-of-the-art data protection.

Isilon's OneFS operating system stripes both data and corresponding metadata across all nodes in a cluster to create a single shared pool of storage. This approach is improves on the traditional method of striping data across a fixed set of disks, i.e. RAID groups. Traditional SAN and NAS architectures use a centralized server to manage a disk array, creating dependencies and multiple points of failure within a storage system.

In an Isilon clustered storage system, each node is a fully-functioning peer, so any node can handle storage Input/Output (I/O) requests. OneFS provides each node with knowledge of the entire filesystem layout, giving users access to all content in one unified namespace with no volumes to manage, no volume-size limits, no downtime required for reconfiguration or expansion of storage, and without having to manage multiple network drives.

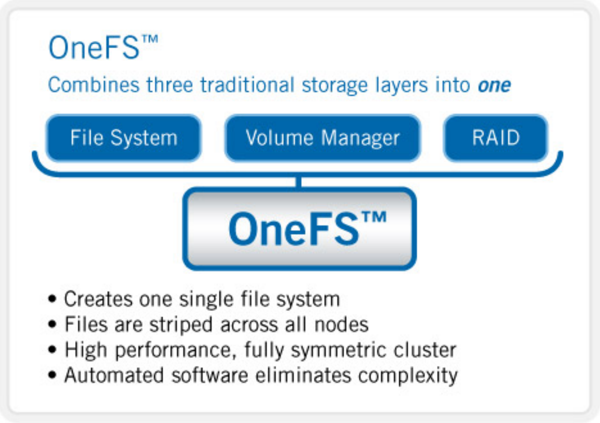
OneFS™
Combines three traditional storage layers into one
File System
Volume Manager
RAID
OneFS™
• Creates one single file system
• Files are striped across all nodes
• High performance, fully symmetric cluster
• Automated software eliminates complexity

Figure 1: Consolidating storage layers with OneFS

End-to-End Storage Solution for Server Virtualization

Isilon Scale-out storage is designed specifically for file-based and virtualized workflows, producing a highly scalable and redundant enterprise storage solution that is simple to manage and simple to scale:

- Integration with vSphere Virtual Center (vCenter) to enable VMware High Availability® (HA), Dynamic Resource Scheduling® (DRS), vMotion®, Storage vMotion®, and VMware Storage API for Data Protection® (VADP)—formerly known as VMware Consolidated Backup / VCB) and Site Recovery Manager (SRM).

- Support is included for both iSCSI and NFS-based datastores for ease of use and to maximize available capacity on your Isilon cluster.
- The Isilon SyncIQ® and SnapshotIQ® applications, combined with FlexProtect in OneFS, provide flexible, high-performance VM data-protection capabilities.
- SmartConnect® and AutoBalance in OneFS ensure continual efficient usage of storage resources.

Isilon Is VMware-Ready

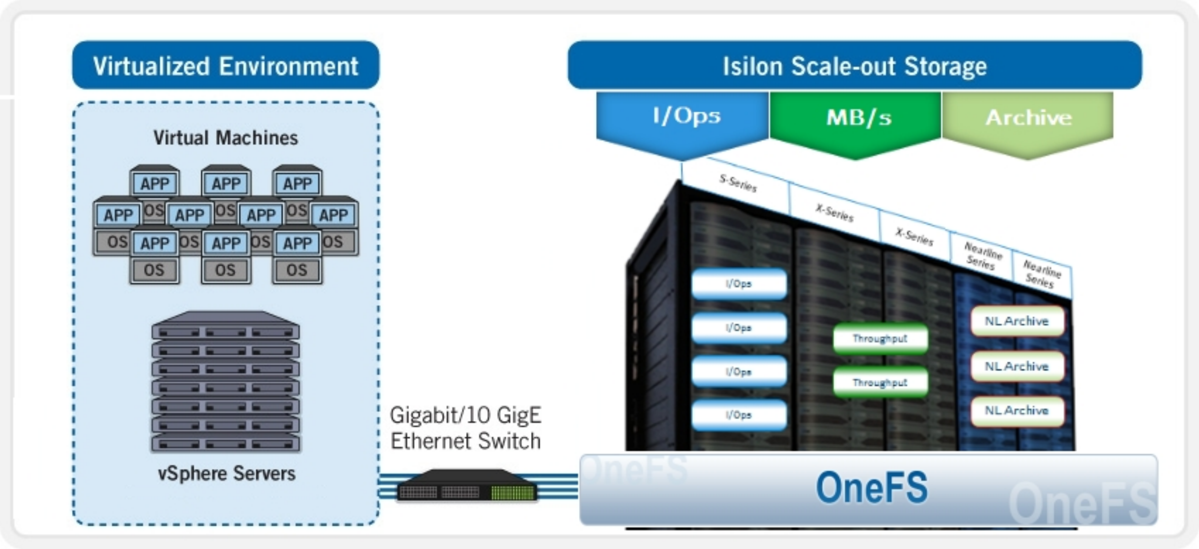

Figure 2: Virtual environment using Isilon scale-out storage

Isilon from EMC is VMware Ready-certified for VMware ESX® 3.x, vSphere 4.x and vSphere 5. VMware certification is covered across all Isilon product lines, including the X-Series, S-Series, and NL-Series product families. This ensures compatibility with VMware products, and certifies that Isilon is ready for deployment in customer environments.

vSphere Support for Network Storage

A vSphere virtual machine (VM) uses a virtual hard disk to store the operating system, program files, and other files associated with its applications and user data. Each virtual disk that a VM can access resides in either a Virtual Machine File System® (VMFS) datastore, a Network File System (NFS) datastore, or on a Raw-Device-Map (RDM) disk. From the perspective of the VM, each virtual disk appears as if it were a physical disk, connected to a SCSI controller. Regardless of whether the underlying physical storage is accessed through parallel SCSI, iSCSI, network, or Fibre Channel adapters on the host, the actual manner of access is transparent to the guest operating system and the applications running in it.

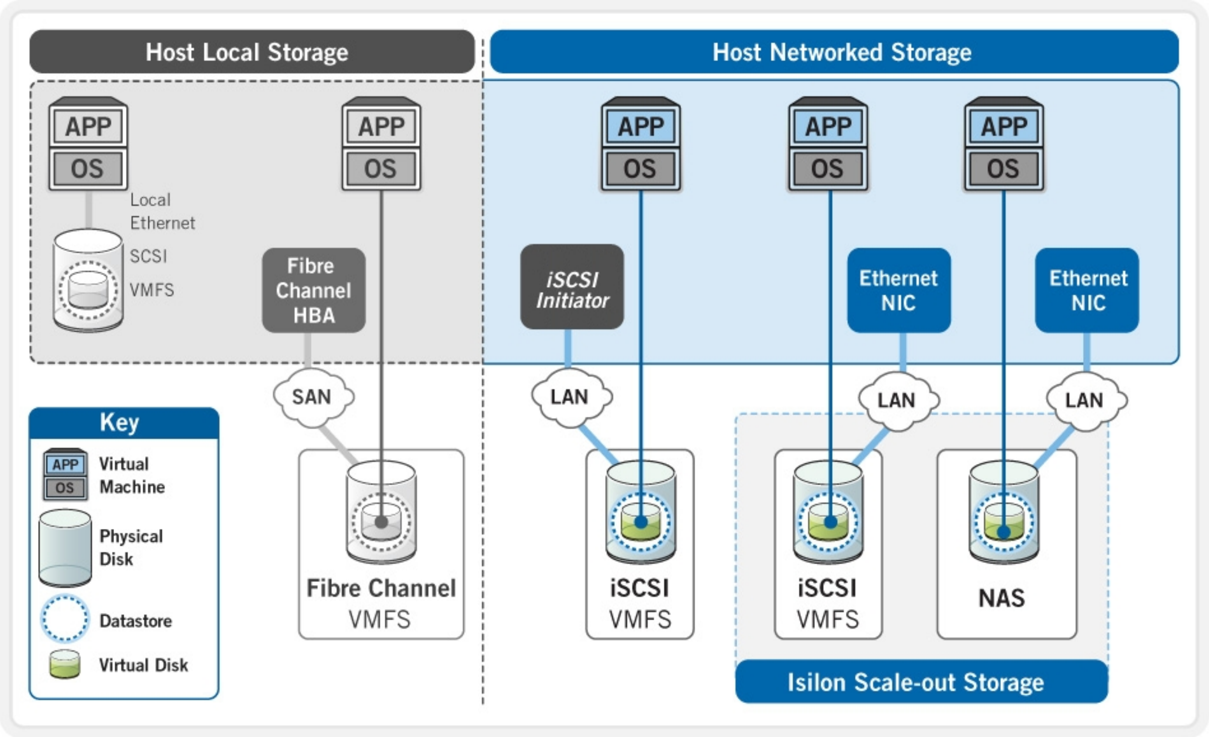

Figure 3: VMware-supported storage types

VMware vSphere supports the following types of storage:

- **Local Storage:** Stores virtual machine files on internal or external storage disks or arrays attached to the host through a direct connection.
- **Networked Storage:** Stores virtual machine files on external shared storage systems located outside of the vSphere host. The host communicates with the networked devices through a high-speed network.

This best-practices document will focus on Isilon Ethernet-based network storage, specifically NFS datastores and VMFS datastores over iSCSI.

Network Attached Storage for vSphere

Fundamentally, NAS stores large VM datastores on a Network File System (an industry-standard file-sharing protocol) export rather than block-based VMFS datastores. The storage system is presented to each vSphere host through an *NFS mount*, and vSphere hosts access VM data using the NFS protocol. Among the advantages of NAS-based datastores:

- **Rapid and simple storage provisioning:** Instead of managing LUNs for individual virtual machines, all VMDK files may be stored on a common file export.
- **Higher storage utilization rates:** VMware disk files (VMDK files) are thin-provisioned by default when using NAS-based datastores, and filesystem storage behind an NFS export is not allocated up front. Additionally, NFS file systems are typically much larger than individual LUNs, thereby allowing more virtual

machines to share a larger storage pool and reduce the overhead of unused storage.

- **Simplified backup scenarios:** All VM files may be backed up through a single, central mount point. Individual virtual machines can be protected and backed up simply by backing up the filesystem directories of those virtual machines.

iSCSI Storage Area Networking (SAN) for vSphere

The Internet Small Computer Systems Interface (iSCSI) protocol packages traditional SCSI commands into IP packets and transmits them across an Ethernet network. In this manner, industry-standard network interface cards (NICs) act as host bus adapter (HBA) devices, without the level of complexity or cost often associated with Fibre-Channel storage area networking.

Advantages to iSCSI-based storage include:

- IT administrators can quickly transition to a virtual environment using block-level storage infrastructure without changes to their applications.
- vSphere can leverage iSCSI multipathing to increase aggregate bandwidth to, and provide network redundancy for, a single iSCSI Logical Unit (LUN). This improves performance and availability to support the aggregate I/O stream from a large mix of virtual machines across multiple vSphere hosts.
- For specific use cases, vSphere can provide VMs direct access to iSCSI LUNs, including: physical-to-virtual-machine host-based clustering; Microsoft Cluster Server (MSCS); and/or applications that need direct access to a block storage device. The feature that enables direct LUN access to a VM is referred to as Raw Device Mapping (RDM®) over iSCSI. In this design, the vSphere hypervisor acts as a connection proxy between the VM and the storage array.
- A guest OS can also use an internal iSCSI initiator to gain exclusive access to an iSCSI LUN. In this mode, only higher-level networking calls are managed by the virtualization layer while SCSI- and/or iSCSI-level commands are handled by the guest OS.
- The ability to clone LUNs either as full clones, shadow clones, or read-only-snapshot clones.

While each storage protocol offers its own benefits and limitations, the choice of using iSCSI- or NFS-based datastores often comes down to the infrastructure available and what types of access and administration an organization is already familiar with.

Note: Because NFS and iSCSI are each optimized for different functionality and performance requirements, EMC recommends evaluating your environment’s specific needs when choosing between deploying NFS- or iSCSI-based datastores in your vSphere environment. Absent a sufficiently compelling reason to use iSCSI-based datastores, EMC recommends using NFS-based storage for both management simplicity and optimal performance.

Isilon Scale-Out Storage Protocol Features for vSphere

Table 1 and Table 2 highlight the features and capabilities of each protocol supported by Isilon scale-out storage.

	VMFS	vMotion	Storage vMotion	DRS	HA	Boot VM	Boot ESXi	RDM	MSCS	Multi-Path	VADP	SRM
iSCSI*	Yes	Yes	Yes	Yes	Yes	Yes	No*	Yes	Yes	MPIO **	Yes	Yes
NFS	N/A	Yes	Yes	Yes	Yes	Yes	No	No	No	LAG	Yes	Yes

Table 1: VMware supported features per storage protocol

Capability	iSCSI[†]	NFS
File System	VMFS or RDM	OneFS
Maximum Number of vSphere Datastores	256	256
Max vSphere Datastore Size	60TB‡	15PB
Max LUN/File System Size	2TB/60TB‡	15PB
Max VMDK File Size	2TB	2TB
Recommended Number of VMs per LUN | File System	20/LUN	Thousands
Network Bandwidth	1Gb/10Gb	1Gb/10Gb

Table 2: VMware storage capabilities

General Guidelines for vSphere Datastores on Isilon Storage

To ensure quality of service and the ability to survive network failures, the same design thinking that has traditionally been applied to Fibre Channel infrastructure—in which traffic to the storage array is isolated and redundant—should also be applied to Ethernet/IP storage.

Some important considerations EMC recommends for both iSCSI and NFS datastore storage networking include:

- Separate storage and VM network traffic to different physical network ports. This separation will prevent a busy LAN from overwhelming storage traffic. If traditional vSwitches are used, EMC also recommends having separate vSwitches for VM network, vMotion, and storage area network (VMkernel) traffic.

[*] Isilon currently supports vSphere Software iSCSI initiators only.
[**] iSCSI MPIO was introduced in vSphere 4.0. Before that, LAG was used for multipathing.
[†] Isilon currently supports vSphere Software iSCSI initiators only.
[‡] VMFS3 can only see LUNs up to 2TB in size. For iSCSI-based datastores larger than 2TB, VMFS5 is required

- Keep the number of hops between vSphere hosts and the storage nodes to a minimum. If your network topology allows it, EMC recommends connecting hosts and storage to the same physical switch(es) on the same subnet(s).
- Consider using switches that support cross-stack EtherChannel or virtual port channel trunking, where interfaces on different physical switches are combined into an 802.3ad Link-Aggregate Group (LAG) for network and switch redundancy.
- Because Ethernet switches from different manufacturers vary in their internal architecture, choose a production-quality switch, sized with an appropriate configuration of port buffers, backplane bandwidth, and other internal resources as appropriate for mission-critical, low-latency, network-intensive workloads like VMware datastores (whether using iSCSI or NFS).
- While standard Gigabit infrastructure is sufficient to meet the performance needs of most vSphere network storage environments, EMC recommends 10GbE infrastructure (if your environment will support it) for optimal performance and to reduce the number of network ports needed in your environment.
- For Gigabit infrastructure, use Cat6 cables rather than Cat5/5e.
- While EMC and VMware both fully support the use of Jumbo frames to minimize network latency and improve overall throughput rates, the effectiveness of this practice largely depends on your particular environment. Your infrastructure configuration may fully support the end-to-end use of Jumbo frames (including vSphere host hardware, switches and routers), but overall latency and throughput may in some cases be better using standard-size frames. EMC recommends testing both configurations if applicable to your environment, and then using whichever configuration delivers the best overall results.
- For vSphere hosts that have fewer NICs (such as blade servers), VLANs can group common IP traffic onto separate VLANs for optimal performance and improved security. VMware recommends separating Management Network access and the Virtual Machine Network from the VMkernel Network for IP storage and vMotion traffic.
- EMC supports and recommends segmenting NFS connections to specific nodes and network interfaces into a dedicated VLAN. If using multiple VLANs over the same interface(s), make sure sufficient throughput can be provided for all traffic.

vSphere provides network-storage high availability using multiple network interfaces on the local vSphere host. The actual configuration of those interfaces varies based on the storage protocol used and the capabilities of the switches connecting the vSphere hosts to the Isilon storage nodes. Those configuration options are discussed in more detail within the iSCSI and NFS multipathing network configuration sections below.

Isilon Cluster Configuration for NFS Datastores

The Isilon storage paradigm is a “Scale-out” approach to NAS storage, and offers significant reductions in management overhead by managing the cluster through a single management point—either a browser-based WebUI, or via command-line interface. With a single file system and a single NFS export, Isilon scale-out storage can accommodate any number of vSphere datastores and virtual machines. This section highlights how using NFS datastores can greatly simplify vSphere storage management.

Provisioning NFS-Based Datastores

Once an Isilon cluster is built (with a minimum of three nodes), all capacity is fully, thinly, and immediately provisioned and available for NFS access from vSphere hosts. vSphere hosts can access the shared storage from any assigned IP address on the cluster and the default root /ifs directory—or any other subdirectory—can be mounted. Because each datastore can access the OneFS file system across different network paths, a single NFS export can support any number of datastores from the vSphere cluster.

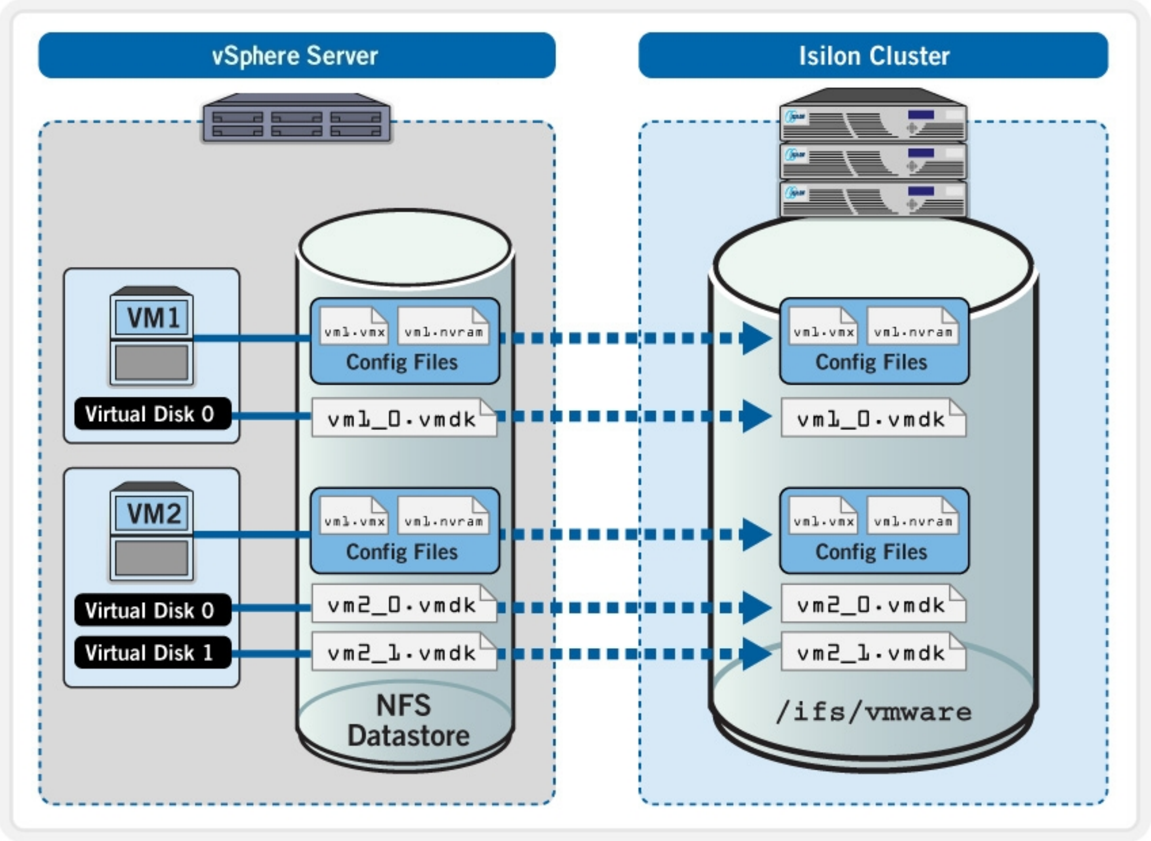

Figure 4: vSphere datastore on Isilon cluster

OneFS NFS Exports

Before creating a vSphere NFS datastore in OneFS, the directory to which the datastore will be mounted must already have been created in the OneFS File System Explorer. Additionally, root-level access to the cluster over NFS is by default prevented by OneFS, which maps the 'root' user to the user 'nobody'. If the exported directory is created by 'root', and its ownership isn't changed to 'nobody' before mounting the export from vSphere, then the vSphere host(s) will not be able to write to the directory.

To avoid complications from this issue, EMC recommends creating a specific administrative account in OneFS for creating and managing vSphere NFS datastores, rather than the default 'root' user.

Isilon Network Configuration Best Practices for vSphere NFS Datastores

Since iSCSI and NFS protocols both work over Ethernet/IP infrastructure, they share some common design guidelines and best practices. Those best practices were described in General Networking Guidelines for vSphere Datastores on Isilon Storage above.

However, NFS and iSCSI have different mechanisms when using multiple network paths to achieve network redundancy and increase aggregate network throughput.

Maximizing performance, flexibility and availability for NFS datastores

1. Consider assigning at least one dynamic IP for each member node interface in a dynamic IP pool. The more unique IP addresses that can be mounted by vSphere to an NFS datastore—and the more NFS datastores are created—the more TCP connections a vSphere host can leverage to balance VM I/O streams.
2. If your network switches support it, Cisco cross-stack EtherChannel link aggregation can be combined with OneFS Dynamic IPs to provide both node-to-interface redundancy and switch redundancy.
3. You can choose to assign a single IP address to an aggregated pair of network interfaces (using LAG 802.3d link aggregation), or to assign both interfaces as independent members of a SmartConnect dynamic IP pool. Either option enables multiple network interfaces on each node to be used for redundancy, and/or to aggregate I/O.

Note: High-availability using dynamic IP addresses is described later in this chapter. Nodes that have four interfaces can be first link-aggregated by pairs (and each linked pair can be assigned its own static IP), or can be members of a dynamic IP pool.

Optimal Configuration for NFS datastores

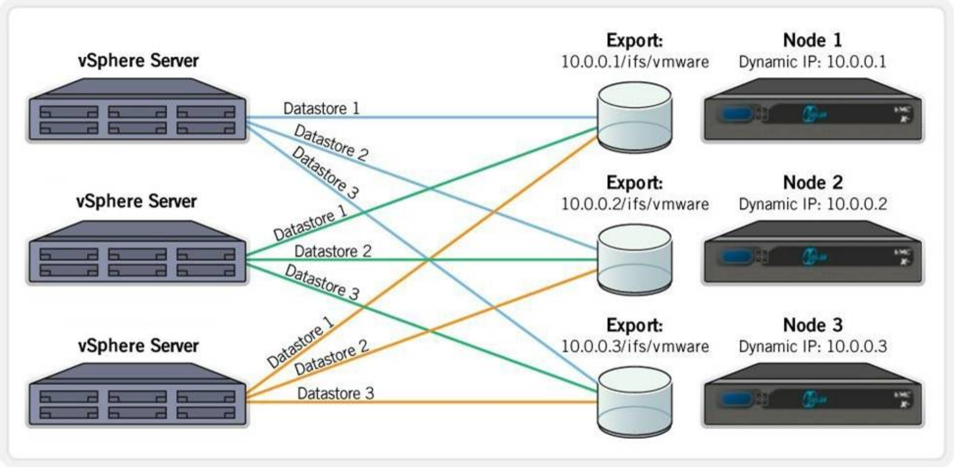

Figure 5: Multiple datastores on a single OneFS NFS export

As Figure 6 illustrates, the best design is a mesh-connectivity topology, in which every vSphere host is connected to every IP address on a cluster-dynamic IP pool (limited only by the maximum amount of NFS datastores as described in Table 2 above). Connecting "everything to everything" enables the following:

- By default, all vSphere hosts are connected to all datastores, enabling vMotion between all vSphere hosts.
- Virtual machines can be created on any datastore to balance the I/O load between vSphere servers and the Isilon cluster.
- Virtual machines can be easily moved between datastores to eliminate "hot spots" without having to migrate VM data.

Note: Further configuration for resiliency and performance is available by using SmartConnect, which is described in further detail elsewhere in this Guide.

Basic Network Configuration for NFS Datastores

When initially configuring a cluster, the first external network interface (typically External-1) is set up by the configuration wizard process. In order for this process to complete successfully, the following information is required: netmask, IP address range, default gateway. You can also provide a Domain Name Server (DNS) list (optional) and a DNS search list (optional).

Once this information has been provided, the following occurs:

- A default external subnet is created, named ***subnet0***, with the specified netmask.
- A default IP address pool is created, named ***pool0***, with the specified IP address range
- The first node in the cluster is assigned an IP from the address pool and set with the specified default gateway and DNS settings.

- A default network provisioning rule is created, named *rule0*, which automatically assigns the first external interface for all newly added nodes to *pool0*.

After the initial cluster configuration is committed, additional subnets and pools can be added or edited.

VLAN Tagging

Virtual LANs (VLANs) are used to logically group together network endpoints and to partition network traffic, e.g., for security purposes, or to separate IP storage traffic from other network traffic.

VLANs are tagged with a unique identifier to segregate traffic. SmartConnect supports VLAN tagging for use in external networks using VLANs. In SmartConnect, VLANs are configured at the subnet level.

High-Availability Using Link Aggregation

Isilon OneFS supports the use of redundant NICs to provide Layer-2 failover. OneFS link aggregation supports the IEEE 802.3ad static LAG protocol, and works with switches and clients that support this protocol.

Note: EMC does not currently support dynamic LACP on Isilon.

Link Aggregation Switch Support

Isilon network link aggregation requires 802.3ad LAG support and proper configuration on the switch.

Note: Cisco switches support using this EtherChannel feature. If you are using Cisco switches in the storage fabric, EMC strongly recommends configuring cross-stack EtherChannel for protection against switch failures as well as NIC failures.

Link Aggregation Cluster Configuration

Link aggregation provides protection against NIC failures but does not increase performance. If configured properly on the vSphere host, both NICs in the aggregate can be used for vSphere datastore I/O, but the two 1Gb channels are not 'bonded' into a single 2Gb link. Each NIC is serving a separate IP connection.

High Availability Using SmartConnect Dynamic IP Address Pools

SmartConnect Advanced implements NFS failover by assigning one or more dynamic IP addresses to each node's member interface in the cluster from a configured range of addresses. If a single interface or an entire node experiences a failure, SmartConnect reassigns the dynamic IP addresses to the remaining pool member interfaces. All datastore I/O is immediately re-routed to the newly assigned member interface and VM virtual disk access continues without interruption.

Note: SmartConnect requires a separate license purchase for the Isilon cluster.

When the failed interface or the failed node is brought back online, the dynamic IP addresses in the pool are redistributed across the updated list of available pool members. This failback mechanism can occur automatically or manually, and happens either way without datastore traffic interruption.

In the event that SmartConnect is not available on your cluster, EMC recommends setting VMware NFS datastores with **dynamic** IP addresses assigned to the Isilon storage cluster, not the fixed IP addresses bound to the individual cluster interfaces, because static IPs on member-node interfaces will not be reassigned in the event of an interface or node failure.

More detailed information on SmartConnect best practices and configuration can be found in the Links section at the bottom of this Guide.

Figure 7 below illustrates NFS datastore redundancy using dynamic IPs for NFS failover.

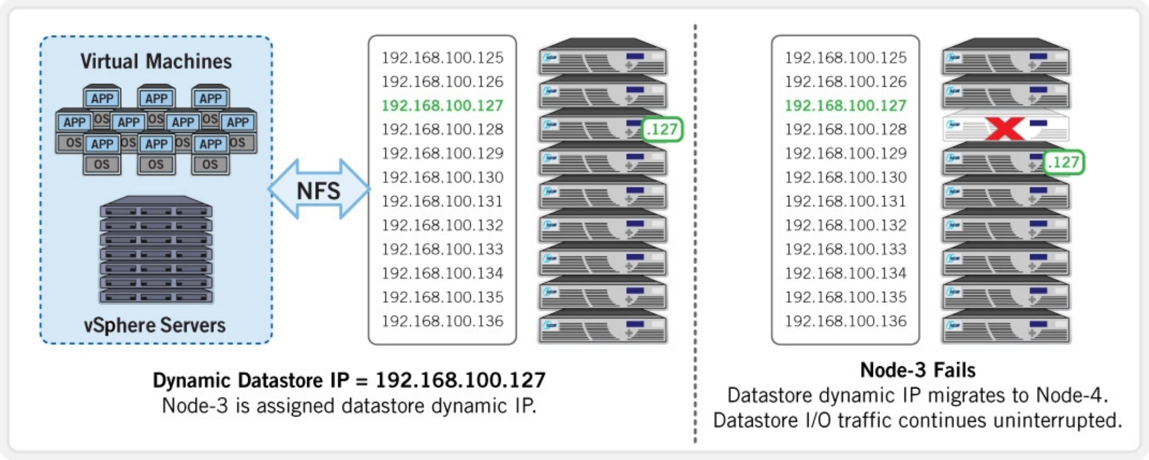

Figure 6: Example of NFS failover process

Note: This method only provides for high availability in the event of a failure between a cluster node interface and the vSphere host(s). It does not enable automatic load-balancing across all available interfaces on the Isilon cluster.

High Availability and Load-Balancing Using Smart-Connect Advanced with DNS

When combined with new features in vSphere 5, SmartConnect Advanced enables a more sophisticated method of NFS failover that now includes automatic Load Balancing as well. By coupling intelligent NFS-client load-balancing and failover capabilities with your existing DNS infrastructure, SmartConnect Advanced automatically balances NFS connections across all available network interfaces on the Isilon cluster. The performance of both the Isilon cluster and the vSphere hosts can be enhanced when connections are more evenly distributed using this method. In addition to dynamic load balancing, SmartConnect also provides NFS datastore failover and failback capabilities to your vSphere hosts.

SmartConnect Configuration Options

SmartConnect provides the flexibility to implement different connection balancing strategies to meet various needs of the enterprise. The two enabling options include:

- **Zoning Strategies:** SmartConnect enables the ability to create zones by node and interface groupings. Zoning based on specific node interfaces gives you the flexibility to apply connection-balancing policies across different workloads, so you have the flexibility to dedicate more Isilon cluster resources to your vSphere host(s) than to other NFS clients, if necessary.
- **Inclusion/Exclusion:** Another feature of SmartConnect is the ability to decide which nodes in the Isilon cluster should participate in a specific connection balancing configuration strategy. In other words, any specific node(s) can be selected to be excluded or included from any or all balancing schemes for the Isilon cluster.

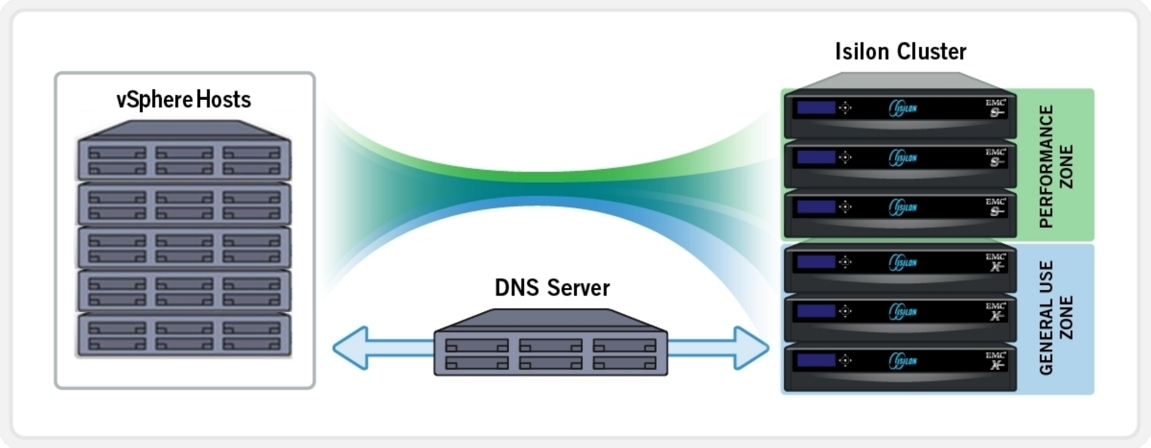

Figure 7: vSphere NFS connectivity using SmartConnect

vSphere Network Redundancy Options for NFS if SmartConnect Is Not Used

For NFS datastores, redundancy is achieved using the link-aggregation capabilities provided by the switch and end points (vSphere hosts and Isilon cluster nodes):

- If the switches support “cross-stack EtherChannel” (or “virtual port channel” trunking), in which two ports from two different switches can be aggregated to form a single network trunk (within a single subnet), vSphere servers can be configured using one VMkernel port with multiple active network interfaces in a team. To use both interfaces concurrently, create multiple NFS datastore IPs on the same subnet to the target Isilon cluster and set vSphere to use source-destination IP load balancing.
- If the switches only support EtherChannel (or “virtual port channel” trunking) within a single switch, and two switches (with two separate subnets)are used for redundancy, then two or more separate VMkernel ports are required in separate vSwitches. Each VMkernel port has one active and one (or more) passive network

interface(s). Each VMkernel port is on a separate subnet, one for each of the redundant switches.

Dynamic NFS Failover and Failback with Performance Rebalance

Traditional storage systems with two-way failover typically sustain at least 50% degradation in performance when a storage controller fails as all clients must fail over to the remaining controller. With SmartConnect, in the event of a failover, clients are evenly redistributed across all remaining nodes in the cluster, minimizing performance impact.

vSphere Configuration for NFS Datastores

The NFS client built into vSphere 5 enables access to the NFS server's file system to store VM images, ISO images, and templates. To accomplish this, use the vSphere Client to configure NFS-mounted volumes as datastores. Configured NFS datastores appear in the vSphere client and can be used to store virtual disk files in the same manner as VMFS-based datastores.

Best Practices for vSphere Network Configuration

EMC recommends the following vSphere network configuration for NFS datastores:

1. Whether using vSphere standard or distributed switching, reserve at least one dedicated network interface card (NIC) for network storage traffic. This will ensure optimal storage I/O performance, as well as isolate potential problems caused by other network traffic.
2. For network redundancy on the vSphere host, EMC recommends using multiple network interfaces for NFS storage. The use of multiple interfaces can also increase the aggregate bandwidth through the VMkernel NFS storage stack (as long as each connected datastore is mapped using a different IP address). The VMkernel will use multiple TCP/IP connections across multiple network interfaces to provide parallel Virtual Machine I/O access.
3. If the vSphere load-balancing option is set to **Route based on IP hash**, the switch-load-balancing method should also be set to IP hash for aggregated port groups on both the storage and the vSphere VMkernel switch ports.

Optimizing OneFS for NFS Datastores with I/O Intensive Virtual Machines

1. By default, OneFS protects calculates and writes parity blocks for all filesystem data, including NFS directories. Depending on their overall I/O profile, some virtual machines whose workload includes small (e.g. less than 32KB per transaction), random I/O operations may exhibit performance degradation when using the default protection. These virtual machines may benefit from having their VM data files configured to use a mirrored (2x) data layout by decreasing write overhead. This setting can be applied on a per-VM-directory or per-file basis, leaving other VM directories and files to continue using the more space-efficient parity-protection layout.

2. OneFS ‘streaming’ mode on virtual machines with high read and write I/O requirements should be enabled. This setting applies to random and sequential I/O workloads, and can be set in conjunction with mirroring layout.
3. Disable OneFS read caching (prefetch) if many virtual machines require a high ratio of small random read operations. Disabling **read prefetch** instructs OneFS to avoid prefetching adjacent file blocks, and will eliminate prefetch latency overhead.

Increasing Performance with Multiple NFS Datastores

Every NFS datastore mounted by a vSphere host uses two TCP/IP connections—one for NFS control information and the other for NFS data flow. Accordingly, the vast majority of the traffic to a single NFS datastore will use a single TCP/IP connection. As a result, the maximum throughput achievable for a single datastore (regardless of link aggregation) will be bound to a single link for traffic to that datastore.

All virtual machines assigned to a single datastore on a vSphere host share this NFS mount and TCP/IP connection. If using 1Gb-Ethernet connections, the effective maximum aggregate bandwidth of all virtual machines within a single datastore is 80MB/s reads and 80MB/s writes (or 160MB/s combined). This can be translated to a maximum of 4,000 IOPs (32K per I/O).

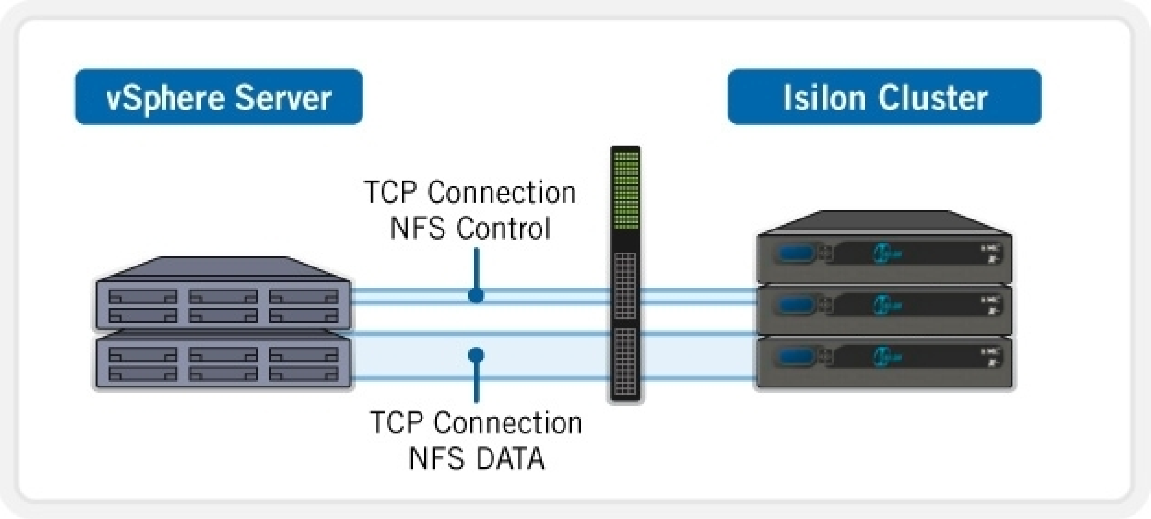

Figure 8: Per-NFS datastore TCP connections

Link aggregation can increase aggregate throughput to the storage by using multiple network interfaces via a load-distribution mechanism called *path selection*. Path selection works on pairs of source-to-destination IP addresses, so vSphere will route traffic from multiple network interfaces to multiple datastores, with each interface pointing to a different IP address on the Isilon cluster.

To achieve higher aggregate I/O, EMC recommends creating multiple datastores and mounting each datastore using a separate IP address and network interface on the Isilon cluster. While this increases the number of datastores that need to be managed, the fact is that any VM can belong to any datastore because they all share a single NFS export and file system location within OneFS. When a virtual machine is reassigned to a different datastore, its data does not need to be moved on the Isilon cluster. A VM can simply be powered down, unregistered from one datastore, and added to the inventory of another datastore.

Isilon dynamic IPs further simplify configuration of multiple datastores because it is not necessary to know the specific node and network interface through which a particular datastore is mounted. Since the Isilon cluster's dynamic IP addresses can be rebalanced at any given moment across all available member nodes on the Isilon cluster, NFS datastore connections can also be redistributed accordingly. When a datastore connection changes from one node to another (via dynamic IP rebalancing) all disk I/O of the virtual machines assigned to that datastore will be rerouted to the new node.

By default, the maximum number of NFS datastores in a single vSphere host is limited to eight, but in vSphere 5 this limit can be easily extended to 256 NFS datastores. EMC recommends increasing the number of datastores, up to the number of network interfaces available across all storage nodes in the Isilon cluster, to provide better load distribution across all nodes.

Note: If increasing the maximum number of NFS datastores in the vSphere cluster, VMware recommends adjusting the TCP/IP Heap Size value on each vSphere host to scale with the number of planned datastores. This procedure requires a reboot of the host to take effect.

Isilon Cluster Configuration for iSCSI Datastores

This section reviews best practices directly related to the management of iSCSI Logical Units (LUNs) with Isilon OneFS and integration with VMware vSphere.

Isilon iSCSI LUNs are constructed as files that reside within OneFS. Each iSCSI LUN is composed of eight extent files within a directory. These extent files can be laid out and protected using OneFS policies, like any other files on OneFS, for optimal protection and performance capabilities. By default, LUN directories are created under the iSCSI target to which the LUN is assigned, although LUNs may be moved or placed anywhere within the directory hierarchy for convenience: e.g., to enforce a single set of SmartQuotas or to aid in SnapshotIQ and SyncIQ replication.

iSCSI initiators can access LUNs from their respective targets through any of the Isilon cluster nodes, providing a high level of performance through aggregate access, and high reliability against both disk and node failures.

Note: A OneFS iSCSI license is required to manage and use iSCSI on Isilon clusters.

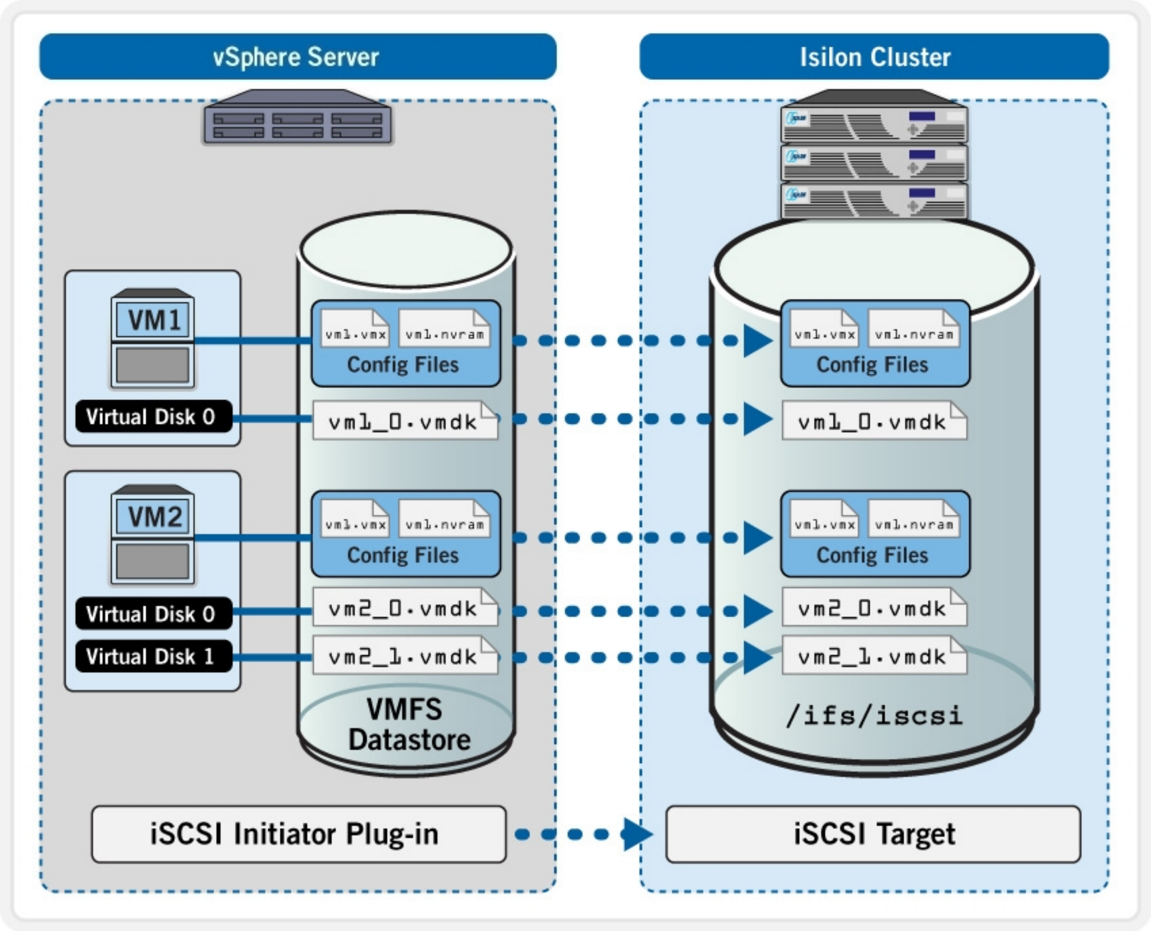

Figure 9: vSphere iSCSI VMFS datastore on Isilon cluster

The Isilon iSCSI implementation features the following capabilities:

- Full implementation of RFCs 3720 and 5048
- Active/Passive multipathing supports using vSphere Native Multi-Pathing (NMP).
- Support for Thin-Provisioned and Fully-Allocated LUNs
- Support for One-Way CHAP Authentication and Initiator Access Control
- Support for dynamically growing LUNs
- Support for Raw Device Mapping (RDM) devices.
- Support for creating LUN clones using full normal copies, snapshot-read-only copies and shadow-copy writable snapshots.
- Support for VSS (Volume Shadow Service) when utilizing the Microsoft iSCSI Initiator inside virtual machines.
- The number of LUNs per target is restricted to 256 by vSphere. Isilon does not enforce a limit on the number of targets.

OneFS iSCSI Target Configuration

OneFS supports an unlimited number of targets and 256 LUNs (numbered 0-255) within each target. Each target can be accessed through any of the cluster storage nodes IP addresses that belong to a SmartConnect static IP pool.

Each target can have its own unique CHAP authentication and target initiator access control listing.

Target CHAP Authentication

EMC recommends the use of Challenge Handshake Authentication Protocol (CHAP) for optimal security. EMC supports per-target, one-way CHAP authentication using username / password pairs. Usernames follow the same rules as target names (arbitrary strings following standard hostname rules—colons and periods are okay).

Target Initiator Access Control

EMC supports per-target access control. Targets can be in one of two states: open to all initiators, or closed to all initiators except for those on the **allowed-initiator** list. The initiator list can be empty, leaving three possible configurations:

1. Open to all
2. Closed to all
3. Open to a specified few

If a target is closed to a given initiator, that target will not be revealed to that initiator upon a **SendTargets** discovery query, and access will be denied to that initiator upon any attempt to log in or scan for LUNs.

Note: EMC recommends restricting access to all LUNs used for VMFS datastores by listing only the specific vSphere iSCSI initiators that share VMFS volumes in the virtual environment. Taking this action ensures no other iSCSI initiators can access or corrupt the VMFS datastores.

OneFS iSCSI LUN Configuration

LUN Type

EMC supports provisioning LUNs as *Thin* or *Thick*. Provisioning a LUN as thin will take less time at creation, and will not claim new space on the storage cluster until blocks needing that space are written by the vSphere host. Thinly-provisioned LUNs may initially yield higher performance than thickly-provisioned LUNs, as every initial write in a thinly provisioned LUN does not result in a read-modify-write operation within OneFS unless those LUNs are mirrored. This improvement goes away, however, once a block within the thin LUN needs to be modified. Every rewrite within a thin LUN could potentially result in the same read-modify-write operation as a thickly provisioned LUN, at which point there should be no discernible performance difference between the two LUN types.

OneFS LUN settings for Optimizing vSphere iSCSI VMFS I/O Performance

LUN layout settings determine how LUNs are protected and striped across disks on OneFS. Layout settings affect performance and utilization, and should be set according to the type of I/O load generated by the vSphere hosts and their underlying virtual machines. In most cases I/O load generated by a set of virtual machines in a vSphere host is mixed-read-and-write, and relatively random in nature. However, in some cases I/O load can be dominated by one type of access or another. Below are LUN layout recommendations for specific types of I/O-intensive load:

1. Very random writes within a large LUN have been shown to benefit from setting the **Access Pattern** to **Streaming** in the OneFS Administration WebUI. Testing of Concurrency versus Streaming has demonstrated no performance increase or decrease except for this one scenario.
2. In most cases, 2x mirroring protection will result in better performance since parity reads aren't required during the write process. However, if the vSphere workload is primarily read-based, then 2x mirroring will not provide any significant benefit. If space is a primary concern, then sacrificing some write performance to avoid mirroring space overhead is a possible option.
3. Write Caching on Isilon iSCSI LUNs is turned off by default. Turning on Write Caching can result in write-performance improvements, but there is a risk of corruption if a node loses power or crashes while uncommitted data is in the write cache. One exception would be to turn write caching on during the creation of a thick-allocated LUN, and then turn write caching off once LUN creation is completed.
4. Disable OneFS read prefetch if many virtual machines require a high ratio of small, random read operations. Disabling read prefetch instructs OneFS to avoid prefetching adjacent file blocks to eliminate prefetch latency overhead. Read prefetch is only valuable for sequential read workflows.

Note: Before setting 2X mirroring protection on a LUN the FlexProtect global setting must be changed to **Advanced**.

OneFS Networking Best Practices for iSCSI

Since iSCSI and NFS protocols both work over Ethernet/IP infrastructure they share some common network design guidelines and best practices. Those common best practices are described in General Guidelines for Networked vSphere Datastores above. But NFS and iSCSI have different mechanisms when using multiple network paths to achieve network redundancy and increase aggregate network throughput.

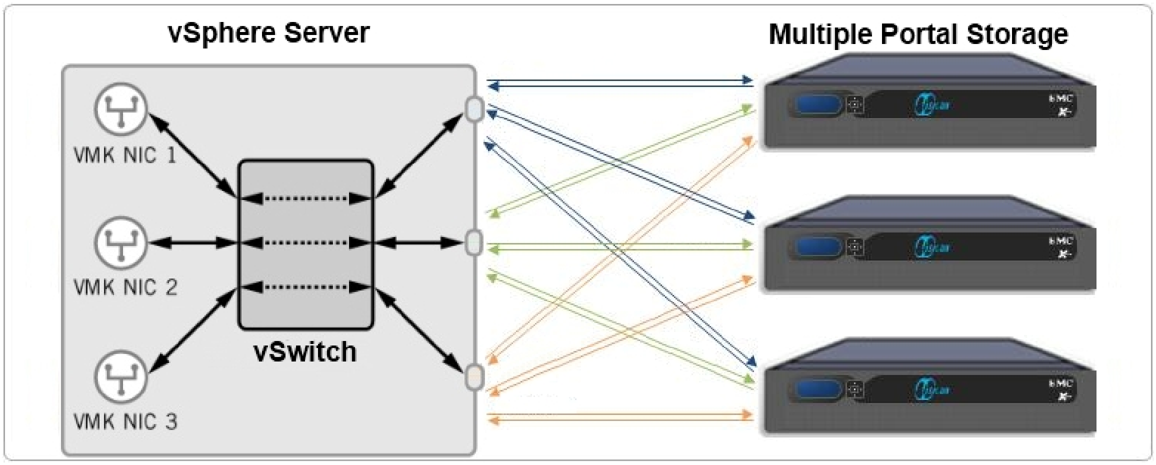

Figure 10: Multipath using iSCSI port binding in vSphere

vSphere Multipathing for iSCSI

Although it is possible, VMware does not recommend using network-link-aggregation to configure iSCSI network redundancy for vSphere. VMware instead recommends using iSCSI built-in **Multipath I/O (MPIO)** which defines end-to-end paths from iSCSI initiator to target. This is because, unlike NFS, block-storage architecture uses MPIO capabilities that VMware bundled into the storage stack, not the networking stack for multipathing behavior.

iSCSI MPIO relies on the ability of the iSCSI initiator to discover multiple paths to storage targets in order to establish paths between each of the VMkernel ports that are bound to the iSCSI initiator and each of the target portals defined by a unique target IP and port.

To ensure the VMkernel ports used by the iSCSI initiator are actual paths to storage, vSphere requires that they are connected to a portgroup that only has one active uplink interface and no standby uplink interfaces. This way, if the interface is unavailable, then vSphere considers the storage path to be down, and the iSCSI multipathing mechanism can choose a different path.

On the Isilon storage end, iSCSI targets can be represented by IP addresses assigned to a link-aggregation trunk on individual nodes for ESX 3.5 and below, or IP addresses assigned to individual interfaces across all nodes in the cluster on later versions of vSphere.

EMC recommends the following steps to maximize performance and availability for vSphere iSCSI datastores:

- Dedicate a static IP pool on the Isilon cluster for managing iSCSI target IPs. EMC does not support dynamic IP addresses for iSCSI targets and OneFS will not publish targets on dynamic IP addresses in **SendTarget** discovery responses.
- Create multiple VMkernel port groups on the vSphere host, with a single active network interface and no standby interfaces. Use **iSCSI port binding** to associate those VMkernel port groups with the iSCSI initiator.

- Ensure the selected **Storage Array Type Plugin (SATP)** is **Active/Active** and the **Path Selection Policy (PSP)** is set to **Fixed**. These are normally set by default.
- To avoid lock contention as a result of multiple vSphere hosts accessing the same LUN from different target nodes, configure all vSphere hosts to use the same preferred path to the same LUN.

Cloning LUNs for vSphere Datastores

Using the vSphere client, VMware allows iSCSI VMFS datastores to be created with thin-provisioned VMDKs. OneFS offers further improvements in storage utilization and faster provisioning by supporting thin-provisioning of new iSCSI LUNs (described above) and creating Shadow clones of existing iSCSI LUNs on the storage cluster. Shadow-cloned LUNs act as space-efficient, writeable copies of iSCSI LUNs. They are created instantaneously by selecting the “Shadow” type when cloning a LUN, and only consume space on the OneFS file system when an iSCSI initiator modifies the cloned LUN data. A Shadow-cloned LUN can be used to store different versions of gold images or templates of virtual machines by cloning an entire VMFS datastore, or to clone RDM LUNs associated with individual VMs.

Note: Prior to creating shadow clones, it might be necessary to shut down all VMs in the datastore stored on the LUN. Otherwise, the VM data in the LUN may be inconsistent at the time it is cloned. Snapshot consistency is covered in more detail in Working with Snapshots below.

The cloned datastore can be re-signed and the new virtual machines (that are clones of the original ones) can be powered on. However, certain limitations need to be taken into account when using shadow clones:

- Shadow-cloned LUNs are linked to the original base LUN from which they were cloned, but changes to settings on the cloned LUN do not affect the base LUN. Since any blocks changed on the cloned LUN are written to a new location on the clone, LUN extents protection and LUN layout settings are only applied to the modified blocks.
- The only LUNs that can be fully backed-up, restored or replicated using Isilon SyncIQ from within an Isilon cluster are normal LUNs and normal clones. This is because cluster-based backup, restore and replication applications do not have knowledge of the base LUNs and snapshot data associated with them. For this reason, EMC recommends that backups be performed at the virtual machine level, or using VMware array data protection (VADP, discussed in Backups below).

Note: iSCSI shadow clones, as well as other types of clones, require that Isilon SnapshotIQ be licensed.

If a cloned LUN contains a VMFS datastore copy, there will be two datastores with the same VMFS Unique Universal Identifier (UUID) signature. You can now choose whether to mount the LUN copy with the original VMFS signature or create a new VMFS signature. Since you cannot have two VMFS datastores with the same signature mounted at the same time, you will have to unmount the original VMFS datastore before mounting the cloned VMFS datastore with the same signature. Otherwise you can choose to create a new signature for the clone while keeping the original datastore mounted.

Note: Datastore "re-signaturing" is irreversible. Once you perform a datastore resignature on a LUN clone it is no longer treated as a copy, but rather becomes a unique LUN hosting a unique VMFS datastore.

vSphere Configuration for iSCSI Datastores

Much of the information in this section is paraphrased from the vSphere 5 Documentation Center, available online from VMware. This Guide includes a selection of key considerations and configuration guidelines, but EMC encourages a review of VMware's online vSphere storage documentation for a more comprehensive review of iSCSI setup and other advanced settings.

LUN Settings and Sizing Considerations

An iSCSI LUN can only host one VMFS datastore. vSphere can create a single VMFS datastore from multiple iSCSI LUNs as extents to the VMFS volume. Each LUN should have the correct protection and layout settings for applications in virtual machines in that LUN. If multiple virtual machines will access the same VMFS, use VMDK disk shares to prioritize your virtual machines' I/O access appropriately.

If VMFS3 datastores are used, the maximum LUN size is 2TB. If larger datastores are required, VMFS5 (introduced with vSphere 5) enables VMFS volumes of up to 64TB in size. Because LUN-size planning and consideration must take into account the I/O load that the virtual machines in that datastore will generate, a commonly-deployed size for a VMFS datastore is between 300GB and 700GB, and / or 5-10 virtual machines. In the event that a single datastore becomes overloaded in terms of either capacity or workload, VMware provides Storage vMotion as a means to re-balance VMFS files to alternative datastores without disruption to the dependent VMs.

Each underlying LUN can also be tuned using more granular settings to satisfy specific application I/O and data protection requirements for the VMs hosted on that datastore. As more datastores are accessed concurrently, the aggregate I/O increases because more paths can be used in parallel, and therefore contention per VMFS volume is reduced.

Generally, because EMC recommends configuring all vSphere hosts to use the same preferred path to a given datastore, care should be taken to ensure that the total network bandwidth for all vSphere hosts using the datastore should be less than the throughput rating of the network path. So a datastore that will be accessed concurrently by five vSphere hosts over a 1Gb network link should be sized to ensure that each vSphere host uses no more than about 100-200Mb/s of that bandwidth.

There are also benefits to using fewer, larger VMFS datastores, however, including greater flexibility in creating virtual machines or resizing virtual. Having fewer VMFS datastores to manage also reduces the administrative overhead associated with balancing free space across all available datastores. If this approach is used, EMC recommends monitoring datastore I/O traffic regularly to ensure per-datastore bandwidth usage remains within manageable and acceptable levels.

iSCSI Initiator Multipathing Configuration

General networking best practices were provided in the section OneFS Networking Best Practices for iSCSI in the previous chapter on OneFS iSCSI Configuration. This section discusses multipathing configuration guidelines specific to the vSphere iSCSI software initiator.

iSCSI datastores and RDMs can both use multiple paths to provide redundancy across all network components in the iSCSI data path, from vSphere host network interfaces to switches, Isilon cluster nodes, and interfaces. An Isilon cluster in turn provides a multi-node active/active storage cluster with multiple storage targets that can be used for both redundancy and aggregate I/O.

Because iSCSI multipathing does not have direct access to physical network interfaces, you must first connect each physical interface to a separate VMkernel port, and then associate all VMkernel ports with the software iSCSI initiator using **port binding**. As a result, each VMkernel port connected to a separate network interface becomes a different path that the iSCSI software initiator can use.

Note: Hardware iSCSI adaptors have not been tested and are not currently supported for use with Isilon storage.

Path Management for Load Balancing

Balancing workload across available paths increases availability and can also improve performance by distributing the I/O workload load from one or more LUNs across multiple parallel paths. With Active/Active storage systems, you can configure your vSphere hosts to load balance traffic across multiple adapters by assigning preferred paths to your LUNs. EMC Path selection policy must be set to **"Fixed (VMware)"** in order to be able to specify a preferred path and fail back to it if and when it becomes available again after a failover event.

An alternative method is to use an **Active/Passive** storage system with path selection set to **"MRU" (most recently used)**, but this type of path selection will not fail back to the preferred path, resulting in multiple disparate paths being used to connect to the datastore, and negatively affecting overall storage performance.

Note: While EMC supports the Active/Active storage plug-in type with fixed-path selection policy, EMC recommends configuring all vSphere hosts accessing the same datastore to use the same preferred path in order to eliminate any cross-node lock contention. This, in turn, should be factored in when sizing individual iSCSI datastores for optimal VM placement and load balancing

Virtual Machine Configuration and Management

Once NFS or iSCSI datastores are created on an Isilon storage cluster, virtual machines can be created or migrated to or from other storage systems using the vSphere client.

VMDKs and Other VM Files on Isilon Clusters

Virtual Machines on NFS datastores

When using NFS datastores on an Isilon cluster, virtual machines are managed directly by OneFS and can be located by navigating to the virtual machine directory path on OneFS. Each of the datastore subdirectories represents a single virtual machine instance along with the **.vmx** config file, **.vmdk** virtual disk files, and other files.

Creating Virtual Disks in a Datastore

When creating a new virtual machine, the first virtual hard disk is created along with—and stored in the same datastore as—the virtual machine. The datastore can be either an NFS mount or an iSCSI volume. Additional virtual hard disks can be added to existing virtual machines, and their corresponding VMDK files can be stored in the same datastore as the virtual machine or on other datastores accessible by the vSphere host.

Using iSCSI Datastores and iSCSI LUNs for Raw Device Mapping

Raw Device Mapping (RDM) allows a special file in a VMFS volume to act as a proxy for a raw LUN. The mapping file contains metadata used to manage and redirect disk accesses to the physical device. The mapping file, which can be thought of as a symbolic link from a VMFS datastore to a raw LUN, merges VMFS manageability with raw LUN access.

RDMs are useful when you want to dedicate a raw LUN to a single virtual machine. RDMs are also useful for native clustering technologies such as Microsoft Clustering Services (MSCS), or clustering between physical and virtual machines.

RDM disks are creating by selecting **the Edit Settings option** of a VM and selecting the **Raw Device Mapping** in the **Add Hard Disk** wizard. When adding an RDM you must choose a datastore in which to keep the mapping file, and one of two compatibility modes:

- **Physical compatibility mode -** the virtual machine can access the LUN directly. This is generally used from the application inside VM wants to directly access LUN. However using physical compatibility mode you lose the ability to clone the virtual machine, creating virtual machines from a template, or storage migration of virtual disks. Some documentation refers to this as **Pass-Thru Mode**.
- **Virtual compatibility mode**—allows the LUN to behave as a VMDK, which enables the use of features like cloning-to-template, cloning-to-VM or storage migrations of virtual disks. Some documentation refers to this as **Non-Pass-Thru Mode**

Note: When you wish to implement Microsoft clustering, you must create any shared-drive RDMs using Physical Compatibility mode.

Best Practices for Availability and Performance

Path failover occurs when the active iSCSI path to a LUN is changed from one path to another, usually because of some network component failure along the current active path.

Setting Guest Operating System Timeout for iSCSI Failover

During the failover process, I/O to and from the storage array might pause for 30 to 60 seconds until the vSphere host determines that the link is unavailable and until failover to another path is complete. As a result, the virtual machines (with their virtual disks installed on the Isilon cluster) may appear unresponsive. If you attempt to connect to a vSphere host, its storage devices, or its adapter, the operation may appear to stall until after failover is complete, at which point I/O resumes normally and the vSphere client will start responding again. If failover takes a long time to complete, some Windows virtual machines might encounter I/O errors on their disks and fail. To avoid the failure, VMware recommends setting the disk timeout value for the Windows virtual machine to at least 60 seconds. This is accomplished by editing the local registry.

VMDK Alignment

Virtual machine hard disks are stored as VMDK files on NFS or iSCSI VMFS datastores. VMs use their virtual hard disks to store data by formatting them into local partitions and file systems. If guest OS partitions are not properly aligned with the shared storage block boundaries, overall performance may suffer. OneFS aligns data at 8K block boundaries, and if VMDK partitions are not aligned with those boundaries then additional I/O operations are required to serve VMDK access requests. EMC and VMware recommend the following to avoid issues with VMDK alignment:

- VMDK misalignment-related performance issues are mostly visible in random I/O operations, so EMC recommends creating separate VMDKs for OS boot volumes (which are mostly used for reading OS and application binaries) and data volumes (which are used for random I/O applications and user data access).
- Boot partitions should be aligned before assigning them to a new virtual machine by using another virtual machine as a surrogate for the new VMDK. This is accomplished by adding a new virtual disk to the surrogate virtual machine and aligning the VMDK, then dissociating the VMDK from the surrogate virtual machine and adding it to the new virtual machine by choosing the newly-aligned VMDK for the new virtual machine hard disk.
- Another way to address boot partition VMDK alignment is by creating a template virtual machine with an aligned boot VMDK and cloning it to create new virtual machines.

Creating Aligned Disk Partitions in an existing Windows Virtual Machine

Once you have created a new Windows virtual machine with a hard disk or added hard disks to an existing Windows virtual machine, you should be able to see each disk as a unique device in the **Windows Disk Management (WDM)** tool. However, if you create partitions and format the partitions in WDM you will not achieve the best performance possible. There are two reasons for this:

1. **Disk alignment and partitioning**. In versions of Windows® prior to Windows 2008, when creating the first partition of a disk, the first sector of that partition is not, by default, aligned with the physical disk. There are also situations when Windows 2008 partitions are not properly aligned. Therefore, EMC highly recommends using the 'diskpart' utility with the ‘align=64’ option when creating new partitions.
2. **Block size**. Isilon uses an 8KB block size. Windows typically formats in a 4KB block size. To get the best performance you will want your disks also formatted with an 8K block size (also called the unit allocation size). When formatting the partition, EMC recommends using the command line interface to do so in order to ensure proper block alignment before turning the new drive over to the VM for use.

Creating Aligned Disk Partitions Prior to Installing a Windows Virtual Machine

Virtual disks can be formatted with the correct offset at the time of creation by simply booting the VM before installing an operating system and manually setting the partition offset. For Windows guest operating systems, consider using the Windows Preinstall Environment boot CD or alternative "**live dvd**" tools.

VMDKs can also be aligned by using the **‘fdisk’** utility from a command-line session on the vSphere host. This utility can be used for VMDKs stored both in iSCSI VMFS and NFS datastores, regardless of the guest OS to be installed on them

Creating Virtual Machine Templates with Aligned VMDK Partitions

To avoid having to align each VM’s VMDK files individually, EMC recommends creating a set of VM templates with pre-aligned partitions during the initial deployment of the vSphere cluster, then deploying all guest VMs using these templates.

Migrating Virtual Machines between vSphere Hosts

Once the Isilon cluster is configured and available to vSphere using iSCSI VMFS and NFS datastores, virtual machines located on those hosts can be migrated without further configuration.

vSphere 5 provides the option of either migrating the virtual machine to a different vSphere host (vMotion), a different datastore (Storage vMotion), or simultaneous migration to both a different vSphere host and a different datastore.

Migration Types

Cold Migrations

A cold migration involves moving a virtual machine between hosts while the VM is powered off. This type of migration consists primarily of ‘relocating’ the VM files located on shared storage and changes to the VM config files to reflect the new host where it resides.

Migrating with vMotion

vMotion allows a running VM to be migrated between vSphere hosts. vMotion requires the hardware of the two vSphere servers to be compatible, and the two hosts must be able to access the same shared storage via the same paths, as with cold migrations.

When using vMotion with Isilon storage, the two vSphere hosts participating in the migration must use the same datastore to access the VM:

- For NFS datastores the node IP address and directory of the datastore must be the same. For example, if HOST1 points to 192.168.0.1/ifs/vmware, then HOST2 must also point to the same datastore via the same path.
- For iSCSI VMFS datastores, HOST2’s software iSCSI initiator must have access to the same iSCSI LUN (CHAP and target access restrictions), and must have both the source and destination VMFS datastores mounted via the same target paths.

Following best practices for network design (as detailed above) will ensure all hosts in the vSphere cluster see the same datastore.

If NFS failover is used, and the hosts involved in the migration are using the dynamic IP address for the shared datastore, then a migration can’t be interrupted or impacted by a failure on the node to which the hosts are connected.

Migration with Storage vMotion

vSphere 5 has the ability to migrate all the components of a virtual machine (VMDKs and configuration files) to another shared storage location while the virtual machine is still running by using *Storage vMotion*. This simplifies the process of migrating from one networked storage device (i.e. datastore) to another. Storage vMotion also enables live (i.e., with no downtime for the VM) migration of VMs between various types of datastores (SAN, iSCSI and NFS) and across different types of storage systems, as well as changing the configuration of the VMDKs if necessary (e.g., thin-to thick-provisioning).

Note: If simultaneous migration of a VM to another host and another shared datastore are necessary, then the VM must be shut down before initiating the migration.

Working with Snapshots

A snapshot of a virtual machine is simply a point-in-time copy of the state of the virtual machine, including its memory, disk contents, and virtual machine settings. Because a virtual machine stored in an NFS datastore consists of a set of files located on an Isilon cluster, it is possible to take snapshots using either the vCenter Snapshot Manager or EMC's SnapshotIQ feature. Virtual machines hosted in an iSCSI datastore can either use Raw Device Mapping LUNs to take snapshots of raw LUNs on an Isilon cluster, or use the vCenter Snapshot Manager.

Snapshot Data-Consistency Considerations

When taking a snapshot of a virtual machine, it is critical to consider the state of the VM and the application activity before initiating. Taking a snapshot of a running VM—especially a VM running applications that depend on data consistency, such as a database or Active Directory directory services—can result in data corruption or lost transactions.

For example, if a VM is reading or writing a large file, and a snapshot is taken before this transaction is complete, then if the VM is reverted to its pre-snapshot state the file that was being read or written during the snapshot process will be corrupted by the revert operation. This data-consistency issue must be considered regardless of which snapshot process is used: any storage- or host-based snapshot technology can result in data corruption if not done with consideration for the VM's application state. Snapshot consistency is also highly dependent on the OS type and version. Depending on the process used to quiesce the VM, the snapshot can be:

- **Crash-consistent**: this snapshot relies on the operating system's ability to survive sudden crashes or reboots and allows the operating system to restart without corruption. Most modern operating systems such as Windows and Linux are crash-consistent.

- **Filesystem-consistent:** a snapshot that integrates with filesystem capabilities to flush its cache before a snapshot is taken. VMware provides a synchronization driver as part of its VMware tools package. To take advantage of the VMware sync driver a virtual machine's state should be quiesced before a snapshot is captured in OneFS.

- **Application consistent**: similar to filesystem-consistent snapshots; only in this case the OneFS snapshot needs some mechanism to communicate with the applications running inside the virtual machine(s) to complete any outstanding application I/O and clear any application-specific cache data. A filesystem-consistent snapshot is not enough in this case because applications often either have their own data-consistency logic or do not use a file system at all and access block devices directly (in this case the virtual hard disk or RDM). vSphere 5 integrates with Window Virtual Shadow Service (VSS) to allow application providers to supply VSS writers that are triggered by the VSS quiescing mechanism. In this way a vCenter server (or client) initiating a vSphere snapshot can trigger the VSS mechanism in a Windows VM and provide application-level consistency. In this manner OneFS snapshots can be integrated in the same way as described for filesystem-consistency snapshots.

Note: Care must be taken to ensure that VSS-based applications are properly installed inside the virtual machine in addition to the VMware-supplied VMware tools.

A good resource for fully understanding VMware snapshot capabilities can be found in VMware's "Using Snapshots to Manage Virtual Machines".

Snapshot Types

Online Snapshots (Hot)

An online snapshot is taken while the VM is running, and possibly has applications running inside it. Online snapshots allow the VM and its applications to continue running, so there is no interruption of service; however, as noted above, if applications are reading or writing data, there is no guarantee of data consistency if the VM is reverted to an online snapshot.

Near-line Snapshots (Warm)

A near-line snapshot is taken while the VM is in the suspended state. Near-line snapshots, like offline snapshots, generally take less than a second to complete, and when the VM is returned to a powered on state, processing resumes from the point in time at which the VM was suspended. Near-line snapshots can have the same potential issues with data consistency as online snapshots.

Offline Snapshots (Cold)

An offline snapshot is taken with the VM guest OS shut down and the VM powered off. Taking snapshots in this manner ensures data consistency for applications inside the VM. The trade-off is that for production virtual machines, it may not be possible to completely shut down operations in order to take a snapshot. Offline snapshots are nearly instantaneous, however, and a VM is typically much faster to boot than an equivalent physical server, so downtime is minimized.

Isilon Plugin 1.0 for vCenter

The Isilon Plugin for vCenter comprises two components:

- Virtual appliance running an Isilon virtualization server
- VMware vCenter WebUI plugin module.

The virtual appliance uses VMware APIs to interface with VMware vCenter, and uses the Isilon Platform API to interface with the cluster in order to orchestrate backups and restores across Isilon storage environments. Finally, the Plugin for vCenter hooks into vCenter and presents a management screen to control VM backup and restore functionality.

Plugin for vCenter 1.0 Considerations

The Plugin for vCenter is accessed via a tab in the vCenter console, or by right clicking on the appropriate Virtual machine in the vCenter server resource pane.

- In order for full application consistency (for instance, when backing up a database server), a **'cold backup'** should always be selected.
- **'Hot backup'** is the default selection.
- RDM-snapshot and -restore functionality is not available in this version of the Plugin for vCenter.
- Isilon SnapshotIQ must be licensed and enabled in order to use the Plugin for vCenter functionality.

Integrating SnapshotIQ and vCenter Snapshots via Plugin for vCenter 1.0

It is often beneficial to integrate Isilon SnapshotIQ snapshots with vCenter snapshots for the following reasons:

1. vCenter Snapshots, combined with Isilon SnapshotIQ snapshots provide more flexibility in data management for vSphere environments.
2. SnapshotIQ snapshots can be tied into a storage backup application which enables you to capture the vCenter snapshot to tape. They are often tied in as a pre-executable process to the backup application and give you the ability to mirror VM files from the storage snapshot to a different location.
3. Having a large number of snapshots in vCenter can get unwieldy, particularly if they are not purged on a regular basis. SnapshotIQ snapshots allow you to keep a larger number of snapshots.
4. SnapshotIQ snapshots can be taken automatically at any interval, and retention/deletion policies can be applied to purge SnapshotIQ snapshots at user-defined intervals. It is not currently possible to schedule snapshots in vSphere or to automatically purge snapshots without customized code using the vStorage API.

Restoring from a Previous Plugin for vCenter Backup.

The use of the Plugin for vCenter 1.0 is mostly suited for NFS datastores where each VM is housed in its own directory on the Isilon cluster, or for iSCSI environments in which each iSCSI LUN hosts only one VM. In these instances an Isilon snapshot can be taken for individual VMs.

For iSCSI datastores housing multiple virtual machines, an Isilon snapshot can only be taken on the entire datastore. In this case, all VMs on a shared iSCSI datastore must be quiesced before the snapshot is taken. For this reason, if iSCSI storage will be used in conjunction with the Isilon Plugin for vCenter, EMC recommends no more than one VM per iSCSI LUN.

Using SyncIQ

Using Isilon SyncIQ, virtual machines can be replicated to a secondary Isilon cluster and vSphere host for disaster recovery, testing or other uses.

Replicating VMs and Datastores using SyncIQ

Similar to Isilon snapshots, the granularity of the replication data set depends on whether virtual machines are stored in iSCSI VMFS or NFS datastores:

1. NFS datastores allow replicating individual virtual machines because each VM is stored in a separate directory on the Isilon cluster.
2. iSCSI VMFS datastores map to iSCSI LUN directories on the Isilon cluster so all VMs in that LUN are replicated at once.

SyncIQ uses Isilon snapshots to replicate asynchronously a point-in-time view of VM directories and LUNs between clusters. Scheduling of this replication can be automated by creating SyncIQ replication policies. SyncIQ optimizes data transfer over the network by replicating only changed blocks after the initial replication is completed.

On the target SyncIQ replication cluster, Isilon snapshots can be taken after each replication event to keep multiple versions of each replication sequence and revert back in time to the selected version.

Note: Isilon SyncIQ does not provide VM file-level restores. Please contact one of the following EMC partners in virtualization-data-protection solutions to address file-level restores: Commvault, Symantec, Vizioncore, and/or Veeam.

Restoring Virtual Machines from SyncIQ Replication Sets

Once replicated, the remote cluster can be restored and accessed from the same vSphere host or a secondary vSphere host. Restoring virtual machines from the target replication target can be done in several ways:

1. Stopping all replications to the target cluster and mounting the target cluster nodes as datastores. Once datastores are mounted VMs stored on those datastores can be registered and powered on by the same or remote vSphere hosts.
2. Copying VMs or LUNs from the target cluster to another location locally. In this way the target cluster can continue to operate as a SyncIQ replication target while also servicing vSphere host datastore storage.
3. Copying VMs from or LUNs from the target cluster back to the primary cluster and re-register those VMs by the original vSphere host. The target cluster continues to operation only as a SyncIQ replication target (or serving a separate vSphere set of hosts).

Note: When the VM is powered on, a dialog may appear indicating that a duplicate *UUID (universal unique identifier)* was found. The UUID is a unique machine identifier that is copied with the VM configuration. If you want to use the original VM on the primary vSphere cluster, then you will need to create a new VM UUID on the target vSphere cluster.

Performance Tuning for SyncIQ Replication Jobs

The following options can improve SyncIQ replication jobs.

- Filtering out .vswp files from the replication set. These files are virtual machine files used to store swapped-out memory data from virtual machines. They represent run-time data not required when restoring a virtual machine from a target replication set. Often times .vswp files comprise the majority of changed data between replication jobs, and can add considerably to the replication set for transfer to the remote target. They can be filtered out of the replication set when defining a SyncIQ replication policy.
- Disabling block-hash comparisons during replication jobs. By default, SyncIQ jobs compare every block of existing data between the source and target replication sets in order to ensure they have not changed since the last replication job. This can significantly increase replication time and add considerable CPU load on the nodes conducting the hash calculations (thereby also potentially disrupting other storage activity), particularly in environments where there is a fast network link between source and target cluster. In these environments, CPU load can slow down network transfer. EMC recommends disabling hash calculations on the source cluster when defining the SyncIQ policy.

Note: For more information on SyncIQ for Isilon replication, please refer to the document Best Practices for Data Replication with EMC Isilon SyncIQ.

Backups

Most enterprises include tape backup as part of their data protection process. Tape backup is an important element of the picture; however, faster methods to restore lost data, such as snapshots and remote synchronization of data, are a first line of defense.

In general, there are three options for backing up virtual machines:

Guest OS VM backups using backup client software (e.g. CommVault Galaxy or Symantec NetBackup) installed on each VM. Guest-level backups using VM-installed software have the primary advantage of application awareness. The main disadvantage of this method is that as the number of virtual machines grows, resource utilization and scheduling for a growing number of backup agents become complex to manage.

1. *vStorage API for Data Protection (VADP),* formerly known as *VMware Consolidated Backup (VCB)*, centralizes the management of VM backups and integrates with leading data protection vendors to provide additional flexibility in how virtual machines are backed up and restored.

2. *Network Data Management Protocol (NDMP)* backup from a NAS device using NDMP-compliant backup software. While NDMP backups are prevalent in NAS environments, they do not provide application-consistent snapshots of the virtual machines if such snapshots are required.

Note: For EMC recommendations and best practices for data protection please see the Data Protection for Isilon Scale-out NAS Best Practices Guide.

vStorage API for Data Protection (VADP)

With support for both NAS and iSCSI datastores, VADP is available for backups in an Isilon environment. VADP can perform image-level backups of entire virtual machines or file-level backups of Windows and Linux virtual machines. VADP is supported by most backup applications, allowing a familiar interface to be used for managing backups. VADP also utilizes VMware snapshot and Windows VSS integration to ensure data consistency with applications running inside Windows virtual machines.

Since all virtual machines are represented by files on the NFS datastore, using NDMP is a viable option. In this use case, all VM virtual disks, config files and snapshots (if they exist) are backed up using NDMP-compliant backup software. This method has the further advantage of taking the backup load off the ESX server machine(s). The main limitation of this method is similar to that of taking snapshots: VM and application state must be considered; and that in order to back up a consistent copy of the VM, its applications and data, the VM should be shut down or suspended.

Note: For more details on this architecture and process, please visit the VMware website for VADP.

NDMP Backup of Virtual Machines

As noted, NDMP based backups only take crash consistent snapshots of the virtual machines and do not provide application consistent snapshots (such as databases driven applications).

NDMP backups are managed by backup software certified for compatibility with Isilon clustered storage. Isilon is certified with these products:

- Symantec NetBackup
- Commvault Simpana
- EMC Networker
- IBM Tivoli Storage Manager
- BakBone NetVault
- Atempo Time Navigator

For current information on certified OneFS and NDMP backup software versions, please consult the Isilon NDMP compatibility guide.

Managing Virtual Machine Performance and Protection

The combination of Isilon's SmartPools technology and VMware vSphere 5 enables real-time or policy-driven management of a virtual machine's performance. The SmartPools product applies storage-tiering concepts to disk pools, facilitating storage alignment according to file policies or attributes. For example, virtual machines with a high I/O requirement may be placed in a fast storage pool, whereas less subscribed virtual machines can reside in a less expensive pool.

Isilon SmartPools

Isilon SmartPools includes the following features:

1. **Disk pools:** Dynamic groups of disks associated in a single pool of storage, for example "all disks of all S-series nodes on the cluster." Disk pool membership changes through the addition or removal of nodes and drives.
2. **Disk pool provisioning:** Rules to automatically allocate new hardware to disk pools as it is added.
3. **Virtual hot spares:** Reserved space in a disk pool (up to four full drives) which can be used for data re-protection in the event of a drive failure.
4. **File pools:** Logical collections of files and associated policies governing attributes such as file size, file type, location, and file creation, change, modification, and access times.
5. **Disk pool spillover management:** Rules governing handling of write operations to a full disk pool.

By using Isilon SmartPools, it is possible to tier virtual machine files, datastores and RDMs dynamically and automatically. These can be placed on the appropriate class of storage (SSD, SAS or SATA) via disk pools.

And with SmartPools file pools, protection (parity or mirroring) and IO optimization can be applied per-file, per-VM or per-Datastore.

For more information on SmartPools for Isilon, please refer to the SmartPools Best Practices Guide.

Virtual Machine Storage Resource Balancing

vSphere 5 includes *Storage I/O Control (SIOC)* functionality. Using SIOC®, it is now possible to balance virtual machine I/O across asymmetrically-loaded vSphere hosts. To achieve this, SIOC uses an I/O-throttling mechanism to automatically equalize disk shares, in terms of contention (latency), for all VMs.

Storage I/O Control

Enabling SIOC on an iSCSI or NFS datastore will prevent any one VM from monopolizing the storage resources for that datastore by equalizing I/O requests from all VMs on that datastore, across all hosts in the vSphere cluster.

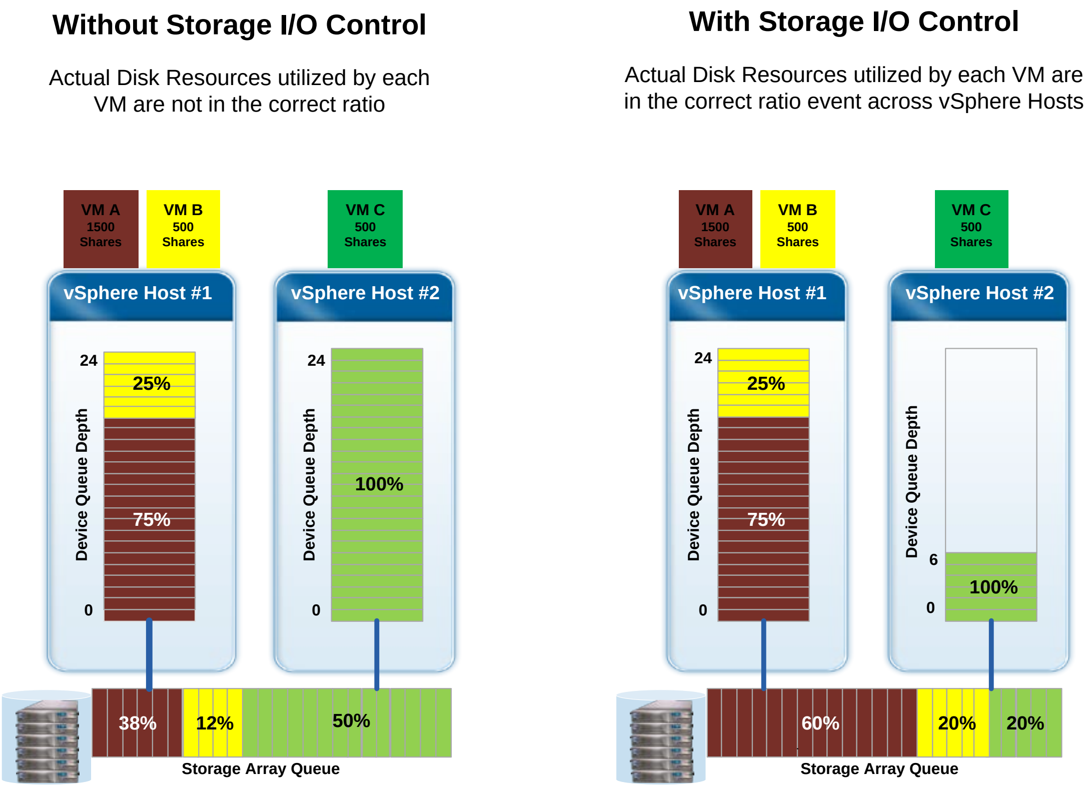
Without Storage I/O Control
Actual Disk Resources utilized by each VM are not in the correct ratio
With Storage I/O Control
Actual Disk Resources utilized by each VM are in the correct ratio event across vSphere Hosts
VM A 1500 Shares
VM B 500 Shares
VM C 500 Shares
vSphere Host #1
vSphere Host #2
Device Queue Depth
24
0
6
25%
75%
100%
38%
12%
50%
60%
20%
20%
Storage Array Queue

Figure 11: Storage resource equalization using Storage I/O Control

Storage I/O Control is disabled by default. It is enabled through vCenter on a per-datastore basis and configured with a default latency trigger of 30ms for a sustained period of 4 seconds. EMC recommends changing the congestion threshold value to 20ms.

More information on SIOC configuration is available here.

Note: Storage I/O Control in vSphere 4.x could be enabled on VMFS volumes only. vSphere 5 extends SIOC support to NFS-based datastores as well.

Additional Reference Links

SmartConnect documentation

VMware vSphere 5 Documentation Center

Adjusting TCP/IP Heap Size to accommodate more NFS datastores

Using Snapshots to Manage Virtual Machines

Best Practices for Managing Replication with SyncIQ

Data Protection for Isilon Scale-out NAS Best Practices Guide

VMware VADP Information

Isilon NDMP Compatibility Documentation

VMware Storage I/O Control Documentation

About EMC

EMC Corporation (NYSE: EMC) is the world’s leading developer and provider of information infrastructure technology and solutions that enable organizations of all sizes to transform the way they compete and create value from their information. Information about EMC’s products and services can be found at www.EMC.com.

About EMC Isilon

EMC Isilon, a division of EMC, is the global leader in scale-out NAS. We deliver powerful yet simple solutions for enterprises that want to manage their data, not their storage. Isilon’s products are simple to install, manage and scale, at any size and, unlike traditional enterprise storage, Isilon stays simple no matter how much storage is added, how much performance is required, or how business needs change in the future. We're challenging enterprises to think differently about their storage, because when they do, they'll recognize there’s a better, simpler way. Learn what we mean at www.isilon.com.

Contact EMC Isilon

505 1st Avenue South, Seattle, WA 98104

Toll-free: 877-2-ISILON | Phone: +1-206-315-7602

Fax +1-206-315-7501 | Email: sales@isilon.com