ESX Server 3.0 Tips and Tricks

Mostafa Khalil, VCP
VMware Product Support Engineering
Agenda

- Storage Area Network
- VMFS-3
- iSCSI Storage
- NFS
Storage Area Network (SAN)

- Components
- Topology
- Identification
SAN Components

- Initiators (host HBAs)
  - QLogic 1/2/4Gb
  - Emulex 1/2/4Gb
  - See HCL for supportability
  - Set HBA’s BIOS to “Point-to-Point” or similar to that

- Targets (Storage Processors’ Ports)
  - Active/Active Arrays
  - Active/Passive Arrays
  - Check the HCL for supportability

- Fabrics (Switches and Fibre Connections)
  - No specific listing on HCL
  - 1/2/4Gb switches
  - May need to set switch port to “Fx” or equivalent (e.g. F-Port)
SAN Topology – Point-to-Point

- AKA Direct Connect
- Not supported
- Except CX100
SAN Topology – Absolute Minimum

- Switch allows additional connections of more storage or more servers
- No redundancy
  - Many single points of failure
- Supported by VMware for ESX Server as a bare minimum configuration
  - If the Storage Arrays are not listed on the hardware compatibility list (HCL), VMware only supports them with this simple configuration
SAN Topology – Arbitrated Loop

- Not Supported
SAN Topology – Fully Redundant Fabrics

- Recommended

Storage

SP1

SP2

FC Switch

HBA

HBA

Server

HBA

HBA

Server
Identifying SAN Configuration

- VI Client
- Console
- **Multipath Analysis**

  - **Canonical Name**: `vmhbaC:T:L`
  - **In this example**: `Vmhb0a:0:0`
  - **Policy**: MRU or Fixed
    - **In this Example**: MRU
  - **Path States**:
    - **On**
    - **Off**
    - **Standby**
    - **Dead**

  ![Manage Paths Window]

  - **Device** | **SAN Identifier** | **Status** | **Preferred**
  - `vmhba0:0:0` | `20:06:00:a0:b8:17:4e:dd` | Active |
  - `vmhba0:1:0` | `20:07:00:a0:b8:17:4e:dc` | Standby |
  - `vmhba1:0:0` | `20:06:00:a0:b8:17:4e:dc` | Standby |
  - `vmhba1:1:0` | `20:07:00:a0:b8:17:4e:dd` | Standby |

  ![Command Output]

  ```
  [root@giza root]# esxconf-mpath -l
  Disk `vmhba0:0:0' /dev/sda (512MB) has 4 paths and policy of Most Recently Used
  FC 2:2.0 21000012593a86a<->20060a0b8174edd vmhba0:0:0 On active preferred
  FC 2:2.0 21000012593a86a<->20070a0b8174edd vmhba0:1:0 Standby
  FC 2:2.1 21000012593a86b<->20060a0b8174edd vmhba1:0:0 On
  FC 2:2.1 21000012593a86b<->20070a0b8174edd vmhba1:1:0 Standby
  ```
### Multipath Analysis - Identifying Targets

<table>
<thead>
<tr>
<th>SAN Type</th>
<th>Port ID</th>
<th>WWN</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>IBM FAStT</strong></td>
<td>N/A</td>
<td>20:0X:00:00:00:00:xx</td>
</tr>
<tr>
<td></td>
<td>N/A</td>
<td>20:0Z:00:00:00:00:zz</td>
</tr>
</tbody>
</table>

For **IBM FAStT**:
- **Compare X and Z**: Higher number is the secondary processor.
- **Compare xx and zz**: Higher number is the higher number port.
VI Client – Multipath Analysis (Target Failover)

- Canonical name not equal to active path → Failover
- Target Failover example

<table>
<thead>
<tr>
<th>Device</th>
<th>SAN Identifier</th>
<th>Status</th>
<th>Preferred</th>
</tr>
</thead>
<tbody>
<tr>
<td>vmhba0:0.0</td>
<td>20:06:00:a0:b8:17:4e:dd</td>
<td>Active</td>
<td></td>
</tr>
<tr>
<td>vmhba0:1.0</td>
<td>20:07:00:a0:b8:17:4e:dc</td>
<td>Standby</td>
<td></td>
</tr>
<tr>
<td>vmhba1:0.0</td>
<td>20:06:00:a0:b8:17:4e:dc</td>
<td>Standby</td>
<td></td>
</tr>
<tr>
<td>vmhba1:1.0</td>
<td>20:07:00:a0:b8:17:4e:dd</td>
<td>Disabled</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Device</th>
<th>SAN Identifier</th>
<th>Status</th>
<th>Preferred</th>
</tr>
</thead>
<tbody>
<tr>
<td>vmhba0:0.0</td>
<td>20:06:00:a0:b8:17:4e:dd</td>
<td>Standby</td>
<td></td>
</tr>
<tr>
<td>vmhba0:1.0</td>
<td>20:07:00:a0:b8:17:4e:dc</td>
<td>Active</td>
<td></td>
</tr>
<tr>
<td>vmhba1:0.0</td>
<td>20:06:00:a0:b8:17:4e:dc</td>
<td>Standby</td>
<td></td>
</tr>
<tr>
<td>vmhba1:1.0</td>
<td>20:07:00:a0:b8:17:4e:dd</td>
<td>Disabled</td>
<td></td>
</tr>
</tbody>
</table>
VI Client – Multipath Analysis (HBA Failover)

- Canonical name not equal to active path -> Failover
- HBA Failover example
Console – Multipath Analysis (Target Failover)

- esxcfg-mpath -l
- Console View of the target failover event
Console – Multipath Analysis (HBA Failover)

- `esxCFG-mpath -l`
- Console View of the HBA failover event

```
[root@giza root]# esxCFG-mpath -l
Disk vmhba0:0:0 /dev/sda (512MB) has 4 paths and policy of Most Recently Used
FC 2:2.0 210000112593a86a<->200600a0b8174edd vmhba0:0:0 On active preferred
FC 2:2.0 210000112593a86a<->200700a0b8174edc vmhba0:1:0 Standby
FC 2:2.1 210000112593a86b<->200600a0b8174edc vmhba1:0:0 On
FC 2:2.1 210000112593a86b<->200700a0b8174edd vmhba1:1:0 Standby
```

```
[root@giza root]# esxCFG-mpath -l
Disk vmhba0:0:0 /dev/sda (512MB) has 4 paths and policy of Most Recently Used
FC 2:2.0 210000112593a86a<->200600a0b8174edd vmhba0:0:0 On preferred
FC 2:2.0 210000112593a86a<->200700a0b8174edc vmhba0:1:0 Standby
FC 2:2.1 210000112593a86b<->200600a0b8174edc vmhba1:0:0 On active
FC 2:2.1 210000112593a86b<->200700a0b8174edd vmhba1:1:0 Standby
```
VMFS-3

- Journaling
- Hierarchical
- Metadata
- LVM
- Snapshot handling
- VMotion with Storage Migration
- New Virtual Disk Types
Journaling

- VMFS-3 is now a journal based file system
- Better recovery from crashes
- Online File System Check
Hierarchical Structure

- VMFS-3 volumes support directory structure
- Virtual Machine files stored in the same directory
  - Configuration file (VMX)
  - VM Swap File
  - Virtual Disks
  - Logs
  - Monitor dump
VMFS-3 Metadata

- Copies of metadata stored at the root level of each volume
- Backup these files on a regular basis
- Run vm-support on a regular basis
  - Collects Partition Table info

```
root@giza:/vmfs/volumes/IBM-80-FS3
[root@giza IBM-80-FS3]# ls -a *.sf
.fbb.sf .fdc.sf .pbc.sf .sbc.sf .vh.sf
[root@giza IBM-80-FS3]# 
```
Spanned VMFS volumes
  • Volume still valid with missing extent(s)
Snapshot LUN handling
VMFS file system modules
  • vmfs2
  • vmfs3
  • fsaux
A volume UUID looks like this:
42263200-74382e04-b9bf-009c06010000

Also Known As “Signature”

Created as a response to non-persistent vmhba names

Volume names are supposed to be used instead of UUIDs

Not to be confused with host UUIDs
Snapshot LUN Handling

- Basic rules:
  - A LUN shared by multiple hosts MUST be presented with the SAME LUN ID to all hosts
  - A VMFS volume signature is associated with the LUN ID and the UUID among a few other elements (this may change at a later release)
- AX100 and IBM ESS (Shark) may not meet these rules.
- Symptoms
  - Volume hidden
  - Volume identified as a snapshot
Snapshot LUN Handling

- Sample log entries

Jul 18 10:58:31 ServerB vmkernel: 0:14:17:59.787 cpu13:1046)LVM: 5676: disk ID: <type 2, len 22, lun 5, devType 0, scsi 5, h(id) 10179760818951437974>
Jul 18 10:58:31 ServerB vmkernel: 0:14:17:59.787 cpu13:1046)LVM: 5678: m/d disk ID: <type 2, len 22, lun 1, devType 0, scsi 5, h(id) 10179760818951437974>
Jul 18 10:58:31 ServerB vmkernel: 0:14:17:59.790 cpu13:1046)LVM: 5676: disk ID: <type 2, len 22, lun 6, devType 0, scsi 5, h(id) 11552037668126695191>
Jul 18 10:58:31 ServerB vmkernel: 0:14:17:59.790 cpu13:1046)LVM: 5678: m/d disk ID: <type 2, len 22, lun 2, devType 0, scsi 5, h(id) 11552037668126695191>
Jul 18 10:58:31 ServerB vmkernel: 0:14:17:59.794 cpu13:1046)LVM: 5676: disk ID: <type 2, len 22, lun 7, devType 0, scsi 5, h(id) 13372428508588014685>
Jul 18 10:58:31 ServerB vmkernel: 0:14:17:59.794 cpu13:1046)LVM: 5678: m/d disk ID: <type 2, len 22, lun 3, devType 0, scsi 5, h(id) 13372428508588014685
Correcting Hidden Volumes

- See KB 6482648 at: http://kb.vmware.com/kb/6482648

- Correct LUN ID so that all hosts “see it” with the same LUN number
  - Symmetrix: Assign the LUNs to the FAs using the same LUN number
  - Clariion: Assign the LUNs to the same Storage Group
  - IBM DS4000/FAStT family: Assign the LUNs to the same Host Group (Logical Partitioning option enabled)
  - Other Arrays: use equivalent features to present the LUN with the same Host LUN ID to all hosts sharing it

- If the above is not possible:
  - Set advanced option LVM.DisallowSnapshotLUN to “0” then rescan
  - **DO NOT** present actual snapshot LUNs to that server with this option setting in place. Data Corruption may result
AutoResignature

- Auto Resignature feature is disabled by default
- To enable it set advanced option LVM.EnableResignature to “1” then rescan
- This will apply to ALL volumes visible to this host.
- This will affect ALL hosts sharing this volume
- All VMs configurations need to be edited to reflect the new volume ID
- This will be addressed in a future release
VMFS Kernel Modules

- VMFS2
  - Loaded for “Read-Only” by default for all VMFS2 volumes
  - Can be unloaded after all VMFS2 volumes have been upgraded
- VMFS3
  - All operations related to VMFS3 volumes
- FSAUX
  - Auxillary File System functions
  - Upgrading VMFS2 to VMFS3
  - Other functions internal to VMware
VMFS3 – Virtual Disk Structure Changes

- Virtual machine virtual disk now has a descriptor file which is in ASCII format
- ESX 2.x contained this information in last 512 bytes of vmdk

```
root@giza:/vmfs/volumes/ax100i-100GB/Windows2000-RDM$
# cat Windows2000-RDM.vmdk

# Disk DescriptorFile
version=1
CID=877e825d
parentCID=ffffffff
createType="vmfsParststhroughRawDeviceMap"

# Extent description
RW 52428800 VMFSRDM "Windows2000-RDM-rdmp.vmdk"

# The Disk Data Base
#DDB

ddb.virtualHWVersion = "4"
ddb.geometry.cylinders = "3263"
ddb.geometry.heads = "255"
ddb.geometry.sectors = "63"
ddb.adapterType = "buslogic"
ddb.toolsVersion = "7172"
[root@giza Windows2000-RDM]$
```
## VMFS3 – Virtual Disk Structure Changes

<table>
<thead>
<tr>
<th>Format</th>
<th>Example</th>
</tr>
</thead>
</table>
| createType="<type>  
# Extent description  
RW <size in 512bytes blocks> <extent>" | createType="vmfsPassthroughRawDeviceMap"  
# Extent description  
RW 52428800 VMFSRDM "Win2K-rdmp.vmdk" |

<table>
<thead>
<tr>
<th>Virtual Disk Type</th>
<th>Extent</th>
</tr>
</thead>
<tbody>
<tr>
<td>vmfsPassthroughRawDeviceMap</td>
<td>VMFSRDM “&lt;name&gt;-rdmp.vmdk”</td>
</tr>
<tr>
<td>vmfsRawDeviceMap</td>
<td>VMFSRDM “&lt;name&gt;-rdm.vmdk”</td>
</tr>
<tr>
<td>vmfs (Thin / Thick)</td>
<td>VMFS “&lt;name&gt;-flat.vmdk”</td>
</tr>
</tbody>
</table>
Sample Virtual Disk

```
[root@giza Windows2000-RDM]# vmkfstools -c lg -d thin -a lsilologic thin.vmdk
```

```
[root@giza Windows2000-RDM]# cat thin.vmdk
# Disk DescriptorFile
version=1
CID=4d4662f9
parentCID=ffffffff
createTime="vmfs"

# Extent description
RW 2097152 VMFS "thin-flat.vmdk"

# The Disk Data Base
#DDB

ddb.virtualHWVersion = "4"
ddb.geometry.cylinders = "512"
ddb.geometry.heads = "128"
ddb.geometry.sectors = "32"
ddb.adapterType = "lsilologic"
ddb.thinProvisioned = "1"
```
Question?
**iSCSI Storage**

- Block-level I/O over TCP/IP using SCSI-3 protocols

**Pros**
- Direct access to random blocks on disks or LUNs for performance
- Block and file traffic on single Ethernet infrastructure
- Enabled by adoption of Gigabit Ethernet
- Can use normal IP authentication, encryption, routing features

**Cons**
- Protocol overhead added by TCP/IP
  - Small packets, connection-oriented chatter
  - Gigabit traffic can saturate a 2.4-GHz CPU
- Slower than Fibre Channel (for now)

**TCP Offload Engines (TOE)**
- Reduce CPU overhead, but expensive
- Plain Gigabit NICs catching up with better drivers
iSCSI Requirements

- Install a VMotion and IP Storage licenses on VC
- Host and storage ports on the same subnet
- A supported Ethernet NIC OR
- A QLogic 4050 or other card on the HCL (Experimental)
- A supported iSCSI Storage
iSCSI Configuration

- **Authentication**
  - CHAP
    - Per HBA or per target
  - None
- **Target discovery**
  - Static targets
  - SendTargets discovery (Dynamic)
  - No SLP discovery
  - No iSNS discovery
- **Administration**
  - SNIA-based IMA library
  - Configured through VMware Virtual Infrastructure Client
  - Common approach for all initiators
  - No need for vendor-specific tools
iSCSI – Configuring HW Initiator

- Select: Configuration – Storage Adapters
- Select the HBA and click Properties (in details pane)
- Click “Configure”
iSCSI – Configuring SW Initiator

- COS port group (for Authentication)
- VMKernel Port Group (for Data Traffic)
- Initiator IQN
- Security
- Discovery
- Multipathing
  - Only with both ports on the same switch and subnet
Configuring Network

Virtual Switch: vSwitch1

VMkernel Port
iSCSI
10.16.86.10

Service Console Port
Service Console for iSCSI
vswif1: 10.16.86.9

Physical Adapters
vmnic1 1000 Full
Enable software iSCSI client in firewall
Enable "Software iSCSI Client" in firewall
iSCSI – SW Initiator

- Enable the Software Initiator
  - Storage Adapters – Select “iSCSI” Software Adapter
Enable the Software Initiator

> Storage Adapters – Select “iSCSI” Software Adapter

> Select “properties”
iSCSI – SW Initiator

- Enable the Software Initiator
  > Storage Adapters – Select “iSCSI” Software Adapter

- Select “properties”
- Select “Configure”
- Click “Enabled” then Click “OK”
iSCSI – SW Initiator

- iSCSI properties get filled automatically
- Select “Dynamic Discovery”
- Select “Add” then input the iSCSI Server’s IP address
- Repeat for each port on the iSCSI storage
Select “Add” then input the iSCSI Server’s IP address
iSCSI – Dynamic Discovery

- Select “Add” then input the iSCSI Server’s IP address
- Repeat for each port on the iSCSI storage
iSCSI – Static Discovery

- Same as Dynamic but select the “Static” tab instead
- Works with HW initiators ONLY
iSCSI – Authentication

- Select the “CHAP Authentication” Tab then “Configure”
iSCSI – Troubleshooting

- vmkiscsi-tool
  - Usage: vmkiscsi-tool <command> <subcommand> adapterName

shows the Initiators properties
Shows Discovery Properties
If this were a Hardware Initiator, Static Targets would have been listed
iSCSI – Troubleshooting

- Shows iSCSI Node Name (IQN)
Lists Targets

[root@giza root]# vmkiscsi-tool -T -l vmhba40
### iSCSI – Troubleshooting

<table>
<thead>
<tr>
<th>Name</th>
<th>IQN: 1992-04.com.emc:ax.apm00054207419.b0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alias</td>
<td>7419.b0</td>
</tr>
<tr>
<td>Discovery Method Flags</td>
<td>0</td>
</tr>
<tr>
<td>Send Targets Discovery Settable</td>
<td>0</td>
</tr>
<tr>
<td>Send Targets Discovery Enabled</td>
<td>0</td>
</tr>
<tr>
<td>Portal 0</td>
<td>10.16.92.37:3260</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Name</th>
<th>IQN: 1992-04.com.emc:ax.apm00054207419.a0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alias</td>
<td>7419.a0</td>
</tr>
<tr>
<td>Discovery Method Flags</td>
<td>0</td>
</tr>
<tr>
<td>Send Targets Discovery Settable</td>
<td>0</td>
</tr>
<tr>
<td>Send Targets Discovery Enabled</td>
<td>0</td>
</tr>
<tr>
<td>Portal 0</td>
<td>10.16.92.36:3260</td>
</tr>
</tbody>
</table>

- Lists Targets
Lists LUNs
iSCSI – Troubleshooting

Lists LUNs
Shows Node Alias then shows Authentication Method
Q&A
NAS and NFS – Overview

- Use Network FileSystem Protocol
- NFS 3 TCP only (no UDP)
- Supported NAS filers only but also works with Linux NFS 3.x
- No CIFS/SMB yet
- NFS Naming Convention
  - `nfs.remote.com:/remote/filesystem`
- Locking Handled by VMkernel
- Lease-based locks
- 8 NFS mounts by default. Can be increased to 32
NFS
Configuring datastore

- Configure the network
NFS
Configuring datastore (cont.)

- In “Storage” pane select “add” then “Network File System”
NFS
Configuring datastore (cont.)

- Fill in the NFS info
NFS
Configuring datastore (cont.)

- Now the storage shows the new DataStore
Mounted in the VMkernel NOT on the Service Console
No need to modify /etc/fstab
Recognized by vmkfstools as an NFS volume
Virtual Machines can be stored there
NAS Tips and Tricks

- ESX Server needs full access to NFS datastores to create directories, set permissions
  - turn off root squash
- 8 NFS mounts per ESX Server allowed by default. To increase
  - Select host from inventory, right-click “Advanced Settings” and select “NFS”
  - Adjust “NFS.MaxVolumes”

- Avoid VM swapping to NFS volumes
  - Edit VM config file to add:
    - `sched.swap.dir = /vmfs/volumes/<volume_name>/<dir_name>`
  - Where <volume_name> is a VMFS3 volume
Troubleshooting NAS

- Verify NAS filer/host configuration
- Make sure ESX Server is on the (Read/Write) host list on the filer
- Use ethereal to monitor NFS traffic
- On ESX Server use `tcpdump` to get a trace
- Verify the NFS connection properties on ESX Server
- To list Linux NFS host throughput, use:
  
  ```
  > hdparm -tT /dev/sd<x>
  ```
Questions?
Please remember to complete your **session evaluation form** and return it to the room monitors as you exit the session.

The presentation for this session can be downloaded at [http://www.vmware.com/vmtn/vmworld/sessions/](http://www.vmware.com/vmtn/vmworld/sessions/)

Enter the following to download (case-sensitive):

- **Username:** cbv_rep
- **Password:** cbvfor9v9r
Some or all of the features in this document may be representative of feature areas under development. Feature commitments must not be included in contracts, purchase orders, or sales agreements of any kind. Technical feasibility and market demand will affect final delivery.