



An Oracle White Paper
April 2011

Sun ZFS Storage Appliance Reference Architecture for VMware vSphere

Introduction	2
About the Sun ZFS Storage Appliance	3
About VMware vSphere	3
Reference Architecture Overview	4
Architectural Components.....	5
Reference Architecture Validation	11
Space Savings with Thin Provisioning and De-Duplication	11
Application Performance Validation	13
Microsoft Exchange Server Jetstress Validation.....	13
OLTP Database Validation	14
DSS Database Validation	16
.....	16
Conclusion	17

Introduction

The reference architecture described in this paper demonstrates the design and testing of a Sun ZFS Storage Appliance configuration featuring the VMware vSphere 4 virtualization platform. It is intended to help IT administrators plan a virtualization deployment with confidence that the configuration will meet their IT and business needs.

This reference architecture highlights a multi-application deployment of mail servers, database servers, web servers, and various development and test servers. A multi-pool approach is used to help consolidate and scale differing workloads on a single storage platform while maintaining performance levels and meeting service level agreement (SLA) requirements for each application.

Highlighted in this paper are:

A multi-pool design with multiple data store repositories in VMware vSphere 4

Use of the Sun ZFS Storage Appliance Hybrid Storage Pool with flash media

Space savings using the de-duplication and thin provisioning features of the Sun ZFS Storage Appliance

A high availability design for storage access and performance

Test validation of the reference architecture

The key components of the reference architecture described in this paper are:

Sun ZFS Storage 7420 cluster

VMware vSphere 4

About the Sun ZFS Storage Appliance

The Sun ZFS Storage Appliance offers innovations in storage that include a DTrace-based storage analytics tool, flash-based Hybrid Storage Pools, enterprise-class data services, massive scalability, and a choice of storage protocols, while delivering significant cost savings.. The Sun ZFS Storage Appliance features a common, easy-to-use management interface that has the industry's most comprehensive analytics environment to help isolate and resolve issues to minimize business impact.

Oracle offers four models of the Sun ZFS Storage Appliance to meet the scalability and availability needs of today's demanding applications. These models include the Sun ZFS Storage 7120, 7320 and 7320C, 7420 and 7420C, and the 7720. All of these utilize a common storage software foundation. Three of these models (excluding the Sun ZFS Storage 7120) offer up to 2 TB of read cache, enabling a typical appliance response time in the low single digit milliseconds. Write flash, available on all four platforms, enables response times of less than 1 ms for synchronous writes.

The new Sun ZFS Storage Appliance platforms offer faster CPUs, bigger flash cache, larger storage capacity, and better throughput to meet the storage requirements of mission critical applications.

About VMware vSphere

VMware vSphere™, the industry's first cloud operating system, leverages the power of virtualization to transform datacenters into dramatically simplified cloud computing infrastructures. VMware vSphere helps preserve business-critical application availability by enabling transparent migration of applications and files from one storage array to another. IT organizations can deliver the next generation of flexible and reliable IT services, using internal and external resources securely and with low risk. Key benefits of vSphere include:

Broad interoperability across servers, storage, operating systems, and applications

Robust, reliable foundation

Established install base

Reference Architecture Overview

Figure 1 shows a high level overview of the physical components of the reference architecture. The reference configuration consists of two physical VMware ESX 4.1 servers, a 10 GbE network infrastructure, and a Sun ZFS Storage 7420 with 6 disk shelves. A total of 24 virtual machines are configured in the architecture and are running a mail server workload, an OLTP database workload and a DSS database workload.

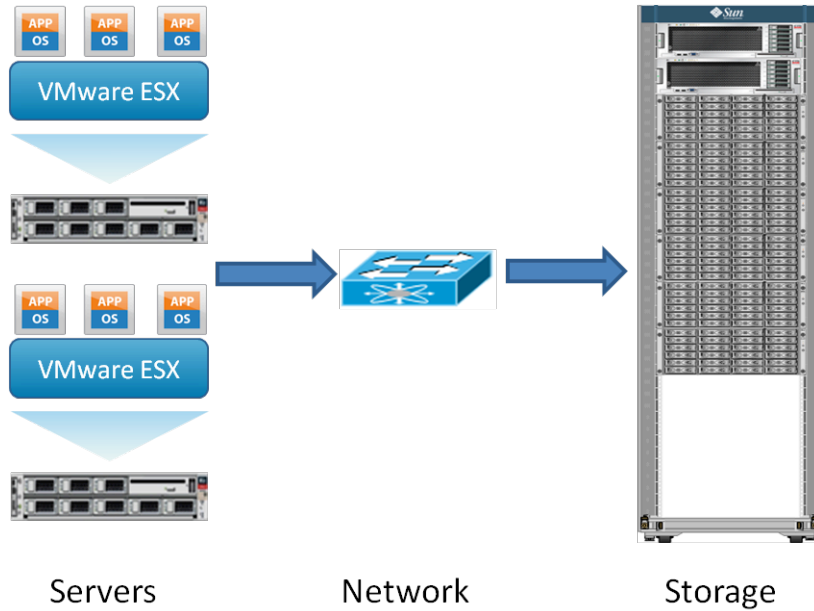


Figure 1. Physical components of the reference architecture.

Architectural Components

The tables below describe the hardware, virtual machine, and software components of the reference architecture.

Table 1 shows the hardware components used.

TABLE 1. HARDWARE COMPONENTS USED IN REFERENCE ARCHITECTURE

EQUIPMENT	QUANTITY	CONFIGURATION
Primary 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 2 TB SAS-2 disk trays 4 x 24 2 TB SAS-2 disk trays 8 x 18 GB write cache SSD 2 x dual port 10 GbE NIC
Network	2	10 GbE Network Switch
Server	2	Sun Fire X4170 M2 Server 72 GB DRAM 2 internal HDDs 1 x dual port 10 GbE NIC

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)	12	2 Microsoft Exchange Servers 4 Mail Utility servers 5 Windows development and test servers 1 domain controller
Oracle Enterprise Linux 5.4	12	2 OLTP database servers 2 DSS database servers 4 database utility servers 4 development and test servers

Table 3 shows the software components used.

TABLE 3. SOFTWARE COMPONENTS USED IN REFERENCE ARCHITECTURE

SOFTWARE	VERSION
Sun ZFS Storage Appliance software	2010.Q3.3
Microsoft Exchange Server Jetstress verification tool	2010 (x64)
Oracle ORION I/O calibration tool	11.1.0.7.0
VMware vCenter virtualization management software	4.1.0 (Build 258902)
VMware ESX hypervisor software	4.1.0 (Build 260247)

Reference Architecture Design

In the reference architecture, the VMware ESX hypervisor accesses the Sun ZFS Storage 7420 using Network File System (NFS) protocol over a 10 GbE interface. The Sun ZFS Storage 7420 cluster provides two separate controllers that can be configured in an Active/Active cluster implementation to provide simultaneous access by both controllers to the workload.

A primary consideration in performance and capacity planning is the storage pool layout of the disk trays. Six disk trays are used in the reference configuration, with three trays of disks configured for each controller. The six trays were configured as shown in Table 4.

TABLE 4. POOL LAYOUT OF DISK TRAYS

POOL NAME / CONTROLLER ASSIGNMENT	POOL CONFIGURATION	POOL USE
Pool1 – Controller1	Double-parity RAID 6 data disks, 1 write SSD, 0 read cache SSD	Microsoft Windows virtual machine boot drive virtual disks
Pool2 – Controller1	Mirrored 44 data disks, 2 write SSD, 2 read cache SSD	Microsoft Exchange Server database virtual disks
Pool3 – Controller1	Mirrored 18 data disks, 1 write SSD, 0 read cache SSD	MS Exchange Server log virtual disks
Pool4 – Controller2	Double-parity RAID 6 data disks, 1 write SSD, 0 read cache SSD	Linux virtual machine boot drive virtual disks
Pool5 – Controller2	Mirrored 44 data disks, 2 write SSD, 2 read cache SSD	OLTP database virtual disks
Pool6 – Controller2	Double-parity RAID 18 data disks, 1 write SSD, 0 read cache SSD	DSS database virtual disks

As shown in Table 4, the virtual boot disks for both the Windows virtual machines and Linux virtual machines and the DSS database virtual disks are stored in pools configured as double-parity RAID. These virtual disks do not require high random read performance but instead must be configured to maximize sequential, large block I/O performance.

The Microsoft Exchange Server database files, log files and the OLTP database files are stored in pools that are configured for mirrored protection. Mirrored pools are preferred when the application or virtual machine requires a high degree of small block, random read I/O with low latency. Such a layout may not be necessary in all deployments as multiple workloads could be serviced from a single large pool. However, this layout is used to demonstrate a segregated configuration to facilitate guaranteed quality of service and performance requirements.

Additionally, the Windows virtual boot disks are stored in pools owned by the same controller as the pools for the Windows application virtual disks and the Linux virtual boot disks are stored in pools owned by the same controller as the Linux application virtual disks. This layout, while seemingly

redundant, facilitates the deployment of the VMware vCenter Site Recovery Manager, which has a requirement that all the virtual disks of a virtual machine reside on the same controller.

Once the pools have been created, the next step is to lay out the projects and underlying NFS file system shares. The projects and shares shown in Table 5 were created for this configuration.

TABLE 5. PROJECTS AND FILE SYSTEM SHARES CREATED FOR THE REFERENCE ARCHITECTURE

POOL NAME	PROJECTS	FILE SYSTEMS	FILE SYSTEM DATABASE RECORDSIZE
Pool1	winboot	/export/winboot	64kb
	vswap	/export/vswap	64kb
Pool2	ms-exchgdb	/export/ms-exchgdb1	32kb
Pool3	ms-log	/export/ms-log1	128kb
Pool4	linuxboot	/export/linuxboot	64kb
Pool5	oltp-db	/export/oltp-db1	8kb
Pool6	dss-db	/export/dss-db1	128kb

The following were taken into account when the projects and shares were created:

All projects were configured with “Update Access Time on Read” disabled.

The winboot share and the linuxboot share were configured to enable de-duplication.

All virtual machines were configured to use a centralized vswap location.

The layout for the Microsoft Exchange datastore is shown in Figure 2. Each Microsoft Exchange virtual machine is booted from a single virtual boot disk configured with 40 GB of space. The virtual machine is attached to four 256 GB mail-database virtual disks and four 30 GB mail-log virtual disks. Thus, each Exchange Server datastore is configured with 1 TB of disk space for this test.

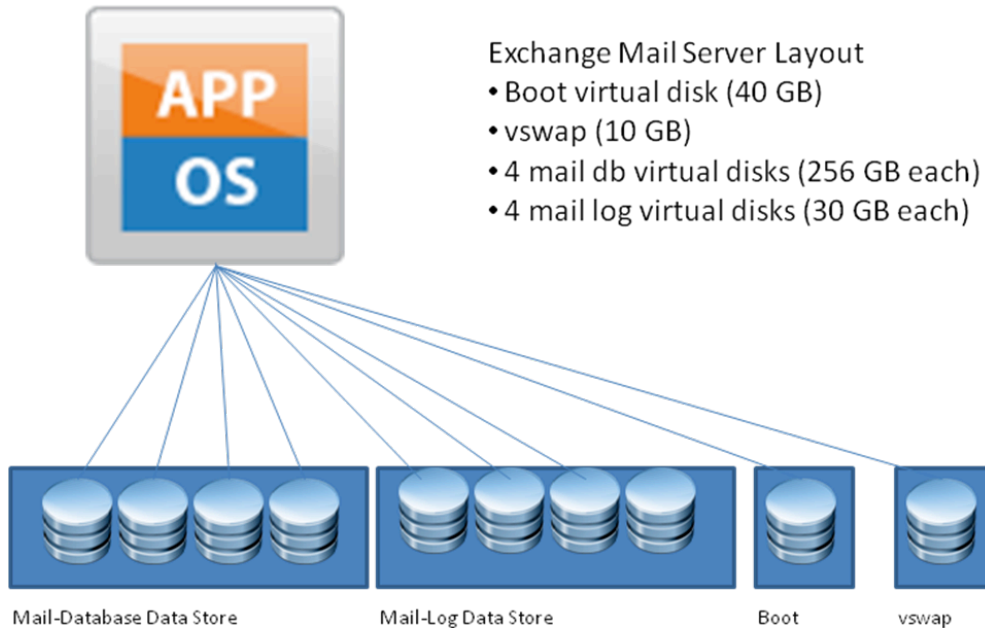


Figure 2. Microsoft Exchange datastore layout.

The layout for the OLTP database is shown in Figure 3. Each OLTP virtual machine is booted from a single virtual boot disk configured with 40 GB of space. The virtual machine is attached to four 256 GB database virtual disks. Thus, 1 TB of disk space is dedicated to the database load test for each OLTP server.

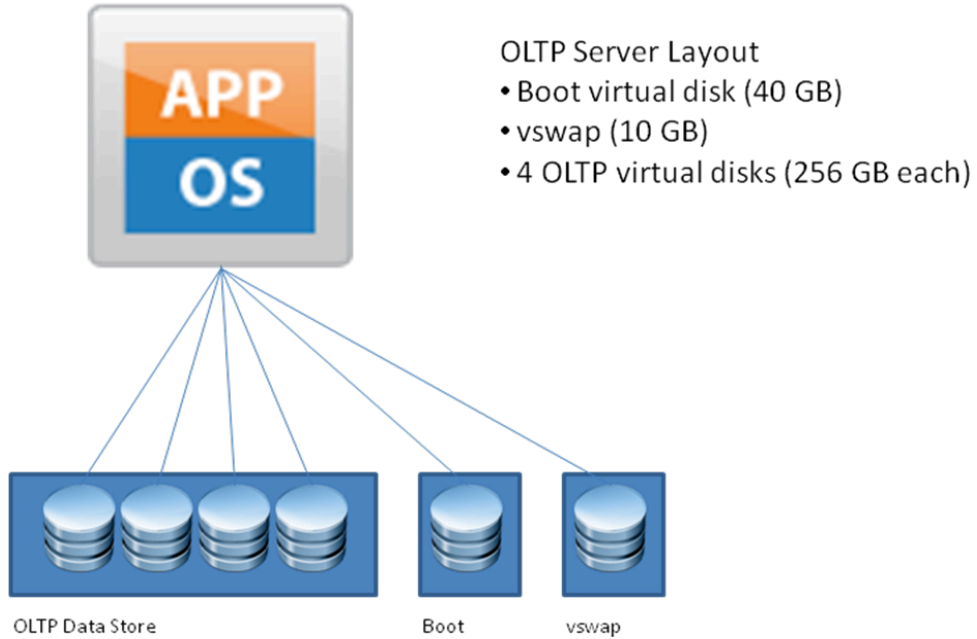


Figure 3. OLTP database layout.

The layout for the DSS database is shown in Figure 4. Each DSS virtual machine is booted from a single virtual boot disk configured with 40 GB of space. The virtual machine is attached to four 256 GB database virtual disks. Thus, 1 TB of disk space is dedicated to the database load test for each DSS server.

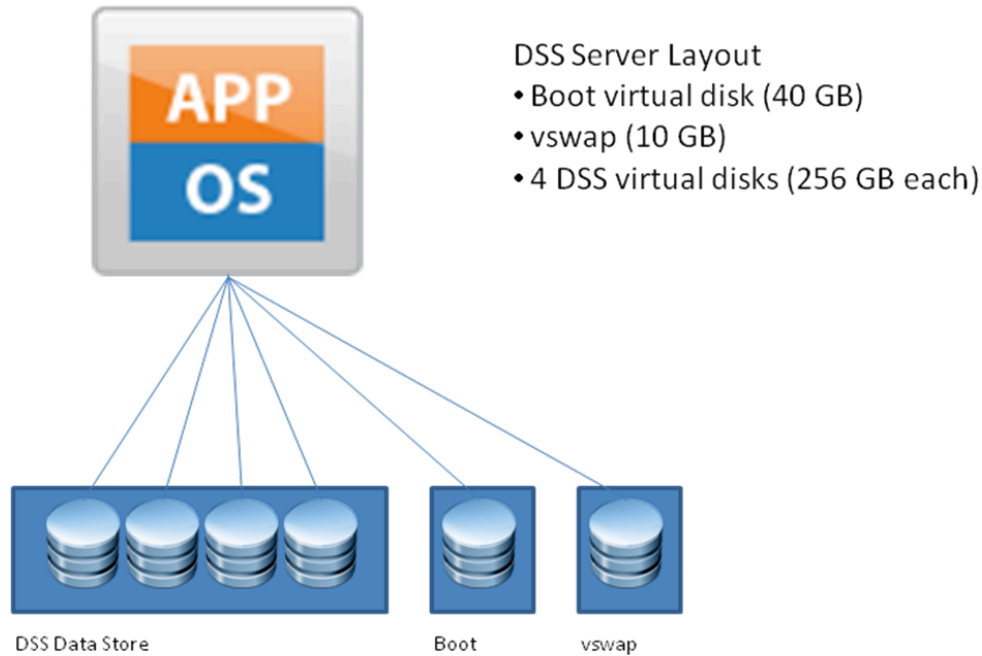


Figure 4. DSS database layout.

Reference Architecture Validation

This section describes how the reference configuration was validated to ensure it meets acceptable performance metrics while delivering storage space savings through the use of de-duplication and thin provisioning.

While not explicitly shown in this paper, the solution was also validated with VMware vSphere virtualization features such as vMotion, Storage vMotion and Distributed Resource Scheduling (DRS). All features worked as expected in the overall reference architecture and no impact on the overall performance was seen.

Space Savings with Thin Provisioning and De-Duplication

A combination of thin provisioning and de-duplication was used to achieve a space savings of over 95 percent for the pool containing the 12 Windows virtual machine boot disks. Similar space savings were found on the pool containing the Linux virtual machine boot disks.

Figure 6 shows that, initially, 12 virtual machines configured with 40 GB C: drives used 484 GB of provisioned space.

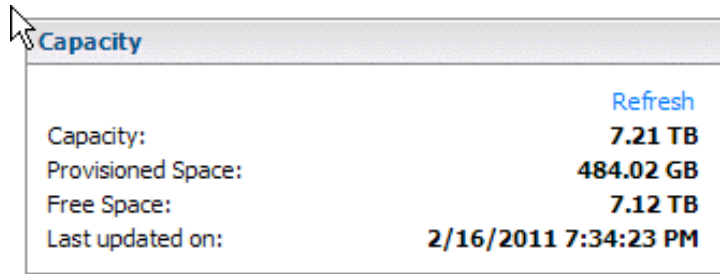


Figure 6. VMware vCenter screenshot showing space usage for Windows virtual boot disks.

Figure 7 shows that after implementing thin provisioning, the space consumed by the VMware environment on the NFS datastore was 100 GB, for a space savings of 80 percent.

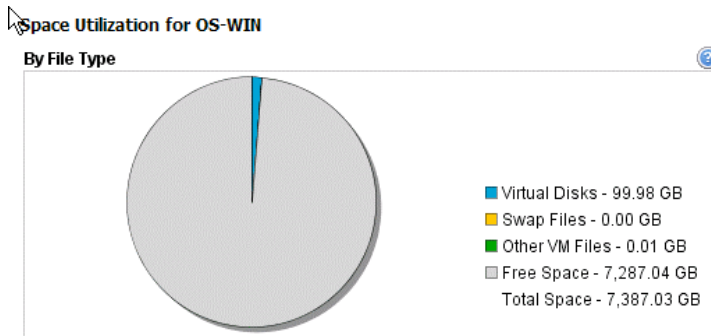


Figure 7. VMware vCenter screenshot showing space usage after implementation of thin provisioning.

The de-duplication feature in the Sun ZFS Storage Appliance was then enabled, reducing the total space consumed in the storage pool to 18.7 GB as shown in Figure 8. The de-duplication ratio was 8.16X.

The overall space savings with thin provisioning implemented and de-duplication enabled was 95% percent.

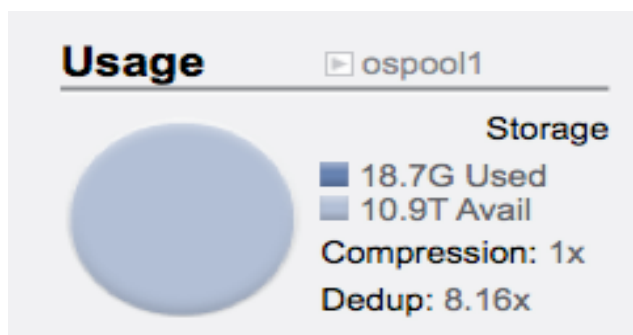


Figure 8. Sun ZFS Storage Appliance pool usage for Windows virtual boot disks after thin provisioning and de-duplication were implemented.

Application Performance Validation

Next, an application workload was applied to the system to verify that the system as configured could deliver the needed performance in terms of I/O operations, latency, and bandwidth. Application workloads included a Microsoft Exchange mail system, an OLTP database, and a DSS database.

For each application, the application virtual disks were modified to be aligned on a 64-block offset. This ensures the most efficient use of I/O in the configuration. The Exchange database and log virtual disks, as well as all disks used for the OLTP and DSS databases were aligned using either the Microsoft diskpart or the Linux fdisk utility. See the appropriate operating system documentation for a more details about disk alignment.

Microsoft Exchange Server Jetstress Validation

Two Microsoft Exchange mail servers were configured with a total of 6000 mail users. Each user’s mailbox held 250 MB of data for a total of 1.5 TB of mail data under test. The workload profile was 0.5 IOPs per user which corresponds to a “heavy” email user according to Microsoft guidelines.

TABLE 6.M ICROSOFT EXCHANGE SERVER JESTRESS PERFORMANCE RESULTS

METRIC	RESULT
IOPs Achieved	3080 IOPs
Avg. DB Read Latency (target < 20ms)	14.79 ms
Avg. Log Write Latency (target < 10ms)	1.78 ms

Using the Sun ZFS Storage Appliance DTrace Analytics, the administrator can see exactly how the mail systems are behaving down to the virtual disk level. Figure 9 shows the number of NFS operations per virtual disk in the mail system.

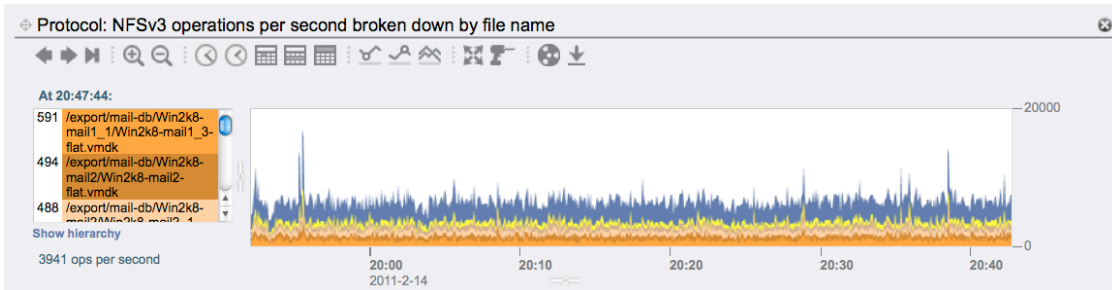


Figure 9. DTrace Analytics graph showing the number of NFS operations per second for each virtual disk in the mail system.

Figure 10 shows the number of NFS operations per second broken down into read operations and write operations.

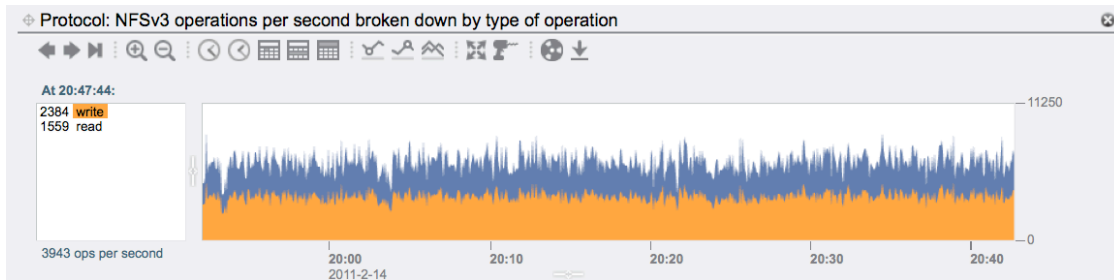


Figure 10. DTrace Analytics graph showing the number of NFS read and write operations for mail servers.

These results validate that this configuration can support 6000 MS Exchange users using 44 mirrored data disks for the Exchange database files and 18 mirrored data disks for the Exchange log files.

OLTP Database Validation

Two Oracle Enterprise Linux virtual machines were configured with the ORION benchmark tool to simulate an OLTP database workload against a total of 8 virtual disks. The capacity of each virtual disk was 256 GB resulting in a total size of 2 TB for the databases under test.

TABLE 7. ORION OLTP PERFORMANCE RESULTS

PARAMETER	VALUE
Block size	8KB
RAID level	Mirror
Write/read ratio	60/40
Number of outstanding I/Os per virtual machine	64
Total IOPs achieved	11,887
Average latency	10.8ms

Using the Sun ZFS Storage Appliance DTrace Analytics, the administrator can see exactly how the OLTP database systems are behaving, down to the virtual disk level. Figure 11 shows the number of NFS operations per second for each database virtual disk.



Figure 11. DTrace Analytics graph showing the number of NFS operations per second for the OLTP database virtual disks.

Figure 12 shows the number of NFS operations per second broken down into read operations and write operations.

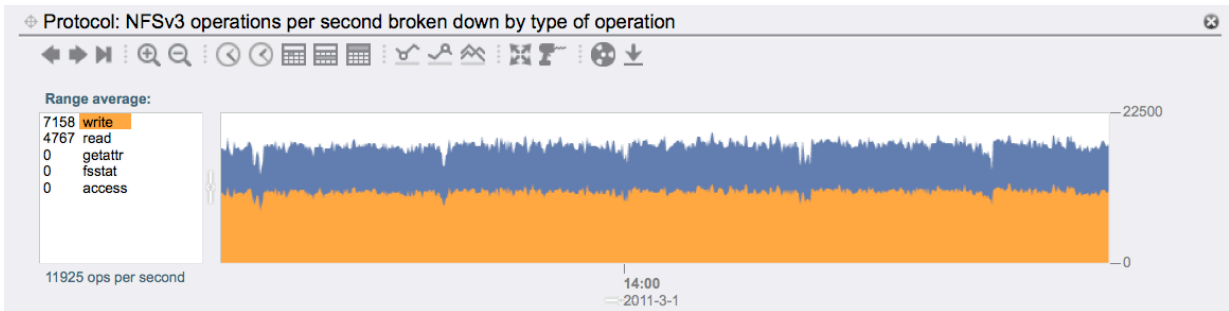


Figure 12. DTrace Analytics graph showing the number of NFS read and write operations for the OLTP databases.

These results validate that this configuration can sustain a large number of I/Os typical of an OLTP database, while maintaining acceptable service levels using 44 mirrored data disks.

DSS Database Validation

Two Oracle Enterprise Linux virtual machines were configured with the ORION benchmark tool to simulate a DSS database workload against a total of 8 virtual disks. The capacity of each virtual disk was 256 GB, resulting in a total size of 2 TB for the databases under test. The object of this test was to demonstrate sustained bandwidth throughput for sequential reads of large data blocks.

TABLE 8. ORION DSS PERFORMANCE RESULTS

PARAMETER	VALUE
Block Size	1024KB
RAID Level	Mirror
Write/Read Ratio	0/100
Number of Outstanding I/Os per virtual machine	24
Total Throughput	100 % Cache Hit – 645 MB/s 60% Cache Hit – 277 MB/s <40% Cache Hit – 133 MB/s

Using the Sun ZFS Storage Appliance DTrace Analytics, the administrator can see how the database systems are behaving, down to the virtual disk level.

Figure 13 shows the NFS operations per second for the DSS database virtual disks.

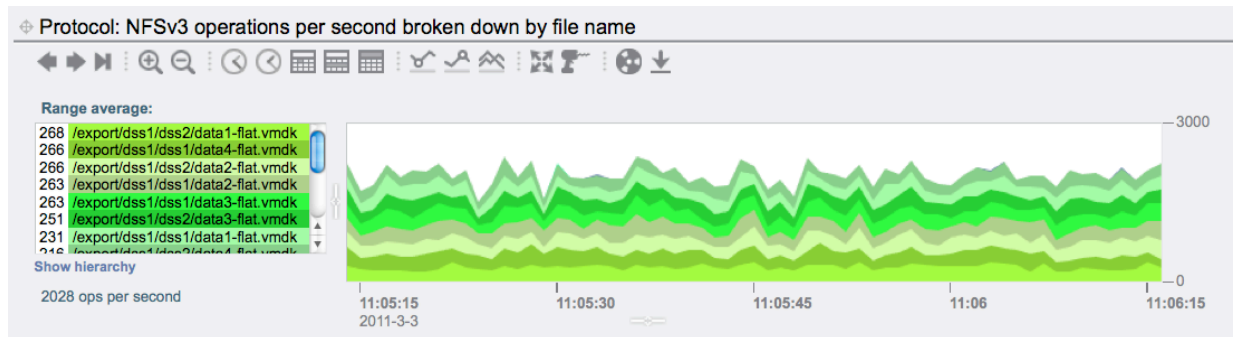


Figure 13. DTrace Analytics graph showing the number of NFS operations per second for the DSS database virtual disks.

Figure 14 shows the network interface throughput in bytes per second.

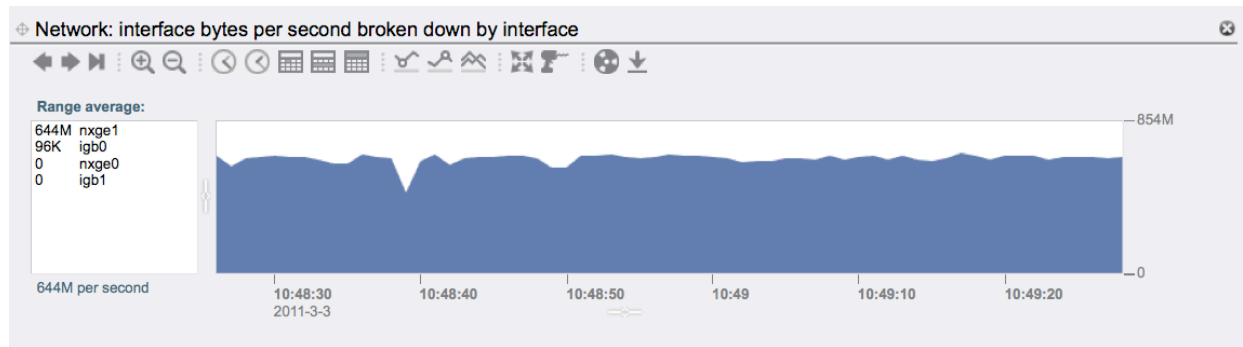


Figure 14. DTrace Analytics graph showing the network interface throughput.

Conclusion

The reference architecture described in this paper comprises a Sun ZFS Storage Appliance and VMware vSphere configuration that can be deployed for highly demanding IT application workloads. Mixed application workloads can be consolidated on a robust and flexible Sun ZFS Storage platform to improve performance, increase operational visibility, and reduce management costs.

This configuration was validated to deliver the performance needed for three applications, a Microsoft Exchange mail system, an OLTP database, and a DSS database, with workloads distributed across multiple operating systems and application workload profiles. No degradation in performance was seen while running all applications and workloads simultaneously.

The use of the built-in de-duplication feature on the Sun ZFS Storage Appliance in conjunction with thin provisioning resulted in significant space savings and efficiencies.



Oracle Sun ZFS Storage Appliance Reference
Architecture for VMware vSphere
April 2011, Version 1.0
Author: Ryan Arneson
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

Copyright © 2011, Oracle and/or its affiliates. All rights reserved. This document is provided for information purposes only and the contents hereof are subject to change without notice. This document is not warranted to be error-free, nor subject to any other warranties or conditions, whether expressed orally or implied in law, including implied warranties and conditions of merchantability or fitness for a particular purpose. We specifically disclaim any liability with respect to this document and no contractual obligations are formed either directly or indirectly by this document. This document may not be reproduced or transmitted in any form or by any means, electronic or mechanical, for any purpose, without our prior written permission.

Oracle and Java are registered trademarks of Oracle and/or its affiliates. Other names may be trademarks of their respective owners.

AMD, Opteron, the AMD logo, and the AMD Opteron logo are trademarks or registered trademarks of Advanced Micro Devices. Intel and Intel Xeon are trademarks or registered trademarks of Intel Corporation. All SPARC trademarks are used under license and are trademarks or registered trademarks of SPARC International, Inc. UNIX is a registered trademark licensed through X/Open Company, Ltd. 1010

Hardware and Software, Engineered to Work Together