An Oracle Technical White Paper
March 2014, v1.2

Best Practices for Oracle ZFS Storage Appliance and VMware vSphere 5.x

ORACLE

Introduction

This white paper provides best practices and recommendations for configuring VMware vSphere 5.x with Oracle ZFS Storage Appliance to reach the optimal I/O performance and throughput.

The outlined best practices and recommendations highlight configuration and tuning options for Fibre Channel, NFS and iSCSI protocols for a VMware vSphere 5.x environment working with an Oracle ZFS Storage Appliance. The paper also includes recommendations for correct design of network infrastructure for VMware cluster as well as multi-pool configurations, and recommended data layout for virtual machines. The paper demonstrates the utilization of VMware linked clone technology with Oracle ZFS Storage Appliance.

Highlighted in this paper are:

- Best practices and recommendations for employing VMware vSphere 5 with Oracle ZFS Storage Appliance
- Tuning options for Fibre Channel, iSCSI and NFS protocols in production environments
- IP network design for NFS storage as well as Fibre Channel and iSCSI protocols
- VMware cluster recommendations for high availability and load balancing
- Clone operation using VMware linked clone and Oracle ZFS Storage Appliance
- VMware virtual machine data layout
- Monitoring options using VMware's esxtop tool and Oracle ZFS Storage Appliance's DTrace Analytics

NOTE: References to Sun ZFS Storage Appliance, Sun ZFS Storage 7000, and ZFS Storage Appliance all refer to the same family of Oracle ZFS Storage Appliances.

About Oracle ZFS Storage Appliance

The basic architectural features of Oracle ZFS Storage Appliance are designed to provide high performance, flexibility and scalability. The Oracle ZFS Storage Appliance provides multiple connectivity protocols for data access, including: Network File System (NFS), Common Internet File System (CIFS), Internet Small Computer System Interface (iSCSI), InfiniBand (IB), and Fibre Channel (FC). It also supports the Network Data Management Protocol (NDMP) for backing up and restoring data. The Oracle ZFS Storage Appliance architecture also offers the Hybrid Storage Pool (HSP) feature, in which direct random access memory (DRAM), flash and physical disks are seamlessly integrated for efficient data placement (see Figure 1). A powerful performance monitoring tool called DTrace Analytics provides details about the performance of the various components, including network, storage, file systems, and client access. The tool also offers plenty of drill-down options that allow administrators to monitor specific rates of latency, size of transfer, and utilization of resources. The Oracle ZFS Storage Appliance provides a variety of RAID protections to balance the capacity, protection, and performance requirements of the applications, databases, and virtualized environments.

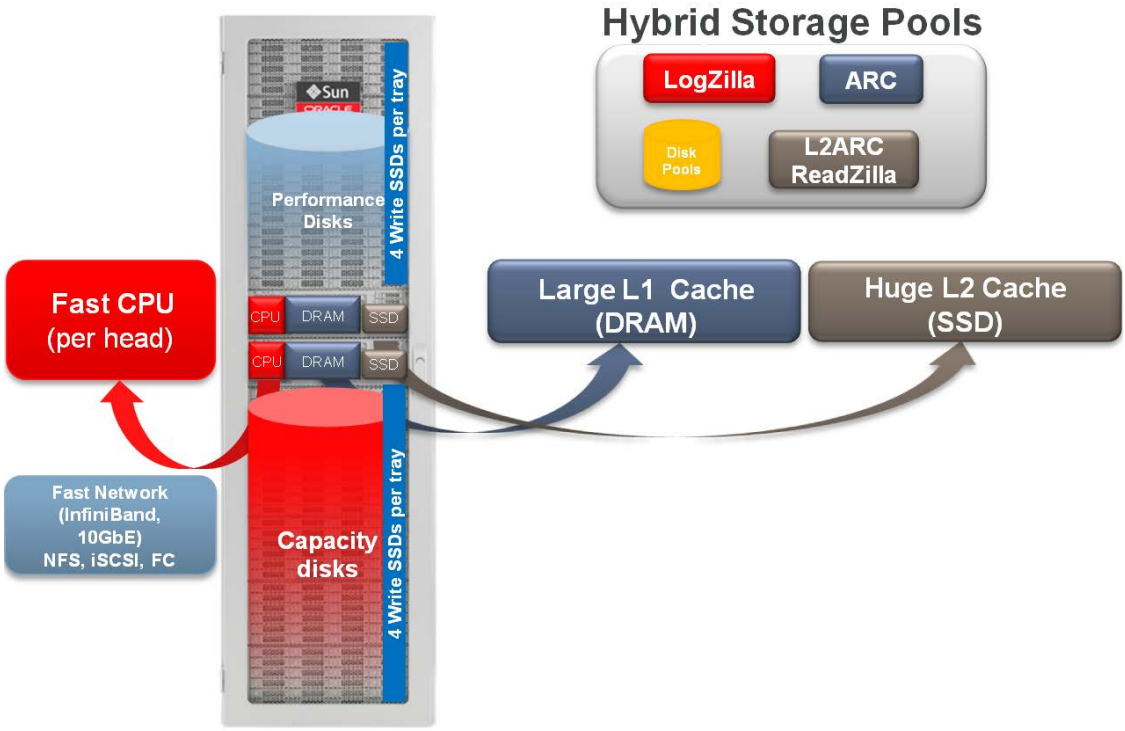

Figure 1. Oracle ZFS Storage Appliance – architecture overview

Overview of Example System Components

The following tables describe the hardware configuration, operating systems, and software releases utilized by this white paper.

Table 1 shows the hardware used.

TABLE 1. HARDWARE USED IN REFERENCE ARCHITECTURE

EQUIPMENT	QUANTITY	CONFIGURATION
Storage	1 cluster (2 controllers)	Sun ZFS Storage 7420 cluster 256 GB DRAM per controller 2 x 512 GB read cache SSD per controller 2 x 20 2TB SAS-2 disk trays 2 x dual port 10GbE NIC 2 x dual port 8Gbps FC HBA 2 x 17GB log device
Network	2	10 GbE network switch
Server	2	Sun Fire X4440 Server 256 GB DRAM 2 internal HDDs 1 x dual port 10GbE NIC 1 x dual 8Gbps FC HBA

Table 2 shows the virtual machine components used.

TABLE 2. VIRTUAL MACHINE COMPONENTS USED IN REFERENCE ARCHITECTURE

OPERATING SYSTEM	QUANTITY	CONFIGURATION
Microsoft Windows 2008 R2 (x64)	1	Microsoft Exchange Server
Oracle Linux 6.2	1	ORION: Oracle I/O Numbers Calibration Tool

Table 3 shows the software used.

TABLE 3. SOFTWARE USED IN REFERENCE ARCHITECTURE

SOFTWARE	VERSION
Oracle ZFS Storage Appliance Software	2011.04.24.4.0,1-1.21
Microsoft Exchange Server Jetstress verification tool	2010 (x64)
ORION: Oracle I/O Numbers Calibration Tool	11.1.0.7.0

VMware vCenter Server	5.1.0 (Build 880146)
VMware ESX hypervisor software	5.1.0 (Build 799733)

Best Practices for VMware vSphere 5 NFS

This section provides best practices and recommendations for VMware vSphere 5 using NFS protocol and Oracle ZFS Storage Appliance.

Oracle ZFS Storage Appliance Settings

The following configurations for the Oracle ZFS Storage Appliance are recommended to optimize performance with VMware vSphere 5.

Controllers, Software Release and Disk Pools

Virtual desktop infrastructures produce high random I/O patterns and need high storage performance as well as availability, low latency, and fast response time. To meet these demands, use a **mirrored** data profile. This configuration duplicates copies as well as produces fast and reliable storage by dividing access and redundancy usually between two sets of disks. In combination with write SSDs' log devices and the Oracle ZFS Storage Appliance architecture, this profile can produce a large amount of input/output operations per second (IOPS) to attend to the demand of critical virtual desktop environments.

The recommended minimum disk storage configuration for VMware vSphere 5.x includes:

- A mirrored disk pool of (at least) 20x300/600 or 900GB (10000 or 15000 RPM performance disks) or 44x3TB SAS-2 (7200 RPM capacity disk drives) with at least two 73GB SSD devices for LogZilla working with a stripped log profile.
- At least 2x512GB for L2 cache (L2ARC) – Striped cache.

Note: The example demonstrates 44x3TB SAS-2 7200 RPM disks. See figures 2, 3 and 4.

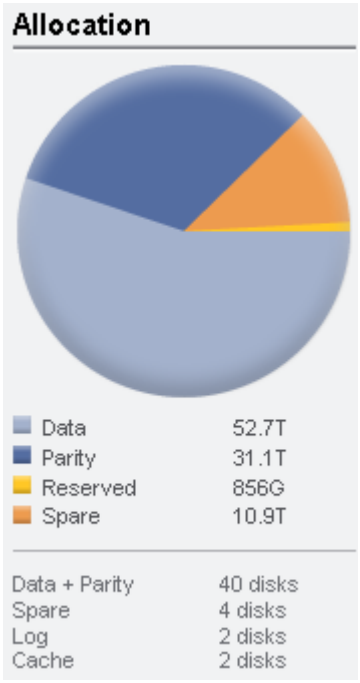

Figure 2. Oracle ZFS Storage Appliance – disk pools configuration

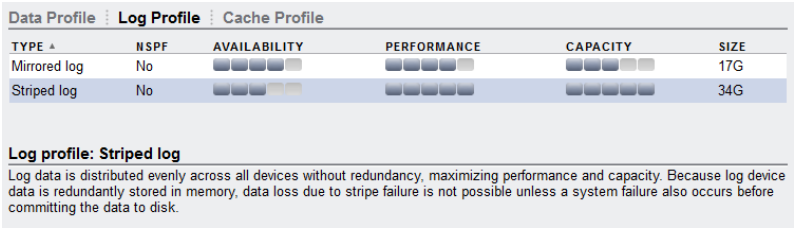

Figure 3. Oracle ZFS Storage Appliance – log profile configuration

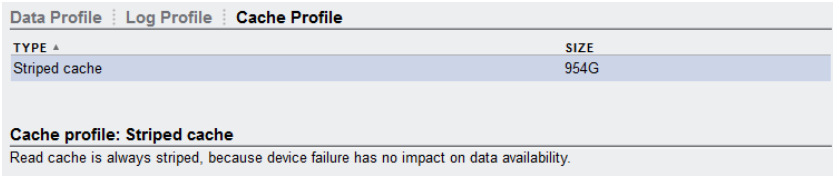

Figure 4. Oracle ZFS Storage Appliance – cache profile configuration

Note: For high availability and proper load balancing for a virtual desktop infrastructure, use an Oracle ZFS Storage Appliance model that supports clustering. Configure the cluster in active/active mode and use Oracle ZFS Storage Appliance software release 2011.1.4.2.x or greater.

The Oracle ZFS Storage Appliance software releases can be downloaded from the following URL:

http://www.oracle.com/technetwork/server-storage/sun-unified-storage/downloads/index.html

If you are working with DE2-24C/P drive enclosure models, ensure that the system is working with Oracle ZFS Storage Appliance software release 2011.1.5.0.x or greater. Please refer to the following link for additional information:

https://wikis.oracle.com/display/FishWorks/ak-2011.04.24.5.0+Release+Notes

Also, additional information on an Oracle ZFS Storage Appliance cluster configuration can be found in the *Sun ZFS Storage 7000 System Administration Guide* at:

http://docs.oracle.com/cd/E26765_01/html/E26397/index.html

CPU, L1 and L2 Cache

The following combination and sizing of CPU, L1 (ARC) and L2 (L2ARC) is critical to meet the demands of compression and deduplication operations as well as overall performance in large deployments of a virtual desktop infrastructure. The minimum recommended configuration is:

- At least two 2GHz Intel® Xeon CPUs (X7550 @ 2.00GHz) per Oracle ZFS Storage Appliance head
- At least 512GB of DRAM memory (L1 cache) per head
- At least two 512GB SSD for ReadZilla cache (L2 cache) per head

Network Settings

To ensure that the network configuration where the NFS and iSCSI traffic will run has been designed to archive high availability and no single point of failure:

- Isolate the storage traffic from other networking traffic. You can configure this utilizing VLAN, network segmentation, or dedicated switches for NFS and iSCSI traffic only.
- On the Oracle ZFS Storage Appliance, configure at least two physical 10GbE (dual-port) NICs per head, bundled into a single channel using the IEEE 802.3ad Link Aggregation Control Protocol (LACP) with a large maximum transmission unit (MTU) jumbo frame (9000 bytes). If you are working with a cluster configuration, configure at least two 10GbE (dual-port) NICs per head, and also use an IP network multipathing (IPMP) configuration in combination with LACP.
- With IPMP configuration you will achieve network high availability, and with link aggregation you will obtain a better network performance. These two technologies complement each other and can be deployed together to provide benefits for network performance and availability for virtual desktop environments.
- For picking an outbound port based on source and IP addresses, utilize LACP policy L3.
- For switch communication mode, use the LACP active mode, which will send and receive LACP messages to negotiate connections and monitor the link status.
- Use an LACP short timer interval between LACP messages, as seen in the configuration in Figure 5.

Note: Some network switch vendors do not support LACP protocol. In this situation, set the LACP mode to "Off." Please refer to your switch vendor documentation for more information.

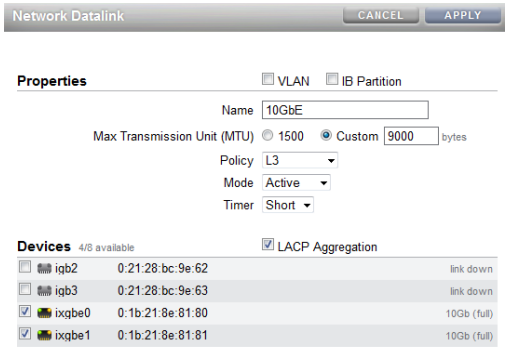

Figure 5. LACP, jumbo frame and MTU configurations on the Oracle ZFS Storage Appliance

NFS, Projects and Shares

When working with an Oracle ZFS Storage Appliance with more than one disk shelf, try to split the workload across different disk pools and use the 'no single point of failure' (NSPF) feature. This design will provide you with more storage resources as well as better I/O load balancing, performance, and throughput for your virtualized environment.

The following example/performance test uses only one disk shelf, a mirrored storage pool, one project and six different NFS shares. Table 4 lists the pool's projects and filesystem shares.

TABLE 4. PROJECTS AND FILESYSTEM SHARES CREATED FOR PERFORMANCE TEST

POOL NAME	PROJECTS	FILESYSTEMS
Pool1	winboot	/export/winboot
	vswap	/export/vswap
	ms-exchangedb	/export/ms-exchangedb
	ms-log	/export/ms-log
	linux-os	/export/linux-os
	oltp-db	/export/oltp-db

Figure 6 shows the share configuration and figure 7 shows the filesystem and mountpoint configurations on the Oracle ZFS Storage Appliance browser user interface (BUI) for the performance tests. Details for the configuration choices follow.

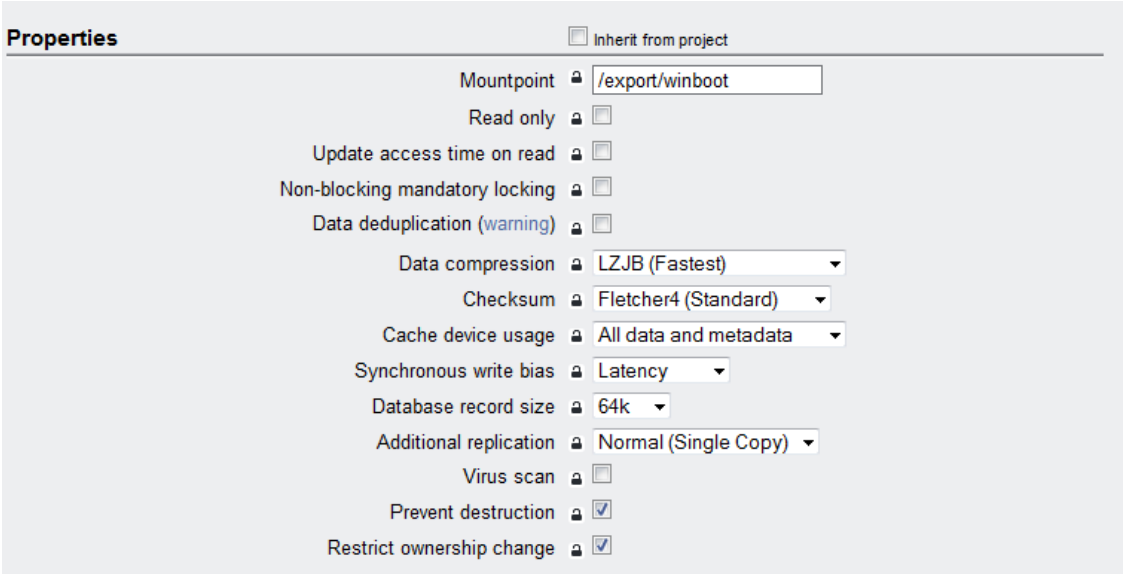

Figure 6. Share configuration shown in Oracle ZFS Storage Appliance BUI

- Under **Space Usage** settings, Quota Reservation or even User or Group configuration details for the Oracle ZFS Storage Appliance side are beyond the scope of this white paper. For more information on these settings, please refer to the Oracle ZFS Storage Appliance documentation (URLs are listed in Appendix B: Resources at the end of this document). However, as a best practice that includes security considerations, set up NFS ACL permitting the NFS shares to be mounted only by VMware ESXi5.x hosts.
- **Read-only** option: Leave unchecked.
- **Update access time on read**: Uncheck this option. This option is only valid for filesystems, and controls whether the access time for files is updated upon read. Under heavy loads consisting primarily of reads, and also over a large number of files, turning this option off may improve performance.
- **Non-blocking mandatory locking**: Do not check the option. This option is only valid for filesystems for which the primary protocol is SMB. SMB is not covered by this white paper.
- **Data deduplication** option: Do not check this option.
- **Data compression**: **S**elect the LZJB algorithm of data compression. Before writing data to the storage pool, shares can optionally compress data utilizing different algorithms of compression.

  Note: The LZJB algorithm is considered the fastest algorithm and it does not consume much CPU. The LZJB algorithm is recommended for virtualized environments.
- **Checksum:** Select the Fletcher4 (Standard) checksum algorithm. This feature controls the checksum algorithm used for data blocks and also allows the system to detect invalid data returned from devices. Working with the Fletcher4 algorithm, which is the default checksum algorithm, is sufficient for normal operations and can help avoid additional CPU load.

- **Cache device usage:** The 'All data and metadata' option is recommended. With this option, all files, LUNs, and any metadata will be cached.
- **Synchronous write bias:** To provide fast response time, select the Latency option**.**
- **Database record size:** Configure this setting according to the following table:

TABLE 5. DATABASE RECORD SIZE FOR PERFORMANCE TEST

POOL NAME	PROJECTS	FILE SYSTEMS	DATABASE RECORDSIZE
Pool1	vswap	/export/vswap	64k
	ms-exchangedb	/export/ms-exchangedb	32k
	ms-log	/export/ms-log	128k
	linux-os	/export/linux-os	64k
	oltp-db	/export/oltp-db	8k
	winboot	/export/winboot	64k

- **Additional replication:** To store a single copy of data blocks, select the normal (single copy) option.
- **Virus scan:** Assuming that each virtual machine will have its own anti-virus software up and running, enabling virus scan here is not recommended. However, in a virtualized environment, this option can be enabled for an additional NFS share that hosts a Windows home directory or shared folder for all users.

  Refer to the following document for more information if you want to enable this feature at the appliance level:

  http://www.oracle.com/technetwork/server-storage/sun-unified-storage/documentation/mcafee-antivirus-final-41712-1614883.pdf
- **Prevent destruction:** By default this option is turned off. Enabling this option to prevent the NFS share from accidental destruction is recommended.
- **Restrict ownership change:** By default this option is turned on. Also, for this test, changing the ownership of virtual machines files was not recommended.

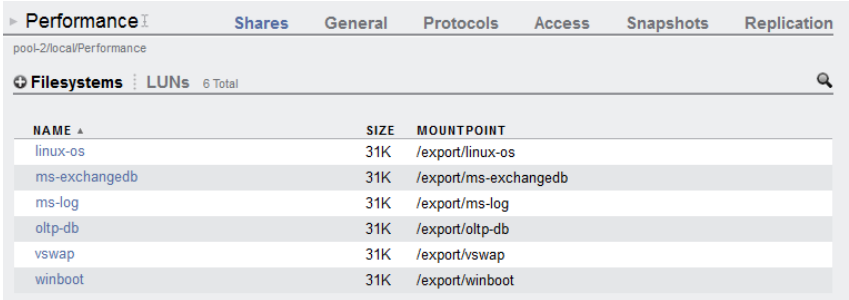

Figure 7. File systems and mountpoint configuration shown in Oracle ZFS Storage Appliance BUI

Figure 8 shows the minimum recommended configuration of the Oracle ZFS Storage Appliance for VMware vSphere 5.x working with NFS protocol. Note that "Oracle ZFS head" in the graphic refers to the Oracle ZFS Storage Appliance head.

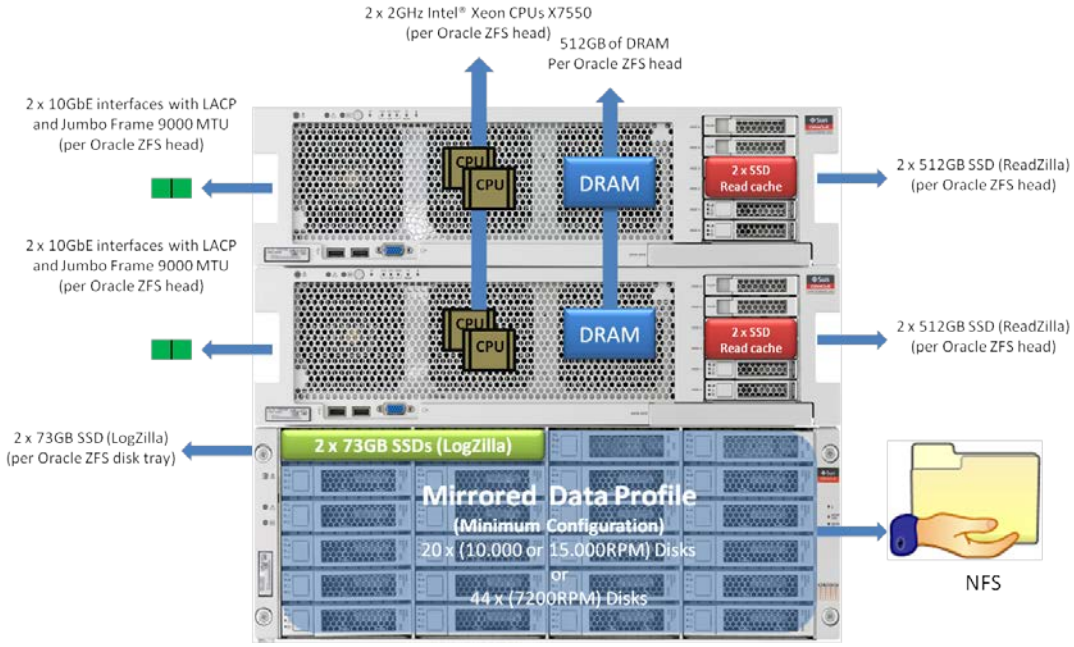

Figure 8. Oracle ZFS Storage Appliance – minimum recommended configuration for VMware vSphere 5 working with NFS protocol

IP Network Infrastructure

The example that follows employs two Cisco Nexus 5010 10GbE IP switches with all interfaces working in 10GbE speed in full duplex mode. Also, the IP switches' ports connected to the Oracle ZFS Storage Appliance are grouped in a Cisco EtherChannel working with port groups configuration 9000 MTU (jumbo frame) and 802.3ad Link Aggregation Control Protocol (LACP). On VMware, the default NIC teaming configuration is using active and standby interfaces mode.

Note: If you are working with more than one physical network card that is a member of a port-channel group, use the VMware NIC teaming configuration shown in figure 9 that includes:

- Load Balancing: Route based on IP hash
- Network Failover Detection: Link status only
- Notify Switches: Yes
- Failback: Yes

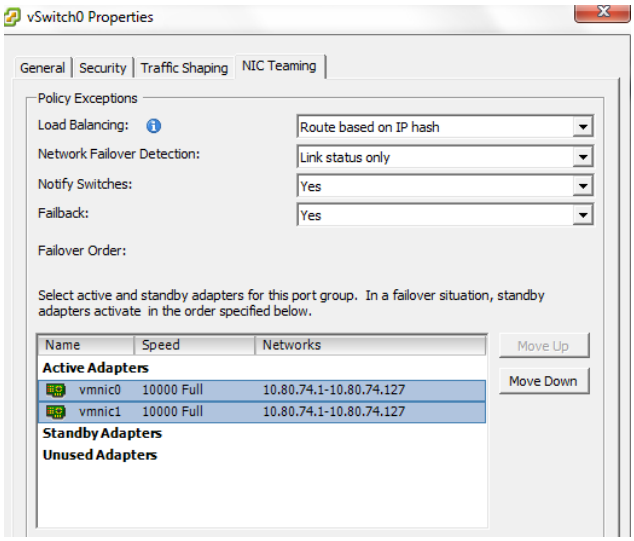

Figure 9. VMware vSphere – NIC teaming configuration

Note: VMware has added the LACP feature on VMware ESXi5.1 hosts utilizing vSphere Distributed Switch. However, the vSphere Distributed Switch configuration is beyond the scope of this paper and the featured examples do not use LACP with VMware. LACP configuration has been enabled only on port-group 100 and Oracle ZFS Storage Appliance 10GbE interfaces.

On the VMware side, work with at least 4x10GbE interfaces and two virtual switches. Configure two physical 10GbE for the management and virtual machine network, and also two 10GbE for NFS and vMotion operations. All 10GbE must be configured with 9000 MTU. Figures 10, 11 and 12 reflect these settings.

Note: Use of VMware vSphere distributed switch (VDS) in combination with VMware direct I/O technology and pass-through-capable hardware is recommended. Performance gains have been reported when using the combination of these technologies. However, the examples do not use pass-through-capable hardware, and so configuration of these features is beyond the scope of this paper. For more information about this technology refer to VMware's official documentation.

Figures 10, 11, and 12 show three different network environments that are supported with the Oracle ZFS Storage Appliance.

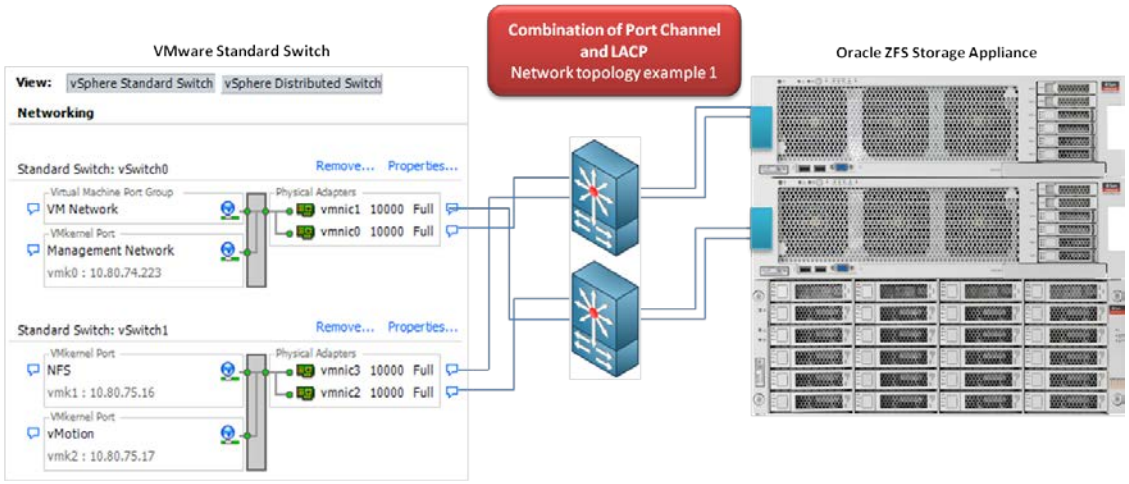

Figure 10. Example 1. Oracle ZFS Storage Appliance and VMware ESXi5.1 network infrastructure for NFS

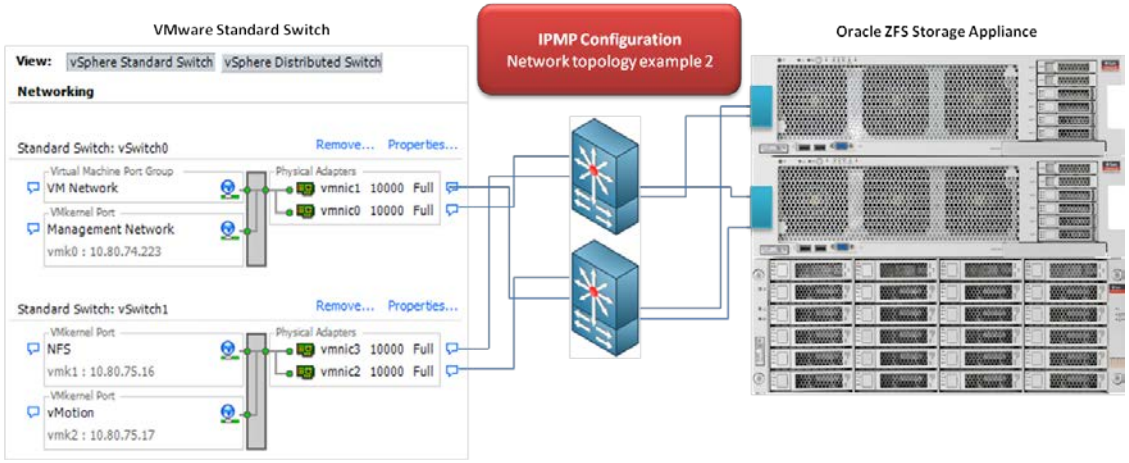

Figure 11. Example 2. Oracle ZFS Storage Appliance and VMware ESXi5.1 network infrastructure for NFS

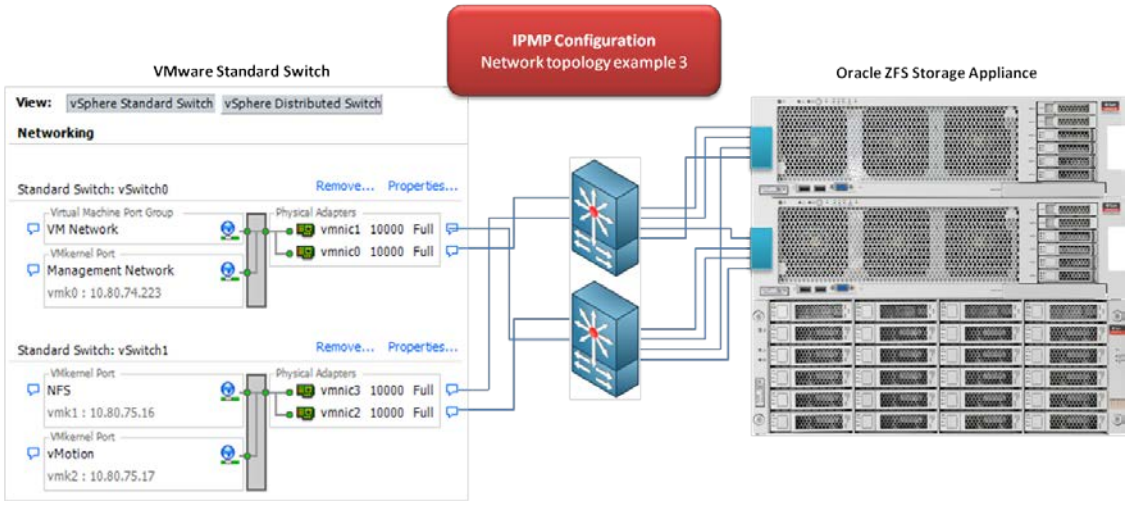

Figure 12. Example 3. Oracle ZFS Storage Appliance and VMware ESXi5.1 network infrastructure for NFS

The following steps show how to configure port-channels as well as LACP and 9000 MTU jumbo frame on a Cisco NEXUS 5010 switch. Before starting, ensure that your IP switches have the LACP feature enabled. To do this, open an SSH session with your switches and run the following listed commands.

Note: The following steps must be performed in all IP switch members of this solution. The example reflects two physical Cisco Nexus IP switches, so perform the Cisco EtherChannel, LACP and jumbo frame configuration in both switches.

```
nexus_ip_sw_01# show feature
Feature Name          Instance  State
--------------------  --------  --------
cimserver             1         disabled
fabric-binding        1         disabled
fc-port-security      1         disabled
fcoe                  1         enabled
fcsp                  1         disabled
fex                   1         disabled
fport-channel-trunk   1         disabled
http-server           1         enabled
interface-vlan        1         disabled
lacp                  1         disabled
lldp                  1         enabled
npiv                  1         enabled
npv                   1         disabled
port_track            1         disabled
private-vlan          1         disabled
sshServer             1         disabled
tacacs                1         disabled
telnetServer          1         enabled
udld                  1         disabled
vpc                   1         disabled
vtp                   1         disabled
```

If you do not have the LACP feature enabled, use the following steps to enable this feature.

```
nexus_ip_sw_01# configure terminal
```

```
Enter configuration commands, one per line.  End with CNTL/Z.
nexus_ip_sw_01 (config)# feature lacp
nexus_ip_sw_01 (config)# end

nexus_ip_sw_01# show feature
Feature Name          Instance  State
--------------------  --------  --------
cimserver             1         disabled
fabric-binding        1         disabled
fc-port-security      1         disabled
fcoe                  1         enabled
fcsp                  1         disabled
fex                   1         disabled
fport-channel-trunk   1         disabled
http-server           1         enabled
interface-vlan        1         disabled
lacp                  1         enabled
lldp                  1         enabled
npiv                  1         enabled
npv                   1         disabled
port_track            1         disabled
private-vlan          1         disabled
sshServer             1         disabled
tacacs                1         disabled
telnetServer          1         enabled
udld                  1         disabled
vpc                   1         disabled
vtp                   1         disabled
```

Creating a Port-Channel

Follow these steps to create the port-channel 100.

```
nexus_ip_sw_01# configure terminal
nexus_ip_sw_01 (config)# interface port-channel 100
nexus_ip_sw_01 (config-if)# interface ethernet 1/9-10
nexus_ip_sw_01 (config-if-range)# channel-group 100 mode active

nexus_ip_sw_01# show interface port-channel 100
port-channel 100 is down (No operational members)
  Hardware: Port-Channel, address: 0000.0000.0000 (bia 0000.0000.0000)
  MTU 1500 bytes, BW 100000 Kbit, DLY 10 usec,
     reliability 255/255, txload 1/255, rxload 1/255
  Encapsulation ARPA
  Port mode is access
  auto-duplex, auto-speed
  Beacon is turned off
  Input flow-control is on, output flow-control is on*
```

*NOTE: In a VMware ESXi or ESX environment, the flow control feature is enabled on all network interfaces by default. In a VMware environment with Oracle ZFS Storage Appliance, flow control is the preferred configuration and should be enabled.

Refer to the following VMware URL for additional information about flow control: http://kb.vmware.com/selfservice/microsites/search.do?language=en_US&cmd=displayKC&externalId=1013413

```
Switchport monitor is off
  No members
  Last clearing of "show interface" counters never
```

```
  0 seconds input rate 0 bits/sec, 0 packets/sec
  0 seconds output rate 0 bits/sec, 0 packets/sec
  Load-Interval #2: 0 seconds
    input rate 0 bps, 0 pps; output rate 0 bps, 0 pps
  RX
    0 unicast packets  0 multicast packets  0 broadcast packets
    0 input packets  0 bytes
    0 jumbo packets  0 storm suppression packets
    0 runts  0 giants  0 CRC  0 no buffer
    0 input error  0 short frame  0 overrun   0 underrun  0 ignored
    0 watchdog  0 bad etype drop  0 bad proto drop  0 if down drop
    0 input with dribble  0 input discard
    0 Rx pause
  TX
    0 unicast packets  0 multicast packets  0 broadcast packets
    0 output packets  0 bytes
    0 jumbo packets
    0 output errors  0 collision  0 deferred  0 late collision
    0 lost carrier  0 no carrier  0 babble
    0 Tx pause
  0 interface resets
```

Now that the port-channel 100 has been created, you need to add network interfaces to this channel group. To accomplish this, use the following steps.

```
nexus_ip_sw_01# configure terminal
Enter configuration commands, one per line.  End with CNTL/Z.
nexus_ip_sw_01 (config)# interface ethernet 1/9-10
nexus_ip_sw_01 (config-if-range)# channel-group 100
nexus_ip_sw_01 (config-if-range)# end

nexus_ip_sw_01# show port-channel summary
Flags:  D - Down        P - Up in port-channel (members)
        I - Individual  H - Hot-standby (LACP only)
        s - Suspended   r - Module-removed
        S - Switched    R - Routed
        U - Up (port-channel)
--------------------------------------------------------------------------------
Group Port-       Type     Protocol  Member Ports
      Channel
--------------------------------------------------------------------------------
100   Po100(SU)   Eth      LACP      Eth1/9(P)    Eth1/10(P)
```

The next task is to enable the port-channel load balancing feature. Use the following steps.

Enabling Port-Channel Load Balance

```
nexus_ip_sw_01# configure terminal
Enter configuration commands, one per line.  End with CNTL/Z.
nexus_ip_sw_01 (config)# port-channel load-balance ethernet source-dest-ip
nexus_ip_sw_01 (config)# show port-channel load-balance

Port Channel Load-Balancing Configuration:
System: source-dest-ip

Port Channel Load-Balancing Addresses Used Per-Protocol:
Non-IP: source-dest-mac
IP: source-dest-ip source-dest-mac
```

The Cisco EtherChannel configuration is now completed and the network interfaces are grouped into a channel-group 100 using LACP protocol. To ensure that the port-channel is up and running as well as utilizing LACP protocol and load balance features, perform the following command.

```
nexus_ip_sw_01# show port-channel summary
Flags:  D - Down        P - Up in port-channel (members)
        I - Individual  H - Hot-standby (LACP only)
        s - Suspended   r - Module-removed
        S - Switched    R - Routed
        U - Up (port-channel)
--------------------------------------------------------------------------------
Group Port-       Type     Protocol  Member Ports
      Channel
--------------------------------------------------------------------------------
100   Po100(SU)   Eth      LACP      Eth1/9(P)    Eth1/10(P)


nexus_ip_sw_01# show port-channel usage
Total 1 port-channel numbers used
============================================
Used  :   100
Unused:   1 - 99 , 101 - 4096
          (some numbers may be in use by SAN port channels)


nexus_ip_sw_01# show port-channel traffic
ChanId      Port Rx-Ucst Tx-Ucst Rx-Mcst Tx-Mcst Rx-Bcst Tx-Bcst
------ --------- ------- ------- ------- ------- ------- -------
   100    Eth1/9  48.22%  94.51%  57.80%  37.29%  32.35%  51.93%
   100   Eth1/10  51.77%   5.48%  42.19%  62.70%  67.64%  48.06%
```

Save your configuration running the following command:

```
nexus_ip_sw_01# copy running-config startup-config
[########################################] 100%
```

Enabling Jumbo Frame 9000 MTU

Based on Cisco official documentation, the Cisco Nexus 5000 Series switch only supports system-level MTU, which means the MTU attribute cannot be changed on an individual port basis. However, you can still modify MTU size by setting QoS policy and class maps.

To enable jumbo frame for the whole switch, perform these steps:

```
nexus_ip_sw_01# configure terminal
Enter configuration commands, one per line.  End with CNTL/Z.
nexus_ip_sw_01 (config)# policy-map type network-qos jumbo
nexus_ip_sw_01 (config-pmap-nq)# class type network-qos class-default
nexus_ip_sw_01 (config-pmap-nq-c)# mtu 9000
nexus_ip_sw_01 (config-pmap-nq-c)# end
nexus_ip_sw_01# configure terminal
Enter configuration commands, one per line.  End with CNTL/Z.
nexus_ip_sw_01 (config)# system qos
nexus_ip_sw_01 (config-sys-qos)# service-policy type network-qos jumbo
nexus_ip_sw_01 (config-sys-qos)# end
```

Check your configuration to ensure the IP switch Ethernet interfaces are carrying traffic with jumbo MTU. Perform the following commands to validate that information:

```
nexus_ip_sw_01# show interface ethernet 1/9 counters detailed
Ethernet1/9
```

```
Rx Packets:                                1503095493
Rx Unicast Packets:                        1503070519
Rx Multicast Packets:                           14499
Rx Broadcast Packets:                           10475
Rx Jumbo Packets:                              210539
Rx Bytes:                                919451945239
Rx Packets from 0 to 64 bytes:              823994390
Rx Packets from 65 to 127 bytes:             60266586
Rx Packets from 128 to 255 bytes:            41809329
Rx Packets from 256 to 511 bytes:             7941051
Rx Packets from 512 to 1023 bytes:            7991931
Rx Packets from 1024 to 1518 bytes:         561092203
Tx Packets:                               59232316116
Tx Unicast Packets:                       59196278214
Tx Multicast Packets:                        14618899
Tx Broadcast Packets:                        21418053
Tx Jumbo Packets:                              251642
Tx Bytes:                              70304189240915
Tx Packets from 0 to 64 bytes:               54643893
Tx Packets from 65 to 127 bytes:          11529933522
Tx Packets from 128 to 255 bytes:          1166365207
Tx Packets from 256 to 511 bytes:           460593642
Tx Packets from 512 to 1023 bytes:          816852512
Tx Packets from 1024 to 1518 bytes:       45203675698
Tx Trunk Packets:                             5045352
Output Errors:                                      3
```

Note: Cisco Nexus 5000 series switches do not support packet fragmentation, so an incorrect MTU configuration may result in packets being truncated. Ensure that your network interfaces have the right duplex and speed configuration, and members of Cisco EtherChannel have the LACP feature enabled and are correctly configured.

Refer to the following URL for additional information about the Cisco Nexus IP switch:

http://www.cisco.com/en/US/docs/switches/datacenter/nexus5000/sw/configuration/guide/cli_rel_4_0_1a/EtherChannel.html

Recommendations for NFS Protocol

Be sure to alter the NFS and TCP/IP advanced settings prior to starting the tests or putting your VMware servers in production. These options are extremely important to ensure high availability of the NFS datastores in a failover/failback situation with the Oracle ZFS Storage Appliance. Table 6 shows the Advanced Settings. To alter the parameters listed in table 6, go to the VMware vCenter 5.x server and select a VMware server. Select the software tab and click on **Advanced Settings.** See figures 13 and 14.

Note: This configuration must be performed in all VMware host members of the cluster. A reboot of each VMware host will be needed in order to activate the new settings.

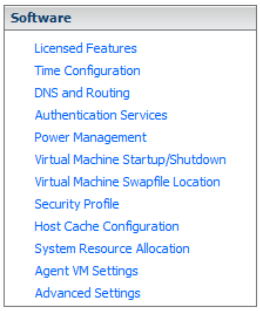

Figure 13. VMware Advanced Settings category shown on VMware vCenter 5.x server

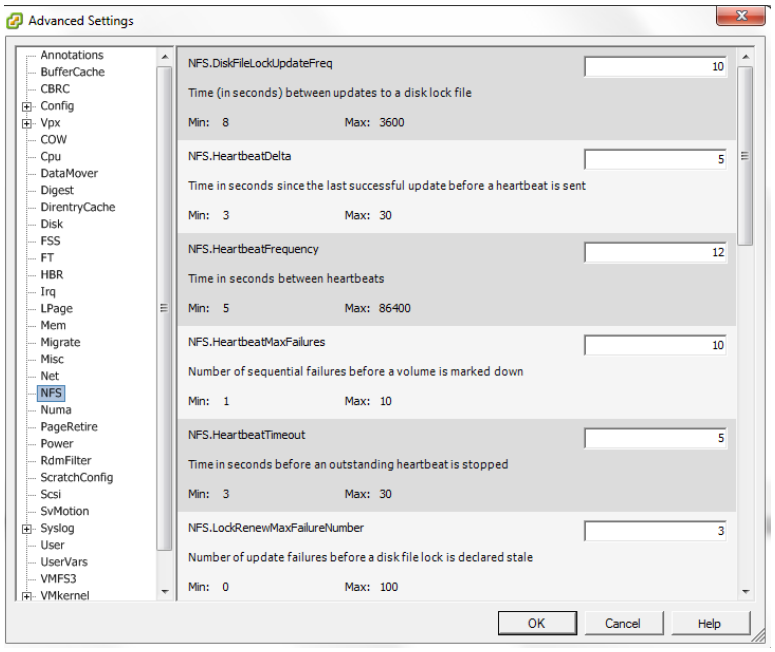

Figure 14. VMware advanced settings shown configured on VMware vCenter 5.x server

TABLE 6. RECOMMENDED NFS AND TCP/IP ADVANCED SETTINGS FOR VMWARE VSPHERE 5.1 DATASTORES ON ORACLE ZFS STORAGE APPLIANCE	
OPTION	**VALUE**
NFS.HeartbeatTimeout	5
Nfs.Sendbuffersize	264
Nfs.Receivebuffersize	256
Nfs.MaxVolumes	256
Net.TcpipHeapMax	128

Net.TcpipHeapsize	32
Nfs.heartbeatfrequency	20
Nfs.heartbeatdelta	12
Nfs.heartbeatmaxfailures	10

Recommendations for Fibre Channel Protocol

Follow these best practices and recommendations when working with Fibre Channel protocol and VMware vSphere 5.x.

- Update the Fibre Channel host bus adapters' (HBAs') firmware and drivers to their latest version and also ensure that the HBA is on the VMware HCL.
- Ensure you have only one VMware VMFS (virtual Machine File System) volume per LUN.
- For raw devices, use RDM (Raw Device Mapping).
- Work with at least two Fibre Channel switches and one dual port 8Gbps HBA per Oracle ZFS Storage Appliance controller and VMware ESXi5.x host.
- Ensure that your Storage Area Network (SAN) has been designed for high availability and load balance without critical points of failure. See figure 15.

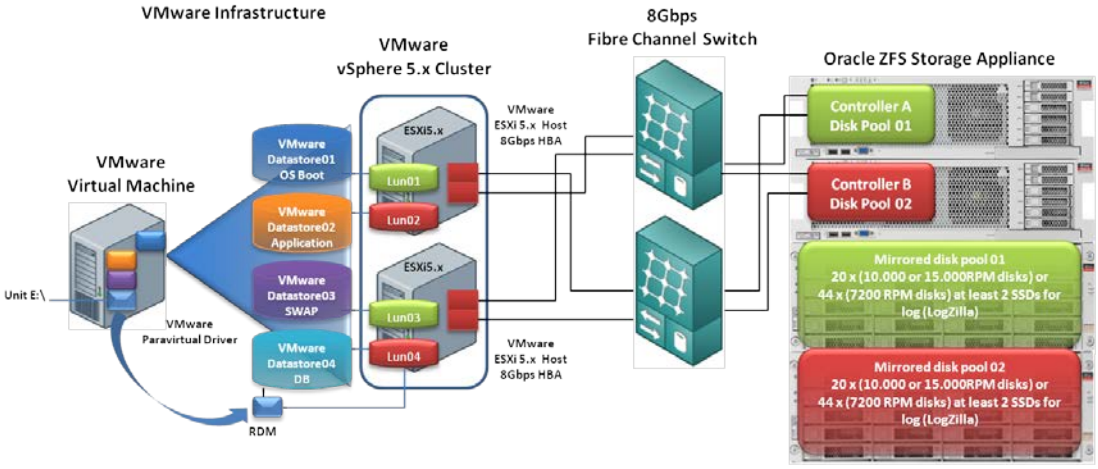

Figure 15. Oracle ZFS Storage Appliance and VMware vSphere 5.x Fibre Channel environment

When working with Fibre Channel protocol with VMware vSphere 5.x and Oracle ZFS Storage Appliance, change the default storage array type as well as path selection policy and round robin I/O operation limit prior to putting the servers into production. Follow the steps shown in the next several code examples to perform this change.

For changing the round robin I/O operation limit, use the steps shown in the following ESXi command lines. Identify all Oracle ZFS Storage Appliance disks that will be utilized by your virtualized server.

```
# esxcli storage nmp device list | egrep -i "SUN Fibre Channel Disk"

   Device Display Name: SUN Fibre Channel Disk (naa.600144f0c36f708b0000509aa8780005)
   Device Display Name: SUN Fibre Channel Disk (naa.600144f0c36f708b0000509aa94f000b)
   Device Display Name: SUN Fibre Channel Disk (naa.600144f0c36f708b0000509aa8ff0009)
   Device Display Name: SUN Fibre Channel Disk (naa.600144f0c36f708b0000509aab40000d)
   Device Display Name: SUN Fibre Channel Disk (naa.600144f0c36f708b0000509aa8d70008)
   Device Display Name: SUN Fibre Channel Disk (naa.600144f0c36f708b0000509aa8930006)
   Device Display Name: SUN Fibre Channel Disk (naa.600144f0c36f708b0000509aa8b50007)
   Device Display Name: SUN Fibre Channel Disk (naa.600144f0c36f708b0000509aaa77000c)
   Device Display Name: SUN Fibre Channel Disk (naa.600144f0c36f708b0000509aa92a000a)
```

Change the storage array type VMW_PSP_RR to VMW_SATP_ALUA and the path selection policy VMW_PSP_MRU to VMW_PSP_RR:

```
#esxcli storage nmp satp set --default-psp=VMW_PSP_RR --satp=VMW_SATP_ALUA

#esxcli storage nmp device list

naa.600144f0c36f708b0000509aa92a000a
   Device Display Name: SUN Fibre Channel Disk (naa.600144f0c36f708b0000509aa92a000a)
   Storage Array Type: VMW_SATP_ALUA
   Storage Array Type Device Config: {implicit_support=on;explicit_support=off;
explicit_allow=on;alua_followover=on;{TPG_id=0,TPG_state=AO}}
   Path Selection Policy: VMW_PSP_MRU
   Path Selection Policy Device Config: Current Path=vmhba7:C0:T0:L6
   Path Selection Policy Device Custom Config:
   Working Paths: vmhba7:C0:T0:L6
```

The example shows the following command line for capturing only Sun ZFS Fibre Channel disks as well as changing the path selection policy.

Note: If needed, adjust the following command line for your particular environment.

```
# esxcli storage nmp device list | egrep -i "SUN Fibre Channel Disk" | awk '{ print $8
}' | cut -c 2-37

naa.600144f0c36f708b0000509aa8780005
naa.600144f0c36f708b0000509aa94f000b
naa.600144f0c36f708b0000509aa8ff0009
naa.600144f0c36f708b0000509aab40000d
naa.600144f0c36f708b0000509aa8d70008
naa.600144f0c36f708b0000509aa8930006
naa.600144f0c36f708b0000509aa8b50007
naa.600144f0c36f708b0000509aaa77000c
naa.600144f0c36f708b0000509aa92a000a
```

Ensure that you are not using round robin path selection before performing changes.

```
# for a in `esxcli storage nmp device list | egrep -i "SUN Fibre Channel Disk" | awk '{
print $8 }' | cut -c 2-37`
> do
> esxcli storage nmp psp roundrobin deviceconfig get -d $a
> done
```

```
Device naa.600144f0c36f708b0000509aa8780005 Does not use the Round Robin path selection policy.
Device naa.600144f0c36f708b0000509aa94f000b Does not use the Round Robin path selection policy.
Device naa.600144f0c36f708b0000509aa8ff0009 Does not use the Round Robin path selection policy.
Device naa.600144f0c36f708b0000509aab40000d Does not use the Round Robin path selection policy.
Device naa.600144f0c36f708b0000509aa8d70008 Does not use the Round Robin path selection policy.
Device naa.600144f0c36f708b0000509aa8930006 Does not use the Round Robin path selection policy.
Device naa.600144f0c36f708b0000509aa8b50007 Does not use the Round Robin path selection policy.
Device naa.600144f0c36f708b0000509aaa77000c Does not use the Round Robin path selection policy.
Device naa.600144f0c36f708b0000509aa92a000a Does not use the Round Robin path selection policy.
```

Run the following command to change the path selection policy VMW_PSP_MRU to VMW_PSP_RR:

```
~ # for a in `esxcli storage nmp device list | egrep -i "SUN Fibre Channel Disk" | awk
'{ print $8 }' | cut -c 2-37`
> do
> esxcli storage nmp device set -d $a --psp=VMW_PSP_RR
> done
```

Run the following command to ensure that the new path selection policy has been updated:

```
~ # esxcli storage nmp device list
naa.600144f0c36f708b0000509aa92a000a
   Device Display Name: SUN Fibre Channel Disk (naa.600144f0c36f708b0000509aa92a000a)
   Storage Array Type: VMW_SATP_ALUA
   Storage Array Type Device Config: {implicit_support=on;explicit_support=off;
explicit_allow=on;alua_followover=on;{TPG_id=0,TPG_state=AO}}
   Path Selection Policy: VMW_PSP_RR
   Path Selection Policy Device Config:
{policy=rr,iops=1000,bytes=10485760,useANO=0;lastPathIndex=1:
NumIOsPending=0,numBytesPending=0}
   Path Selection Policy Device Custom Config:
   Working Paths: vmhba6:C0:T0:L6, vmhba7:C0:T0:L6
```

Change the I/O operation limit value to 1 and also the type of the round robin path switching to iops for all Fibre Channel disks on the Oracle ZFS Storage Appliance. List the device configuration before changing.

```
~ # esxcli storage nmp psp roundrobin deviceconfig get -d
naa.600144f0c36f708b0000509aa92a000a
   Byte Limit: 10485760
   Device: naa.600144f0c36f708b0000509aa92a000a
   IOOperation Limit: 1000
   Limit Type: Default
   Use Active Unoptimized Paths: false
```

Perform the configuration:

```
# for a in `esxcli storage nmp device list | egrep -i "SUN Fibre Channel Disk" | awk '{
print $8 }' | cut -c 2-37`
> do
> esxcli storage nmp psp roundrobin deviceconfig set -d $a -I 1 -t iops
> done
```

Run the following command to ensure that the new values for operation limit and also round robin path switching have been updated:

```
# for a in `esxcli storage nmp device list | egrep -i "SUN Fibre Channel Disk" | awk '{
print $8 }' | cut -c 2-37`
```

```
> do
> esxcli storage nmp psp roundrobin deviceconfig get -d $a
> done

Device: naa.600144f0c36f708b0000509aa92a000a
IOOperation Limit: 1
Limit Type: Iops
Use Active Unoptimized Paths: false
```

To check the same information on the VMware vSphere 5.x client, go to the **Configuration** tab, select **Storage adapters**, click on the vmhba that is attached with your Oracle ZFS Storage Appliance, and then right-click on the disk that you wish to validate for configuration. Select **Manage Paths**, as seen in figure 16. Figure 17 shows the result.

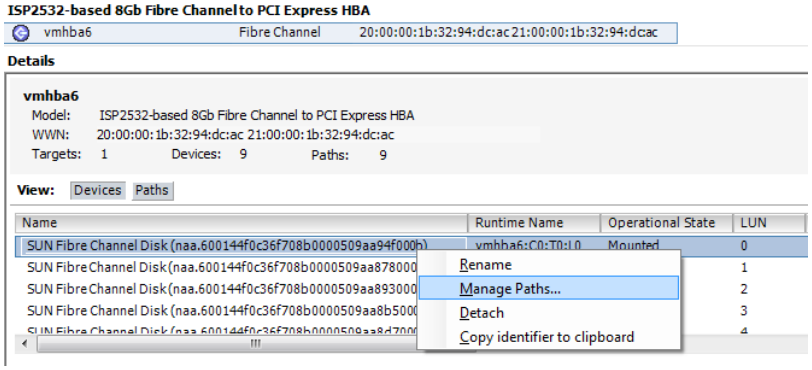

Figure 16. Managing VMware LUN paths shown in VMware vSphere 5.x client

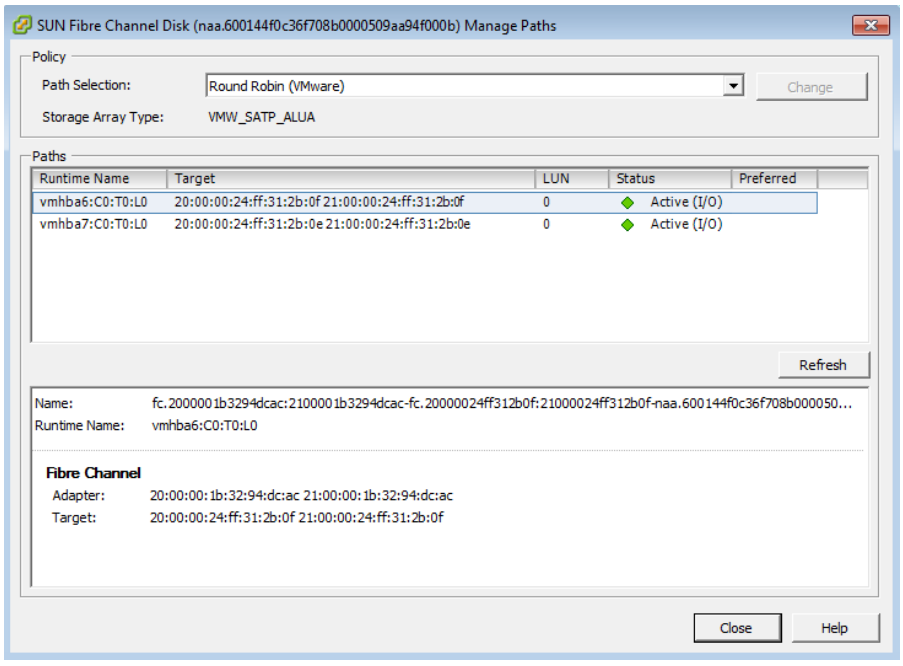

Figure 17. VMware path selection and storage array type overview shown in VMware vSphere 5.x client

Changing Queue Depth – QLogic and Emulex HBAs

As a best practice for VMware vSphere 5.x and Oracle ZFS Storage Appliance, adjust the queue depth option for all HBAs attached with the system.

Use the following steps to accomplish this task.

1. Identify which HBA module is currently loaded on the VMware hypervisor using the following commands.

For QLogic HBAs, run:

```
# esxcli system module list | grep qla*
qla2xxx                              true        true
```

For Emulex HBAs, run:

```
# esxcli system module list | grep lpfc*
```

Note: The example uses QLogic HBAs (module `qla2xxx`).

2. Use the following commands to set a new queue depth value.

For QLogic HBAs, run:

```
# esxcli system module parameters set -p ql2xmaxqdepth=64 -m qla2xxx
```

For Emulex HBAs, run:

```
# esxcli system module parameters set -p lpfc0_lun_queue_depth=64 -m lpfc820
```

3. Reboot your host and run the following command to confirm that the new queue depth value has been applied.

```
# esxcli system module parameters list -m qla2xxx
```

The following is the output for QLogic HBAs:

```
Name                         Type  Value  Description
---------------------------  ----  -----  ----------------------------------------
ql2xmaxqdepth                int   64     Maximum queue depth to report for target
devices.
```

Recommendations for iSCSI Protocol

The following best practices and recommendations apply for VMware vSphere 5.x using iSCSI protocol with Oracle ZFS Storage Appliance.

- On VMware ESXi5.x hosts, ensure that you have at least one dual 10GbE NIC working with 9000 MTU jumbo frame.
- Use at least two physical IP network switches.
- On the Oracle ZFS Storage Appliance side, ensure that you have at minimum a link aggregation of two or more 10GbE NICs attached with a physical IP network switch, configured and working with port-channel group or even IPMP technologies.
- Ensure that your 10GbE IP network is properly configured and working with high availability and load balance (without point of failure).
- Ensure that your physical IP switches or routers are not congested or saturated.
- Ensure that your iSCSI network provides adequate throughout as well as low latency between initiators and targets.
- Isolate the iSCSI traffic through different VLANs or even network segmentation. Also, work with a different VMware vSwitch for iSCSI traffic.
- To achieve best performance, as well as load balance the I/O traffic between paths and failover, configure VMware iSCSI to work in port binding mode.
- Change the storage array type to VMW_SATP_ALUA and also path selection policy VMW_PSP_MRU to VMW_PSP_RR, and ensure that all physical NIC members of the port binding are balancing the I/O traffic.

Figure 18 shows the high-level architecture, suitable for a production environment, of two different iSCSI topologies working with LACP, port-channel, and IPMP configuration with VMware vSphere 5.x and Oracle ZFS Storage Appliance.

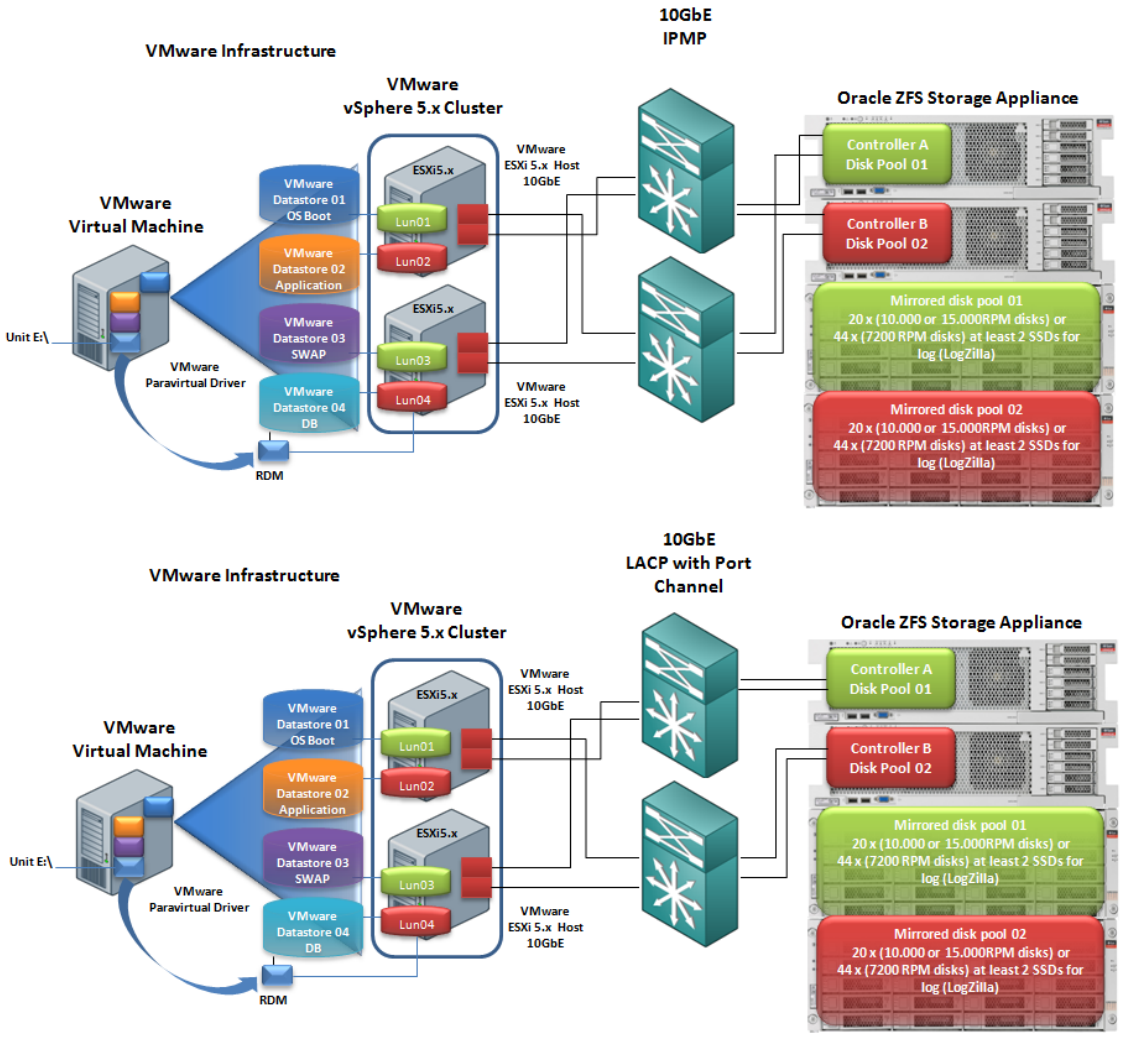

Figure 18. Oracle ZFS Storage Appliance and VMware vSphere 5.x iSCSI environments configuring iSCSI for vSphere 5.x

The following steps show how to configure VMware vSphere 5 iSCSI in port binding mode with Oracle ZFS Storage Appliance.

1. Create a new vSwitch with at least two VMkernel ports and two 10GbE interfaces, each one working with 9000 MTU (jumbo frame) and VMware port binding configuration. The example in figure 19 shows `iSCSI01` and `iSCSI02` VMkernel ports, while the 10GbE interfaces are `vmnic2` and `vmnic3`.

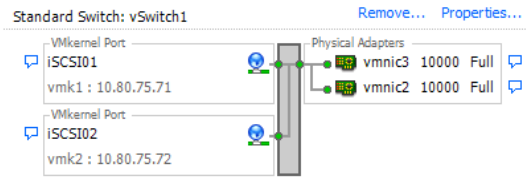

Figure 19. VMware vSwitch configuration screen shown in VMware vSphere 5.x client

2. For each VMkernel port, enable the **Override switch failover mode** option, as seen in figure 20. Ensure that only one 10GbE adapter is enabled per port group. Additional cards must be moved to Unused Adapters. To perform this task, select the **ESXi5.x host**, then select the **Configuration** tab, **Networking** and **Properties** of your iSCSI vSwitch. Select the **iSCSI port group**, click on **Edit**, and then select the **NIC Teaming** tab. See figure 20.

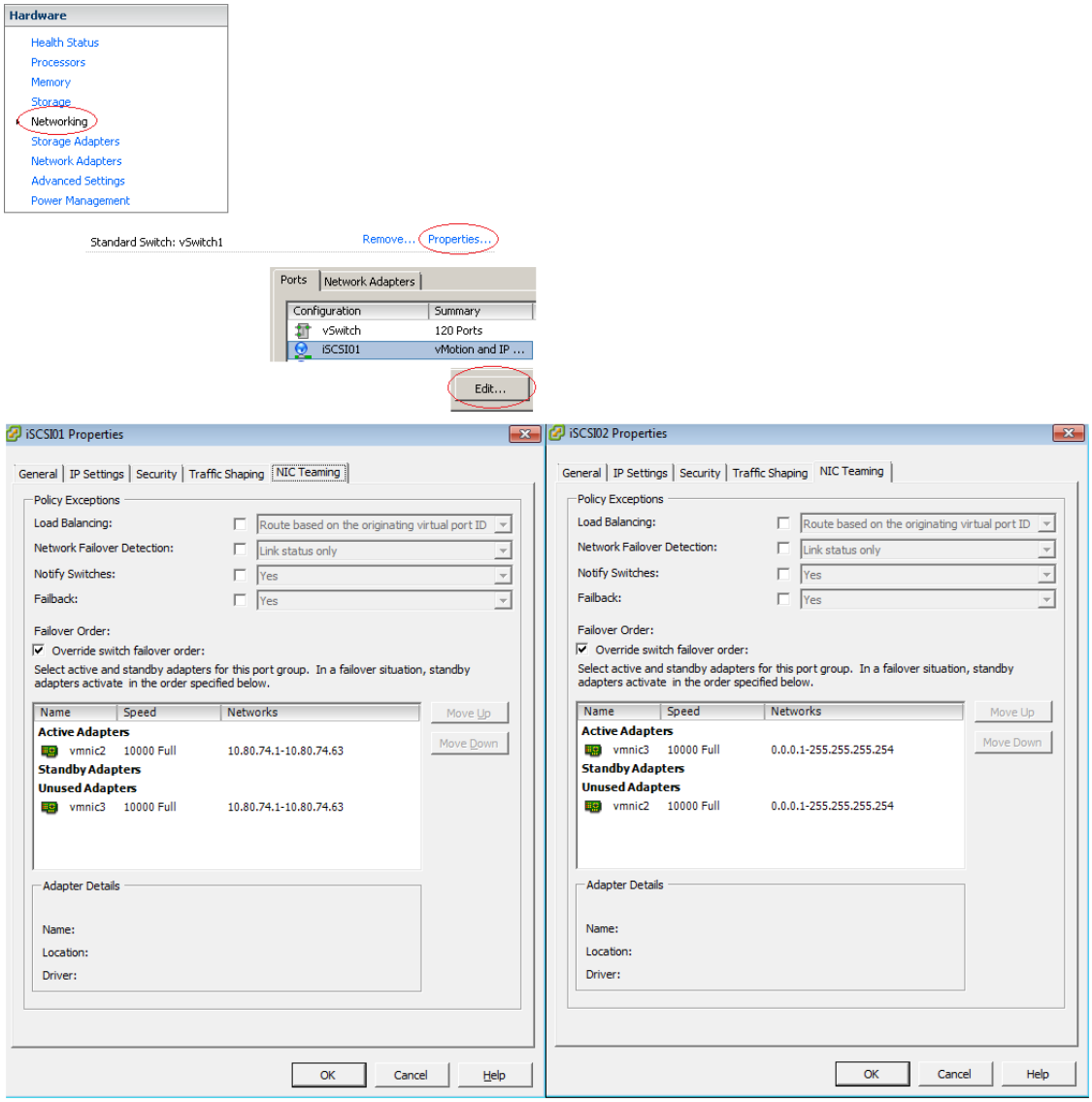

Figure 20. VMware iSCSI vSwitch NIC teaming and configuration screen shown in VMware vSphere 5.x client

The example in figure 20 shows two 10GbE adapters and two different VMkernel ports. Both 10GbE adapters (`vmnic2` and `vmnic3`) are being balanced across two different port groups, with the following configuration:

- The `iSCSI01` port group has `vmnic2` adapter enabled and `vmnic3` adapter unused.
- The `iSCSI02` port group has `vminc3` adapter enabled and `vmnic2` adapter unused.

Important: When working with a port binding configuration, each port group must have only one active adapter. All other adapters must be moved to Unused Adapters. Do not use standby mode. Refer to figure 20.

Once the port group configuration is ready, add the VMware iSCSI software using the following steps.

1. Open a connection with your VMware vCenter server, select the **ESXi5.x** host and select **Configuration.**
2. Under the **Hardware** option, select **Storage Adapter**, then **Add.**
3. Select **Add Software iSCSI Adapter**. Click **OK**. A new iSCSI vHBA will be created, as seen in figure 21.

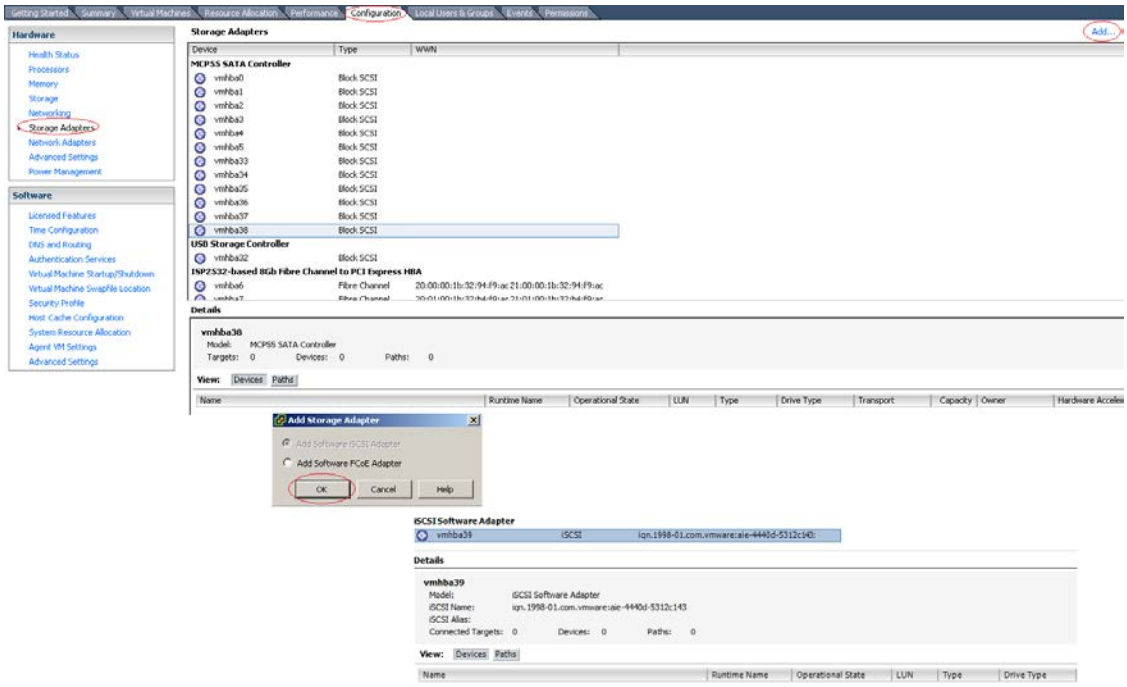

Figure 21. iSCSI Software Adapter screen shown in VMware vSphere 5.x client

4. Under **Hardware** and **Storage Adapter**, select your new **iSCSI vHBA**, then **Properties**. The **iSCSI Initiator Properties** screen will open.
5. Select **Configure** and enter an iSCSI alias name for this vHBA. Click on **OK**.

The figure 22 example shows the iSCSI alias name `ESXi5.x`. Choose an alias that best fits your environment. Also, take note of the IQN name of your ESXi5.x host, which in the example is `iqn.1998-01.com.vmware:aie-4440d-5312c143`**.** This information will be required for registering a new iSCSI initiator on the Oracle ZFS Storage Appliance, as shown in figure 22.

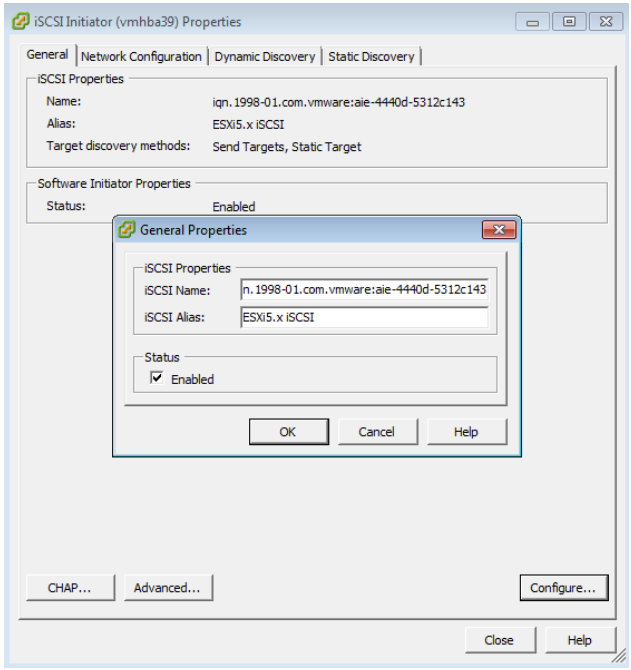

Figure 22. iSCSI Initiator Properties screen shown in VMware vSphere 5.x client

6. On the same screen, for binding the port groups to the software iSCSI adapter as well as active vmknic-based multipathing for iSCSI software, select the **Network Configuration** tab, click on **Add,** and then select `iSCSI01` and `iSCSI02` port groups. Click **OK.** Figure 23 shows the port binding details.

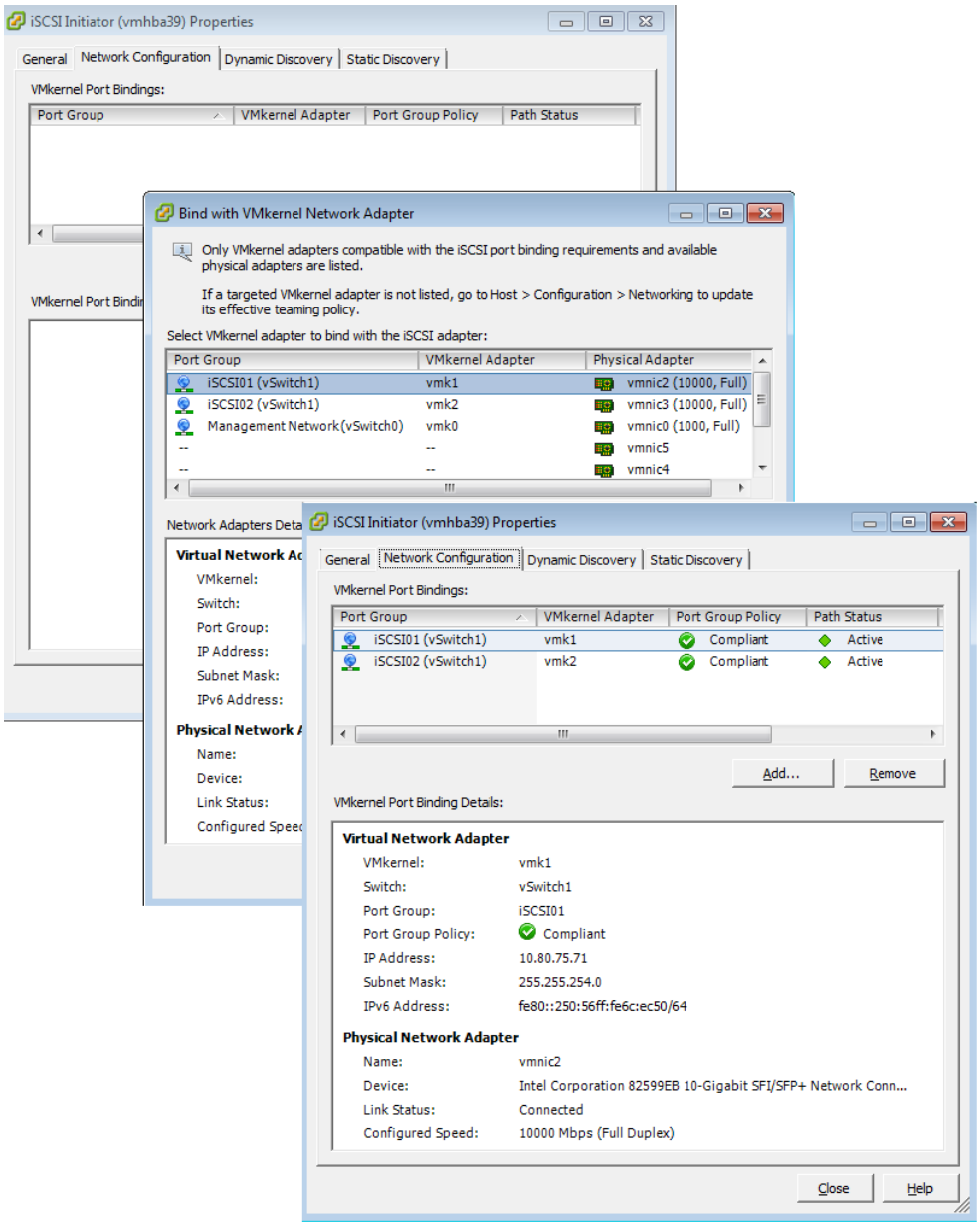

Figure 23. iSCSI Initiator Properties screen showing port binding details

7. Create a new iSCSI target on the Oracle ZFS Storage Appliance. To perform this, log in to the Oracle ZFS Storage Appliance BUI, click on **Configuration**, **SAN**, and then the **iSCSI Targets** option. Select the **Target IQN Auto-assign** option; enter an alias name that best fits your environment, select a **network interface** and click on **OK**. See figures 24.

The example shows the interface `aggr1`, which is a link aggregation of two 10GbE interfaces.

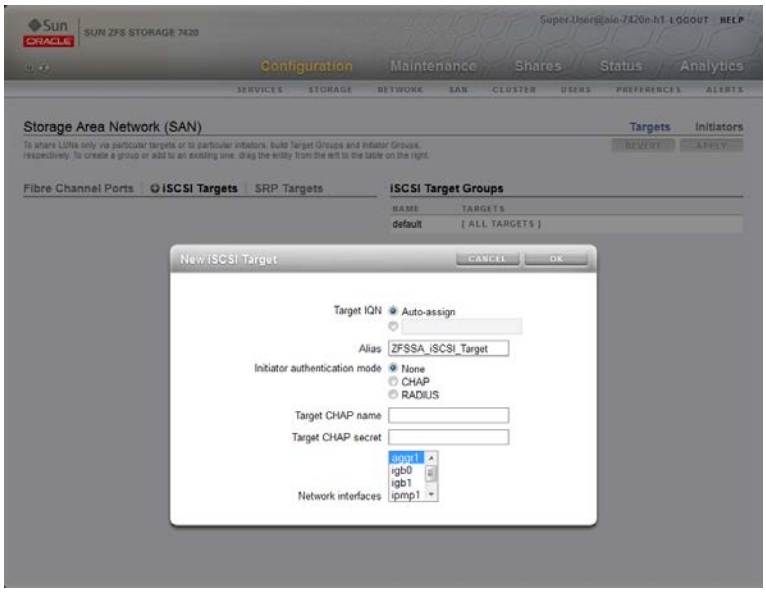

Figure 24. iSCSI Target configuration shown in Oracle ZFS Storage Appliance BUI

8. The iSCSI target is created. Select your new iSCSI target and drop it down to the **iSCSI Target Groups**, select **Edit** and change the name**.** Click on **OK** and **APPLY.** Figure 25 shows the edit window for the iSCSI target**.**

Note: As previously mentioned in this paper, a best practice is to work with at least two 10GbE in link aggregation mode LACP interfaces per Oracle ZFS Storage Appliance controller. CHAP authentication is not being used in this example, so you do not need to enter CHAP information.

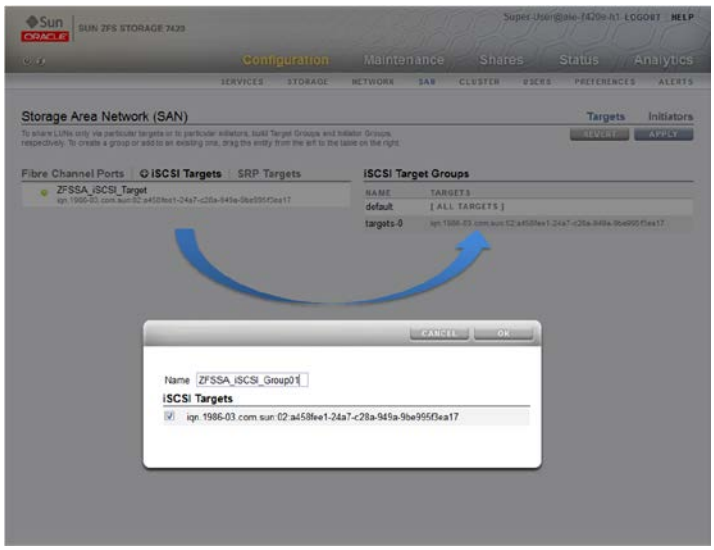

Figure 25. iSCSI Target Groups configuration shown in Oracle ZFS Storage Appliance BUI

9. On the same screen, click on **Initiators**, and then **iSCSI Initiators** to create a new iSCSI initiator. Enter the IQN initiator shown in figure 22. In the example the IQN initiator is `iqn.1998-01.com.vmware:aie-4440d-5312c143`. Enter an alias name and click on **OK**.

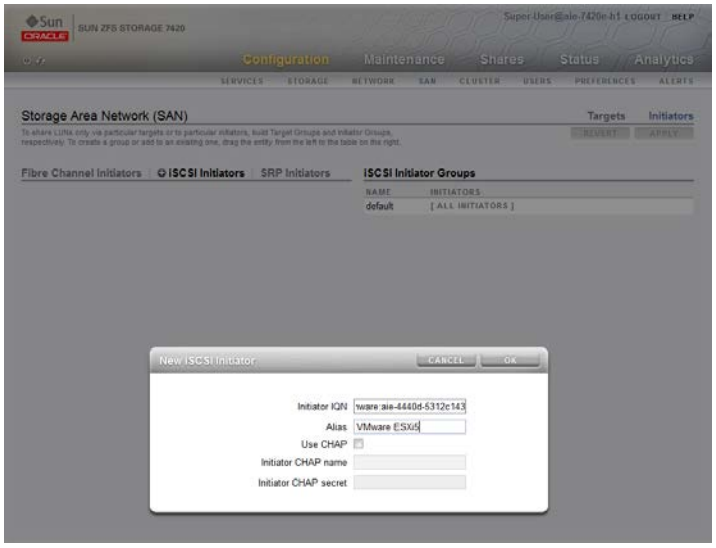

Figure 26. iSCSI Initiators configuration shown in the Oracle ZFS Storage Appliance BUI

10. Now that the new iSCSI Initiator is created, select it and drop it down to iSCSI Initiator Groups. Select **Edit** and change the iSCSI initiator name. Click on **OK** and **Apply.** Figure 27 shows the iSCSI initiator group edit window.

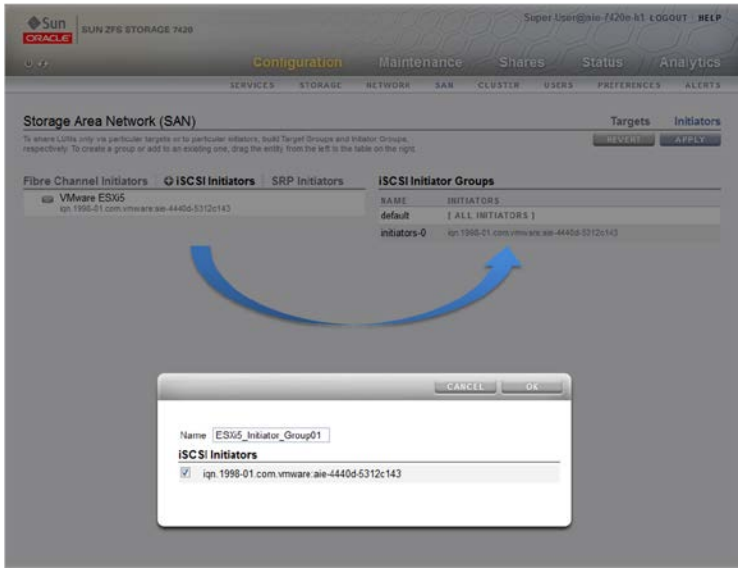

Figure 27. iSCSI Initiator Groups configuration shown in the Oracle ZFS Storage Appliance BUI

11. Next, you will create a LUN to map to the target and initiator group you have just created. Click on **Shares**, select your project and create the LUN. Figure 28 shows the Create LUN dialog window where you can map the LUN to the target and initiator group.

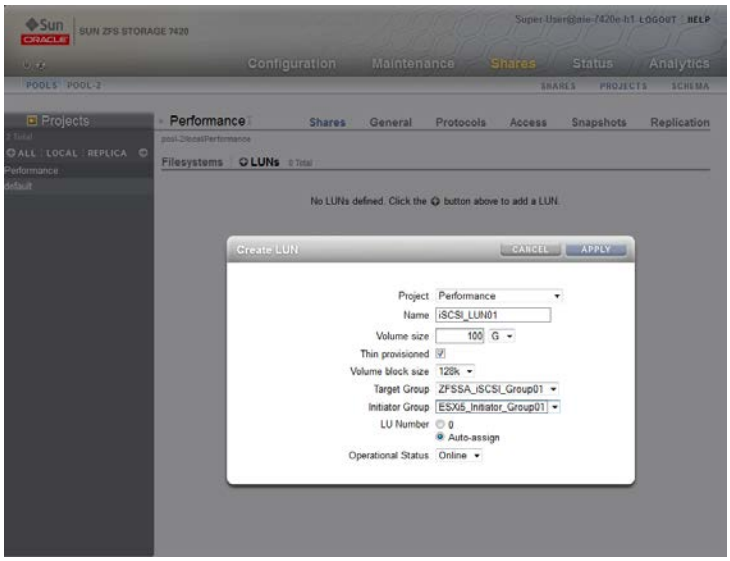

Figure 28. iSCSI LUN provisioning shown in the Oracle ZFS Storage Appliance BUI

12. On the VMware ESXi5.x host on which you have created the iSCSI configuration, open **iSCSI Initiatior Properties**, select the **Dynamic Discovery** tab, and click **Add**.
13. In the **Add Send Target Server** screen seen in figure 29, add the iSCSI IP address of the 10GbE link aggregation interface for the Oracle ZFS Storage Appliance. Click **OK** and close.

A rescan of the adapters will be needed in order to discover the new iSCSI LUN that you have just created.

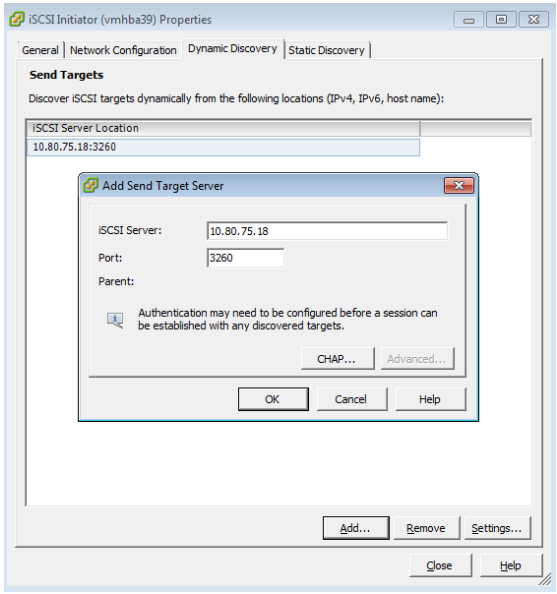

Figure 29. Adding an iSCSI server in the VMware vSphere 5.x client

After a rescan of the iSCSI HBA, the new LUN will be available to the ESXi5.x host as well as connected to two active paths' members of the port binding configuration, as seen in figure 30.

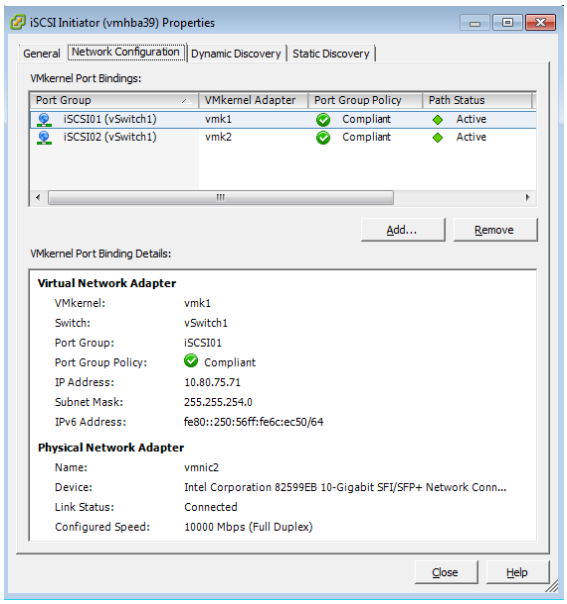

Figure 30. Overview of VMware vSphere 5.x iSCSI network configuration shown in VMware vSphere 5.x client

Ensure that the new iSCSI LUN is visible and also accessible by the ESXi5.x host. Also, validate and ensure that the multipath configuration is working properly by using the following commands.

14. Open a SSH connection to the ESXi5.x host and run the `esxcfg-mpath –l` command to list all LUNs attached to the ESXi5.x host. Identify the new iSCSI LUN(s).

```
# esxcfg-mpath -l
iqn.1998-01.com.vmware:aie-4440d-5312c143-00023d000002,iqn.1986-03.com.sun:02:a458fee1-
24a7-c28a-949a-9be995f3ea17,t,2-naa.600144f0a9b12ec6000050b93e310002
   Runtime Name: vmhba39:C1:T0:L0
   Device: naa.600144f0a9b12ec6000050b93e310002
   Device Display Name: SUN iSCSI Disk (naa.600144f0a9b12ec6000050b93e310002)
   Adapter: vmhba39 Channel: 1 Target: 0 LUN: 0
   Adapter Identifier: iqn.1998-01.com.vmware:aie-4440d-5312c143
   Target Identifier: 00023d000002,iqn.1986-03.com.sun:02:a458fee1-24a7-c28a-949a-
9be995f3ea17,t,2
   Plugin: NMP
   State: active
   Transport: iscsi
   Adapter Transport Details: iqn.1998-01.com.vmware:aie-4440d-5312c143
   Target Transport Details: IQN=iqn.1986-03.com.sun:02:a458fee1-24a7-c28a-949a-
9be995f3ea17 Alias= Session=00023d000002 PortalTag=2

iqn.1998-01.com.vmware:aie-4440d-5312c143-00023d000001,iqn.1986-03.com.sun:02:a458fee1-
24a7-c28a-949a-9be995f3ea17,t,2-naa.600144f0a9b12ec6000050b93e310002
   Runtime Name: vmhba39:C0:T0:L0
   Device: naa.600144f0a9b12ec6000050b93e310002
   Device Display Name: SUN iSCSI Disk (naa.600144f0a9b12ec6000050b93e310002)
   Adapter: vmhba39 Channel: 0 Target: 0 LUN: 0
   Adapter Identifier: iqn.1998-01.com.vmware:aie-4440d-5312c143
   Target Identifier: 00023d000001,iqn.1986-03.com.sun:02:a458fee1-24a7-c28a-949a-
9be995f3ea17,t,2
   Plugin: NMP
   State: active
   Transport: iscsi
   Adapter Transport Details: iqn.1998-01.com.vmware:aie-4440d-5312c143
   Target Transport Details: IQN=iqn.1986-03.com.sun:02:a458fee1-24a7-c28a-949a-
9be995f3ea17 Alias= Session=00023d000001 PortalTag=2
```

Note: The following command line can be filtered by iSCSI only.

```
# esxcfg-mpath -l | grep -i iSCSI
   Device Display Name: SUN iSCSI Disk (naa.600144f0a9b12ec6000050b93e310002)
   Transport: iscsi
   Device Display Name: SUN iSCSI Disk (naa.600144f0a9b12ec6000050b93e310002)
   Transport: iscsi
```

15. Once you have identified the right iSCSI LUN, run the following command to validate the multipath configuration.

```
# esxcfg-mpath -bd naa.600144f0a9b12ec6000050b93e310002
naa.600144f0a9b12ec6000050b93e310002 : SUN iSCSI Disk
(naa.600144f0a9b12ec6000050b93e310002)
   vmhba39:C0:T0:L0 LUN:0 state:active iscsi Adapter: iqn.1998-01.com.vmware:aie-4440d-
5312c143  Target: IQN=iqn.1986-03.com.sun:02:a458fee1-24a7-c28a-949a-9be995f3ea17
Alias= Session=00023d000001 PortalTag=2
   vmhba39:C1:T0:L0 LUN:0 state:active iscsi Adapter: iqn.1998-01.com.vmware:aie-4440d-
5312c143  Target: IQN=iqn.1986-03.com.sun:02:a458fee1-24a7-c28a-949a-9be995f3ea17
Alias= Session=00023d000002 PortalTag=2
```

16. Similarly to the previous instructions for Fibre Channel protocol and as part of the tuning options for iSCSI protocol, change the default storage array type as well as path selection policy and round robin I/O operation limit prior to putting the servers into production. Follow the steps shown in the next several code examples to perform this change. For changing the round robin I/O operation limit, use the steps shown in the following ESXi command lines. Identify all Oracle ZFS Storage Appliance iSCSI disks that will be utilized by your virtualized server.

Identify the Oracle ZFS Storage Appliance iSCSI disks:

```
esxcli storage nmp device list | egrep -i "SUN iSCSI Disk"
   Device Display Name: SUN iSCSI Disk (naa.600144f0fe9845750000513f7c570001)
   Device Display Name: SUN iSCSI Disk (naa.600144f0fe9845750000513f9b580002)
```

Using `for`, `egrep` and `awk` instructions as filters, get information for the devices for which the path selection policy and round robin I/O operation limit will be changed:

```
esxcli storage nmp device list | egrep -i "SUN iSCSI Disk" | awk '{ print $7 }' | cut
-c 2-37

naa.600144f0fe9845750000513f7c570001
naa.600144f0fe9845750000513f9b580002
```

```
for a in `esxcli storage nmp device list | egrep -i "SUN iSCSI Disk" | awk '{ print $7
}' | cut -c 2-37`
do
esxcli storage nmp psp roundrobin deviceconfig get -d $a
done
```

IMPORTANT: The recommended VMware Storage Array Type Plug-in (SATP) policy for iSCSI protocol with Oracle ZFS Storage Appliance is VMW_SATP_DEFAULT_AA. The following `esxcli` command presents the correct options for changing the path selection policy, and the storage array type plugin for the iSCSI LUNs attached to an ESXi host. This `esxcli` command will change the SATP policy of *all* of the iSCSI LUNs attached to the ESXi host. This change will only take effect after a reboot of the ESXi host.

Change the patch selection policy and the storage array type plug-in.

```
esxcli storage nmp satp rule add --transport=iscsi --satp=VMW_SATP_DEFAULT_AA --
psp=VMW_PSP_RR
```

Change the I/O operation limit and limit type of policy of iSCSI disks only:

```
for a in `esxcli storage nmp device list | egrep -i "SUN iSCSI Disk" | awk '{ print $7
}' | cut -c 2-37`
do
esxcli storage nmp psp roundrobin deviceconfig set -d $a -I 1 -t iops
```

```
done
```

Run the following command to ensure that the new values for operation limit and also round robin path switching have been updated:

```
for a in `esxcli storage nmp device list | egrep -i "SUN iSCSI Disk" | awk '{ print $7
}' | cut -c 2-37`
do
esxcli storage nmp psp roundrobin deviceconfig get -d $a
done
```

```
Device: naa.600144f0fe9845750000513f9b580002
IOOperation Limit: 1
Limit Type: Iops
Use Active Unoptimized Paths: false
```

17. Alter the following iSCSI software parameters listed in the following table.

TABLE 7. ISCSI SOFTWARE PARAMETERS	
ISCSI ADVANCED SETTINGS OPTION	**VALUE**
MaxOutstandingR2T	8
FirstBurstLength	16777215
MaxBurstLength	16777215
MaxRecvDataSegLen	16777215

To perform this task, right click on your iSCSI interface, then click on Properties and Advanced options, as seen in the following figure:

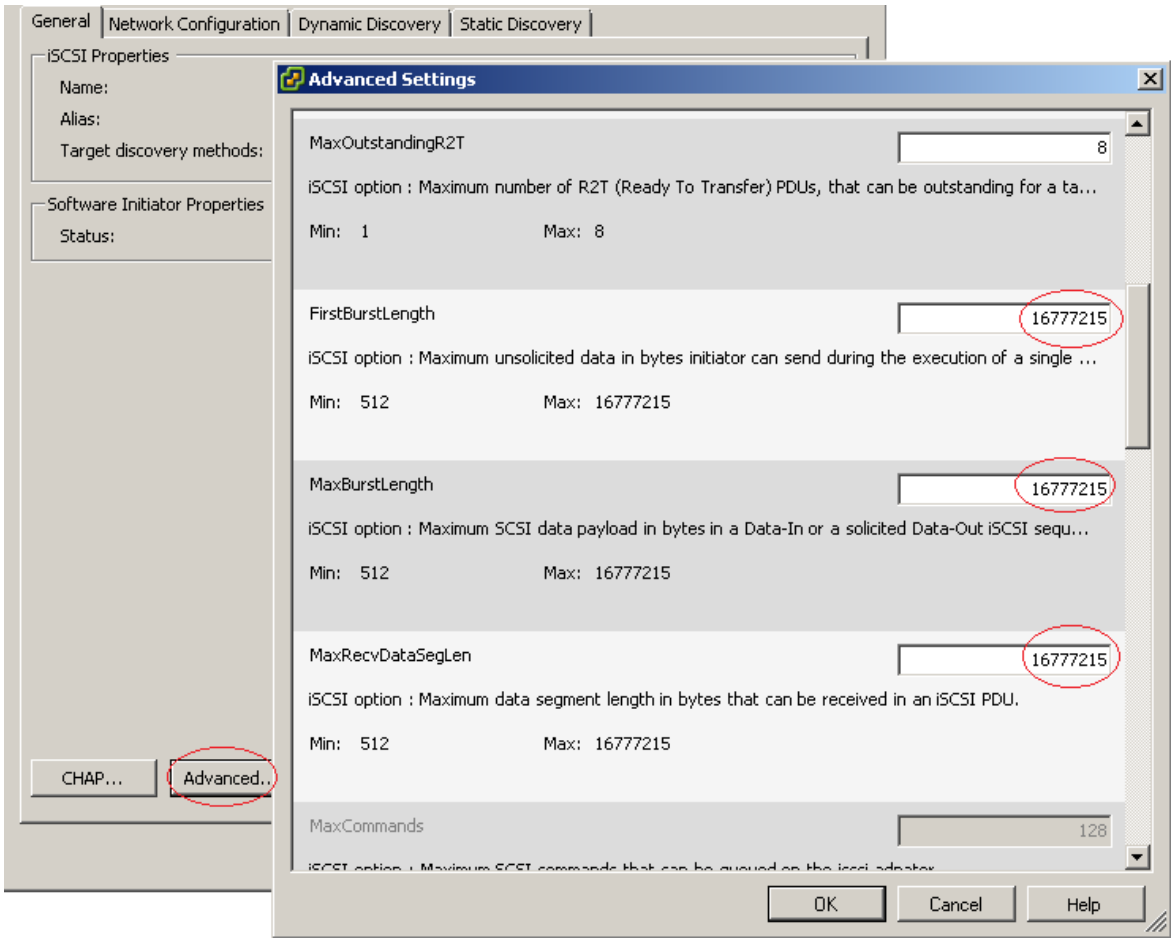

Figure 31. Changing iSCSI parameters in Advanced Settings

VMware Cluster Recommendations

VMware vSphere 5.x cluster configuration is beyond the scope of this white paper; however, when working with the Oracle ZFS Storage Appliance, the following options are recommended:

- Work with vSphere high availability (HA) and vSphere Distributed Resources Scheduler (DRS) cluster options.
- At the cluster automation level, use the 'fully automated' option, and choose the priority level that best fits your virtual environment.
- For the power management cluster (DPM), choose the automatic option and select the DPM threshold that best fits your virtualized environment.
- Enable host monitoring options and admission control.

- Choose the virtual machine restart option for your cluster. The example reflects a 'VM restart medium priority' and 'Powered on for host isolation response' option.
- Enable the VM monitoring option and choose the sensitivity that best fits your virtualized environment.
- Enable the 'Enhanced vMotion Compatibility' option for your cluster. Choose the right VMware EVC mode for your CPU (AMD or Intel).
- For swap file, select the option 'Store a swapfile in the same directory as the virtual machine'. Use a central datastore for swap files.

Using the Datastore Heartbeating Feature

For better HA management and also to avoid false positives due to network problems, VMware vSphere 5.0 added a new HA feature called Datastore Heartbeating. A heartbeating datastore can be any shared datastore across VMware hosts. With this feature, VMware hosts are able to exchange heartbeats utilizing shared VMFS datastores.

Note: Datastore heartbeating configuration needs to be performed after VMware datastore configuration.

To enable the datastore heartbeating feature for a VMware HA cluster with two nodes, you will need at least two shared datastores. Right-click on your VMware cluster profile. In the example the cluster name is `ESXi5`. Select the **Datastore Heartbeating** option and choose S**elect any of the cluster datastores** as shown in figure 32.

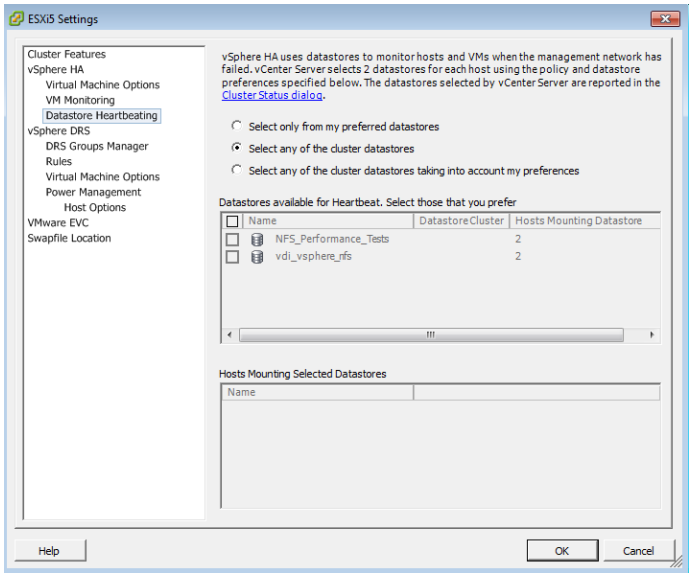

Figure 32. Enabling datastore heartbeating for `ESXi5` in the VMware vSphere 5.x client

Virtual Machine Data Layout

Recommendations for virtual machine data layout as well as best practices for a VMware virtual machine working with the Oracle ZFS Storage Appliance are:

- Work with VMware virtual machine version 8.
- To improve storage efficiency and performance, configure your virtual machine with a thin provisioning virtual disk drive using a VMware paravirtual SCSI controller.
- For raw devices and for LUNs with more than 2TB, use raw device mapping (RDM).
- When working with ZFS Storage Appliance Provider for Volume Shadow Copy Service Software, use RDM in physical compatibility mode.
- Note: Sun ZFS Storage Appliance Provider for Volume Shadow Copy Service Software (Microsoft Visual SourceSafe [VSS] Plug-in for Sun ZFS Storage Appliance) is not supported in a virtualized environment (VMware) using Fibre Channel or NFS protocol, only iSCSI using Microsoft iSCSI initiator software.
- To improve network performance, use a VMXNET3 network adapter.
- Install the VMware Client Tools. For more information on these tools and to install them, use the following link:

  http://www.vmware.com/pdf/vmware-tools-installation-configuration.pdf

- When working with the Microsoft Windows platform, ensure that you have the latest service pack as well as all recommended patches installed.
- Ensure that your virtual machine is working with the right partition alignment.
- Work with a central swap datastore for all virtual machines. By default VMware creates a virtual swapfile that usually is equal to the amount of memory allocated to each virtual machine. Reallocate the virtual machine swap file to a central VMware datastore.

  To configure the central swap datastore, select `ESXi5.1` in the VMware vSphere 5.x client. Select the **Configuration** tab, select **Virtual Machine Swapfile Location**, and then **Edit**. Select the vswap datastore that was previously configured for this purpose, as seen in figure 33.

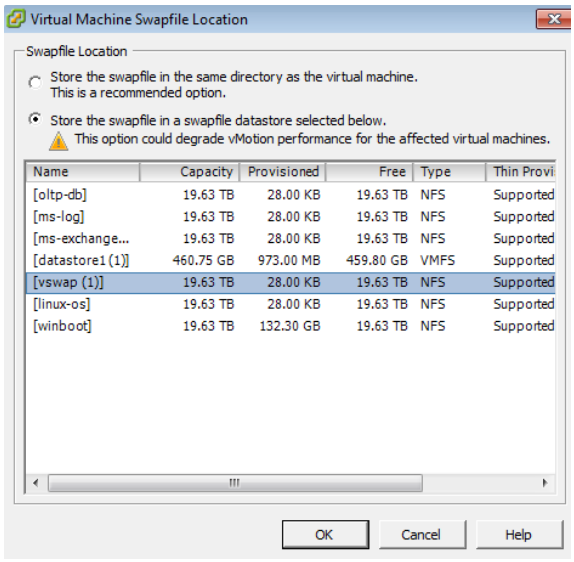

Figure 33. VMware ESXi5 host swapfile configuration

Right-click on the virtual machine which will have the swap file relocated to a different datastore. Select **Options**, **Swapfile Location**, and choose 'Store in the host's swapfile datastore' as seen in figure 34.

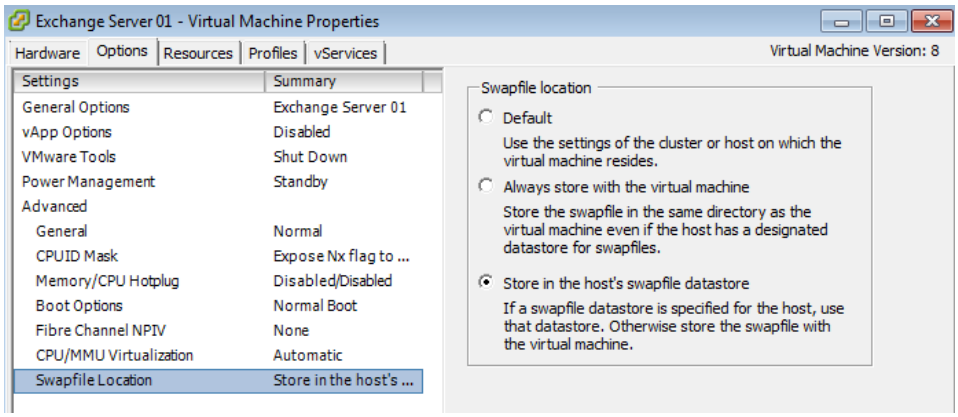

Figure 34. VMware virtual machine swapfile configuration

- As a best practice to achieve better performance for virtualized applications as well as management of your virtual environment, work with a multi-pool design with multiple datastore repositories in VMware vSphere 5.x. Figure 35 shows the high-level view of a virtual machine layout with a multi-pool design.

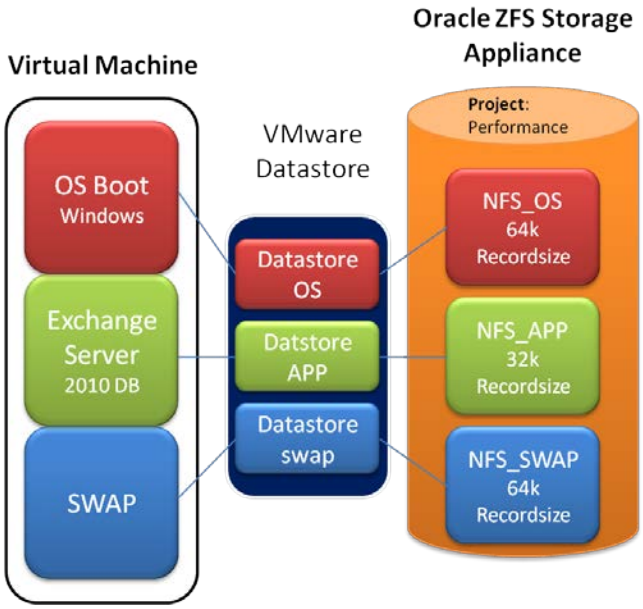

Figure 35. Recommended data layout for a VMware virtual machine

In this approach, the virtual machines are deployed in multiple datastore repositories, each with a different configuration. The example shows a single virtual machine configured with three different datastores. The first datastore is configured with a 64k database record size and is designed to host virtual machines' operating system disk images. The second datastore is configured with a 32k database record size and is designed to host all binaries of the virtualized applications Lastly, the third datastore is configured with a 64k database record size and is designed as a central swap area for all virtual machines.

The example in figure 36 shows a layout for a Microsoft Exchange Server that can be used for production environments. The layout consists of four different VMware datastores. The Exchange Server has been configured with 100GB of disk space for the operating system virtual disk, attached to eight 800GB RDM LUNs for the Exchange mail database and eight 150GB RDM LUNs for the mail-log virtual disk.

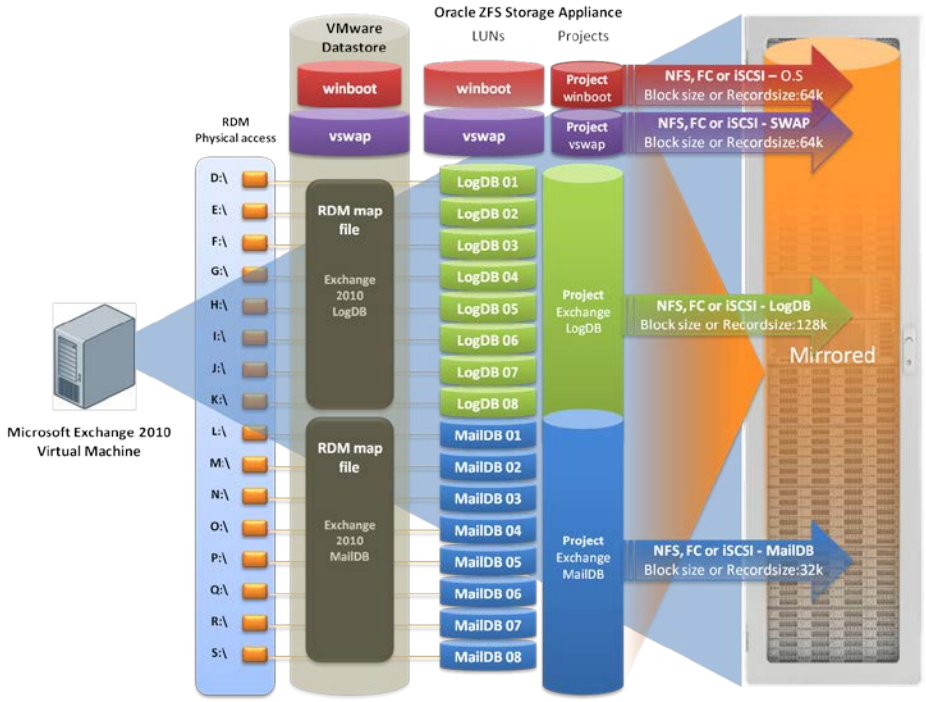

Figure 36. Data layout for a Microsoft Exchange virtual machine

VMware Linked Clone

Linked clone is a technology provided by VMware for cloning virtual machines. This technology allows multiple virtual machines to share virtual disks with a parent image. Linked clone improves the storage efficiency and also performance of cloning operations.

Note: Linked clone is only available through powershell or powerCLI scripts, not the VMware vCenter GUI.

To work with linked clone technology, use the following steps:

1. Use the linked clone script that follows.
2. Create a snapshot of the virtual machine for which you want to create the linked clones before running the script.
3. Edit the highlighted options in red to best fit your production environment. The options are: VMware vCenter host name, virtual machine name for which you want to have the linked clones, number of clones, and the total of concurrent clone operations.
4. Copy the contents of the script and save with the extension `.ps1`., then open PowerCLI and execute the script.

At this point you will be asked to enter the username and password of your VMware vCenter server. After the credentials have been validated, the linked-clone operation will start and you will see a screen similar to the one shown in figure 37.

Note: This operation is not supported with virtual disks in independent mode or raw device mappings in physical compatibility mode.

```
$VMHost="VCenter host name"

Add-PSSnapin VMware.VimAutomation.Core # Add PowerCLI cmdlets.

#Open the Connection to the vCenter Server
Connect-VIServer -Server $VMHost

#Get the VM that you want to clone
$VMs = "Windows 2008 R2"

$vm = Get-VM "Windows 2008 R2" | Get-View
$clonePrefix = "linked_clone_"
$numClones = 100
$concurrentClones = 20

$cloneFolder = $vm.parent
$cloneSpec = new-object Vmware.Vim.VirtualMachineCloneSpec
$cloneSpec.Snapshot = $vm.Snapshot.CurrentSnapshot
$cloneSpec.Location = new-object Vmware.Vim.VirtualMachineRelocateSpec
$cloneSpec.Location.DiskMoveType =
[Vmware.Vim.VirtualMachineRelocateDiskMoveOptions]::createNewChildDiskBacking

#This option is avaialbe to power on each clone immediately after it is created:
$cloneSpec.powerOn = $true

$i = 1
while ($i -le $numClones) {
$taskViewArray = @()
foreach ($j in 1..$concurrentClones) {
$taskViewArray += $vm.CloneVM_Task( $cloneFolder, $clonePrefix+$i, $cloneSpec )
$i++
}
$taskArray = $taskViewArray | Get-VIObjectByVIView
Wait-Task $taskArray
}
```

Figure 37 shows the PowerCLI screen during execution of the linked clone script.

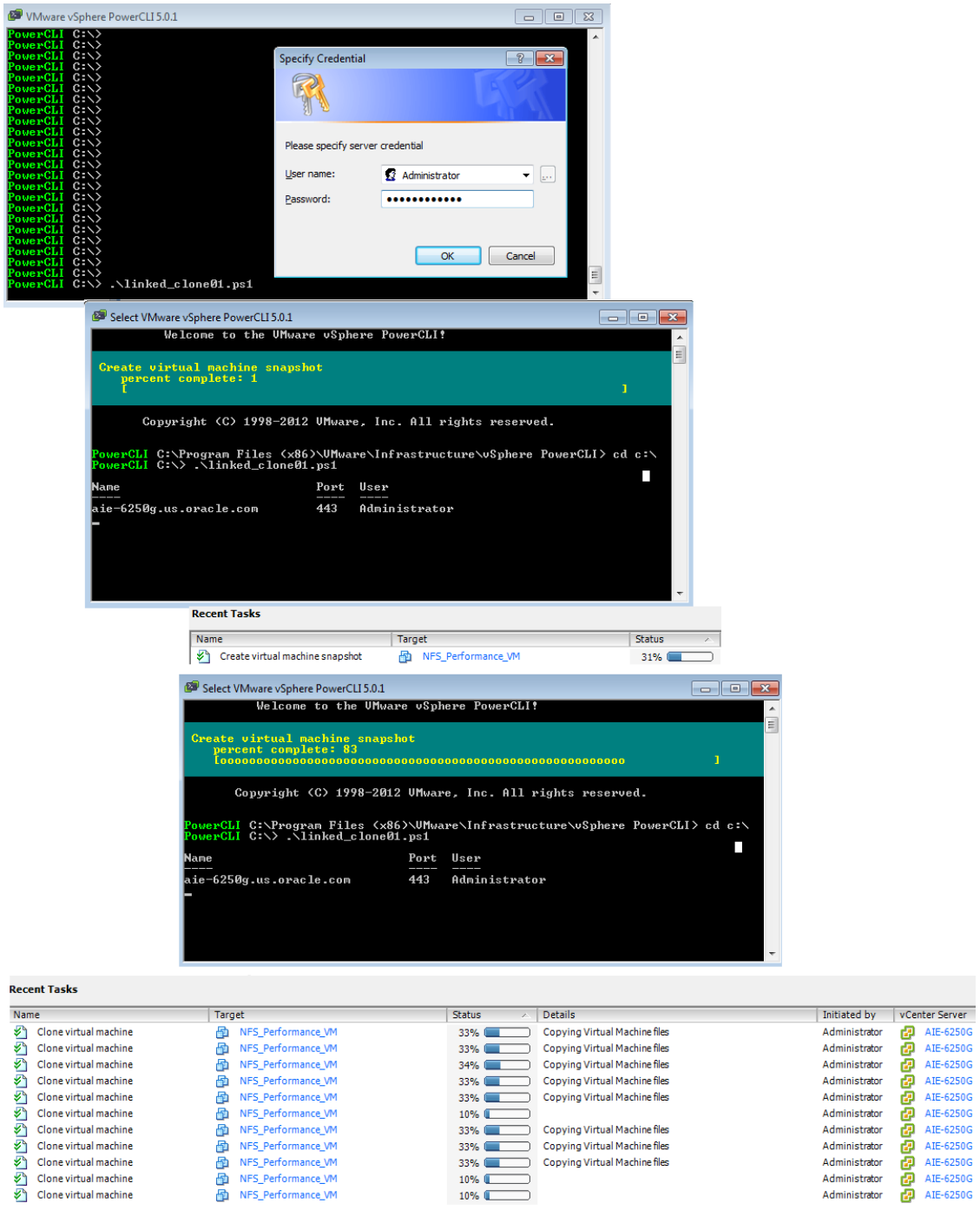

Figure 37. VMware linked clone script execution

Monitoring VMware with DTrace Analytics and ESXTOP

DTrace Analytics is a powerful and advanced monitoring tool provided by the ZFS Storage Appliance. DTrace Analytics gives the storage administrator unique visibility into the entire system: monitoring in real-time of different statistics for the operating system stack, storage resources, and protocols used as well as I/O throughput and performance of the virtualized environment.

VMware provides a powerful monitoring tool called ESXTOP that is used at the VMware ESXi host level to monitor performance and resource usage of the virtual environment. With this tool, you can identify possible bottlenecks, I/O performance issues, and network degradation as well as throughput levels.

VMware ESXTOP and DTrace Analytics should always be used together to validate as well as monitor your entire VMware storage performance and throughput for the most realistic report. To ensure that your VMware's NFS configuration is properly working, use the following DTrace Analytics and ESXTOP options.

Monitoring Fibre Channel Performance

The following examples show how to use ESXTOP and DTrace Analytics to monitor VMware Fibre Channel and iSCSI LUNs as well as datastore and HBA performance and throughput.

For VMware ESXTOP, open a SSH connection with your ESXi5.x host and run the following commands:

1. Type **`esxtop`** and then press **`n`** for monitoring VMware Fibre Channel or iSCSI LUNs.
2. Press **`s 2`** to alter the update time to every 2 seconds, and press **Enter**.

Figure 38 shows the VMware ESXTOP output of the "`n`" option.

Note: For interpreting VMware ESXTOP statistics, read VMware DOC-9279 at the following URL: http://communities.vmware.com/docs/DOC-9279

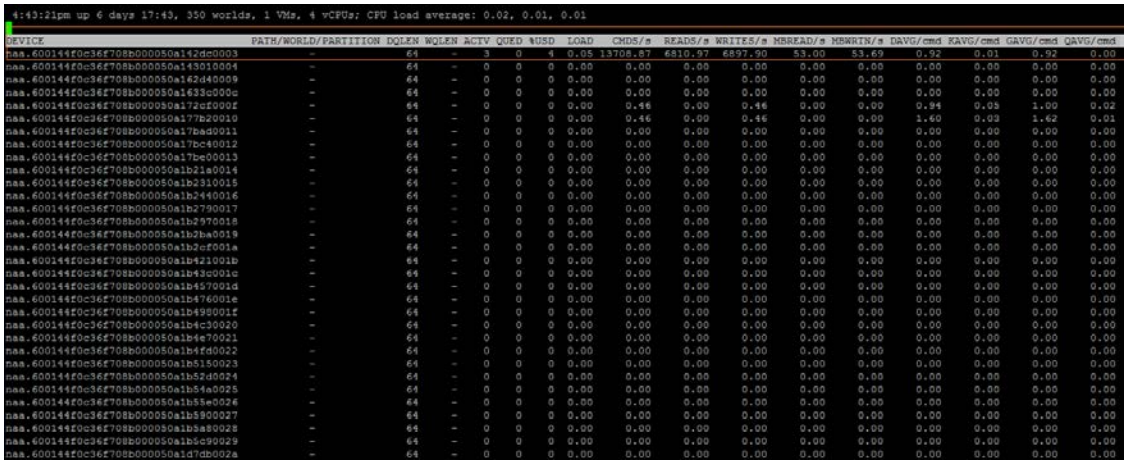

Figure 38. Monitoring Fibre Channel and iSCSI LUNs with VMware ESXTOP

Option `d` or disk view option (HBA mode) can be used for monitoring virtual HBAs. Figure 39 shows output for this option.

To perform this task, use the `esxtop` command, then type `f` to choose different monitoring options. Type `s 2` to alter the update time to every 2 seconds and then press Enter.

Note: Ensure that the virtual HBAs (vmhbas) are correctly balancing the I/O. The example highlights `vmhba6` and `vmhba7`. Watch all options available on the screen shown in figure 39, and then compare with DTrace Analytics outputs.

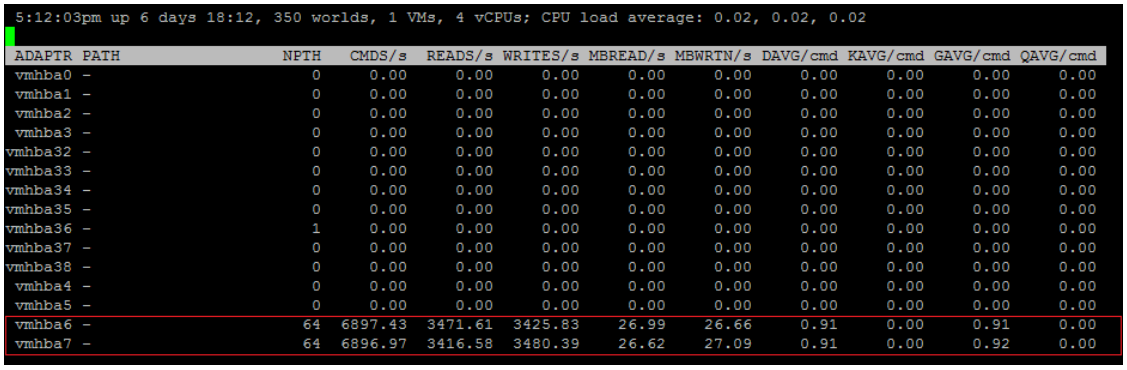

Figure 39. Monitoring VMware HBAs with ESXTOP

The following figures show different examples of DTrace Analytics that can be used in combination with VMware ESXTOP for monitoring Fibre Channel performance and throughput.

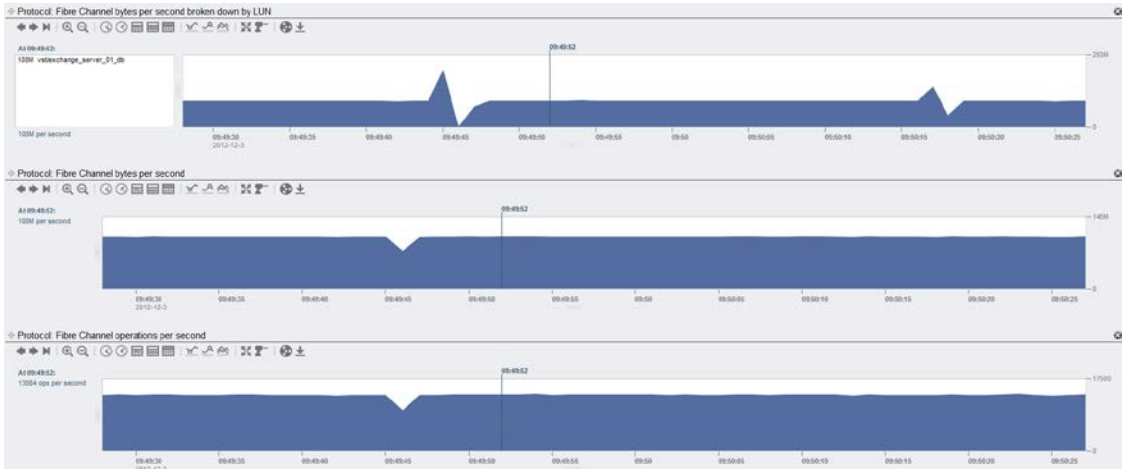

Figure 40. Example 1 – Monitoring Fibre Channel protocol using DTrace Analytics

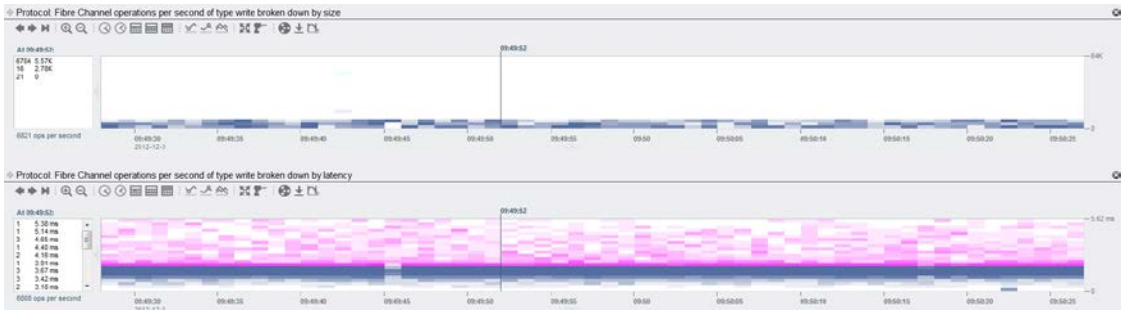

Figure 41. Example 2 – Monitoring Fibre Channel protocol using DTrace Analytics

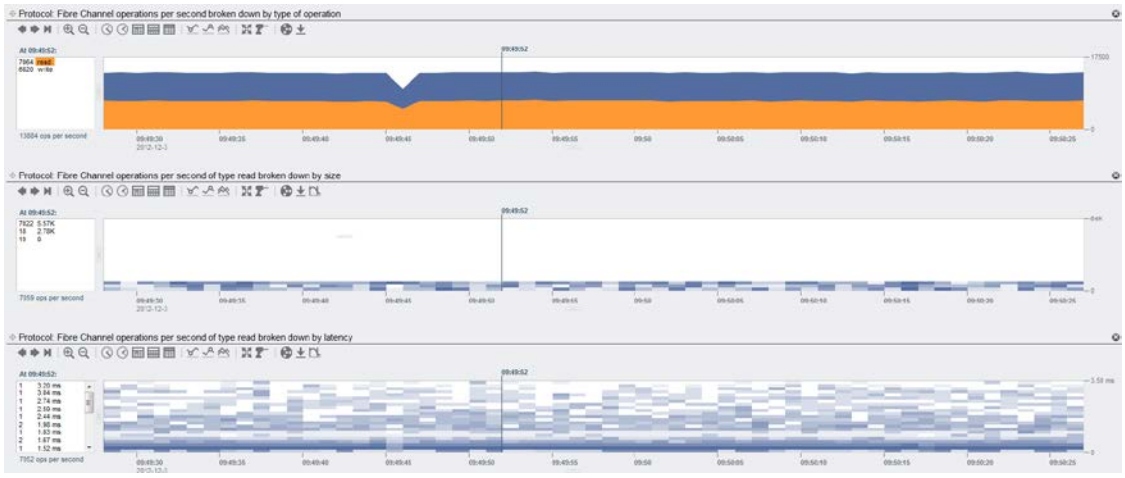

Figure 42. Example 3 – Monitoring Fibre Channel protocol using DTrace Analytics

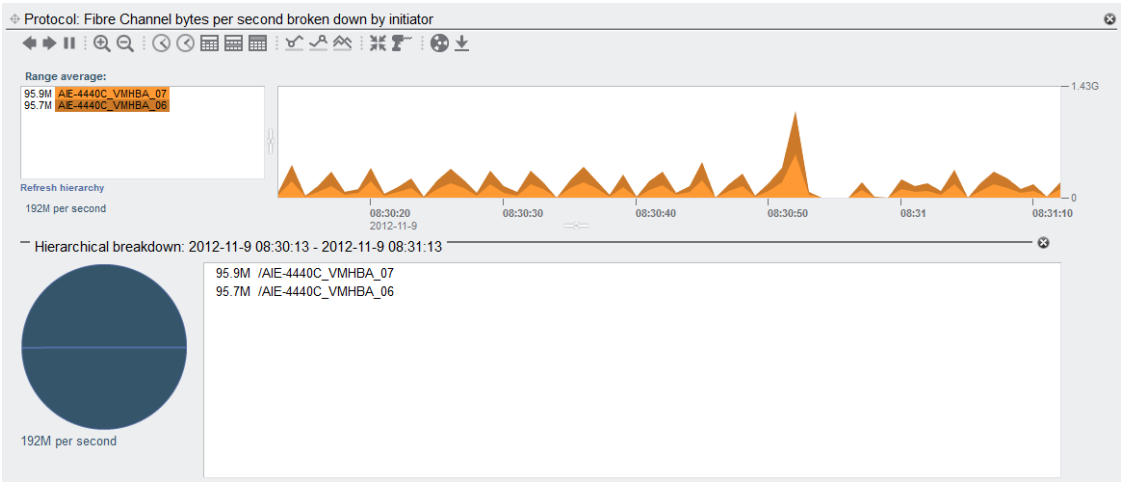

Figure 43. Example 4 – Monitoring Fibre Channel protocol using DTrace Analytics

Monitoring NFS Performance

The following figures show examples of VMware ESXTOP and DTrace Analytics output used to monitor NFS datastores' utilization and performance as well as the IP network.

Figures 44 to 48 show that the NFSv3 protocol is used for virtual machines' disk datastore. In this approach, DTrace Analytics is monitoring the virtual machines' disk usage in IOPS for each vmdk file.

Figure 44 shows the VMware ESXTOP option for monitoring NFS datastores. To perform this, run `esxtop`, then type `u`. Type `s 2` to alter the update time to every 2 seconds and press **Enter**.

```
DEVICE                                PATH/WORLD/PARTITION  DQLEN  WQLEN  ACTV  QUED  %USD  LOAD  CMDS/s  READS/s  WRITES/s  MBREAD/s  MBWRTN/s  DAVG/cmd  KAVG/cmd
{NFS}NFS_Performance_Tests            -                     -      -      2     -     -     -     6713.40 3357.87  3355.54   25.46     25.57     -         -
{NFS}linux-os                         -                     -      -      0     -     -     -     0.00    0.00     0.00      0.00      0.00      -         -
{NFS}ms-exchangedb                    -                     -      -      0     -     -     -     0.00    0.00     0.00      0.00      0.00      -         -
{NFS}ms-log                           -                     -      -      0     -     -     -     0.00    0.00     0.00      0.00      0.00      -         -
{NFS}oltp-db                          -                     -      -      0     -     -     -     0.00    0.00     0.00      0.00      0.00      -         -
{NFS}vdi_vsphere_nfs                  -                     -      -      0     -     -     -     0.47    0.00     0.47      0.00      0.00      -         -
{NFS}vswap (1)                        -                     -      -      0     -     -     -     0.00    0.00     0.00      0.00      0.00      -         -
{NFS}winboot                          -                     -      -      0     -     -     -     0.00    0.00     0.00      0.00      0.00      -         -
```

Figure 44. Monitoring VMware NFS datastores with VMware ESXTOP

Figure 45 shows the VMware ESXTOP option for monitoring virtual machines. To perform this, run `esxtop`, then type `v`. Type `s 2` to alter the update time to every 2 seconds and press **Enter**.

```
    GID VMNAME          VDEVNAME NVDISK   CMDS/s  READS/s WRITES/s MBREAD/s MBWRTN/s LAT/rd LAT/wr
 273229 NFS_Performance        -      1  7065.17  3523.27  3541.90    26.88    27.00   2.10   2.78
```

Figure 45. Monitoring VMware virtual machine with VMware ESXTOP

Figure 46 shows DTrace Analytics options for monitoring network datalink as well as interface and TCP bytes.

Figure 46. Monitoring network datalinks, interface and TCP bytes with DTrace Analytics

Figure 47 shows DTrace Analytics options for monitoring NFS protocol broken down by type of operation, clients and also per file name – in this case, `.vmdks` files.

Figure 47. Monitoring NFS protocol broken down by type of operation, clients and file name with DTrace Analytics

Figure 48 shows more DTrace Analytics options for monitoring NFS protocol broken down by latency and by size, as well as cache ARC broken down by hit/miss.

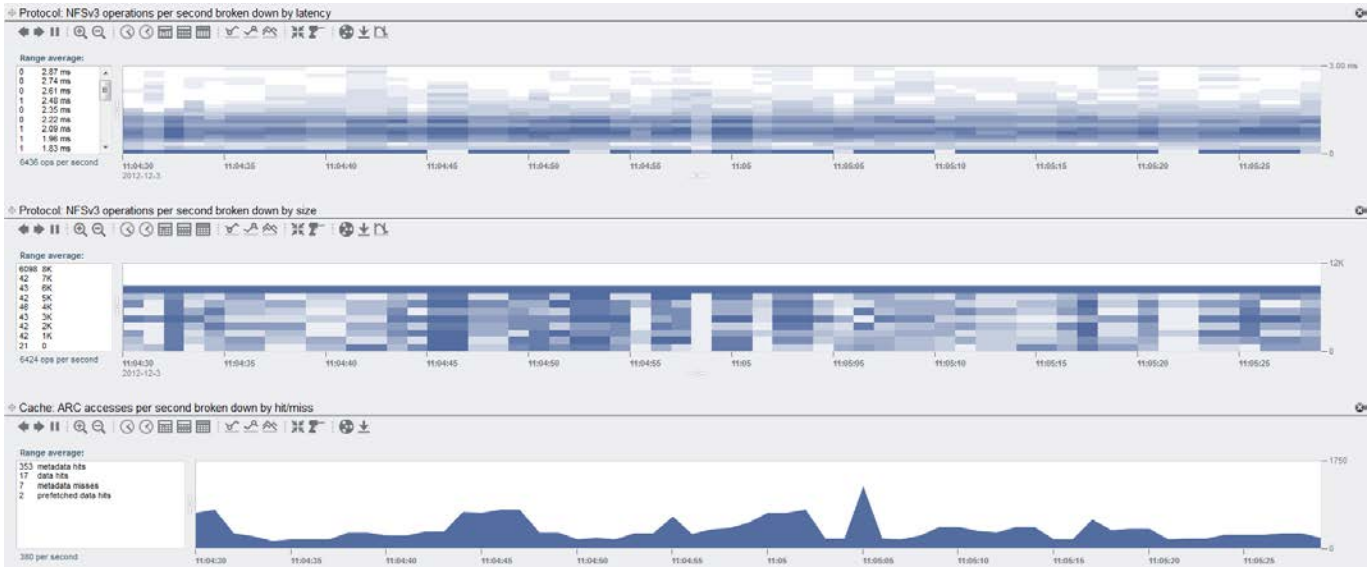

Figure 48. Monitoring NFS protocol broken down by latency and size, and cache ARC broken down by hit/miss with DTrace Analytics

Monitoring iSCSI Performance

The following figures show different examples of VMware ESXTOP and DTrace Analytics monitoring iSCSI protocol utilization and performance.

Note: Some options used to monitor NFS protocol, such as TCP and interfaces, as well as some options used to monitor Fibre Channel protocol can be used for monitoring iSCSI protocol as well.

Figure 49 shows the VMware ESXTOP option for monitoring iSCSI datastores. To perform this, run `esxtop`, then type `u`. Type `s 2` to alter the update time to every 2 seconds and press Enter.

Figure 49. Monitoring iSCSI protocol utilization and performance with VMware ESXTOP

Figure 50 shows the VMware ESXTOP option for monitoring iSCSI virtual HBAs. The example shows the virtual HBA `vmhba39`.

ADAPTR	PATH	NPTH	CMDS/s	READS/s	WRITES/s	MBREAD/s	MBWRTN/s	DAVG/cmd	KAVG/cmd	GAVG/cmd	QAVG/cmd
vmhba0	-	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
vmhba1	-	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
vmhba2	-	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
vmhba3	-	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
vmhba32	-	1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
vmhba33	-	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
vmhba34	-	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
vmhba35	-	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
vmhba36	-	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
vmhba37	-	1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
vmhba38	-	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
vmhba39	-	2	10032.60	4989.04	5043.55	38.83	39.22	1.94	0.00	1.95	0.00
vmhba4	-	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
vmhba5	-	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
vmhba6	-	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
vmhba7	-	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Figure 50. Monitoring iSCSI virtual HBAs with VMware ESXTOP

The following figures show different examples of DTrace Analytics which can be used in combination with VMware ESXTOP for monitoring iSCSI performance and throughput.

Figure 51 shows DTrace Analytics options for monitoring iSCSI protocol broken down by initiators, targets and LUNs.

Figure 51. Monitoring iSCSI protocol broken down by initiators, targets and LUNs with DTrace Analytics

Figure 52 shows more DTrace Analytics options for monitoring iSCSI protocol broken down by client and LUNs (operations per second) as well as network interfaces broken down by interface (bytes per second). The LUNs display appears at the bottom of figure 52.

Figure 52. Monitoring iSCSI protocol broken down by clients, network interfaces, and LUNs with DTrace Analytics

Figure 53 shows more DTrace Analytics options for monitoring iSCSI protocol broken down by latency and devices.

Note: Ensure that all your 10GbE NICs' members of the VMware port binding configuration are correctly balancing the I/O traffic. The example in figure 52 shows two 10GbE NICs (`ixgbe0` and `ixgbe1`) balancing the I/O traffic. That is the expected behavior when you are working with an iSCSI port binding configuration as well as a storage array type VMW_SATP_ALUA and path selection policy VMW_PSP_RR. If you do not see this behavior, review your iSCSI port binding configuration as well as VMware path policy and storage array type. Also review the network configuration of the Oracle ZFS Storage Appliance and the port-channel configurations of your IP switches.

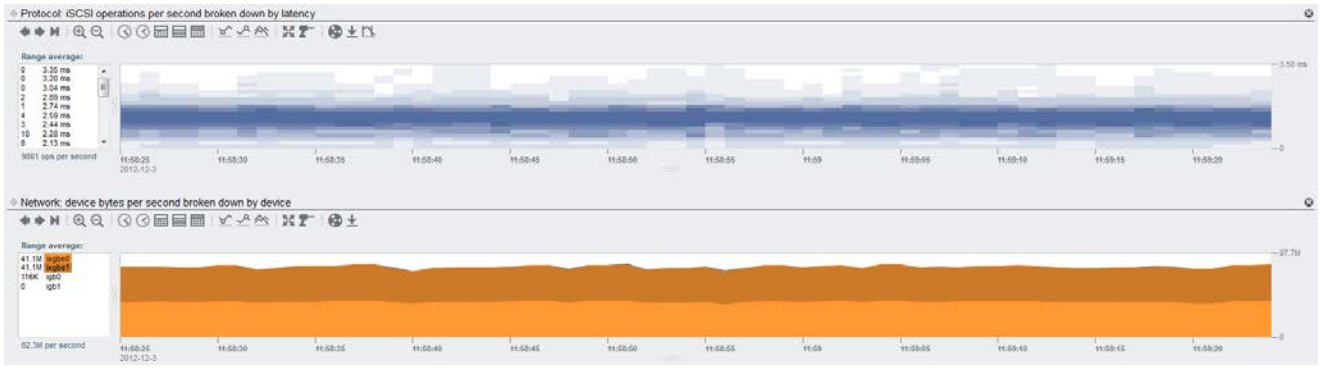

Figure 53. Monitoring iSCSI protocol broken down by latency and network devices with DTrace Analytics

Also, you can work with VMware ESXTOP in batch mode, which permits you to output data into a CSV format, and then use `perfmon` on Windows or even VMware `esxplot` to view the data/results. To run ESXTOP in batch mode, type:

```
esxtop -b > esxtop_whatever.csv.
```

It is always a best practice to use more DTrace Analytics options, such as ARC, L2ARC access by hit/miss and latency as well as disk I/O outputs.

Conclusion

Oracle ZFS Storage Appliance offers outstanding performance for virtualized environments. Its architecture's features and intelligent caching technology are designed to deliver thousands of IOPS for your virtualized environment as well as the best throughput and response time for your virtualized applications and databases.

VMware is a robust hypervisor and also provides an easy way to manage your virtualized infrastructure. In combination, the VMware and Oracle ZFS Storage Appliance platforms and technology are an excellent choice for your virtualized environment.

Appendix A: Benchmark Results

Refer to the following web sites for further information on testing results for the Oracle ZFS Storage Appliance.

SPC-2 Results

http://www.storageperformance.org/benchmark_results_files/SPC-2/Oracle_SPC-2/B00058_Oracle_ZFS-7420/b00058_Oracle_Sun-ZFS_7420_SPC2_executive-summary.pdf

Oracle Quality Awards for NAS

http://www.oracle.com/us/products/servers-storage/storage/nas/storage-quality-awards-jan12-1521728.pdf

Appendix B: References

Oracle ZFS Storage Appliance Documentation

References to Sun ZFS Storage Appliance, Sun ZFS Storage 7000, and ZFS Storage Appliance all refer to the same family of Oracle ZFS Storage Appliance products. Some cited documentation may still carry these legacy naming conventions.

RESOURCES	LOCATION
Sun ZFS Storage Appliance Administration Guide	http://download.oracle.com/docs/cd/E22471_01/index.html
Sun ZFS Storage 7000 Analytics Guide	http://docs.oracle.com/cd/E26765_01/pdf/E26398.pdf
Sun ZFS Storage 7x20 Appliance Installation Guide	http://docs.oracle.com/cd/E26765_01/pdf/E26396.pdf
Sun ZFS Storage 7x20 Appliance Customer Service Manual	http://docs.oracle.com/cd/E26765_01/pdf/E26399.pdf
VMware	http://www.vmware.com
VMware Multipathing policies in ESX/ESXi 4.x and ESXi 5.x	http://kb.vmware.com/selfservice/microsites/search.do?language=en_US&cmd=displayKC&externalId=1011340
VMware Knowledge Base: "Changing the queue depth for QLogic and Emulex HBAs"	http://kb.vmware.com/selfservice/microsites/search.do?language=en_US&cmd=displayKC&externalId=1267
VMware vSphere 5.1 Documentation	http://www.vmware.com/support/pubs/vsphere-esxi-vcenter-server-pubs.html

Best Practices for Oracle ZFS Storage Appliance and VMware vSphere 5.x
March 2014, Version 1.2
Author: Anderson Souza

Oracle Corporation
World Headquarters
500 Oracle Parkway
Redwood Shores, CA 94065
U.S.A.

Worldwide Inquiries:
Phone: +1.650.506.7000
Fax: +1.650.506.7200

oracle.com

Oracle is committed to developing practices and products that help protect the environment

Hardware and Software, Engineered to Work Together