

A Reference Architecture for Implementing an Oracle RAC Database on Oracle Servers, Switches, and SAN Storage Hardware

ORACLE WHITE PAPER | AUGUST 2016





Disclaimer


This white paper is intended to outline our general product direction, is intended for information purposes only, and may not be incorporated into any contract. It is not a commitment to deliver any material, code, or functionality, and should not be relied upon in making purchasing decisions. The development, release, and timing of any features or functionality described for Oracle's products remains at the sole discretion of Oracle.



Table of Contents

Executive Summary	1
Introduction	1
Intended Audience	2
Reference Architecture	2
Design Goals	2
System Designs	3
Hardware Components	4
Software Components	5
Storage and ASM Configuration	5
Oracle FS Storage Profiles	7
Oracle Database 12c Configuration	7
Performance Analysis	7
Performance Measurement Tools	8
Oracle AWR	8
Swingbench	9
Swingbench Download and Setup	10
Database and Schema Setup	10
Swingbench OrderEntry Benchmark	10
Schema Size	11
Workload setup	11

Swingbench Script	12
Swingbench Workloads	12
Assessment of System Balance	13
Balance Charts	13
Load Response Charts	13
Quarter Rack and Half Rack System Charts	13
Determining Configuration Settings	16
Conclusion	17
Related Documentation	17
Oracle Flash Storage Documentation	17
Oracle Technical Support and Resources	17
Appendix A. Installation and Configuration Details	18
Target Audience	18
Hardware Configuration	18
Bill of Materials	19
Configure Server Internal Storage	20
Install Oracle Linux	20
Install and Configure Oracle FS Path Manager (FSPM)	20
Configure Oracle FS Storage	20
Diskgroup and LUN Layout	21
Install Oracle RAC	23
Install ASMLib	23
Create the ASM Instance	23



Creating the Oracle Database	23
Undo Tablespace Considerations	23
Configure the Redo Logs	24
Enable Reliable Datagram Sockets	24

Executive Summary

Implementing the Oracle FS System with the Oracle Real Application Clusters (RAC) software and Oracle servers and switches provides the following key business benefits:

- » Maximized return on Oracle software investments
- » Enterprise-grade storage availability for mission-critical applications
- » Simultaneously optimize efficiency, performance, and cost according to business value
- » Horizontal and vertical scalability for optimized flash \$/TB and \$/IOP
- » Faster deployment—less time spent setting up and optimizing key applications

This white paper describes how to design, configure, and build such optimized solutions of hardware (servers, networking and storage) and software (operating systems and database software).

The system designs described in this white paper provide outstanding performance. The architecture achieves this with "well-balanced" system designs; that is, systems where the components' performance profiles are well matched, the components achieve their maximum performance at similar points, and none of the components bottleneck performance. The architectures are described with enough detail (including part numbers and detailed configuration settings) that the reader can implement identical systems.

However, since many readers' real world situations will differ from those described here, the white paper also describes, in considerable detail, the following:

- » How such systems are designed
- » How, using readily available free software, performance measurements are made
- » What constitutes a "well-balanced system design"


With this knowledge, the reader can make modifications to the presented designs (or even start from scratch) to design systems finely-tuned to their needs and their environments.

Introduction

Designing and building a full stack of hardware and software to support a real-world database is a complex and time-consuming activity. One of the particularly complex parts is the design of the hardware system. That is, what is the set of hardware (servers, networking, and storage) components that will work well together and how should the components be configured? This white paper addresses the task of designing such systems.

The primary objective of this white paper is to describe well-balanced architectures for building Oracle RAC databases on Oracle servers, switches and SAN storage. We define sets of components (software and hardware) that work well together to implement Oracle RAC databases. The contents of this white paper are detailed enough that readers can purchase the individual components and build out these architectures.

A second objective of this white paper is to enable a reader to create modified and yet still well-balanced systems. For example, although this white paper uses all-Oracle hardware components, the reader could change to a different vendor's servers. Such servers would likely have different configurations (number of CPUs, cores, memory, etc.). Hence, other aspects of the system, such as the storage design, would require changes in order to maintain overall system balance. This white paper should assist readers in being able to design and assess their proposed system designs.



A third objective is to outline the Oracle database workload testing and performance measurement environment. The environment we have used is based on freely available software. We will describe this environment with the intent that readers are able to set up their own corresponding performance assessment environments.

Intended Audience

The information presented in this white paper is intended for chief executive officers, IT directors and managers, database architects, and purchasing managers who design and build complete Oracle RAC database systems. The information benefits Oracle database administrators, system administrators, and storage administrators who test system performance.

The complete software stack consists of Oracle Database 12c, RAC clustering software, and the Oracle Linux operating system. The complete set of hardware consists of Oracle servers, switches, and SAN storage. At a minimum, we assume you are familiar with the following concepts:

- » Concepts related to the Oracle Database 12c and RAC clustering software
- » Installation and configuration of the Oracle FS System
- » Configuration and operation of the Oracle servers, switches, and Oracle FS System storage
- » Operating systems, such as Oracle Linux

Reference Architecture

Design Goals

A system can become little more than a collection of arbitrary components chosen in an ad-hoc manner unless a set of design goals and principles are specified and adhered to. The following summarizes the goals that drove the design of the systems presented in this white paper:

- » A well-balanced system design

The Performance Analysis section of this white paper is dedicated to this topic.

- » Adequate storage capacity to support multiple multi-TB databases
- » Superior transactional performance

As was stated in the Objectives section, these systems have been designed and tuned for OLTP-style applications. Example configuration choices that have been made include the QoS parameters assigned to the volumes created in the storage system.

- » Redundant and highly available

The systems should have no single point of failure and they should be built with redundant components.

- » Simple configuration – no proliferation of choice

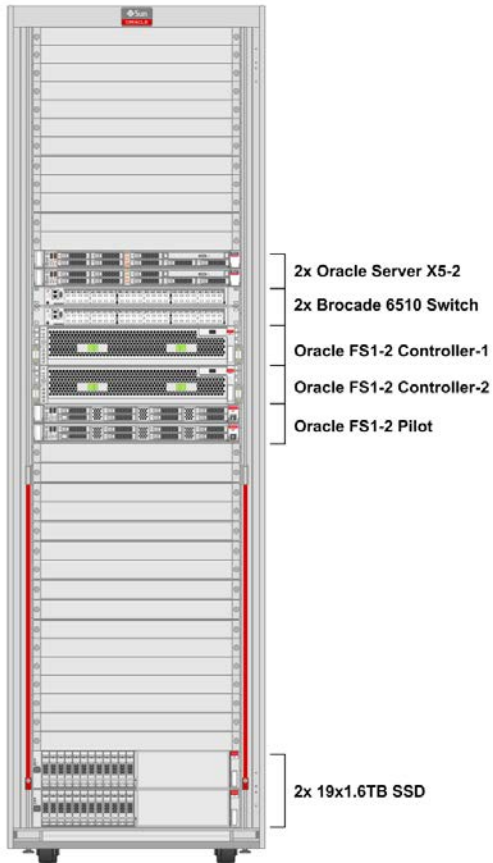
We have specified two distinct systems. While there are potentially many variations that could be made to these systems, the variations are not described in this white paper. Therefore, after reading and fully comprehending the designs, consider whether, in your particular environment, changes are required.

- » Ease of upgrade (migrating from the quarter rack configuration to the half rack configuration)

System Designs

In this white paper, we specify two system designs, as shown in Figure 1. The designs occupy 14U and 20U and we refer to these as the quarter rack and half rack designs, respectively. The designs implement two-node and four-node RAC clusters.

Quarter Rack Configuration



Half Rack Configuration

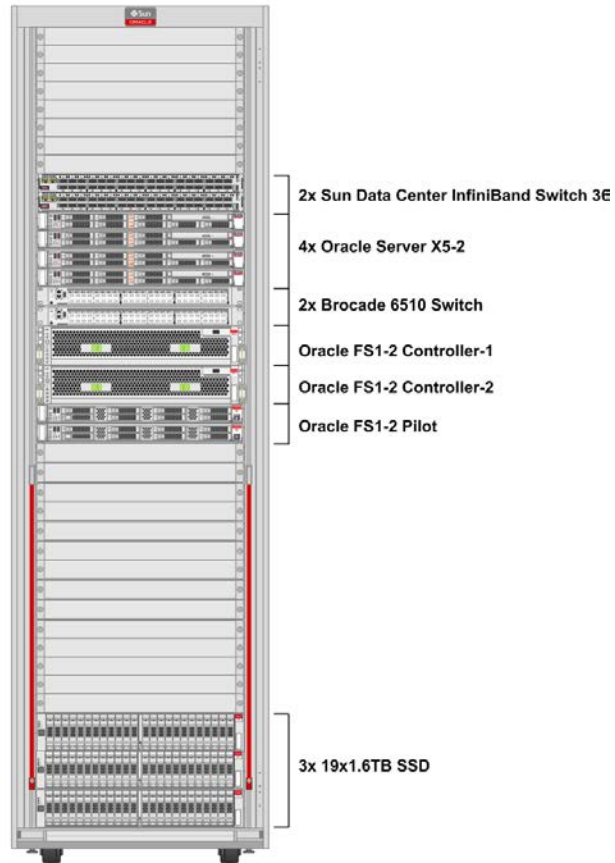


Figure 1. Quarter rack and half rack configurations

A summary for the quarter rack and half rack system designs is shown in Table 1.

Table 1. Configuration specifications

		Quarter rack configuration	Half rack configuration	
Rack	Number of Units (U)	14	20	
Servers	Number of servers	2	4	
	Number of cores	72	144	
	Memory (GB)	512	1024	
Oracle FS Storage	Raw capacity (TB)	34.8	78.3	
	Nominal maximum IOPS (70:30 random < 1ms latency)	189K	248K	
Oracle Database	Number of RAC nodes	2	4	
	SGA size (GB)	256	512	
	Tablespaces (usable TB)	DATA	14.1	32.4
		REDO	0.3	0.6
FRA		3.7	8.2	
Swingbench	TPS @ 20ms response time for 2 TB database	25,800	39,000	

Hardware Components

Key hardware components are listed in Table 2.

Table 2. Hardware components

Hardware component	Component details
Sun X5-2 Server	<ul style="list-style-type: none"> » Two Intel Xeon processors E5-2600 running at 2.3GHz » Each CPU has 18 cores » 256 GB memory » Two internal hard drives used for boot and application binaries » One dual-ported FC 16 Gb HBA » One Dual Port QDR InfiniBand Adapter M3
Brocade 6510 FC switch	<ul style="list-style-type: none"> » 24 x 16 Gb ports
Oracle All Flash FS Storage System	<ul style="list-style-type: none"> » Dual controller SAN storage array » 12 host ports (16 Gb FC) » Quarter rack: Two drive enclosures each of 13 x 1.6 TB 2.5-inch SAS-2 SSDs (41.6 TB total capacity) » Half rack: Three drive enclosures each of 19 x 1.6 TB 2.5-inch SAS-2 SSDs (91.2 TB total capacity)
Sun Data Center InfiniBand Switch 36	

Software Components

Key software components are listed in Table 3.

Table 3. Software components

Software component	Description
Oracle Linux	Oracle Linux version 6.7 is installed on the servers. Note: The servers are configured to use large pages.
Oracle Automatic Storage Management (ASM)	ASM acts as a volume manager and filesystem connecting the volumes (presented by the SAN storage array) to all the database nodes. The details of the storage and ASM configuration are presented in Storage and ASM Configurations.
Oracle RAC	RAC provides server node connectivity and clustering. Dedicated and redundant InfiniBand channels (using the RDS protocol) provide robust and reliable high speed, low latency connectivity between the cluster nodes. In the quarter rack design, the cluster nodes are directly connected. In the half rack design, the cluster nodes are connected by the InfiniBand switch.
Oracle Database 12c	Oracle Database 12c was installed on the RAC nodes. Oracle Database 12c is configured as a container database with multiple pluggable databases. The configuration of the database is presented in Oracle Database 12c Configuration on page 7.

Storage and ASM Configuration

Volumes (LUNs) are created on the Oracle FS System according to the specifications shown in Table 4 for the quarter rack system and Table 5 for the half rack system.

Note: Except for the volume reserved for the Oracle Cluster Registry (OCR), that is a small volume with minimal I/O, all volumes are created in even numbers. Half of the volumes are provisioned on the first Oracle FS System Controller and the other half are provisioned on the second Oracle FS System Controller.

The Oracle FS Storage Profiles assigned to the volumes optimize capacity usage and performance for the I/O patterns. The different I/O patterns characterize the types of Oracle files that they store. For example, the data access pattern typically found in transactional applications is one of many small read and write operations. Storage volumes configured for mirroring and striping benefit from this I/O pattern. A mirroring and striping configuration optimizes both I/O performance and data protection. Details about storage profiles and provisioning are presented in Table 6.

Table 4 and Table 5 present the details of the ASM diskgroups formed from the provisioned storage LUNs.

Note: The diskgroup configurations result in the usable database capacities listed in Table 1.

- » The size of the Fast Recovery Area (FRA) is limited to handle archived redo files and other files required for fast recovery.
- » The FRA is not sized to store backup sets. Backup sets are better suited for cheaper, legacy storage.
- » In all cases, the ASM diskgroups are configured for external redundancy.

Table 4. Storage and ASM configuration for the quarter rack system

ASM diskgroup	Number of LUNs	LUN capacity	Oracle FS storage profile	Assigned FS controller	Purpose
OCR	1	20 GB	Oracle DB ASM: Redo and Control Files	1	Oracle Cluster Registry
DATA	4	1,800 GB	Oracle DB ASM: Data OLTP	1	Data files and indexes
	4			2	
REDO	1	150 GB	Oracle DB ASM: Redo and Control Files	1	Online redo logs
	1			2	
FRA	1	1,900 GB	Oracle DB ASM: FRA	1	Archived Redo Logs, Fast Recovery Area
	1			2	

Table 5. Storage and ASM configuration for the half rack system

ASM diskgroup	Number of LUNs	LUN capacity	Oracle FS storage profile	Assigned Oracle FS controller	Purpose
OCR	1	20 GB	Oracle DB ASM: Redo and Control Files	1	Oracle Cluster Registry
DATA	5	3,320 GB	Oracle DB ASM: Data OLTP	1	Data files and indexes
	5			2	
REDO	1	300 GB	Oracle DB ASM: Redo and Control Files	1	Online redo logs
	1			2	
FRA	2	2,100 GB	Oracle DB ASM: FRA	1	Archived Redo Logs, Fast Recovery Area
	2			2	

Oracle FS Storage Profiles

When configuring a logical volume, you can select a collection of predefined properties to apply to that volume. This collection of properties is called a *storage profile*.

The Oracle FS System has many defined storage profiles available for use by the storage administrator. Table 6 lists the specific storage profiles used in these system designs.

Table 6. Details of Oracle FS storage profiles

Oracle FS storage profile name	RAID level	Read ahead	QoS priority	Stripe width
Oracle DB ASM: Redo and Control Files	Single parity	Conservative	Premium	All
Oracle DB ASM: Data OLTP	Mirrored	Normal	High	Auto-select
Oracle DB ASM: FRA	Double parity	Normal	Archive	Auto-select

Oracle Database 12c Configuration

Oracle Database 12c is configured as a container database with multiple pluggable databases.

- » Each RAC node instance is assigned two online log groups.
- » Each log group comprises a single 33 GB logfile. With this log size, even at the maximum load on the systems, log switches occur once every 12 minutes. At lower, more sustainable workloads, the logs switch less frequently. You can adjust this log size to suit the specific workloads you run on the system.

The Oracle Database 12c configuration enables multiple applications and databases to run against the systems while utilizing the performance efficiencies of the pluggable database. The values of initialization parameters, which were changed from their default values, are shown in Table 7.

Table 7. Database 12c initialization parameters

Parameter name	Value
open_cursors	1,000
processes	16,000
sga_target	128 G
gcs_server_processes	12

Performance Analysis

An IT system consists of computing, networking, and storage components, and each component contains myriad internal components. To optimize performance, availability, and serviceability, the objective is to ensure that all the internal components are configured to work together.

These system designs use a set of off-the-shelf and readily available servers, switches, and storage devices. When configuring these off-the-shelf hardware components, we considered the following:

- » How much server memory is required? The answer determines how best to combine the servers in appropriate quantities.

» How many servers are required for a balanced system design?

In this context, “balanced” means that all the major components reach their performance limits at approximately the same time. If one or more of the components perform poorly, that component throttles the overall system performance and the other components are under-utilized.

Performance Measurement Tools

We used two readily available tools to measure performance: Oracle Automatic Workload Repository (AWR) reports and Swingbench. These tools are discussed in the following sections.

Oracle AWR

The AWR is a built-in repository that exists in the SYSAUX tablespace in every Oracle database. At regular intervals, the Oracle database makes a snapshot of all its vital statistics and workload information and stores them in the AWR. From the AWR, you can generate reports for user-specified time intervals. The AWR reports are extensive and include many different types of statistics.

For the purposes of this white paper, two particular sections of the Oracle AWR Report are highlighted:

- » Wait Event Statistics—An example of the Top Timed Events is shown in Figure 2.
- » I/O Statistics—Per-second I/O statistics, by function and by file type, are shown in Figure 3.

For more details of AWR, refer to the *Database Performance Tuning Guide*:

https://docs.oracle.com/database/121/TGDBA/gather_stats.htm

Top Timed Events

- Instance '*' - cluster wide summary
- '**' Waits, %Timeouts, Wait Time Total(s) : Cluster-wide total for the wait event
- '***' 'Wait Time Avg (ms)' : Cluster-wide average computed as (Wait Time Total / Event Waits) in ms
- '***' Summary 'Avg Wait Time (ms)' : Per-instance 'Wait Time Avg (ms)' used to compute the following statistics
- '***' [Avg/Min/Max/Std Dev] : average/minimum/maximum/standard deviation of per-instance 'Wait Time Avg(ms)'
- '***' Cnt : count of instances with wait times for the event

#	Wait		Event		Wait Time			Summary Avg Wait Time (ms)				
	Class	Event	Waits	%Timeouts	Total(s)	Avg(ms)	%DB time	Avg	Min	Max	Std Dev	Cnt
*		DB CPU			8,103.39		53.86					2
	User I/O	db file sequential read	10,693,915	0.00	6,481.12	0.61	43.08	0.61	0.60	0.61	0.01	2
	Commit	log file sync	2,664,148	0.00	2,229.30	0.84	14.82	0.84	0.83	0.84	0.01	2
	Other	gcs drm freeze in enter server mode	2,776	71.33	1,250.38	450.43	8.31	450.61	446.04	455.17	6.46	2
	Cluster	gc current block 2-way	5,117,053	0.00	1,157.32	0.23	7.69	0.23	0.19	0.26	0.05	2
	Cluster	gc cr block 2-way	4,079,817	0.00	903.99	0.22	6.01	0.22	0.18	0.26	0.05	2
	Cluster	gc cr grant 2-way	2,387,009	0.00	511.52	0.21	3.40	0.21	0.18	0.25	0.05	2
	System I/O	log file parallel write	1,020,448	0.00	485.73	0.48	3.23	0.48	0.47	0.48	0.00	2
	Cluster	gc current grant busy	741,233	0.00	321.43	0.43	2.14	0.51	0.34	0.67	0.23	2
	Cluster	gc current grant 2-way	1,461,499	0.00	306.95	0.21	2.04	0.21	0.18	0.24	0.05	2
1		DB CPU			4,053.14		55.05					

Figure 2. Top Timed Events section of an AWR report

I/O Statistics

- [IOStat by Function \(per Second\)](#)
- [IOStat by File Type \(per Second\)](#)
- [Segment Statistics \(Global\)](#)

[Back to Top](#)

IOStat by Function (per Second)

- Total Reads includes all Functions: Buffer Cache, Direct Reads, ARCH, Data Pump, Others, RMAN, Recovery, Streams/AQ and XDB
- Total Writes includes all Functions: DBWR, Direct Writes, LGWR, ARCH, Data Pump, Others, RMAN, Recovery, Streams/AQ and XDB

#	Reads MB/sec			Writes MB/sec				Reads requests/sec			Writes requests/sec			
	Total	Buffer Cache	Direct Reads	Total	DBWR	Direct Writes	LGWR	Total	Buffer Cache	Direct Reads	Total	DBWR	Direct Writes	LGWR
1	136.36	136.27	0.00	66.78	52.45	0.00	14.32	17,443.95	17,437.51	0.00	8,853.57	5,360.55	0.01	3,492.47
2	140.61	140.52	0.00	58.38	44.16	0.00	14.21	17,986.48	17,980.60	0.00	8,106.08	4,595.68	0.01	3,509.84
Sum	276.97	276.79	0.00	125.16	96.61	0.00	28.53	35,430.43	35,418.11	0.00	16,959.65	9,956.24	0.01	7,002.32
Avg	138.48	138.39	0.00	62.58	48.30	0.00	14.27	17,715.22	17,709.06	0.00	8,479.83	4,978.12	0.01	3,501.16

[Back to I/O Statistics](#)

[Back to Top](#)

IOStat by File Type (per Second)

- Total Reads includes all Filetypes: Data File, Temp File, Archive Log, Backups, Control File, Data Pump Dump File, Flashback Log, Log File, Other, etc
- Total Writes includes all Filetypes: Data File, Temp File, Log File, Archive Log, Backup, Control File, Data Pump Dump File, Flashback Log, Log File, Other, etc

#	Reads MB/sec			Writes MB/sec				Reads requests/sec			Writes requests/sec			
	Total	Data File	Temp File	Total	Data File	Temp File	Log File	Total	Data File	Temp File	Total	Data File	Temp File	Log File
1	136.28	136.20	0.00	66.77	52.44	0.00	14.32	17,427.85	17,422.52	0.00	8,851.99	5,359.00	0.00	3,492.47
2	140.55	140.46	0.00	58.38	44.16	0.00	14.21	17,976.40	17,970.52	0.04	8,104.97	4,594.55	0.05	3,509.85
Sum	276.83	276.66	0.00	125.15	96.60	0.00	28.53	35,404.25	35,393.04	0.05	16,956.96	9,953.55	0.05	7,002.32
Avg	138.42	138.33	0.00	62.58	48.30	0.00	14.27	17,702.12	17,696.52	0.02	8,478.48	4,976.77	0.03	3,501.16

Figure 3. I/O Statistics section of an AWR report

Swingbench

Swingbench is a free load generator and benchmarking tool designed to stress-test Oracle databases.

For the purposes of this white paper, we used Swingbench to run workloads against the designed systems and to take performance measurements. Swingbench can be run in GUI mode (shown in Figure 4), but it also offers various command-line tools.

For more details of Swingbench, refer to <http://dominicgiles.com/swingbench.html>.

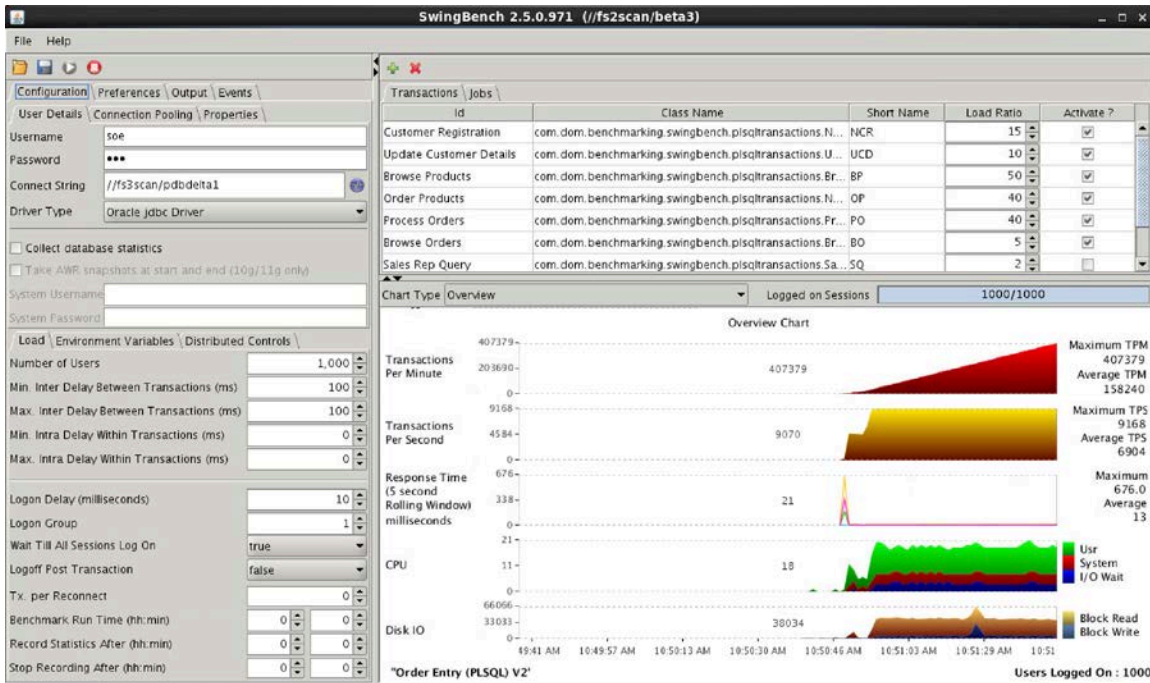


Figure 4. Example Swingbench session

Swingbench Download and Setup

In this configuration, the Swingbench software is installed on a separate Linux server. The Linux server, connected to the RAC cluster by a standard Ethernet network, is used as a workload generator and performance analysis machine.

Database and Schema Setup

The database is configured as a container database with multiple pluggable databases. The container database operates in a manner typical of many enterprise environments, where multiple databases operate simultaneously on the same hardware infrastructure. For the purposes of our testing, we created three pluggable databases and loaded the same database schema into each pluggable database.

Swingbench OrderEntry Benchmark

Swingbench ships with four benchmarks:

- » OrderEntry
- » SalesHistory
- » CallingCircle
- » StressTest

In keeping the system designed focused on transaction workloads, we worked entirely with the OrderEntry benchmark. The OrderEntry benchmark is based on the "oe" example schema that ships with Oracle Database 11g and Oracle Database 12c. The OrderEntry benchmark is similar to the Transaction Performance Council (TPC-C) benchmark

Schema Size

To assess all hardware components of the system design, the required schema size is much larger than the combined SGA sizes of the RAC nodes. 2 TB (aggregate across the three pluggable databases) is enough to generate large amounts of buffer cache reads (and hence storage IO) in all the testing scenarios.

The Swingbench oe wizard is used to prepopulate the pluggable databases. An example script to generate the aggregate 2 TB database is shown in Figure 5. The script uses the oewizard in command-line mode; since generating a schema of this size typically takes many hours. You can monitor progress updates more effectively while operating in command-line mode rather than by using the GUI.

```
#!/bin/bash

cs="//fs3scan/pdbdelta"
scale=400
data_diskgroup="+DATA"

for i in 1 2 3; do
    # drop old schema
    .../oewizard -cl -drop -cs "$cs"$i -v
    # create new schema
    ../oewizard -cl -create -cs "$cs"$i -hashpart -df $data_diskgroup -scale
    $scale -v
done
```

Figure 5. Script for using oewizard to generate an aggregate 2 TB database

Workload Setup

The OrderEntry benchmark includes a set of nine application transactions, as shown in the configuration panel of the Swingbench GUI (see Figure 4). By default, six of the transactions are enabled. The remaining three transactions, which are more analytical than transactional in nature, are disabled. For the purposes of this white paper, the default settings are used. Table 8 lists the transactions and their percentage execution rates.

Swingbench operates as a number of simulated users who are issuing transaction requests to the database. See Table 8 for an example of transaction percentages. After each completed request, the user pauses for a specified time period and then issues the next request. This cycle repeats during the workload run. Hence, the workload can be increased by increasing the number of simulated users or decreasing the delay between each user transaction. You can use more parameters to adjust the workload, but in our tests we only tested with the parameters described in this white paper.

Table 8. OrderEntry benchmark transactions

Transaction name	Percentage execution
Customer registration	9.4%
Update customer details	6.3%
Browse products	31.3%
Order products	25.0%
Process orders	25.0%
Browse orders	3.1%

Swingbench Script

An example script to execute workload runs against the pluggable databases is shown in Figure 6. The script shown takes as input parameter the desired number of simulated users. This same value is applied to each of the workloads operating against the individual pluggable databases.

For a particular system setup and configuration, you can run a series of workloads using the script. The runs use an increasing number of simulated users to assess the impact of an increasing workload.


```
#!/bin/bash
cs="//fs3scan/pdbdelta"
../minibench -cs "$cs"1 -uc $1 -cpuloc co-solsunx5-03 -t "$cs"1 -rt 0:20 -a &
../minibench -cs "$cs"2 -uc $1 -cpuloc co-solsunx5-08 -t "$cs"2 -rt 0:20 -a -pos 600,0 &
../swingbench -cs $cs"3 -uc $1 -cpuloc co-solsunx5-09 -t $cs"3 -rt 0:18 -bs 0:12 -be 0:17 &
#
```

Figure 6. Script to execute Swingbench workloads

Swingbench Workloads

The script launches three instances of Swingbench. Each instance operates against one of the three pluggable databases.

- » Two of the Swingbench instances use Minibench. Minibench produces a smaller version of the GUI, showing the charts for monitoring transaction rates. These two Swingbench instances begin generating application transaction requests immediately upon script execution.
- » The third Swingbench instance is set to produce the full GUI. See Figure 4 for an example Swingbench session.
 - » The Swingbench configuration is set to record statistics and to trigger database AWR snapshots at the beginning and end of the statistics gathering period.
 - » Statistics gathering is set to begin 12 minutes after the workload is started, and to gather statistics for 5 minutes. The delay in gathering statistics ensures that each database starts up satisfactorily and reaches a steady state. A steady state means that the database is beyond the initial period of heavy buffer cache reads and the database writers have begun to issue writes to the storage.



The AWR report and the Swingbench statistics are captured when each workload run completes. Swingbench statistics include the average transactions per second, average CPU usage for the RAC nodes, and average transaction response time for each of the six transaction types.

The various statistics used to assess system performance and system balance are discussed in the following sections.

Assessment of System Balance

At the highest level, the system designs are concerned with the ratio of compute (number of X5-2 servers) to storage (the number and type of SSD drive groups in the Oracle FS Storage System).

Many variations of each configuration were tested. The results for the selected configurations are shown in the following charts. Two types of charts are presented: Balance and Load Response.

Balance Charts

Figure 7 is an example of a balance chart. In this chart, one parameter is selected to proxy for server load performance (host CPU usage – obtained from Swingbench statistics) and one for storage I/O load performance (db file seq read – from the AWR report).

On the chart, shading is used to indicate where the component is reaching its performance limit. For storage I/O on SSD, a db file seq read value greater than 1.5ms is beginning to show storage stress and above about 2ms is indicating serious performance limits. For the servers, the CPU usage should be kept below 70% to 80%.

Hence, a system that is CPU (but not storage) limited would show a line running vertically upwards toward the upper red zone, while a system that is storage limited (but not CPU) would have a line running horizontally toward the right side red zone. A perfectly balanced system would have a line running from bottom left to upper right.

Load Response Charts

Figure 8 is an example of a load response chart for the quarter rack system. Successive Swingbench runs with increasing numbers of simulated users drive the chart results. The Swingbench statistics for each run provide the average transactions per second and average transaction response time.

The load response charts show a characteristic response curve. At first, the response time holds steady or increases moderately. Then, once the performance limit of at least one component is reached, the response time turns sharply upward. Only minimal performance gains (increased transactions per second) occur as further load is applied. The load response charts show the lines for varying numbers of RAC nodes.

Quarter Rack and Half Rack System Charts

The charts for the quarter rack system are shown in Figure 7 and Figure 8. The charts show clearly that having a single RAC (server) node is too little CPU, but going beyond two nodes adds no further performance (on the load response chart the curves for two, three and four servers lie on top of one another showing that the system is storage limited).

The charts for the Half Rack system are shown in Figure 9 and Figure 10. In this case, the additional storage drive groups (adding more capacity and performance) ensure that larger numbers of server nodes lead to increased system performance -- Figure 10 shows increasing performance as increasing numbers of servers are added.

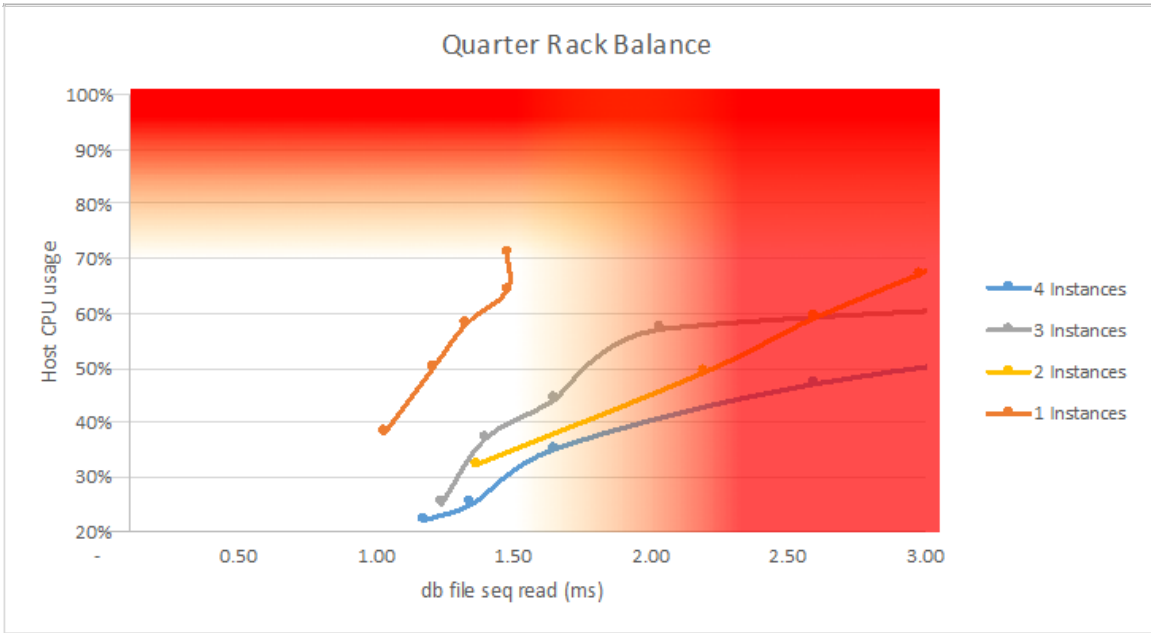


Figure 7. Quarter rack balance chart

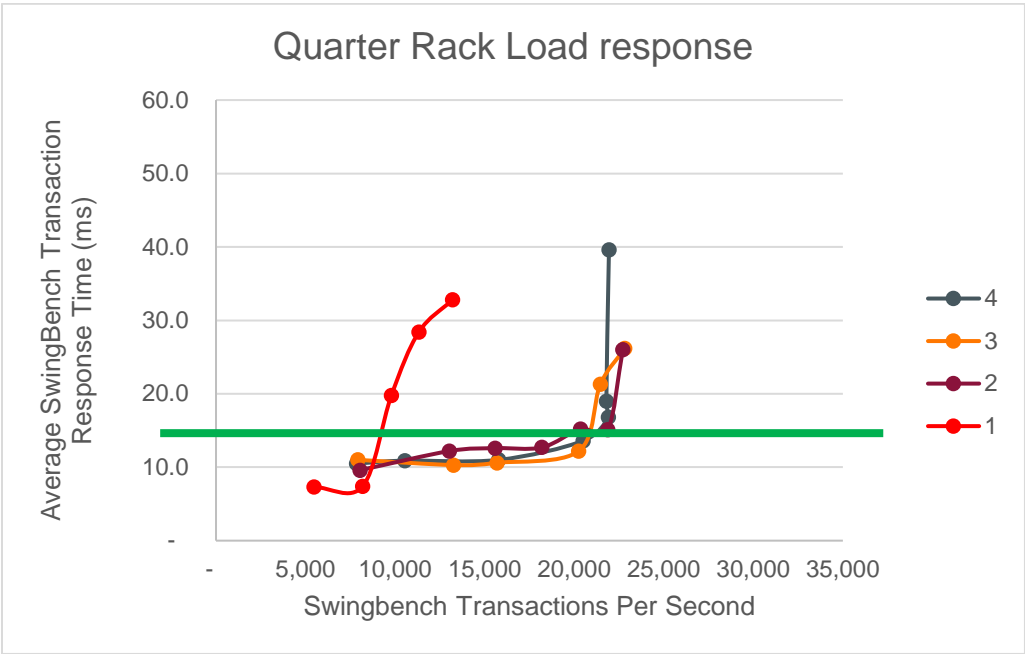


Figure 8. Quarter rack load response chart

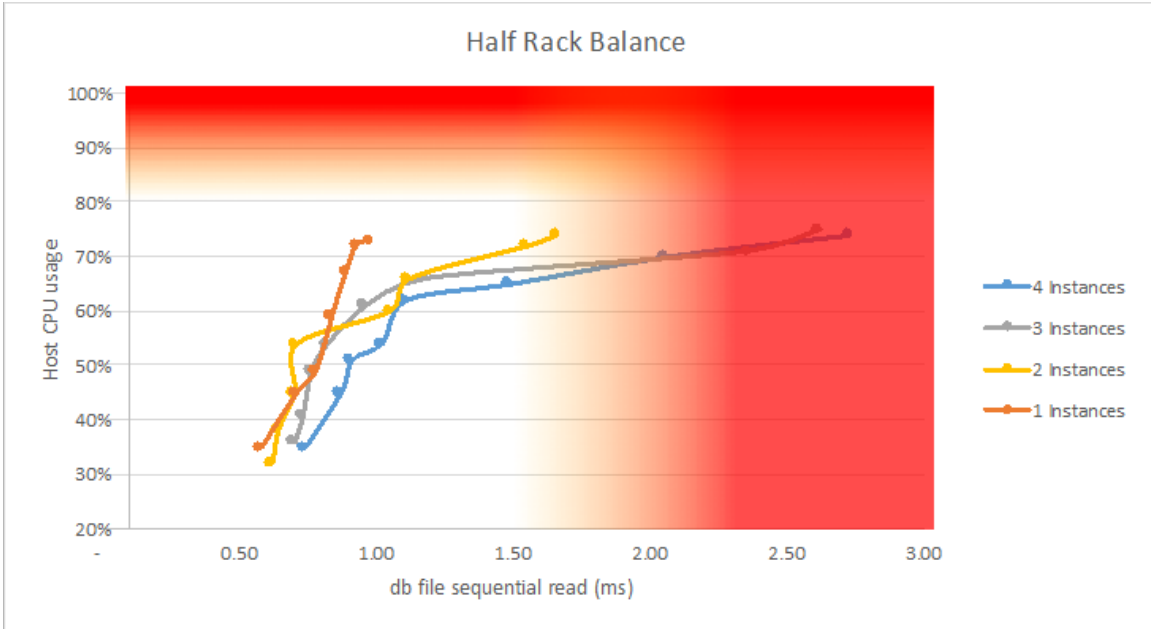


Figure 9. Half rack balance chart

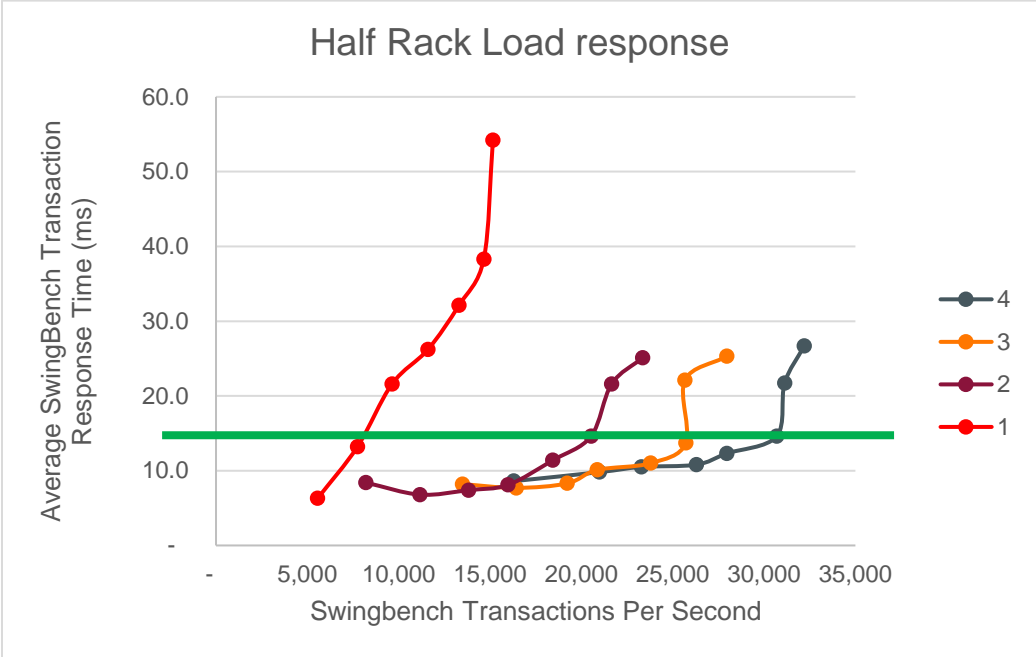


Figure 10. Half rack load response chart

Determining Configuration Settings

Gross system settings and hardware components were assessed by running multiple Swingbench workloads against varying parameter settings. Table 9 lists the results. The parameter settings reflect general transactional workloads. We advise that you reassess these parameters based on your particular application workloads.

Table 9. Configuration settings for application workloads

System component	Parameter	Value	Comment
Servers	Memory	See Table 6 for details about the Oracle FS storage profiles	
	RAID setting		
	QoS priority		
	QoS read ahead		
	Stripe width		
	Number of LUNs per ASM diskgroup	2	
Storage	Storage domains	1	In systems with hard disk drives (HDDs) separating ASM diskgroups into different storage domains mitigates noisy neighbor issues. Since these designs use All Flash storage, separating into separate storage domains is not necessary.
	SGA	128 GB	Per node
Database	gcs_server_processes	12	
	MTTR	0	

Conclusion

The performance of the two reference architectures (with the specifications and configurations as defined in previous sections) is summarized in Figure 11.

For both systems, the charts show a gradually increasing average transaction response time as the workload increases. The response time sharply increases once the system reaches its nominal maximum performance. Using a response time of 20ms as a benchmark, the quarter rack system achieves about 26,000 transactions per second. Using the same benchmark, the half rack system achieves about 40,000 transactions per second.

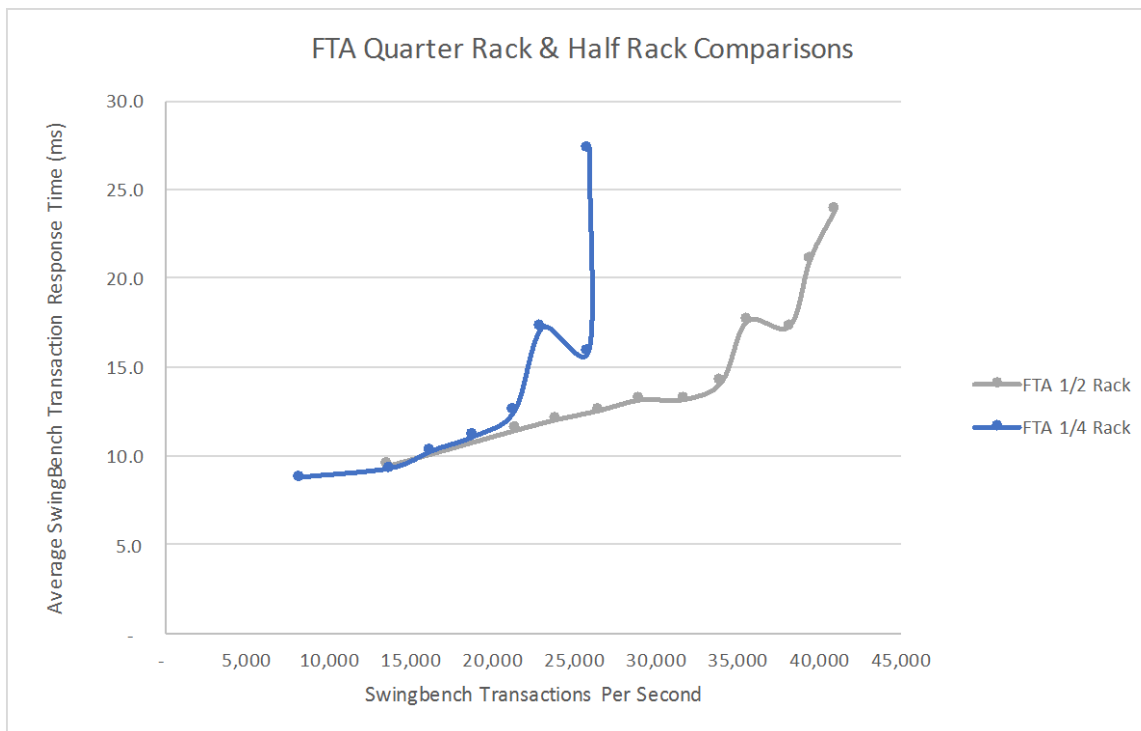


Figure 11. Performance of quarter rack and half rack systems

Related Documentation

In addition to the information contained in this document, refer to Oracle FS System resources listed in this section.

Oracle Flash Storage Documentation

[Oracle Flash Storage System Administrator's Guide](#)

[Oracle Flash Storage System CLI Commands](#)

Oracle Technical Support and Resources

<http://www.oracle.com/support> (non-emergency business hours)

Appendix A. Installation and Configuration Details

The information in this appendix guides you through installing and configuring a two-node Oracle RAC cluster in a quarter rack system and a four-node Oracle RAC cluster in a half rack system. The configuration details of Oracle RAC and Oracle Linux for the Flash Transaction Architecture (FTA) are discussed in the following sections.

Target Audience

This section is for Oracle Database Administrators or Linux System Administrators who are experienced with installing Oracle RAC and Oracle Linux.

- » For information about installing Oracle RAC, consult the *Oracle Real Application Clusters Installation Guide 12c Release 1 (12.1) for Linux and UNIX*.
- » For information about installing Oracle Linux, consult the *Oracle Linux Installation Guide for Release 6*.

Hardware Configuration

Quarter Rack hardware components are detailed in Table 10.

Note: Connect the Infiniband ports from each server to an Infiniband switch. The Infiniband port-to-switch connection is used for the RAC interconnect.

Table 10. Hardware configuration details

Hardware component	Quarter rack component details	Half rack component details
RAC Nodes	2 RAC Nodes	4 RAC Nodes
Oracle Server X5-2	<ul style="list-style-type: none">» 256 GB RAM» 2 x Intel® Xeon® 18-core 2.3-GHz processor» 2 x 600 GB 10,000 rpm SAS-3 HDDs» 1 x Sun Storage Dual 16 Gb Fibre Channel PCIe Universal HBA, Qlogic» 1 x Oracle Dual Port QDR InfiniBand Adapter	<ul style="list-style-type: none">» 256 GB RAM» 2 x Intel® Xeon® 18-core 2.3-GHz processor» 2 x 600 GB 10,000 rpm SAS-3 HDDs» 1 x Sun Storage Dual 16 Gb Fibre Channel PCIe Universal HBA, Qlogic» 1 x Oracle Dual Port QDR InfiniBand Adapter
Oracle All Flash FS1-2 Flash Storage System	<ul style="list-style-type: none">» 2 x Oracle FS1-2 Controller: Performance configuration» 6 x Sun Storage Dual 16 Gb Fibre Channel PCIe Universal HBA, Qlogic» 2 x Oracle Storage Drive Enclosure with thirteen 1.6 TB SAS SSDs	<ul style="list-style-type: none">» 2 x Oracle FS1-2 Controller: Performance configuration» 6 x Sun Storage Dual 16 Gb Fibre Channel PCIe Universal HBA, Qlogic» 3 x Oracle Storage Drive Enclosure with nineteen 1.6 TB SAS SSDs

Bill of Materials

Table 11. Bill of materials for quarter rack and half rack configurations

Part	Description	Quantity	
		Quarter rack	Half rack
7110316	Oracle Server X5-2:Model family	2	4
7101675	2 Sun Storage 16 Gb FC short wave optics, Qlogic (for factory installation)	2	4
2124A	QSFP parallel fiber optics short wave transceiver	2	4
6331A-N	2.5-inch HDD filler panel	8	16
7110346	Intel® Xeon® E5-2699 v3 18-core 2.3 GHz processor	4	8
7110350	Heat sink for 1U	4	8
7110353	16 GB DDR4-2133 DIMM	32	64
7111102	600 GB 10,000-rpm 2.5-inch SAS-3 HDD with marlin bracket	4	8
7104073	Oracle Dual Port QDR InfiniBand Adapter M3	2	4
7101673	Sun Storage Dual 16 Gb Fibre Channel PCIe Universal HBA, Qlogic	4	8
7110337	Oracle Server X5-2: 1U base chassis with motherboard, internal 12 Gb SAS RAID HBA, 2 PSUs, slide rail kit, and cable management arm	2	4
7102748	PCIe filler panel	2	4
333V-10-10-C14	Power cord: Jumper, straight plug-connector, 1.0 meter, IEC60320-2-2 Sheet E (C14) plug, IEC60320-1-C13 connector, 10 A, 250 VAC	4	8
7110360	OSA 8 GB USB stick	2	4
7110339	Eight 2.5-inch drive slots, 1 DVD-RW, and disk cage for 1U	2	4
7110359	DVD filler panel	2	4
Storage			
7113864	Oracle All Flash FS1-2 Flash Storage System: Model family	1	1
7111942	Oracle FS Pilot with 2 Oracle Server X5-2	1	1
7104975	Oracle FS1-2 subsystem: model family	1	1
7104710	Oracle FS1-2 Controller: performance configuration	2	2
7101673	Sun Storage Dual 16 Gb Fibre Channel PCIe Universal HBA, Qlogic	6	6
7101675	2 Sun Storage 16 Gb FC short wave optics, Qlogic	6	6
7112956	Oracle Storage Drive Enclosure DE2-24P with thirteen 1.6 TB SAS SSDs	2	0
7112957	Oracle Storage Drive Enclosure with nineteen 1.6 TB SAS SSDs	0	3
7104928	Cable: 3 meters, mini SAS to mini SAS HD	4	4
7104932	Cable assembly: 3 meters, mini SAS HD to mini SAS HD	6	6
Miscellaneous			
7103553	Brocade 6510 Switch	2	2

Configure Server Internal Storage

1. Consult the *Oracle® Server X5-2 Installation Guide* for instructions on how to configure the two internal 600 GB HDDs into a RAID 1 mirrored pair. You configure the HDDs using the internal RAID configuration utilities for each server.
2. Partition the resulting RAID volume into a swap partition of 16 GB and devote the rest to the root volume.

Install Oracle Linux

1. Install the latest Unbreakable Enterprise Kernel of Oracle Linux version 6.7.
2. Consult the *Oracle Linux Installation Guide for Release 6* and *Oracle Server X5-2 Installation Guide for Linux Operating Systems*. These guides provide detailed steps to install and configure the operating system and apply the latest patches.
3. Install the operating system for each RAC node on the root volume.

Install and Configure Oracle FS Path Manager (FSPM)

1. After Linux is installed, install Oracle FS Path Manager. For detailed instructions, consult the *Oracle FS Path Manager Release 4 Installation Guide for Linux*.
2. Configure HugePages.

Assuming 256 GB RAM, configure Hugepages as root, as shown in the following example.

```
# vi /etc/sysctl.conf
vm.nr_hugepages=66666

# sysctl -p
vm.nr_hugepages=66666
...

# grep -i huge /proc/meminfo
AnonHugePages:          0 kB
HugePages_Total:       66666
HugePages_Free:        1129
HugePages_Rsvd:         0
HugePages_Surp:         0
Hugepagesize:          2048 kB
```

3. Add the following entries to */etc/security/limits.conf*, where the setting is at least the size of the HugePages allocation in KB (*HugePages_Total* * *Hugepagesize*).

```
*                soft    memlock    136531968
*                hard    memlock    136531968
```

Configure Oracle FS Storage

1. Connect all front-end ports of the Oracle FS System to the Fibre Channel switches.
2. Connect the RAC Node HBA ports to the same switches.
3. Perform zoning according to your requirements.

4. After port connection and zoning are finished, proceed to allocate LUNs to the RAC nodes. Details for each LUN are shown in Table 10.

Diskgroup and LUN Layout

Create LUNs using the details in Table 12 (for the quarter rack configuration) and Table 13 (for the half rack configuration). Be sure to assign the LUNs to the Oracle FS Controllers as shown so that the workload is evenly distributed.

Table 12. Diskgroup and LUN layout (quarter rack configuration)

Diskgroup name	LUN name	Storage profile	LUN size	Controller	Purpose
OCR	OCR-01-01	Oracle DB ASM: Redo and Control Files	20 GB	01	Oracle Cluster Registry
DATA	DATA-01-01	Oracle DB ASM: Data OLTP	1,800 GB	01	Data files and indexes
	DATA-02-02			02	
	DATA-01-03			01	
	DATA-02-04			02	
	DATA-01-05			01	
	DATA-02-06			02	
	DATA-01-07			01	
	DATA-02-08			02	
REDO	REDO-01-01	Oracle DB ASM: Redo and Control Files	150 GB	01	Online redo logs
	REDO-02-02			02	
FRA	FRA-01-01	Oracle DB ASM: FRA	1,900 GB	01	Archived redo logs
	FRA-02-02			02	

Table 13. Diskgroup and LUN layout (half rack configuration)

Diskgroup name	LUN name	Storage profile	LUN size	Controller	Purpose
OCR	OCR-01-01	Oracle DB ASM: Redo and Control Files	20 GB	01	Oracle Cluster Registry
DATA	DATA-01-01	Oracle DB ASM: Data OLTP	1,800 GB	01	Data files and indexes
	DATA-02-02			02	
	DATA-01-03			01	
	DATA-02-04			02	
	DATA-01-05			01	
	DATA-02-06			02	
	DATA-01-07			01	
	DATA-02-08			02	
	DATA-01-09			01	
	DATA-02-10			02	
	DATA-01-11			01	
	DATA-02-12			02	
	DATA-01-13			01	
	DATA-02-14			02	
	DATA-01-15			01	
	DATA-02-16			02	
	DATA-01-17			01	
	DATA-02-18			02	
REDO	REDO-01-01	Oracle DB ASM: Redo and Control Files	150 GB	01	Online redo logs
	REDO-02-02			02	
	REDO-01-03			01	
	REDO-02-04			02	
FRA	FRA-01-01	Oracle DB ASM: FRA	1,900 GB	01	Archived redo logs
	FRA-02-02			02	
	FRA-03-03			01	
	FRA-04-04			02	

Install Oracle RAC

For information about installing Oracle RAC, consult the *Oracle Real Application Clusters Installation Guide 12c Release 1 (12.1) for Linux and UNIX*. Refer to the following sections for instructions on how to configure the Oracle RAC for the FTA.

Install ASMLib

1. Install Oracle ASMLib according to the instructions in *Oracle 12c Grid Infrastructure Installation Guide for Linux, Release 1 (12.1)*.
2. Label the disks.

Create the ASM Instance

1. Create all diskgroups with external redundancy.

The Oracle FS System protects the LUN data using RAID. You do not need to protect the data using redundancy at the ASM level.

2. During the Grid Infrastructure installation, ensure that the Oracle Cluster Registry (OCR) LUN is the first LUN used during the ASM installation. In this configuration, all OCR Files are guaranteed to reside on that LUN.

Creating the Oracle Database

For maximum performance, create one or more pluggable databases instead of creating multiple database instances. Use the database parameters in Table 13.

Table 13. Recommended database parameter settings

Parameter	Quarter rack value	Half rack value
dispatchers	(PROTOCOL=TCP)(DISP=10)	(PROTOCOL=TCP)(DISP=10)
shared_servers	800	800
sga_max_size	128 GB	128 GB
sga_target	128 GB	128 GB
db_recovery_file_dest_size	3800 GB	8200 GB
db_recovery_file_dest	+FRA	+FRA
processes	16000	16000
temp_undo_enabled	True	True

Undo Tablespace Considerations

The Temporary Undo feature provides the following benefits:

- » Creates a separate undo segment for transactions against temporary objects
- » Spreads out undo operations across more storage
- » Eliminates redo operations for temporary objects
- » Bases undo retention only on user data and not temporary data



To turn on the Temporary Undo feature in Oracle Database 12c, set TEMP_UNDO_ENABLED to true.

Note: The recommended size for the undo tablespace for each instance is 30 GB.

Configure the Redo Logs

For each redo log instance, create two online redo log groups. Create each group with one 33 GB redo log, for a total of 4 x 33 GB redo logs (quarter rack) and 8 x 33 GB redo logs (half rack)..

Enable Reliable Datagram Sockets

Reliable Datagram Sockets (RDS) is a low-overhead protocol that increases the data transfer performance for the interconnection between Oracle RAC nodes. To enable RDS, enter the following commands from the Linux command line as the Oracle user.

```
$ cd ORACLE_HOME/rdbms/lib
$ make -f ins_rdbms.mk ipc_rds ioracle
```



Oracle Corporation, World Headquarters

500 Oracle Parkway
Redwood Shores, CA 94065, USA

Worldwide Inquiries

Phone: +1.650.506.7000
Fax: +1.650.506.7200

CONNECT WITH US

 blogs.oracle.com/oracle

 facebook.com/oracle

 twitter.com/oracle

 oracle.com

Integrated Cloud Applications & Platform Services

Copyright © 2016, 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.

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. AMD, Opteron, the AMD logo, and the AMD Opteron logo are trademarks or registered trademarks of Advanced Micro Devices. UNIX is a registered trademark of The Open Group. 0116

A Reference Architecture for Implementing an Oracle RAC Database on Oracle Servers, Switches, and SAN Storage Hardware

August 2016