
An Oracle Technical White Paper
July 2014

# Deploying 10,000+ VMware Virtual Machines with Oracle ZFS Storage Appliance

ORACLE

# Introduction

The storage host of a multiple virtual machines infrastructure must drive, in general, four primary loads:

- Provisioning, cloning, snapshots and virtual machine live migration operations – Provisioning, (the process of supplying the OS image) is typically a serial process, so it is more sensitive to I/O latency than other processes. Provisioning through image copy is a heavyweight process, even if used in combination with data deduplication technology, for example. Provisioning using snap/clone is, by contrast, a lightweight process and is vastly preferred when feasible.
- Dump devices – Dump devices are typically used only in the event of a crash, so they do not have much activity on them. While it is rare to have many virtual machines crash at the same time, it is *not* impossible. For example, if Linux guests are configured with the kernel.hung_task_panic set, a common event might cause many Linux guests to panic simultaneously. Obviously, as a rule, this parameter should not be set because when a crash dump happens, it creates a more intensive workload than most virtual machine operaitons.
- Virtual machine workload or footprint – The workload generated by the application that is running inside of each virtual machine. The application can have a variety of diferent I/O patterns, such as online transaction processing (OLTP), mail server operations, web server workloads, and so on.
- Boot process – This is a primary focus in this paper. The boot process is about the same amount of work regardless of whether the protocol is iSCSI, Fibre Channel (FC) or network file system (NFS). For reasons that this paper will cover, experience has shown that NFS is an ideal choice for VMware workloads.

All these different storage I/O workloads can generate potentially thousands of I/O operations per second, which demands an intelligent storage architecture with fast response time, low latency, and performance.

This white paper provides the storage considerations to reach optimal I/O performance and throughput for large-count VMware virtual machine deployments with Oracle ZFS Storage Appliance.

The outlined storage considerations and recommendations highlight configuration and tuning options for VMware NFS protocols, disk layout recommendations, and correct design of IP network infrastructure for a VMware vSphere 5.x environment working with Oracle ZFS Storage Appliance.

Highlighted in this paper are:

- VMware boot storm overview and considerations
- Storage consideration for booting large (over 10,000) VMware virtual machines using NFS protocol and Oracle ZFS Storage Appliance

- VMware 10 GbE network infrastructure and tuning options for NFS protocols
- Oracle ZFS Storage Appliance disk pool layout for over 10,000 (10K+) VMware virtual machines
- VMware simultaneous virtual machines boot and performance results with Oracle ZFS Storage Appliance
- NFS considerations for deploying thousands of VMware virtual machines on the Oracle ZFS Storage Appliance

## Overview of 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<br>(2 controllers) | Oracle ZFS Storage ZS3-2 cluster<br>256 gigabytes (GB) direct random access memory (DRAM) per controller<br>Four x 20 3 terabyte (TB) disk drivers - Oracle Storage Drive Enclosure DE2-24C<br>Four x 10 gigabit Ethernet (GbE) network interface cards (NICs) (per controller)<br>Four x 73 GB log devices |
| IP Network Switch | 2 | 10 GbE network switch |

Table 2 shows the virtual machine components used.

**TABLE 2. VIRTUAL MACHINE COMPONENTS USED IN REFERENCE ARCHITECTURE**

| OPERATING SYSTEM | QUANTITY | CONFIGURATION |
|---|---|---|
| Oracle Linux 6.2 | 10,000+ | Linux Virtual Machine |

Table 3 shows the software used.

**TABLE 3. SOFTWARE USED IN REFERENCE ARCHITECTURE**

| SOFTWARE | VERSION |
|---|---|
| Oracle ZFS Storage Appliance OS | 2013.1.2.0 |
| Vdbench | 50401 |
| Sun StorageTek Workload Analysis Tool (SWAT) | 3.0.2 – MOS Patch 10350687 |
| VMware vCenter Server | 5.1u2 |
| VMware ESX hypervisor software | 5.1u2 |

## VMware Boot Storm and the Importance of Dynamic Caching

A boot storm occurs when a large number of virtual machines or virtual desktops are booted up within the same time window, causing a degradation of performance and service. For virtual machines and virtual desktop infrastructure environments, the initial startup operation generates intensive storage I/O workload while the application and operating system execute a large amount of read operations from disk. This scenario can be devastating for storage devices not able to handle a large amount of unusual and unpredictable read I/O workloads generated by these virtual machines. The storage devices can create a huge drag on network performance, as well as storage I/O degradation and slow or unresponsive virtual desktops.

Oracle ZFS Storage Appliance has been designed to handle extreme I/O workload with low latency, especially for unpredictable I/O workloads generated by virtual desktop infrastructure and virtual machines. Key to its effectiveness is its ability to manage dynamic caching. Hybrid Storage Pools, a feature of the Oracle ZFS Storage Appliance, integrate memory, flash and physical disks for efficient data placement and performance. Using an intelligent and adaptive set of algorithms to manage I/O, the Oracle ZFS Storage Appliance is able to make the most efficient use of hardware resources – dynamic random access memory (DRAM), read- and write-optimized flash-based SSD, and SAS disks – for optimal performance efficiency. Read and write paths are each handled in a distinct manner to address their unique performance and data integrity needs.

DRAM is used as a primary cache to accelerate reads. Because DRAM is a much faster media type than either disk or flash for transactional workloads, having a high proportion of read operations served out of DRAM radically accelerates overall system performance. The portion of DRAM used to serve as a read cache is known as the Adaptive Replacement Cache (ARC). DRAM allocation to ARC is managed by the operating system on a dynamic basis to maximize overall system performance. In the ARC, blocks are classified as most recently used (MRU), most frequently used (MFU), least recently used (LRU), or least frequently used (LFU). This highly adaptive dynamic caching architecture is seen in the following figure.

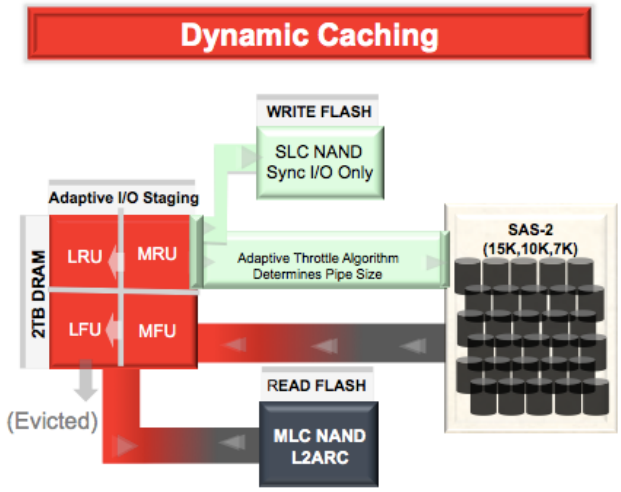


Figure 1. Dynamic caching functionality of the Oracle ZFS Storage Appliance

Note the following abbreviations and their meanings in figure 1: SLC (Single Level Cell), MLC (Multi Level Cell), NAND (Flash technology that offers faster erase, write, and read capabilities).

The concept is to keep the "hottest" portion of the overall data set in DRAM. As ARC becomes saturated and hotter data needs to replace cooler data in the ARC, the Hybrid Storage Pool will evict the coolest data in DRAM to a read flash cache device. This is known as the Level 2 ARC (L2ARC), for which the Oracle ZFS Storage Appliance uses SSDs. Read requests for data that had not been judged hot enough to be placed in either ARC or L2ARC must be served from spinning disk, resulting in a higher latency on those reads. However, in practice, it is common to have ARC hit rates in excess of 80 percent across a wide sampling of installed base systems.

For boot storm events, the Oracle ZFS Storage Appliance provides enough L1 ARC to cache boot images, while also providing excellent service quality for the random I/O workloads originated by virtual machines or virtual desktops environments.

Other important components of the Oracle ZFS Storage Appliance are its multicore CPUs and its operating system based in a symmetric multiprocessing architecture (SMP) that takes full advantage of the multicore CPUs and their multithreading capabilities. Importantly, VMware and other virtualization technologies create a workload that is best suited for SMP architecture.

## Simultaneously Booting 10,000+ VMware Virtual Machines

Due to the intelligent caching architecture, and in a cluster configuration, Oracle ZFS Storage Appliance ZS3-2 can work with a massive amount of virtual machines. How many virtual machines can an Oracle ZFS Storage ZS3-2 cluster system boot and support with optimal performance level? The answer is presented in the following chart.

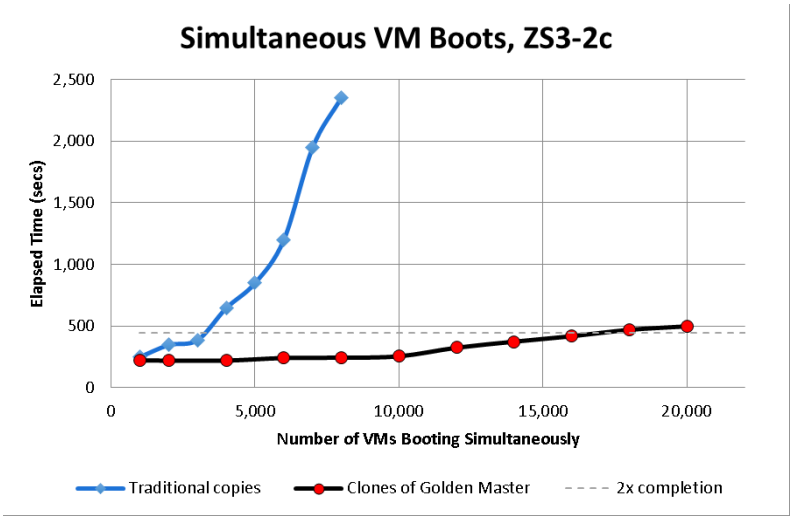


Figure 2. Simultaneous VMware virtual machines boot disks with Oracle ZFS Storage Appliance ZS3-2 (per head)

Basically, for virtual machines that use traditional copy methods, the maximum supported boots for an Oracle ZFS Storage ZS3-2 system is 8,000 virtual machines per head, and 16,000 virtual machines per Oracle ZFS Storage ZS3-2 cluster. If you are working with virtual machine clones, which utilize a golden masters method or VMware linked-clones technologies, the results are even better, as an Oracle ZFS Storage ZS3-2 system will be able to boot up to 20,000 virtual machines in a cluster.

The dashed line presented on the graph represents a conservative cutoff for response time. Using either a traditional copy or golden master clone method, booting 500 VMs takes almost exactly the same amount of time. The cutoff occurs when it starts taking twice as long as it did with the minimal population. That is about 3,500 booting VMs for the traditional method, and more than 16,000 booting VMs when using clones. Even with 20,000 simultaneously booting VMs, the boot process takes only about 8.5 minutes to complete.

These numbers may appear optimistic because they account *only* for booting systems, so there is no capability left in the controllers to handle other workloads while the storm is happening. However, if the boots are truly simultaneous, this is quite reasonable, since the VMs are all booting at once, and once they are finished, the workload on the controller drops to essentially nil.

It is important to note that, while serving up boot images for 20,000 VMs, each of the two controller heads was handling in excess of 150,000 NFS ops/sec, for a system capability of over 300,000 NFS ops/sec, as seen in the following Oracle ZFS Storage analytics screenshot. This performance considerably exceeds the reported benchmark score of 210 k ops/sec because, essentially, all of these reads are coming from L1ARC.

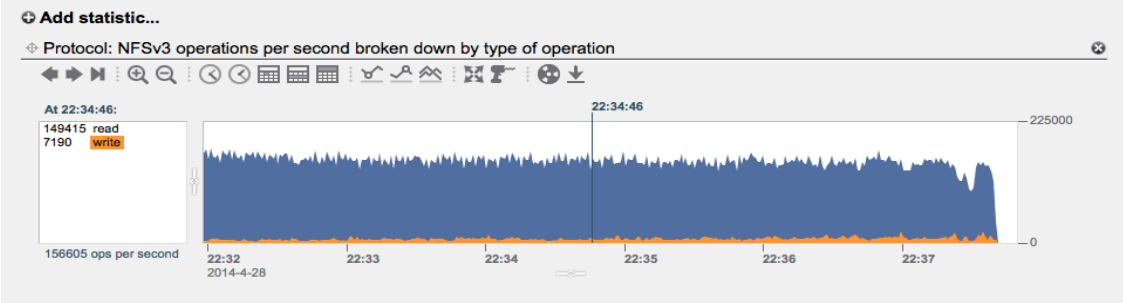


Figure 3. Oracle ZFS Storage Appliance ZS3-2 Analytics performance results of NFS protocol broken down by type of operation

If you are planning a huge deployment of VMware virtual machines with an Oracle ZFS Storage Appliance ZS3-2 system, note the blue line presented on the preceding chart, which is more common for VMware virtual machine servers, where you will be working with traditional (or full) VMware vmdk copies per server.

The black line result would be of more interest for a large deployment of VMware virtual desktops, which allows you to work with different cloning methods such as storage cloning from a golden image and VMware linked-clones.

## Storage Considerations for Large Virtual Machines Deployment

This section provides storage considerations for booting a large amount of VMware virtual machines using NFS protocol and Oracle ZFS Storage Appliance. Ensure that you are working with at minimum the following configuration:

- Oracle ZFS Storage Appliance in Active/Active mode
- Oracle ZFS Storage Appliance software release 2013.1.2.x or greater
- Oracle ZFS Storage Appliance controllers with at least 256 GB of DRAM (L1ARC) cache per controller
- At least two Intel® Xeon® E5-2658 0 2.10 GHz CPUs per Oracle ZFS Storage Appliance controller
- At least 4x dual 10 GbE network interfaces per controller

### Oracle ZFS Storage Appliance Disk Layout

Scenario one presents the minimum recommended disk storage layout for booting a large amount of virtual machines in a VMware vSphere5.x environment attached to an Oracle ZFS Storage Appliance ZS3-2 system. Assuming that you are working with small Linux virtual machines, or even small virtual desktops (which have about 10 GB of thin provisioning OS boot disk), and a footprint of approximately 30 IOPs per VM, the minimum recommended disk configuration is:

- Two mirrored disk pools of (at least) 44 x 4 TB SAS-2 (7200 RPM capacity disk drives)
- At least two 73 GB SSD devices for LogZilla working with a stripped log profile
- Four x 10 GbE network interface cards

**Note:** VMware does not support more than 512 virtual machines per ESXi5.x host, so during the sizing consideration, ensure that you have enough ESXi5.x servers available to host the large amount of virtual machines. The scenarios presented in this paper used a small size memory and virtual CPU configuration per VM (1 GB of memory and one virtual CPU per VMware virtual machine). Following this configuration, for 10,000 VMware virtual machines, you would need at least 20 VMware ESXi 5.x servers, for 16,000 virtual machines you would need 32 VMware ESXi servers, and for 20,000 virtual machines you would need at least 40 VMware ESXi servers.

For VMware sizing considerations, refer to the following document, "VMware vSphere5.5 Configuration Maximums" at:
http://www.vmware.com/pdf/vsphere5/r55/vsphere-55-configuration-maximums.pdf

The example in the following figure demonstrates 44 x 4 TB SAS-2 7200 RPM disks (per head). Note that DE2-24C denotes the Oracle Storage Drive Enclosure DE2-24C.

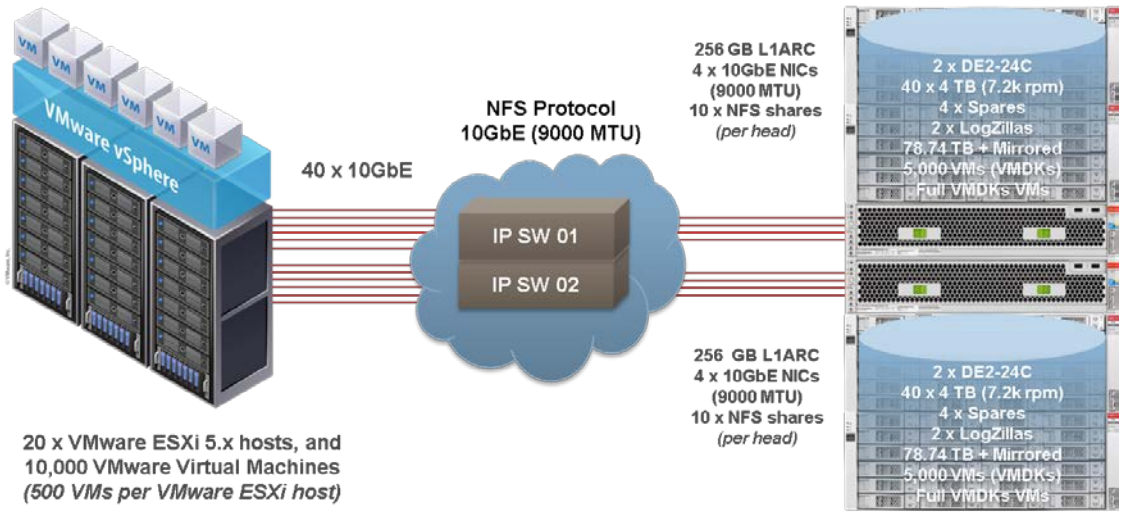


Figure 4. Oracle ZFS Storage Appliance – Disk layout for small virtual machines/virtual desktops with capacity disks for 10,000 VMware virtual machines

Scenario two presents the minimum recommended disk storage layout for booting a midsize to large amount of VMware virtual machines or virtual desktops that are also working with cloning "golden image" or VMware linked-clones technologies. The VMware vSphere5.x environment is attached to an Oracle ZFS Storage ZS3-2 cluster configuration. Further, this option presents a different storage layout

that combines performance, capacity and cloning technologies. This scenario also assumes that you are working with approximately 100 GB to 300 GB (thin provisioning) Windows virtual desktop "clones from golden images," with additional VMware virtual disks for data and additional disk space for ISO, backups and images data stores:

- One mirrored disk pool of (at least) 44 x 900 GB SAS-2 (10 k RPM performance disk drives) with at least one 73 GB SSD device for LogZilla working with a striped log profile (per disk pool and also per head)

- One mirrored disk pool of (at least) 5 x 900 GB SAS-2 (10 k RPM performance disk drives) with at least two 73 GB SSD devices for LogZilla working with a striped log profile (per disk pool and also per head)

- One mirrored disk pool of (at least) 44 x 4 TB SAS-2 (7200 RPM performance disk drives) with at least two 73 GB SSD devices for LogZilla working with a striped log profile (per disk pool and also per head).

Figure 5 shows scenario two.

**Note**: For best performance, 900 GB 10 k performance disks are strongly recommended for virtualization environments, and also should be considered during the disk layout and sizing phase.

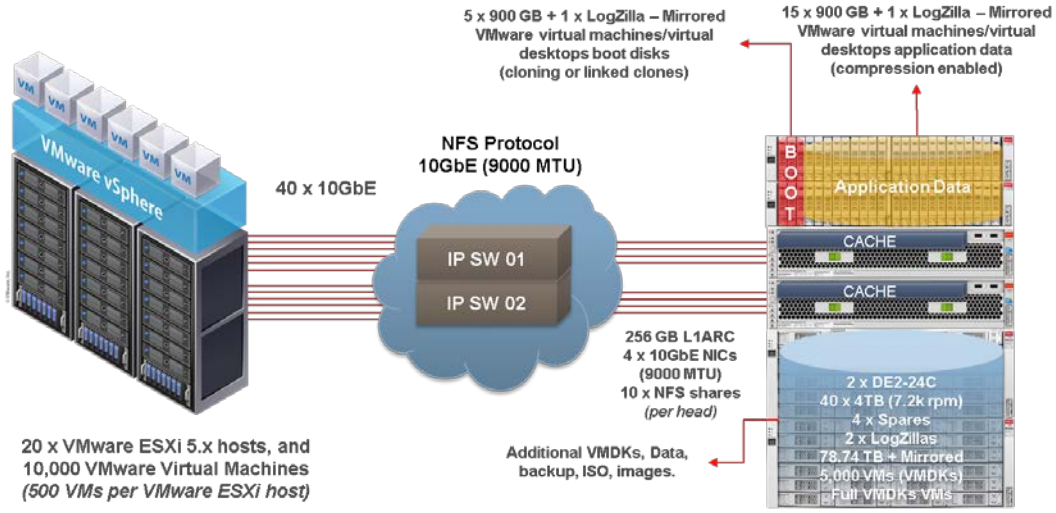


Figure 5. Oracle ZFS Storage Appliance – Disk layout for small virtual machines/virtual desktops with capacity disks for 10,000 VMware virtual machines

Scenario three presents the recommended disk storage layout for booting a midsize to large amount of virtual machines in a VMware vSphere5.x environment attached to an Oracle ZFS Storage Appliance ZS3-2 sytem. Assuming that you are working with midsize to large Linux or Windows virtual

machines/virtual desktops (100 GB to 300 GB or more of thin provisioning OS boot disk), and a footprint of about 50 IOPs per VM, the minimum recommended configuration, presented in the following figure, is:

- Two mirrored disk pools of (at least) 130 x 4 TB SAS-2 (7200 RPM capacity disk drives) with at least four 73 GB SSD devices for LogZilla working with a striped log profile (per disk pool and also per head)
- Two mirrored disk pools of (at least) 40 x 900 GB SAS-2 (10k RPM performance disk drives) with at least four 73 GB SSD devices for LogZilla working with a striped log profile (per disk pool and also per head)
- Four x 10 GbE network interface cards

Note: For larger VMware virtual machines of between 100 GB and 300 GB of thin provisioning OS boot disks, consider using LZJB compression (LZJB Algorithm) for disk space savings. The LZJB algorithm is fast and does not consume much CPU resources. Users who have implements LZJB generally experience better disk utilization and sometimes see improved performance.

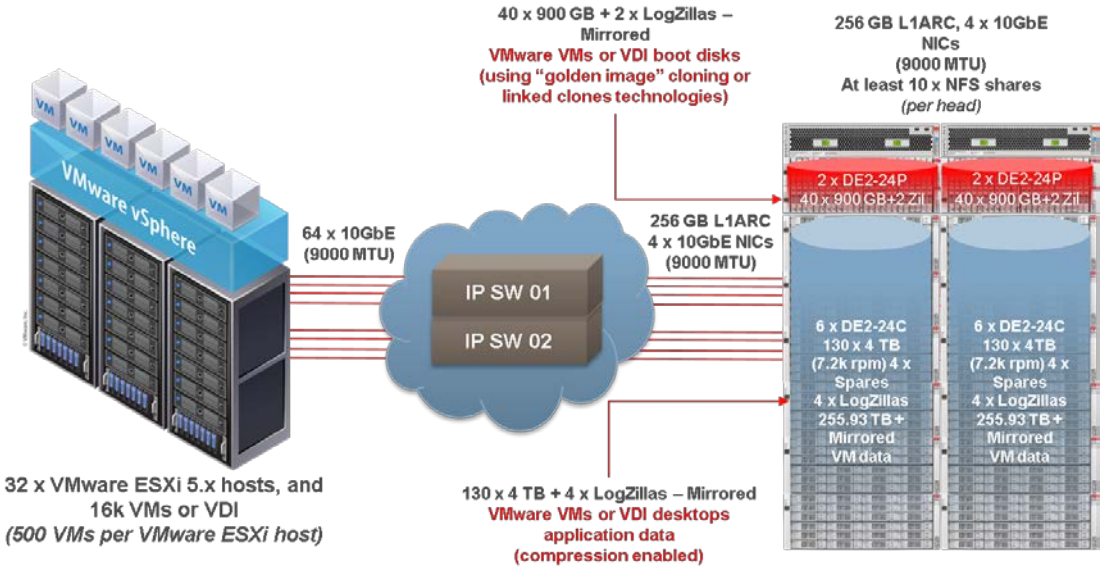


Figure 6. Oracle ZFS Storage Appliance – storage layout with capacity disks

### Network Settings

As described in the white paper titled "Best Practices for Oracle ZFS Storage Appliance and VMware vSphere5.x" (see References at the end of this document), the recommended network settings for NFS are:

- Isolate the storage traffic from other networking traffic. You can do this using VLAN tagging, network segmentation, or dedicated IP switches for NFS traffic only.

- Use at least two dual 10 GbE SFP+ PCIe 2.0 Low Profile adapters – a total of four 10 GbE connections per Oracle ZFS Storage Appliance controller, or 4 x 10 GbE onboard interfaces, which are provided by the standard Oracle ZFS Storage ZS3-2 controllers configuration.
- Enable (very important) 9000 MTU jumbo frames in all 10 GbE NICs and IP network switches.

**Note:** The 10 GbE NICs can be bundled into a single channel (per controller) using the IEEE 802.3ad Link Aggregation Control Protocol (LACP). With IPMP configuration you will achieve network high availability, and with link aggregation you will obtain 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 LACP and for picking an outbound port based on source and IP addresses, use 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 the following figure.

**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.

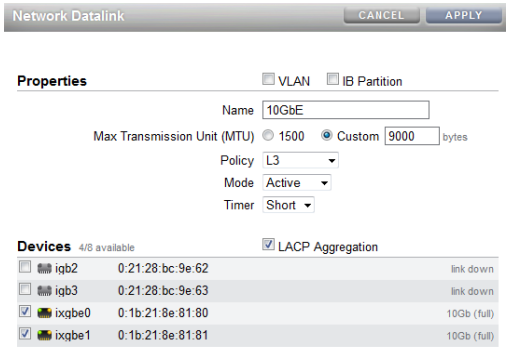


Figure 7. LACP, jumbo frame and MTU configurations on the Oracle ZFS Storage Appliance

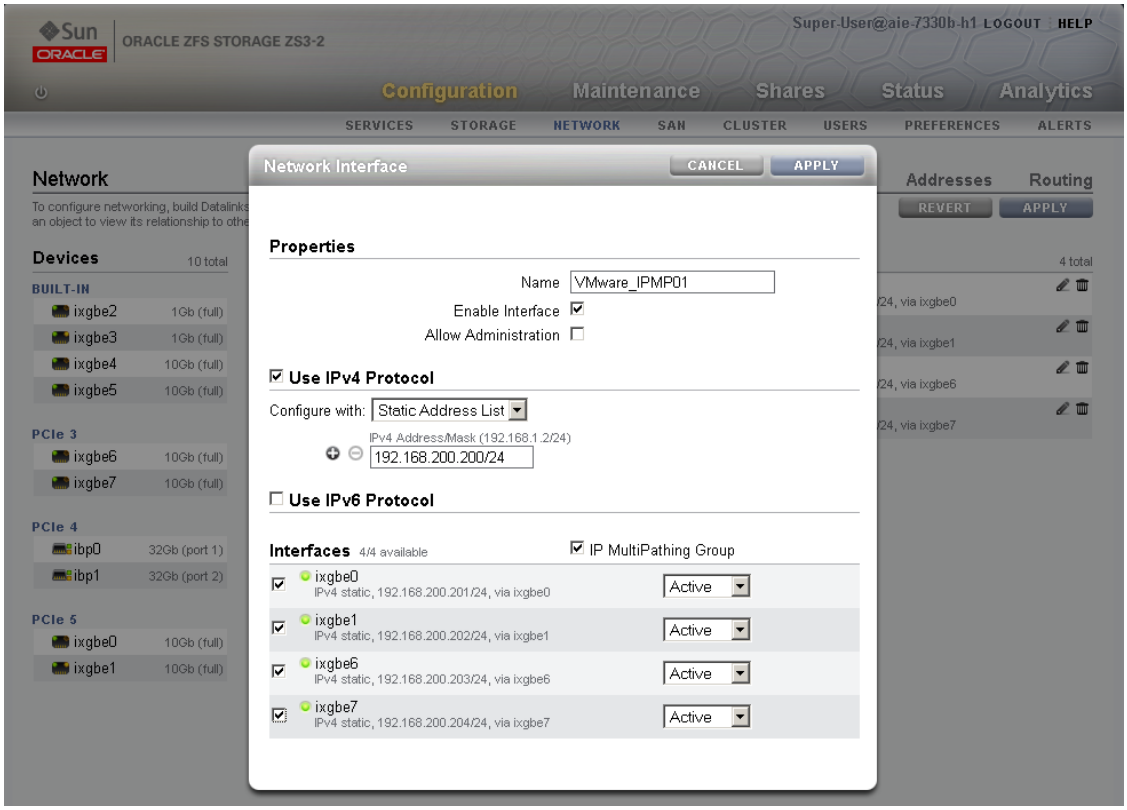


Figure 8. IPMP configuration on the Oracle ZFS Storage Appliance

### NFS, Projects and Shares

The example shows one Oracle ZFS Storage Appliance cluster with active/active mode, one project, and 10 different NFS shares (per Oracle ZFS Storage Appliance head), which are named VMware_01 through VMware_20. A 10 Oracle ZFS Storage Appliance NFS shares configuration per Oracle ZFS Storage head is the minimum recommended configuration for large deployments, such as 10,000 or more VMware virtual machines. In a scenario of 10,000 VMware virtual machines, each Oracle ZFS Storage Appliance NFS share will be hosting 500 virtual machines, totaling 5,000 VMware virtual machines per Oracle ZFS Storage Appliance head.

The following figure presents the Oracle ZFS Storage Appliance (miminum) recommended NFS configuration for hosting 5,000 VMware virtual machines (per head), and figure 10 shows the filesystem and mountpoint configurations on the Oracle ZFS Storage Appliance browser user interface (BUI).

Filesystems | LUNs 10 Total

| NAME ▲ | SIZE | MOUNTPOINT |
|---|---|---|
| VMware_01 | 31K | /export/VMware_01 |
| VMware_02 | 31K | /export/VMware_02 |
| VMware_03 | 31K | /export/VMware_03 |
| VMware_04 | 31K | /export/VMware_04 |
| VMware_05 | 31K | /export/VMware_05 |
| VMware_06 | 31K | /export/VMware_06 |
| VMware_07 | 31K | /export/VMware_07 |
| VMware_08 | 31K | /export/VMware_08 |
| VMware_09 | 31K | /export/VMware_09 |
| VMware_10 | 31K | /export/VMware_10 |

Figure 9. Oracle ZFS Storage Appliance recommended NFS configuration for hosting 10,000 VMware virtual machines

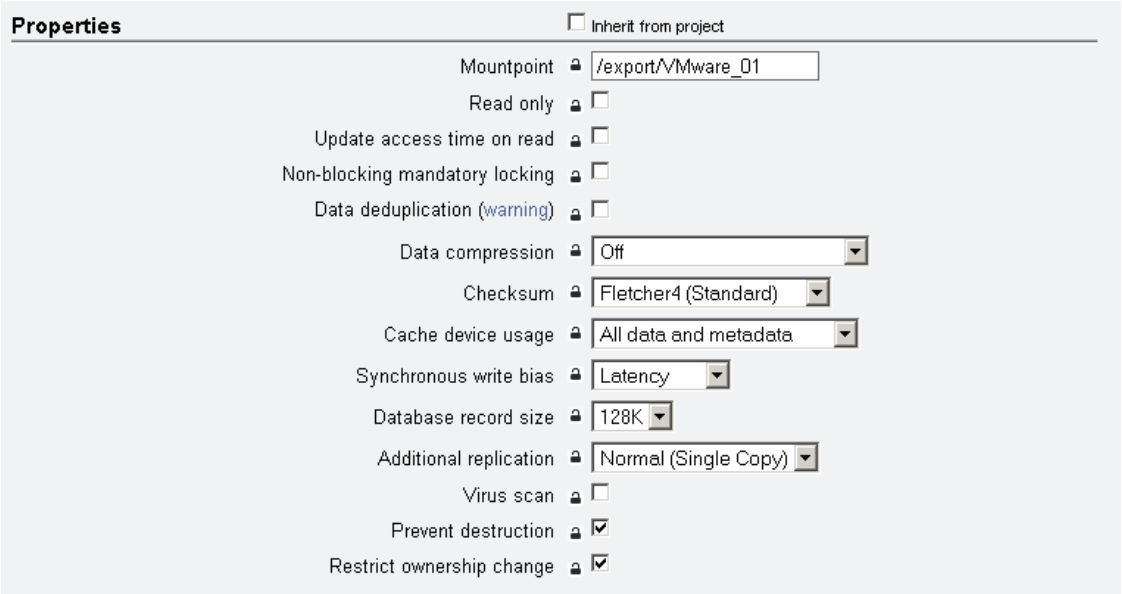


Figure 10. Share configuration shown in the Oracle ZFS Storage Appliance BUI

**Note:** For additional information about NFS properties for VMware, and for VMware IP network, refer to the white paper "Best Practices for Oracle ZFS Storage Appliance and VMware vSphere5" in the References section at the end of this document.

## Recommendations for VMware NFS Protocol

It is important to alter the VMware NFS and TCP/IP advanced settings. These configuration changes will ensure high availability of the VMware NFS data stores during Oracle ZFS Storage Appliance failback or takeover operations. The settings also set the maximum number of allowed NFS datastores per VMware ESXi host, and important changes on the VMware TCP/IP stack.

To alter the parameters listed in the following table, go to the VMware vCenter 5.x server and select a VMware server. Select the software tab and click on **Advanced Settings.** Select **NFS** and **TCP/IP,** and then alter the following options:

**TABLE 4. RECOMMENDED NFS AND TCP/IP ADVANCED SETTINGS FOR VMWARE VSPHERE 5.1 DATA STORES 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 |

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

## Considerations for Virtual Machine Settings

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.
- For storage efficiency, configure your virtual machine with a thin provisioning virtual disk drive, and for performance, configure your VMware virtual machine with VMware paravirtual SCSI controller type.
- For performance improvements, use a VMXNET3 network adapter.
- Install the VMware Client Tools. For more information on these tools and installing them, use the following link:

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

- Pay attention to partition alignment. Ensure that your virtual machine is working with the right partition aligment.

## Conclusion

Oracle ZFS Storage Appliance offers outstanding performance for virtualized environments. Its architecture allows you to deploy a large number of virtual machines with optimal performance, low latency, and without being overwhelmed by boot storm traffic. Oracle ZFS Storage Appliance features and intelligent caching technology are designed to deliver hundreds of thousands of IOPS, and the Hybrid Storage Pool feature provides the horsepower needed to support the maximum number of virtual machines and virtual desktops environemts.

The combination of Oracle ZFS Storage Appliance performance and VMware vSphere hypervisors is 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.spec.org/sfs2008/results/res2013q3/sfs2008-20130819-00227.html

http://www.storageperformance.org/benchmark_results_files/SPC-2/Oracle_SPC-2/B00067_Oracle_ZFS-ZS3-4/b00067_Oracle_ZFS_Storage_ZS3-4_SPC-2_full-disclosure-report.pdf

http://www.storageperformance.org/results/benchmark_results_spc2/#sun_spc2

# Appendix B: References

See the following resources for additional information relating to the products covered in this document.

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.

- Oracle ZFS Storage Appliance Documentation Library, including Installation, Analytics, Customer Service, and Administration guides:
  http://www.oracle.com/technetwork/documentation/oracle-unified-ss-193371.html
- The *Oracle ZFS Storage Appliance Administration Guide* is also available through the Oracle ZFS Storage Appliance help context.
  The Help function in Oracle ZFS Storage Appliance can be accessed through the browser user interface.
- Oracle ZFS Storage Appliance software releases can be downloaded from the following URLs:

  https://wikis.oracle.com/display/FishWorks/Software+Updates

  http://www.oracle.com/technetwork/server-storage/sun-unified-storage/downloads/index.html
- Oracle Support Center
  http://www.oracle.com/support
- Patches and updates downloads from My Oracle Support (MOS)
  (search under Oracle ZFS Storage Software Patches)

- Oracle ZFS Storage Appliance Plug-ins
  http://www.oracle.com/technetwork/server-storage/sun-unified-storage/downloads/zfssa-plugins-1489830.html
- Oracle Storage Product Information
  http://www.oracle.com/us/products/storage/overview/index.html
- Oracle ZFS Storage Appliance Technical White Papers and Solution Briefs,
  including "Best Practices for Oracle ZFS Storage Appliance and VMware vSphere5.x"
  http://www.oracle.com/technetwork/server-storage/sun-unified-storage/documentation/index.html
- VMware
  http://www.vmware.com

Deploying 10,000+ VMware Virtual Machines with Oracle ZFS Storage Appliance
July 2014, Version 1.0
Author: Application Integration Engineering; 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**