I only have three ESXi hosts and no other equipment. What do I do? ..............35
Management Considerations: Multiple VSA ...............................................36
Resilience: Back-End Network Failure ..................................................38
Resilience: Front-End Network Failure ..................................................44
Resilience: ESXi Host Failure ...............................................................49
Resilience: Replacing a Failed ESXi Host ..............................................52
Resilience: VMware vCenter Server Failure .........................................55
Conclusion ..........................................................57
About the Author ..................57
Introduction

In the release of VMware vSphere® 5.0, VMware also released a new software storage appliance called the VMware vSphere Storage Appliance (VSA). VMware VSA provides an alternative shared storage solution to our Small to Medium Business (SMB) customers who might not be in a position to purchase a Storage Area Network (SAN) or Network-Attached Storage (NAS) array for their virtual infrastructure. Without shared storage configured in a vSphere environment, customers have not been able to exploit the unique features available in vSphere 5.0, such as vSphere High Availability (HA), vSphere, vMotion™ and vSphere Distributed Resource Scheduler (DRS). The VSA is designed to provide “Shared Storage for Everyone”.

This paper will provide a deep dive into the VSA architecture.

Acknowledgements

The author appreciates the effort put in by Duncan Epping, James Senicka, Nathan Small, Eileen Hayes, Edward Goggin and Vishal Kher in reviewing this document.
Architectural Overview

VMware VSA can be deployed in a two-node or three-node configuration. Collectively, the two or three nodes in the VSA implementation are known as “VSA Storage Cluster”. Each VMware ESXi™ (ESXi) host in the VSA Storage Cluster has a VSA instance deployed to it as a virtual machine. The VSA instance uses the available space on the local disks of the ESXi hosts to present one mirrored NFS volume per ESXi host. The NFS volume is presented to all ESXi hosts in the VMware vCenter™ Server datacenter.

Each NFS datastore is a mirror, the source residing on one VSA (on an ESXi host), and the target residing on a different VSA (on a different ESXi host). Therefore, should one VSA (or one ESXi) suffer a failure, the NFS datastore can still be presented, albeit from its mirror copy. This means that a failure in any component of the cluster (be it an appliance or ESXi host) is transparent to any virtual machine residing on that datastore, but not running on the failed ESXi host. If the virtual machine was running on the failed ESXi host, vSphere HA, which is installed as part of the VSA deployment, will restart the virtual machine on a remaining ESXi host in the cluster.

The VMware vSphere Storage Appliance can be deployed in two configurations:

• 3x ESXi 5.0 server configuration
• 2x ESXi 5.0 server configuration

The two different configurations are identical from the perspective of how they present storage. A two-node cluster will present two NFS datastores whereas a three-node cluster will present three NFS datastores. The only difference is in how they handle VSA Storage Cluster membership. The following section will cover the details of a three-node and two-node VSA Storage Cluster.
Three-Node VSA Storage Cluster Configuration

In the three-node “standard” configuration, each node runs an instance of VSA. Each node presents a filesystem via NFS that is mirrored to one other filesystem on another VSA. To prevent any sort of “split brain” scenario, the three-node VSA configuration requires at least two nodes to be running to maintain a majority of nodes.

In this illustration, the three VSA datastores in the oval are NFS filesystems. These filesystems are presented as shared storage to the ESXi hosts in the cluster.
Two-Node VSA Storage Cluster Configuration

The two-node VSA storage configuration uses a special “VSA Cluster Service” which runs on the vCenter Server. It behaves like a cluster member and is used to make sure that there is still a majority of members in the cluster should one ESXi host/VSA member fail. There is no storage associated with the VSA Cluster Service.

![Diagram of Two Node VSA Storage Cluster](image)

In the illustration above, the VSA datastores in the oval are NFS filesystems presented as shared storage to the ESXi hosts in the datacenter.

The major difference between the two-node and three-node configurations, apart from the number of datastores that they present, is the fact that the two-node cluster requires an additional service running in the vCenter Server.

Let us look at the purpose of this service next.
VSA Cluster Service

The VSA Cluster Service helps in case of failures in a two-node configuration. For example, if an ESXi host or appliance fails, the VSA Cluster can remain online because more than half of the members of the cluster are still online (i.e. one node and the VSA Cluster Service).

The VSA Cluster Service is installed on the vCenter Server as part of the VSA Manager Installation process. It only plays a role in a two-node cluster configuration. It plays no role in a three-node configuration. It is installed in the following location on the vCenter Server:

![Figure 4. VSA Cluster Service Install Location](image)

As per the description, this service acts as a third cluster member in a two-node configuration. The service keeps the VSA Cluster online in the event of another member failure.
The service mainly embeds the third node clustering VSA component. In addition, it supports a few Remote Method Invocation (RMI) calls so that the VSA can configure the service when necessary. RMI can be thought of as a Java version of Remote Procedure Calls (RPC).

The service is strictly used for clustering purposes and enables the VSA Cluster to tolerate failure of one of the VSAs. In a three-node VSA setup, all clustering traffic is routed on the VSA back-end. However, in two-node VSA setup, since the VSA Cluster Service runs on the vCenter Server (which is assumed to be not connected to the back-end), the clustering traffic needs to be routed on the front-end network. The concept of front-end and back-end networks will be covered in detail later in the document.

In addition, the VSA Cluster Service needs to have a dedicated static IP address to communicate with other VSAs. In the event of a vCenter Server failure, this service can be easily restored to normal operation by first installing VSA Manager on another vCenter Server that runs the VSA Cluster Service IP, and then using “Repair ...” option in the User Interface (UI) to add the service back to the cluster. This procedure will be looked at in greater depth later on in the deep dive document.

The state of the service is monitored by VSA and displayed in the UI.
ESXi Requirements

During the installation of the vSphere Storage Appliance, a number of compatibility checks are carried out on the hosts to make sure they are configured correctly:

• Check that the host installed with ESXi 5.0.

• Verify that this is a green-field deployment of ESXi 5.0 with default management network and virtual machine network portgroups.

• Ensure that the ESXi hosts have the appropriate number of physical network adaptor ports. It requires four ports in total, either two dual-port network adaptors or four single-port network adapters. The ESXi host can have more than four network adapter ports, but only the first four will be used during the VSA installation.

• Check that the ESXi management network gateway is configured to be the same as the vCenter Server which will manage the VSAs (vCenter and appliances on the same subnet).

• Ensure that none of the ESXi hosts already reside in a vSphere High Availability (HA) cluster since the installer will add the ESXi hosts used for the VSA into a newly created vSphere HA cluster.

• Verify that all the ESXi hosts have the same CPU type. The cluster will be configured with Enhanced vMotion Compatibility (EVC).

• Check that the ESXi hosts do not have any running virtual machines.

• Ensure that the ESXi hosts have CPUs which run at a clock speed greater than 2GHz.

The installer will not allow you to select an ESXi host which does not meet the above requirements. There are other requirements which are not checked by the VSA installer, but which must be implemented to run the VSA on a supported ESXi host.

• 4, 6 or 8 physical disk spindles (for supported spindles, please refer to the VMware HCL)

• RAID 5, 6 or 10 configuration for the hosts

The VMware Hardware Compatibility List (HCL) displays both host models and RAID controllers that are supported with the VSA. To see the complete list, which is regularly updated, please go to https://www.vmware.com/resources/compatibility/search.php, and select Product Release Version: ESXi 5.0 and Features: VMware vSphere Storage Appliance (VSA), then click Update and View Results for the supported list.
Appliance Requirements

In the overview, we mentioned that each ESXi host has a VSA instance deployed to it as a virtual machine. Figure 6 lists the requirements for each of the appliances. First, here is a view of the General window taken from the Summary tab of one of the appliances in the VMware vSphere client:

<table>
<thead>
<tr>
<th>General</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Product:</td>
<td>SVA</td>
</tr>
<tr>
<td>Version:</td>
<td>1.0.0.6 (1.0 (SN 6))</td>
</tr>
<tr>
<td>Vendor:</td>
<td>VMware, Inc.</td>
</tr>
<tr>
<td>Guest OS:</td>
<td>Novell SUSE Linux Enterprise 8/9 (64-bit)</td>
</tr>
<tr>
<td>VM Version:</td>
<td>7</td>
</tr>
<tr>
<td>CPU:</td>
<td>1 vCPU</td>
</tr>
<tr>
<td>Memory:</td>
<td>1024 MB</td>
</tr>
<tr>
<td>Memory Overhead:</td>
<td>67.43 MB</td>
</tr>
<tr>
<td>VMware Tools:</td>
<td>Not running (Not installed)</td>
</tr>
<tr>
<td>IP Addresses:</td>
<td></td>
</tr>
<tr>
<td>DNS Name:</td>
<td></td>
</tr>
<tr>
<td>EVC Mode:</td>
<td>Intel® &quot;Merom&quot; Gen. (Xeon® Core™2)</td>
</tr>
<tr>
<td>State:</td>
<td>Powered On</td>
</tr>
<tr>
<td>Host:</td>
<td>10.20.196.24</td>
</tr>
<tr>
<td>Active Tasks:</td>
<td>vSphere HA Protection: Protected</td>
</tr>
</tbody>
</table>

Figure 6. Appliance Summary

The following are the hardware and software features of the appliances:

**Virtual Appliance Hardware**

- Single vCPU with 2GHz CPU reservation
- 1024MB Guest Physical Memory - all reserved
- Storage provisioned across 18x virtual disks on 1 x LSI Logic virtual SAS adapter and 2 x pvSCSI virtual SAS adapters
- 2 x 4GB virtual disks are used by the appliance exclusively for root/boot/swap
- 16 x equally sized virtual disks are evenly spread across two pvSCSI adapters
- The actual size of the virtual disks is configuration dependent
- Two virtual vmxnet3 ethernet adapters

**Virtual Appliance Software**

- Novell SLES11 Service Pack 1 Linux Guest OS
- NFS version 3
- Ext4 File System
- Multi-Disk RAID1 Software RAID
- Linux Open-iSCSI Initiator
- SCST SCSI Target Emulation
Detailed VSA Installer Configuration Steps

During the installation of the VSA, there are a number of configuration steps carried out, which are directly visible via the UI. This section of the deep dive highlights the major configuration steps that are carried out.

1. **Configure networking**
   - Create an additional vSphere Standard Switch (VSS) for the back-end network.
   - Add uplinks to both VSS.
   - Create NIC teams on both VSS.
   - Create virtual machine portgroups for front-end and back-end networks.
   - Create virtual machine kernel portgroup for vMotion.

2. **Configure a VMware vSphere HA Cluster with Enhanced vMotion Compatibility (EVC) processor**
   - EVC ensures vMotion correctly works between all hosts.
   - For a 2-node VSA Cluster, admission control is enabled and percentage of cluster resources reserved as failover spare capacity is set to 50%.
   - For a 3-node VSA Cluster, admission control is enabled and percentage of cluster resources reserved as failover spare capacity is set to 33%.

3. **Deploy VMware vSphere Storage Appliances**
   - This deploys the appliances that are embedded as Open Virtual Machine Format (OVF) in the installer.
   - One appliance is deployed to each ESXi host in the VSA Cluster. If the VSA Cluster has two hosts, then the installer will deploy two appliances; if it is a three-node cluster, then three appliances are deployed.

4. **Install and Configure the VSA Cluster**
   - Each VSA instance will have 18 disks as discussed in the storage deep dive, 2 for the guest OS and 16 for the shared storage.
   - The SUSE Logical Volume Manager (LVM2) creates a volume group.
   - 16 disks are added to the volume group.
   - 8 disks are used to build the primary volume which will be used for shared storage, and 8 disks are used to build the replica volume which will be used as a mirror for a primary volume.

5. **Synchronize the mirror copies of the datastores across different VSA members**
   - iSCSI is configured.
   - The replica volume is presented as a target to the designated mirror appliance over iSCSI.
   - The SUSE LVM2 on each appliances mirrors the local primary to the replica target volume.

6. **Mount the datastores from the VSA Cluster nodes to the ESXi 5.0 hosts**
   - Each appliance is configured as an NFS server.
   - The mirrored primary volumes are exported as an NFS share.
   - Each NFS volume is presented to each ESXi host which was part of the datacenter chosen during the VMware VSA installation.
   - This means that ESXi hosts which are not part of the VSA Cluster can still access the shared storage from the VSA Cluster.
Architecture Deep Dive: Storage

There are three volume managers to consider for each VSA- the hardware RAID adapter’s RAID set volume manager, the ESXi VMFS volume manager, and the appliance’s Guest OS LVM2 volume manager within the VSA Linux OS. Each of these volume managers refers to volumes or logical volumes. In addition, the VSA datastores are referred to as volumes. As we discuss the storage architecture of the VSA, we will try to make it as clear as possible which volume manager and which logical volumes we are referencing.

In this section of the deep dive of the VSA, we first look at the ESXi hosts. The physical hardware must include a RAID Controller and 4, 6 or 8 physical disks. These must be configured into a RAID 5, RAID 6 or RAID 1+0 (commonly referred to as RAID 10) configuration to avoid the situation where a single spindle failure could cause a complete node failure.

![Figure 7. RAID 1+0 Requirement](image)

Here is an example of a RAID 10 configuration on a HP Smart Array P410i:

![Figure 8. RAID 1+0 on HP Smart Array P410i](image)
When installed, ESXi 5.0 recognizes the RAID logical volume (in this example, RAID 10) as a single SCSI disk. ESXi takes enough space on this disk for boot, root, swap and a partition is created on rest of the disk, which is used to construct a VMFS-5 volume.

When the VSA installer runs, it pushes down a single appliance onto the VMFS-5 volume of each ESXi host. The appliance is configured with a single LSI Logic adapter which has two virtual disks attached. These virtual disks are used by the SLES 11 for its own OS partitions. There are also two additional paravirtual SCSI (pvSCSI) adapters configured on the appliance. These are used to create an additional sixteen virtual disks on each appliance. These disks are then used by the VSA to create the redundant NFS volume(s).

The final configuration from a storage perspective looks like this:
If we look at this from a local datastore perspective, the storage configuration of each appliance looks like this. Note the number of virtual disks.

![Figure 11. Appliance Virtual Disk (VMDK) List](image)
A special "EnableSVAVMFS" option is set on VMware VSA deployments. The "EnableSVAVMFS" advanced configuration option forces linear extent allocation across the 16 virtual disks providing storage to the VSA instance on the ESXi host. This is done to optimize the routing of read I/Os by that VSA virtual machine across its dual RAID 1 mirrors for the VSA datastore that it manages/exports. Without this option set, VMFS will not always provision the VMFS extents for a virtual disk using linearly increasing Logical Block Addressing (LBAs) even when it could do so.

Each appliance is responsible for presenting only one shared storage volume. These sixteen virtual disks are aggregated into a single volume group. Two separate volumes from this volume group are then created in the Guest OS. One logical volume is designated a primary volume, and one is designated a secondary volume. The primary volume is mirrored (via RAID 1) to a secondary volume on another appliance. In a two-node configuration, this arrangement is very simple; the primary volume on VSA 1 mirrored to the secondary volume on VSA 2, and vice-versa.
In a three-node configuration, it is a little more complex. Again, there is only a single mirror for each primary volume, so you get something like VSA1 mirrored to VSA2, VSA2 mirrored to VSA3, and VSA3 mirrored to VSA1. Once the primary and secondary volumes have been successfully synchronized, the primary volume is then presented back as an NFS datastore to each ESXi in the datacenter.

A simplified diagram showing the relationship between primary and secondary volumes, and the NFS datastore presentation is shown here:

![Appliance Mirrors](image)

This three-node configuration shows each VSA with a single exported NFS volume which is mounted on each of the ESXi 5.0 servers. Each VSA volume comprises of two mirrored disks, one local and one remote mirror.

**Mirroring**

When the sixteen virtual disks are attached to the SLES 11 appliance, they are assigned Linux SCSI Disk identifiers `sdc` through `sdr`. Disks `sda` and `sdb` are used by the SLES 11 Guest OS for its own root, boot, tmp, var and swap partitions. These sixteen devices are then divided up into two logical volumes, eight disks in each. One of these logical volumes will be used to present the NFS volume and the other will be used as the destination mirror/replica for another NFS volume from a different VSA. The actual volume that is going to be used for NFS is mirrored with another volume on a remote VSA using the LVM2 features of SLES. An EXT4 filesystem is then created on the volume before it is exported as an NFS share. The replication is Synchronous RAID 1 with dual mirroring.

SUSE LVM2 needs to see the remote volume to be able to mirror. To support accessing the remote volume, VSA implements the **Open iSCSI Initiator** ([http://www.open-iscsi.org/](http://www.open-iscsi.org/)) and the generic **SCSI Target (SCST)** for Linux ([http://scst.sourceforge.net/](http://scst.sourceforge.net/)). This allows a VSA instance to access a volume from a remote appliance, and mirror it with a local volume.
The remote iSCSI logical unit is actually accessed as “sdt” since “sds” is treated as a “dummy logical unit” by both the VSA control plane and the iSCSI target emulation in the VSA data plane.

**Maximum Storage Configuration of the VMware vSphere Storage Appliance**

VSA supports 4, 6 or 8 disks in each ESXi host that is going to be a VSA Cluster member. With a full set of 8 disks, then the maximum disk size VSA supports is 2TB. This gives a total physical disk capacity per host of 16TB. Across three hosts, it is 48TB.

Taking into account an example where there is a physical RAID 1+0 configurations across all hosts, total capacity is reduced by 50% since every single disk is mirrored. This provides 24TB of usable storage. Now you must take into account the RAID 1 configuration across all appliances. This is to allow VSA to continue presenting the datastores in the event of an ESXi host, network or appliance failure. This again reduces the usable capacity by 50% since every NFS volume is dual RAID1 mirrored. This gives VSA 12TB of maximum usable storage on a VSA Cluster, presented across 3 NFS datastores, each 4TB in size. Different calculations will need to be made if you use a RAID 5 or a RAID 6 configuration.
Architecture Deep Dive: Networking

To install the VSA, the ESXi 5.0 hosts that are going to be the cluster members must be newly installed with no additional configuration steps carried out. This means that the default networking on each ESXi will contain a stand-alone vSphere Standard Switch (VSS) with a single virtual machine network adapter uplink and two portgroups. One portgroup is a virtual machine kernel portgroup for the management network and the other is a virtual machine network.

The VSA Cluster must share the **same subnet as the vCenter Server**. There are a number of IP addresses for the successful deployment of the VSA. These include:

- VSA Cluster management IP address (for communication with vCenter – one per cluster)
- VSA Cluster service IP address (for 2-node configurations only – one per cluster)

**Front-End Network**
- VSA management IP address (one per appliance)
- VSA NFS server IP address (one per appliance)

**Back-End Network**
- Cluster/Mirroring/iSCSI network IP address(192.168.x.x) (one per appliance)
- vMotion network (Static or DHCP supported) (one per ESXi)

This diagram shows the logical/virtual network configuration including the two VSS, the various ESXi port groups (VSA related and non-VSA related), the virtual machines (VSA and non-VSA) linked to the virtual machine network, VSA front-end, and VSA back-end port groups, and the physical uplinks associated with each VSS.
It should be noted that the VSA installer will automatically configure a vMotion network between all of the ESXi hosts that are VSA Cluster members. This means that virtual machines deployed on the VSA Cluster can be migrated live between each of the ESXi hosts participating in the cluster.

The VSA installer also configures a vSphere HA cluster. This will be discussed in detail later in this document.

As a best practice, VMware recommends separating the front-end and back-end network traffic. To that end, we allow the specifying of different VLANs for both the front-end and back-end networks. The reason we require four network ports is that we create two NIC teams, one for the front-end communication and the other for the back-end communication. In the case of one network adapter port, or even one network adapter failing, there is a backup communication path available.

The management network, VSA Front End, and VSA vMotion port groups for all VSAs in a VSA Cluster must be located on the same VLAN (or all configured to not use a VLAN). The VSA back-end port groups for all VSAs in a VSA Cluster must be located on the same VLAN (or all configured to not use a VLAN).
An optimal configuration is shown here.

Figure 16. Optimal Networking Configuration

In this diagram, you can see that we have installed two dual-port network adapters. We use one port out of each network adapter for each team. If VSA loses a port or a complete network adapter, both the front-end and back-end traffic continue to flow.

Also note that in this example, we are using two physical switches. Should one fail, VSA networking continues to function.
If you examine the networking configuration of an ESXi 5.0 host that has been deployed as a VSA cluster member, you should observe something similar to the following:

![Network Configuration Diagram](image)

**Figure 17. Network Configuration of ESXi host in VSA Cluster**

Here, you can clearly see the NIC teams, the front-end and back-end networks, and the vMotion network. You can also see that the front-end and back-end networks can be placed on separate VLANs.
If you look at the properties of the vSphere Appliances, you can see that they have both a front-end and back-end network connection:

![VSA Properties](image)

The front-end and back-end IP addresses are also clearly displayed in the VSA Manager Appliances view:

![VSA Manager Appliances](image)

In this example, the VSA Management (front-end) IP addresses are 10.21.192.116, 118 and 120 respectively. The back-end IP addresses are 192.168.0.1, 2 and 3 respectively.
The NFS Server IP addresses are displayed in the VSA Manager Datastores view:

![Figure 20. VSA NFS Network Configuration](image)

The NFS datastores are exported to the ESXi hosts using IP addresses 10.20.196.117, 119 and 121 in this example. If you examine any of the ESXi hosts' storage configurations, you clearly see that the NFS datastores mounted to the ESXi hosts are using these IP addresses:

![Figure 21. NFS Datastore Mounted on ESXi hosts](image)

Many questions have arisen around the need for four network adapter ports. It is a tradeoff between project resource constraints, cost, security, reliability and performance.

The guiding principle was simplicity of installation. VSA has a single virtual and physical network configuration rather than support for a more flexible, customizable (and thus complex) networking configuration process during installation.

From a cost perspective, a decision was made to go with 4 network adapter ports. If the decision had been to isolate the VSA back-end cluster network traffic from VSA back-end iSCSI network traffic, in addition to isolating (storage) vMotion network traffic from all VSA back-end network traffic, it would require eight physical network adapters instead of four.

The security aspect primarily involves use of VLANs to protect the VSA virtual machines from denial of service attacks (e.g., ping flood broadcasting ethernet frames to all physical switch ports) by other virtual machines on the same ESXi.

The reliability aspect precludes allowing the use of a single network adapter port in order to allow the VSA datastores to be available in the event of a single network adapter port failure.

This has explained in detail why these requirements are in place and why there is a hard check for it in the VSA installer. You could also run with 4 single-port network adapters rather than 2 dual-port network adapters as shown in this example. However, running with a single quad port network adapter is not advised since it could be considered a single point of failure.
Architecture Deep Dive: Cluster Framework

Business Logic Components of the VSA

A brief overview of the major business logic components of the VSA are as follows:

1. SVA Aggregation Service (SAS)
   - Handle management requests by creating persistent tasks to manage logical components and report back to VSA Manager, for example, add a node.
   - Validate, authenticate and authorize requests from clients, for example, create a datastore.
   - Fault Tolerant - Seamless failover to another cluster member which will resume operations without rollback.
2. Storage Cluster (SC)

- Manage logical components across cluster nodes, for example, create a volume and its replica.
- Fault Tolerant - Seamless failover to another cluster member.

3. Member (Physical)

- Manage physical components (e.g. disks).
- Does not persist across re-starts – does not need to.

4. SVA (Boot-strap / Initialization)

- Instantiated on every node.
- Used for configuring the front-end and back-end network addresses when a node becomes a member.
- Persistent across system restarts (e.g. the IP addresses to be used for data access and when management traffic is persistent).

The business logic components come together as follows in the VSA:
The VSA architecture provides persistent, replicated metadata and an event-driven state machine to provide 'recovery without rollback' upon failure.

To become fault-tolerant, VSA requires the following primitives:

- A **membership management mechanism** to form a cluster, heartbeat with other VSAs to detect failures and rejoins, etc.

- A **reliable, atomic, totally-ordered messaging system** for communication between VSAs and for replication of VSA meta-data state across nodes.

- An **election mechanism** to select a master that will perform VSA meta-data operations for a particular domain, maintain ordering of the requests/messages and replicate the meta-data across nodes to provide fault-tolerance.

VSA uses **ZooKeeper** as the base clustering library for these primitives. ZooKeeper is an open source framework that provides a reliable shared hierarchical name space to its clients. It provides the following guarantees:

- **Cluster Membership Management** - Implements a quorum-based system to maintain a cluster and performs heart-beating to detect failures/rejoins.

- **Single System Image** - The client will see the same view of the service regardless of the server that it connects to.

- **Sequential Consistency** - Updates from a client will be applied in the order that they were sent.

- **Atomicity** - Updates either succeed or fail. No partial results.

- **Reliability** - Once an update has been applied, it will persist from that time forward until a client overwrites the update.

ZooKeeper has two components - the servers and the client library. The ZooKeeper servers provide a fault-tolerant distributed datastore. The client-library provides a way for the applications to interact with the datastore (read/create/update/delete data).

**ZooKeeper Leader Election, Heartbeats and Session Timeouts**

ZooKeeper servers use quorum-based techniques to elect one of the servers as the leader. The leader election ensures that the elected member has the most up-to-date cluster state. The leader heartbeats with followers (rest of the members). If a follower does not hear from the leader, or if the leader is not able to heartbeat with the majority of the followers, then that member triggers a new leader election.

VSA 1.0 has configured the ZooKeeper leader to heartbeat with the followers every second and the leader gives leadership up if it cannot heartbeat to a majority of the followers for four seconds. Similarly, a follower will start looking for a new leader (by starting a new leader election round) if it does not receive a heartbeat from the leader for four seconds. This timeout interval and any intervals described here are subject to change in future releases.

The ZooKeeper client library provides **Application Programming Interfaces (APIs)** so that the application can use them to establish sessions with ZooKeeper servers and read/modify the datastore. The application can create or modify directories and files - called **znodes** - in the ZooKeeper tree. The application can specify a session timeout when it creates a ZooKeeper client. After establishing a session, the application can perform operations on the data tree. In addition, the client (library) automatically starts heartbeating with the ZooKeeper leader. If the ZooKeeper leader does not hear from the client until the session timeout expires, the leader deletes the client’s session, and removes any transient znodes - called as ephemeral znodes - created by the client. Ephemeral znodes are znodes associated with the client session and are deleted from the data tree once the session expires. Other clients can monitor these znodes and the presence or absence of these znodes can be interpreted as an online or offline event for that client.
VSA Cluster Messaging

In the VSA 1.0 release, each VSA (and the VSA Cluster service in a two-node configuration) runs a ZooKeeper server. The servers are configured during cluster creation to use a simple majority for clustering and replication. All domains (and sometimes other components) that need to communicate across VSAs are ZooKeeper clients. These clients use the ZooKeeper client API to send (or receive) messages to (or from) other clients. This communication is performed by creating or updating “znodes” in the ZooKeeper data tree. ZooKeeper ensures that the data tree and the data stored at each znode is read and written atomically. For messaging, each domain creates a parent znode named after its domain ID in ZooKeeper. A domain can send a message to another domain by creating a child znode under the recipient domain’s parent znode.

VSA Health

The health of a VSA is also monitored using znodes. Each domain creates an ephemeral znode named after its domain ID. When a VSA goes offline, its clients will be unable to heart-beat with the leader. As a result, once the VSA session expires, the ZooKeeper leader will delete the ephemeral health znodes and send notification to other VSAs that are watching this znode. Similarly, if the VSA reboots, it will recreate the ephemeral health znode and other VSAs will notice that the VSA has come online. For the VSA 1.0 release, the client session expiry timeout is set to fifteen seconds.

Master Election

In addition to messages, the fault-tolerant domains (SAS and SC) use ZooKeeper to elect a master for the domain and to replicate the domain’s meta-data. The master is responsible for executing state machines for that particular domain and replicating the state machines to other VSAs for fault-tolerance. Replication of state is performed by storing the relevant state in ZooKeeper using the message queues described above.

Master election is performed using sequential ephemeral znodes. Sequential ephemeral znodes are ephemeral znodes with a sequence number appended at the end of the name of the znode. The sequence number is automatically assigned by the ZooKeeper leader when a znode is created with a flag set. Each fault-tolerant domain creates a sequential ephemeral znode. The VSA domain that has the znode with the lowest sequence number is selected as the master. If this node fails, ZooKeeper will delete its znode and inform other domains.

The VSA domain with the least sequence number at the end of its name will be elected as the new master. Thus, there is a master for each fault-tolerant domain that performs state transition for that particular domain.

The master of all SAS domains also starts the VSA Cluster IP that is used by the VSA Manager to interact with the VSA Cluster. The node running this IP is generally referred to as the “VSA Cluster Master”.

Business Logic Interaction during the VSA Cluster Creation

When a three-node cluster is being formed, the following interactions take place in the business logic of the VSA:

1. VSA Manager sends command to Create Cluster.
2. Create a one–node cluster on VSA 1.
3. Request VSA 2 to join cluster.
4. Replica SAS and SC domain, create unique Member Domain.
5. Request VSA 3 to join cluster.
6. Replica SAS and SC domain, create unique Member Domain.
7. Creation completed.
RAID and NFS Volume Creation

The steps that the VSA Business Logic goes through when configuring the mirrored NFS shared storage are:

1. VSA Manager sends “Create DataStore” command to cluster:
   - Cluster creates mgmt task in SAS domain.

2. Task sends “Create Volume” command to SC domain:
   - SC creates volume object.

3. Volume sends “Create Member Disk” commands to member 2 and member 3, and a “Create Member Volume” command to member 2:
   - Member 2 creates member disk object.
   - Member 3 creates member disk object.
   - Member 2 creates member volume object.

4. Member disk objects create logical volume, and on member 3, exports to member 2:
   - Member disks inform volume object in SC domain.

5. Volume object tells member volume that the member disks are online.

6. Member volume imports member disks, creates RAID device and informs volume object in SC domain.

7. Volume object in SC domain informs task in SAS domain that volume is complete.
Now, the mirrored volume is created. The next step is to export the volume as an NFS datastore and setup the Access Control Lists (ACLs). The steps involved are:

8. Task sends "Create NFS Data Access Service" to SC domain:
   - SC domain creates "NFS Data Access Service" object.
9. NFS Data Access Service object sends "Create Member NFS Data Access Service" command to Member 2:
   - Member 2 creates NFS Data Access Service object.
10. Member NFS Data Access Service:
    - Formats volume, creates mount point, mounts volume.
    - Creates NFS export.
    - Brings up vSphere High Availability NFS IP address.
    - Informs NFS Data Access Service object that Member NFS Data Access Service object is complete.
11. NFS Data Access Service object informs task that NFS Service is complete.
12. Task informs the management client that "Create DataStore" command is complete.
13. Task completes and terminates.
Figure 26. NFS Volume Creation – Business Logic Interactions
Install Considerations: vSphere HA Cluster

As already mentioned, the VSA installer also configures a vSphere HA cluster across all ESXi hosts that are participating as cluster members. This makes the cluster very resilient and allows it to handle some really catastrophic failure scenarios.

If, for instance, an ESXi host goes down, this will also bring down the appliance and any virtual machines running on that host. If there are virtual machines running on other ESXi hosts but using the NFS datastore that was presented from the failed appliance, the virtual machines continue to run, as the underlying datastore will be presented via the mirror copy. In the case where virtual machines were running on the failed ESXi host, vSphere HA comes into play. vSphere HA will restart the virtual machines on another ESXi host that is part of the vSphere HA cluster, since their disks and configuration files are still available via the mirrored copy of the NFS datastore.

VSA and vSphere HA provide a very powerful resilient clustering solution for virtual machines.

In this section of the VSA deep dive, we will look at the actual vSphere HA configuration settings, and some considerations what have to be taken into account when deploying the solution. Here is an example of the vSphere HA configuration from a VSA Cluster:

![vSphere HA Configuration](image)

**vSphere HA Configuration Settings**

- vSphere HA is enabled, DRS is disabled.
- Host monitoring is enabled.
- Admission control is enabled.
- Admission control policy: Percentage of cluster resources reserved as failover spare capacity is set to 33% CPU and memory for a three-node VSA Cluster and 50% for a two-node VSA Cluster.
- Virtual machine monitoring is enabled (although the VSA appliances are not monitored by vSphere HA).
- Host isolation response: Leave powered on.
- Application monitoring is disabled.
vSphere HA Considerations

What stands out from the above three-node configuration setting is that the percentage of cluster resources reserved as failover spare capacity is set to 33%. This has a direct impact on the number of virtual machines that you can run on the VSA.

This means that the CPU and memory resources of one whole node have been set aside for vSphere HA to be able to failover virtual machines. Also, you can see that the Admission Control setting above has been set to disallow violations. This means that you will only be able to run virtual machines on each ESXi 5.0 node up to a maximum of 66% of CPU and 66% of memory, assuming all ESXi hosts have the same amount of resources.

In a two-node cluster, this is even more restrictive since it sets the percentage of cluster resource to reserve to 50%, basically one complete node is set aside for failover. If you have an ESXi host failure, you will be able to run all of your non-VSA virtual machines on the remaining host or hosts. But you must keep in mind that this will be a very serious consideration when it comes to sizing the host configurations for your VSA deployment.

Lastly, you might also have noticed that EVC is enabled. This is automatically configured by the VSA installer, and ensures that all hosts are compatible from a vMotion perspective, that is you will be able to successfully move your non-VSA virtual machines between ESXi hosts. The actual EVC mode depends on the type of CPU in the ESXi hosts.
Install Considerations: Memory Over-Commit

The VSA does not provide support for ESXi host memory overcommit. This means that every virtual machine that is deployed on a VSA Cluster must be allocated full reservation of memory, that is 100% of a virtual machine’s memory (and at least the static portion of the virtual machine’s memory overhead) must be reserved. This consideration must also be taken into account when sizing how many virtual machines can be deployed on the VSA Cluster.

You should also consider vSphere HA admission control as discussed previously. The memory reservation requirements use a percentage of cluster wide memory.

As a rule of thumb, consider that each VSA appliance uses 1GB memory, assign 2.5GB to account for ESXi memory, and assign 100MB to account for static virtual machine overhead memory for each non-VSA virtual machine.

The formulas provide only a rough estimate of the likelihood of virtual machine power-on success. They do not take into account dynamic memory growth by ESXi itself above and beyond the 2.5GB estimate or VMM/VMX memory for each virtual machine above and beyond the allotted 100MB per virtual machine.

Management Considerations: VMware vCenter Server

One question that comes up frequently is - “Can I run vCenter Server on an ESXi host that is a VSA Cluster member?”

To deploy VSA, you require two or three vanilla ESXi 5.0 hosts. These hosts are freshly installed with a default 5.0 configuration and can not have any running virtual machines. This means you cannot have vCenter Server in a virtual machine on one of the ESXi hosts to do the install. vCenter Server (either physical or virtual) must reside on a host that is not being used as part of the VSA Cluster framework in order to install the VSA Cluster.

The follow up question is; “What if I build vCenter Server on another host outside of the VSA Cluster just for the install? Will VMware support migrating vCenter Server (in to one of the VSA ESXi hosts and one of shared NFS datastores which make up the VSA Cluster?”

Let us answer that on a per configuration basis.

• Two-Node VSA Configuration

You cannot run vCenter Server in a virtual machine on an NFS datastore in a two–node VSA configuration because in a two–node configuration, vCenter Server runs a special VSA service which behaves like an additional cluster member. There are now three members in the cluster and if one ESXi host fails, the other host is still able to communicate with the special VSA service running on vCenter Server. This means that the remaining node continues to run as there are still two members in the cluster. Now, if vCenter Server is running in a virtual machine and resides on the ESXi 5.0 host that fails, then you have lost two members straight away. Because you have lost a majority of the members, this brings down the whole cluster framework and for this additional reason, VMware only supports having vCenter Server running outside of the VSA Cluster.

• Three-Node VSA Configuration

This is a bit more difficult to explain. First, as there are three nodes, you do not need the special VSA service on the vCenter Server, like we saw in the two-node configuration. A VSA Cluster deployment also configures core vSphere features such as vMotion and vSphere HA. This would appear to be ideal for vCenter Server running in a virtual machine on a shared NFS datastore from the VSA. If one of the ESXi hosts goes down, and vCenter Server was running in a virtual machine on that host, then vSphere HA would restart vCenter Server on another host. And because all the NFS datastores are mirrored, even if vCenter Server was on a datastore presented from the failed host, its mirrored copy would be used from a different host, allowing vCenter Server in a virtual machine to continue to run.
There is one corner case scenario that prevents support of vCenter Server in a virtual machine running on a VSA datastore. The scenario is losing the whole VSA Cluster framework for any reason. If you lose the cluster framework, then VSA loses the ability to present any shared storage from the VSA nodes. This means that no virtual machine, including the vCenter Server, can run. As the VSA Cluster is installed and managed via a VSA plug-in in vCenter Server, without a running vCenter Server, the window into troubleshooting what is happening on the cluster is unavailable. Without vCenter Server, you cannot figure out the root cause of a complete cluster outage scenario. You would not even be able to gather logs from the VSA.

**I only have three ESXi hosts and no other equipment. What do I do?**

There is one possible configuration that you could consider in this case, if you really have no additional equipment available on-site to host the vCenter Server.

- Deploy 3 xESXi 5.0 servers.
- Deploy vCenter Server 5.0 in a virtual machine on the local storage of one of the ESXi 5.0 servers. This ESXi will not be a member of the VSA Cluster.
- Create a datacenter in vCenter Server inventory and put all three ESXi hosts in the same datacenter.
- Install VSA Manager and create a two-node VSA Cluster, omitting the ESXi host that runs vCenter Server in a virtual machine.
- The two NFS datastores from the two VSA member nodes will be automatically presented to all ESXi hosts which are in the same datacenter, including the non-VSA ESXi host, that is all three ESXi nodes will see the two NFS datastores.
- All three nodes can participate in vMotion and vSphere HA, but the non-VSA ESXi must have these features manually configured.

This is one possible solution that you might be able to use in situations where there is no additional equipment for installing the vCenter Server.

VMware hopes that this clarifies our stance on this. While the complete cluster framework failure scenario described above is just one corner case, it is a situation that might arise. Our objective with the VSA is to make it as robust as possible, but we also want to be able to troubleshoot any issues that might occur during its deployment, and to do that we must have vCenter Server available. The only way to guarantee that vCenter Server remains available is to deploy it outside of the VSA. That way, even in the case where the whole cluster is down, we still have the opportunity to fix the issue using the VSA plug-in in vCenter Server. The bottom line is that **when deploying a VSA, the vCenter Server must not be run on an ESXi 5.0 host that is participating in the cluster.** It must be run outside of the cluster framework.
Management Considerations: Multiple VSA

Another common query is how you can manage a VSA Cluster deployed at a remote site. This question came up quite a bit at VMworld® 2011, so quite a few folks are interested in deploying the VSA at various locations, but would like a centralized way of managing it.

As of the VSA 1.0 release, a single vCenter Server can only manage a single VSA Cluster. VMware has received multiple requests to allow a single vCenter Server to manage multiple VSA Clusters. This is being considered for a possible future release. In the meantime, a possible work around exists for the VSA 1.0 release.

The current solution for managing multiple VSA instances involves using linked-mode vCenter Server systems. To enable linked-mode vCenter Server systems, all the vCenter Server systems must be part of an active directory domain. In vSphere 5.0, you can link up to 10 vCenter Server systems together and manage them via a single vSphere client.

In a testing configuration within the technical marketing lab, we were able to set up linked-mode between two vCenter Server systems, each vCenter Server managing its own VSA Cluster. As we selected each vCenter Server, each had its own VSA tab, and we were able to examine their respective VSA Clusters from a single vSphere client.

As an example, start by logging into the vCenter Server VC-1. Next, select VC-1, a VSA tab. This shows the VSA Cluster details for site 1:

![Figure 29. Linked-Mode vCenter Managing VSA – VSA Manager View 1](image)
Staying logged into VC-1, if you now select VC-2 in the inventory view, and then select its VSA tab, you can see the VSA Cluster that VC-2 is managing at site 2:

![Figure 30. Linked-Mode vCenter Managing VSA – VSA Manager View](image)

Alternately, you can log into VC-2. In vSphere 5.0, you can link up to 10 vCenter Server systems together in linked-mode, meaning that 10 remote VSA Clusters could be managed from a single vSphere client back at HQ. While this might not meet every need, it should go some way towards addressing how to manage remote VSA Cluster deployments.
Resilience: Back-End Network Failure

The following section will look at the overall resilience of VSA to network outages and sort of behavior you can expect in the VSA Cluster when there are network problems. Although, the recommended deployment configuration for VSA calls for specific redundancy to be configured for the VSA Cluster (NIC teaming for the networks, redundant physical switches), it is valuable to understand what can happen during specific network outages.

Start by re-examining the VSA Cluster networking. Figure 31 shows a logical diagram detailing the different networks used by the VSA cluster, namely the front-end (appliance management and NFS server) networks and the back-end (cluster communication, volume mirroring and vMotion) networks.

![Figure 31. Appliance Networking Detail Revisited](image)
This example details a three-node VSA Cluster. This means that there are three NFS datastores presented from the cluster. We have deployed a single virtual machine for the purposes of this test. The virtual machine is running on host1, but resides on an NFS datastore (VSADs-2) presented from an appliance that is running on host3. The virtual machine is called WinXP. This diagram shows the network configuration for host3, although all hosts have the same network configuration.

And before beginning testing, take a look at the appliances and datastores from the VMware VSA Manager plug-in.
All three appliances are online, and each is presenting/exporting a datastore. Now look at the datastore view, and see the status of the datastores:

Figure 34. VSA Datastores View

All datastores are online too. The VSA Cluster is functioning optimally.

Next, we demonstrate what happens when the back-end network goes down on one of the nodes. We first bring down the back-end network on host3 (the host which has the appliance that is presenting NFS storage for WinXP). To achieve this, we bring down both network adapters used in the back-end network NIC team using an `esxcli network` command in the ESXi shell.

Figure 35. Downing the First Back End Network Adapter

Now, nothing externally visible happens to the cluster when a single back-end network adapter is removed. However, internally, a significant change occurs. The removal of a single network adapter from the team will cause the two port groups associated with vSphere Standard Switch (that is, VSA-Back End and VSA-vMotion) to utilize the same active uplink since the NIC team for one of these port groups will failover to its previously configured passive uplink. Until the failed network adapter is restored, the network traffic for the two port groups will share the same single active uplink. Next we go ahead and remove the second network adapter from the NIC team, leaving no uplinks for the back-end traffic.
If we check the UI, we can now see that both uplinks are removed from the VSS:

![Standard Switch: vSwitch1](image1)

As you can imagine, this is going to cause some impact in the VSA Cluster. Since the front-end uplinks are still intact, the vSphere HA agent on the ESXi host can continue to communicate with the other hosts. Therefore vSphere HA takes no action. But this scenario causes failover to kick-in in the VSA Cluster. Because the appliance running on this ESXi host can no longer communicate with the other nodes via the cluster communication network (e.g. heartbeats), and because it can no longer mirror to its remote replica, the appliance effectively goes offline as shown in Figure 38.

![Appliance Offline](image2)
And because the appliance is offline, so is its respective datastore. We can see from the above screenshot that the **Exported Datastores** column has changed, so that this appliance is no longer exporting any datastores. Instead, the datastore is now being exported by another appliance, the one which had the mirror copy of the datastore.

This is how datastores view looks in VSA Manager now:

![Figure 39. Datastores Degraded](image)

In the datastores view we can see two datastores degraded. Question arises why two degraded datastores since we only lost one appliance? Well, you must remember that each appliance is responsible for presenting the primary datastore as well as a mirror replica copy of another datastore. Because we have lost an appliance, we have lost both a primary and a replica, which means that we are now running with two un-mirrored datastores. That is why two datastores appear degraded.

The main point to take away from this is that although ‘we’ve had a complete back-end network outage on one cluster node (ESXi), the VSA resilience has allowed another appliance to take over the exporting of the NFS datastore via its mirror copy. This means that the ESXi hosts in the datacenter that have this datastore mounted are not impacted by the fact that the datastore is now being presented from a different appliance (because it is using the same IP address for exporting). Therefore, any virtual machines running on this datastore are also unaffected. In this example, the WinXP virtual machine that was running on the datastore **VSADs-2** continues to run just fine after the seamless failover of the datastore to the replica on another appliance as shown in Figure 40.
The error symbol against host3 in the inventory is to flag that we have lost the network. But we can see that the virtual machine continues to run on datastore VSADs-2, even though this datastore is now being exported from a different appliance in the VSA Cluster.

Finally, let’s bring the networking connections back up and let the appliance come back online.

Once the appliance comes back online, the primary volume and replica volume begin to re-synchronize automatically. There are two synchronization tasks, one for each datastore that was affected by the outage.

When the synchronization is complete, the appliance that was down will take back over the role of exporting the NFS datastore VSADs-2. This fail-back is once again seamless, and goes unnoticed by the ESXi hosts that have the datastore mounted (because the presentation is done using the same IP address), and goes unnoticed by any virtual machines (WinXP) running on the datastore.

This is another demonstration of the resilience of the VMware vSphere Storage Appliance and how it optimizes the availability of your data/virtual machines.
Resilience: Front-End Network Failure

This section of the deep dive will look at what would happen in the event of losing the front-end (management/NFS) network.

To cause this failure, we are going to do the same steps that we carried out in the previous section, namely downing the uplinks on one of the ESXi hosts/VSA cluster nodes. This time we will be downing the network adapters associated with the front-end network. The configuration is exactly the same as before, with a 3-node VSA Cluster presenting 3 distinct NFS datastores. Once again we will have single virtual machines, running on host1, but using the datastore from an appliance running on host3. As before, we will cause the outage on the ESXi host (host3) that hosts the appliance that provides the datastore on which the WinXP virtual machine resides. We will not cause any issues on the ESXi host (host1) where WinXP is running. To begin, let us down the first interface on host3:

![Figure 43. Bring Down Front-End Network Adapters](image)

Once again, while nothing externally visible happens when one of the teamed uplinks is downed, internally, this will cause the three portgroups associated with the vSphere Standard Switch (i.e., VSA-Front End, Virtual Machine Network, and Management Network) to utilize the same active uplink since the NIC team (either the VSA-Front End portgroup or both the Virtual Machine Network and Management Network will failover to its previously configured passive uplink). Until the failed network adapter is restored to health, the network traffic for the three portgroups will share the same single active uplink. Let’s now bring down the second interface on host3.

Recall that the VSA installer places all ESXi hosts that are VSA Cluster members into a vSphere HA Cluster. This provides a way for virtual machines to be automatically restarted if the ESXi host on which they were running goes down. Because you’ve just downed the uplinks for the management interfaces on this ESXi host, the vSphere HA agent that runs on the ESXi will not be able to communicate to the other hosts in the vSphere HA cluster. So, the first thing you see when the front-end network is lost are warnings from vSphere HA that it cannot communicate with the HA agent on that ESXi host:

![Figure 44. vSphere HA Agent Unreachable](image)
This is shortly followed by a **Host Failed** event/status from vSphere HA:

Because the virtual machine (WinXP) is not running on the ESXi host (host3) which failed (it is only running on the NFS datastore exported by the appliance running on the host with the network outage), vSphere HA will not attempt to restart that virtual machine on another host in the cluster.

Eventually, since vCenter Server communication to the **vpx agent** is also via the front-end network, this ESXi and the VSA appliance running on that host become disconnected from vCenter Server.
Now, because cluster communication is handled over the back-end network, and this network is unaffected by the outage, the VSA cluster will not take any corrective action in this case. It continues to export all NFS datastores from all appliances. Therefore, there is no need for another appliance to take over the presentation of the datastore from the appliance that is running on the ESXi host that has the front-end network outage.

Let’s look now at the datastore from the VSA Manager UI:

![Figure 47. VSA Manager shows Appliances Online during FE Network Failure](image)

From a VSA Cluster perspective, the datastores are online. All appliances also remain online:

![Figure 48. VSA Manager shows Datastores Online during FE Network Failure](image)

Because the front-end network is now down, the datastore can no longer be presented to the ESXi hosts in the cluster. This is because the front-end network is used for NFS exporting by the appliances.
It means that all the ESXi hosts in the datacenter have lost access to the datastore (inactive):

![Image of datastore inactive, ESXi View](image1.jpg)

Figure 49. Datastore Inactive, ESXi View

And when we look at the virtual machine, we also see that the datastore is shown as **inactive**:

![Image of datastore inactive, virtual machine view](image2.jpg)

Figure 50. Datastore Inactive, Virtual Machine View

In this network outage, the VSA cluster remains intact and functional (replication and heartbeating continues), but the front-end network outage is preventing it from presenting the NFS datastore(s) to the ESXi hosts. The virtual machine will remain in this state indefinitely until the datastore comes back online.

Essentially, a front-end network outage on an ESXi host in the cluster will mean that that datastore becomes unavailable for the duration of the outage. The virtual machines will retry I/Os for the duration of the outage, and when the network issue is addressed and the datastore comes back online, the virtual machines will resume from the point where the outage occurred.
If we bring the uplinks for the front-end network back up the virtual machine resumes from where it was before the outage:

```
$ esxcli network nic up -n vmnic2
$ esxcli network nic up -n vmnic3
```

Table: Network Adapter Information

<table>
<thead>
<tr>
<th>Name</th>
<th>PCI Device</th>
<th>Driver</th>
<th>Link</th>
<th>Speed</th>
<th>Duplex</th>
<th>MAC Address</th>
<th>MTU</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>vmnic1</td>
<td>0000:08:00.0</td>
<td>bnx2</td>
<td>Up</td>
<td>1000</td>
<td>Full</td>
<td>1c:2d:25:34:94</td>
<td>1500</td>
<td>Broadcom Corporation Broadcom NetXtreme II BCM5719</td>
</tr>
<tr>
<td>vmnic2</td>
<td>0000:08:00.1</td>
<td>bnx2</td>
<td>Up</td>
<td>1000</td>
<td>Full</td>
<td>1c:2d:25:34:96</td>
<td>1500</td>
<td>Broadcom Corporation Broadcom NetXtreme II BCM5719</td>
</tr>
<tr>
<td>vmnic3</td>
<td>0000:08:00.0</td>
<td>bnx2</td>
<td>Up</td>
<td>1000</td>
<td>Full</td>
<td>1c:2d:25:34:98</td>
<td>1500</td>
<td>Broadcom Corporation Broadcom NetXtreme II BCM5719</td>
</tr>
</tbody>
</table>

Figure 51. Bring Front-End Network Adapters Online

Figure 52. Virtual Machine Resumes Without Needing a Reboot
Resilience: ESXi Host Failure

Next, we will look at another resilience feature of the VSA, the loss of an ESXi host. A running virtual machine called WinXP is already deployed on the VSA storage. We will see how an outage to one of the ESXi hosts in the cluster (the one which hosts the VSA storage on which the virtual machine is deployed) has no impact on the running virtual machine. This is achieved because another VSA member who maintains the secondary mirror of the datastore takes over the responsibility of presenting the datastore when the primary mirror has a failure.

This particular VSA is a three-node configuration. This means there are three ESXi hosts running three VSA appliances (one each) and therefore three NFS datastores presented by the cluster to each of the ESXi hosts in the cluster.

In this particular Datastores view of the VSA Cluster, you can see that each datastore is being exported by a unique VSA member. For example, VSADs-0 is being exported by the VSA member VSA-1.

![Figure 53. Three–Node VMware VSA Configuration](image-url)
Now, we will run a command which is going to initiate an 'uncontrolled shutdown' on the ESXi host to simulate a hardware failure. The host that we are going to fail (10.20.196.26) is the host that has appliance VSA-2 running on it. By bringing down this ESXi host, we also bring down VSA-2 in an uncontrolled fashion. Since WinXP is running on an NFS datastore from this VSA (VSADs-2), we will see how the cluster handles the condition. This will cause an outage on the ESXi host on which it is run. We will then see how VSA handles this outage. What you should observe is a slight degradation in performance while the VSA cluster switches from the primary to the mirror datastore, but once the failover has successfully taken place, I/O should return to its previous performance.

From the ESXi shell on VSA-2, run the following command

```
# vsish -e set /reliability/crashMe/Panic 1
```

We can see that the host has faulted and that two of the datastores in our VSA have become degraded. This is because we have lost a VSA, and since each VSA provides both a primary and a replica, the outage will affect two datastores.

![Figure 54. Host Failure](image)

Note the red warning symbol against the host 10.201.96.26. Because the ESXi hosts are in a vSphere HA cluster, HA has detected a configuration issue - a possible host failure. This is a very similar condition to that experienced when the front-end network fails.

Even though we have lost a VSA member, which has caused degradation on two datastores in the cluster, one of which our virtual machine was running on, our virtual machine continues to operate. This is because the VSA member which was previously hosting the replica part of that datastore has now been promoted to primary, and presents the datastore back to the ESXi hosts using the same IP address which was used by the failed cluster member. Therefore the failover is transparent to the ESXi hosts, and does not impact the virtual machines running on the datastore.
Note that VSA-2 is now offline because the ESXi host on which it resides has failed. You can also see that the appliance VSA-0 is now exporting two NFS datastores, one of which is VSADs-2, which earlier were exported from the VSA on the failed ESXi, VSA-2.

So even though we have had a major server failure with one node experiencing an uncontrolled outage, the VSA is resilient enough to survive a failure of this nature and continue to do the NFS exports using mirrored volumes. That is, one of the VSA members takes the responsibility for exporting two NFS datastores in the event of another member failure. The significant business benefit of the resilience of the VSA is that it can prevent unplanned outages.
Resilience: Replacing a Failed ESXi Host

By now you should be aware that VSA can handle failures at both the ESXi host and appliance level and continue to present the full complement of NFS datastores. This means that if the ESXi host on which the appliance is running goes down, the cluster will seamlessly present that NFS datastore from another node in the cluster. This is transparent to the ESXi hosts that have the NFS datastore mounted and is transparent to any virtual machines running on that datastore.

Let us discuss what happens if you have a hardware failure on one of your ESXi hosts and the server vendor is going to take a while to ship you the replacement part. One of the features of the VSA is that it will allow you to replace an offline/failed node with a brand new ESXi host. Look at a sample two–node configuration here:

![Figure 57. ESXi Host Failure also Impacts VSA Appliance on the Host](image)

In this case, we have lost one of the nodes in a two–node cluster (and of course the appliance VSA-0 running on that node). In this case, VSA-1 will take over the presentation of the NFS datastore from VSA-0. This places both the NFS datastores into a degraded state, but the datastores are still presented to the ESXi hosts and the virtual machines on those datastores are unaffected and continue to run. The term ‘degraded’ means that datastores have no mirror copy/replica. The only issue is that both NFS datastores are now being presented from the same appliance on the same ESXi host. The VSA Manager will show the appliance as offline in the Appliances view in VSA Manager:

![Figure 58. Appliance Offline in VSA Manager](image)
To replace this node with a brand new node, and to bring the cluster out of the degraded state, select the offline appliance, right-click it and select the option to do a ‘Replace Appliance’:

![Figure 59. Replace Appliance Initiated](image)

Now, follow the wizard-driven steps to replace the offline appliance with a new appliance on a new ESXi host. Just like the installation process, the UI will show you all available ESXi 5 hosts in the datacenter. Two of these hosts are already used by the VSA Cluster (one of which is failed) and are not available for selection, but as shown in the example a third host that is not in the cluster can be used:

![Figure 60. Select a Replacement Host for the New Appliance](image)

When the networking has been configured on the replacement ESXi and the VSA appliance deployed (all of which is done automatically), the volume and replica are created on the new appliance and synchronized with the NFS volumes already in the VSA Cluster.
When all this is completed, the **Replace Appliance** wizard will display the following:

![Replace Appliance][1]

*Figure 61. Appliance is Now Replaced*

The VSA Cluster is now back to its optimal state. So, even though a node in the VSA Cluster may suffer a hardware failure, procedures have been built into the VSA Cluster to help customers keep it highly available, allowing a failed node to be swapped out of the cluster for a new healthy server. And this can be done while the VSA Cluster continues to present a full complement of NFS datastores.
Resilience: VMware vCenter Server Failure

You have seen how the VSA handles failures of a host or appliance in the cluster. But there are a number of other features and configuration settings that make the VSA even more resilient. These include:

- **RAID** configured at the physical disk level to avoid single spindle failures impacting the cluster.
- **RAID1** at the appliance volume level, so that if an appliance fails, the NFS datastore that it was presenting can still be presented by another appliance.
- Ability to **replace** a broken cluster host with a brand new ESXi host.

The one thing we didn’t mention is vCenter. What if your vCenter Server had a major failure and you didn’t have a backup? Do I really have to tear down and rebuild the VSA Cluster? Will I have to destroy my NFS datastores and recreate them? Is vCenter Server a single point of failure?

The answer is **no** in all cases. Again, this scenario has been thought of and a built-in procedure can pull the VSA Cluster configuration back into vCenter Server in a few steps.

The VSA Cluster continues to run during this time - it is unaffected by the loss of the vCenter Server. The only thing we have lost from a VSA Cluster perspective is the ability to manage and monitor it. The NFS datastores continue to be presented.

To begin the recovery, you need to start off with a re-install of vCenter Server and the VSA Manager. You then enable the VSA Manager plugin like you do during a standard installation (this will need a datacenter object created in the vCenter Server inventory), but this time you will select the **Recover VSA Cluster** option rather than new installation as shown below:

![Figure 62. vCenter Server Reinstalled, Recover VSA Cluster](image)

Some information is required such as the VSA Cluster IP address, username and password (by default svaadmin and svapass respectively), as well as the root password of the ESXi hosts in the VSA Cluster (they must all have the same root password).
You populate this info as shown below:

The procedure will create a new datacenter called **VSADC**, and will add back in the ESXi nodes to the inventory, along with VSA appliances.
When you click **Close** button, the VSA Manager will be launched and everything from a VSA Cluster perspective will be similar to the way it was before the vCenter Server instance was lost.

One final point to note; the initial VSA Installation also created a vSphere HA cluster, and added the hosts to it. This step is not done by the **Recover VSA Cluster** procedure. You will need to do this step manually afterwards.

## Conclusion

VMware vSphere Storage Appliance allows users to get the full range of vSphere features, including vSphere HA, vMotion and DRS without having to purchase a physical storage array to provide shared storage, making VSA a very cost-effective solution.

VSA is very **easy to deploy**, with many of the configuration tasks such as network setup and vSphere HA deployment being done by the installer. The benefit here is that this product can be deployed by customers who may not be well versed in vSphere and give them a good first time user experience.

VSA is very **resilient**. If an ESXi host which is hosting one of the VSAs goes down or one VSA member suffers an outage, with the **redundancy** built into the VSA, the NFS share presented from that VSA will be automatically and seamlessly failed over to another VSA in the cluster.

## About the Author

Cormac Hogan is a Senior Technical Marketing Manager responsible for storage in the cloud infrastructure product marketing group at VMware. His focus is on core VMware vSphere storage technologies and virtual storage in general, including the VMware vSphere Storage Appliance. He was one of the first VMware employees at our EMEA HQ in Cork, Ireland, back in April 2005. He spent two years as the Technical Support Escalation Engineer for Storage before moving into a Support Readiness Training role, where he developed training materials and delivered training to Technical Support and our Support Partners. He has been in Technical Marketing since 2011.

- Follow Cormac on Twitter: [@VMwareStorage](http://twitter.com/VMwareStorage)