Hewlett Packard Enterprise

Reference Architecture

HPE Reference Architecture for Oracle RAC 19C Database on HPE Superdome Flex 280 Server with HPE Primera

Oracle RAC 19c database deployment best practices and scalability

CONTENTS

EXECUTIVE SUMMARY

Enterprise businesses are growing exponentially year-over-year and demanding faster transaction processing speeds and highly scalable infrastructure. Oracle Real Application Clusters (RAC) database architecture enables IT organizations to increase the transaction processing speed by adding more servers to the existing deployment, in a scale-out model. Though adding more servers meets the ever-growing business needs, it also increases the hardware footprint in the data center over a period of time. In such a case, a combination of scale-up and scale-out models would help to cut the hardware footprint, operational cost, and bring down the management overhead.

The HPE Superdome Flex 280 server is a new mission-critical x86 server platform from the HPE Superdome Flex family of servers, it supports the processor scaling from two to eight 3rd generation Intel® Xeon® Scalable processors in a single system with up to 28-cores per socket for a maximum of 224 cores. This highly scalable hardware architecture with the combination of Oracle RAC is an ideal solution for fast-growing enterprise database platforms.

The testing featured in this Reference Architecture highlights capabilities, best practices, and optimal settings for Oracle transaction processing workloads running in a Red Hat® Enterprise Linux® environment on the HPE Superdome Flex 280 server with HPE Primera storage. This document demonstrates the following:

- **Server scale-up:** Transactional throughput increases linearly as the cluster nodes are upgraded from 2-socket to 4-socket configurations. In the testing, transactional throughput increased up to 81% by moving the workload from the 2-socket to the 4-socket server.
- **Server scale-out:** Transaction throughput increases linearly as Oracle RAC nodes are added to the cluster. The scale-out comparison was done on two separate three-node Oracle RAC cluster deployments with 2-socket and 4-socket configurations.
  - With a 2-socket configuration, two (2) nodes provided up to 1.84 times the number of transactions, and three (3) nodes provided up to 2.44 times the number of transactions, as compared to the single (1) RAC node instance.
  - With a 4-socket configuration, two (2) nodes provided up to 1.86 times the number of transactions, and three (3) nodes provided up to 2.62 times the number of transactions, as compared to the single (1) RAC node instance.

With the help of the HPE Application Tuner Express (HPE-ATX) tool, transaction throughput was improved. The HPE-ATX tool aligns the Oracle processes with their data in memory and evenly spreads them across NUMA nodes on the HPE Superdome Flex 280 server.

Target audience: This Reference Architecture (RA) is designed for IT professionals, who use the program, manage, or administer large databases that require high performance. Specifically, this information is intended for those who evaluate, recommend or design new, and existing high-performance IT architectures. Additionally, CIOs may be interested in this document as an aid to guide their organizations in determining when to implement an Oracle OLTP environment alongside the performance characteristics associated with those implementations.

Document purpose: The purpose of this document is to describe a Reference Architecture demonstrating the benefits of running the Oracle RAC 19c Database on a set of HPE Superdome Flex 280 servers and an HPE Primera A650 all-flash array.

This Reference Architecture describes solution testing performed in June 2021.

INTRODUCTION

Faster transaction processing speeds, capacity-based scaling, increased flexibility, high availability, and business continuity are required to meet the needs of a 24/7 business. On the other hand, IT organizations are constantly looking for cost-saving opportunities. Deploying Oracle RAC on the HPE Superdome Flex 280 servers enable IT organizations to meet these requirements in the following ways:

- An Oracle RAC environment enables all active database instances to concurrently execute transactions against a shared database, with data consistency and integrity. This inherently provides high availability (HA) and helps to increase the transaction processing capacity just by adding servers into the existing cluster.
- The HPE Superdome Flex 280 server scales from 2 to 8 sockets with 3rd generation Intel® Xeon® Scalable processors and enables customers to increase the transaction processing capacity just by adding more processors and memory, instead of adding new server hardware.
- HPE Superdome Flex 280 servers are ideal for mission-critical workloads in data-driven enterprises.
- The Reliability, Availability, and Serviceability (RAS) capabilities of the HPE Superdome Flex 280 server ensures the ability to operate continuously, maintain data integrity, and quickly recover back to an operational state after a failure, with minimal impact. Refer to the HPE Superdome Flex Server 280 Architecture and RAS for more details.

SOLUTION OVERVIEW

This solution included the HPE Superdome Flex 280 server with HPE Primera A650 all-flash storage, running on Oracle Database 19c. The HPE Application Tuner Express (HPE-ATX) software was utilized to achieve maximum performance in the NUMA environment.

HPE Superdome Flex 280 Server

The HPE Superdome Flex 280 server is a new model in the HPE Superdome Flex family of servers. It is a highly scalable, reliable, and secure server that starts at two (2) and scales up to eight (8) powerful 3rd generation Intel Xeon Scalable Processors.

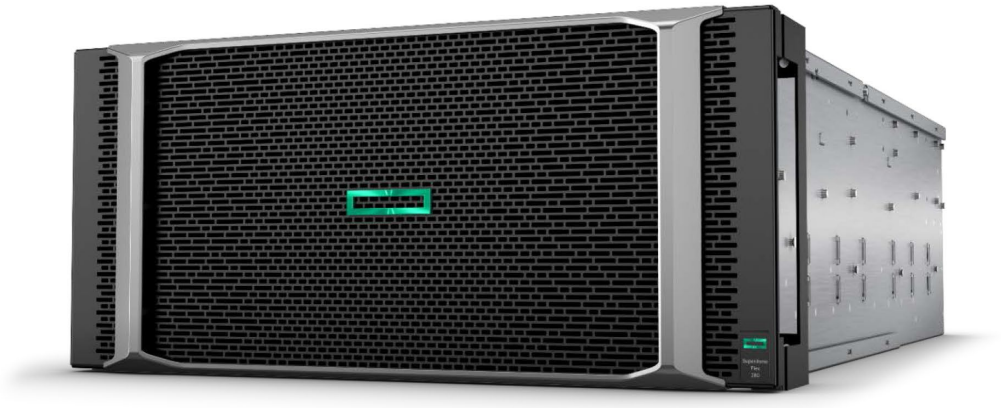

FIGURE 1. HPE Superdome Flex 280 server – Single chassis

With the HPE Superdome Flex 280 server, the Oracle Database infrastructure can extend along with data needs. The HPE Superdome Flex 280 server offers:

- Support for 2 to 8 sockets of Intel Xeon Scalable processors in a single system with up to 28-cores per socket for a maximum of 224 cores.
- Six (6) UPI links per socket providing unparalleled bandwidth and performance.
- 48 DIMM slots of DDR4 memory per chassis.
- 64 GB – 24 TB of shared memory.
- 16 half-height IO slots, or 8 full-height + 4 half-height IO slots, per 4-socket chassis.
- Base IO includes support for 8 SATA drives via VROC, two 1GbE NIC ports, four USB ports, Management LAN.
- Internal storage up to 10 drive bays, or 8 drives and optional DVD.
- HPE Superdome Flex 280 server Error Analysis Engine for better diagnostics and mission-critical reliability.

The HPE Superdome Flex 280 server is configured with an embedded rack management controller (eRMC) for hardware management purposes. The eRMC is included with the base chassis and allows you to perform the following:

- System inventory, health, and configuration
- Launch vMedia, vKVM (HTML5)
- Configure UEFI, reboot, and power on/off
- eRMC security settings
- eRMC LAN settings, etc.

For more details, refer to the HPE Superdome Flex 280 Server QuickSpecs.

HPE Primera

HPE Primera 600 Storage is a Tier-0 enterprise storage solution designed for simplicity, resiliency, and performance when running on mission-critical applications and workloads. Built upon proven resiliency and powered by the intelligence of HPE InfoSight, HPE Primera 600 Storage delivers instant access to data with storage that sets up in minutes, upgrades transparently, and is delivered as a service. It ensures fast storage for all mission-critical applications. HPE Primera 600 Storage is a parallel, multi-node, and all-active platform that achieves predictably high performance at extremely low latency.

The HPE Primera 600 storage systems include the HPE Primera 630, HPE Primera 650, and HPE Primera 670. Each model is available in an all-flash configuration or a converged flash configuration. All models come with factory-installed HPE Primera OS and HPE Primera UI.

The following summarizes the basic configurations. It does not include information about available add-on drive enclosures and other options. For more information on the number of host ports and maximum storage capacities, see the HPE Primera 600 Storage QuickSpecs. For Oracle Database workloads, a four-storage controller-based configuration is recommended.

- **HPE Primera A630 and C630:** The HPE Primera A630 is an all-flash configuration and the HPE Primera C630 is a converged flash configuration. The 630 includes a 2U base enclosure that contains two (2) storage controllers. The 2U base enclosure can hold up to 24 small form factor physical drives.
- **HPE Primera A650 and C650:** The HPE Primera A650 is an all-flash configuration and the HPE Primera C650 is a converged flash configuration. The 650 is available with the 2U or 4U base enclosure that contains either two (2) or four (4) storage controllers. The 2U and 4U base enclosure can hold up to 24 or 48 small form factor physical drives. In this Reference Architecture, the HPE Primera A650 all-flash configuration was used.
- **HPE Primera A670 and C670:** The HPE Primera A670 is an all-flash configuration and the HPE Primera C670 is a converged flash configuration. The 670 is available with the 2U or 4U base enclosure that contains either two (2) or four (4) storage controllers. The 2U and 4U base enclosures can hold up to 24 or 48 small form factor physical drives.

FIGURE 2. HPE Primera 600 Storage series

Management interfaces for HPE Primera are:

- **HPE Primera UI:** The HPE Primera UI is a graphical user interface for managing a single HPE Primera 600 storage system. HPE Primera UI software is included in each HPE Primera 600 storage system and does not require installation on a server.
- **HPE Primera CLI:** HPE Primera CLI is a text-based command-line interface for managing one HPE Primera 600 storage system at a time. The functionality is included in the HPE Primera OS on HPE Primera 600 storage systems. HPE Primera CLI client software can be installed on hosts (servers), running on various computer operating systems.
- **HPE 3PAR StoreServ Management Console (SSMC):** HPE 3PAR SSMC is a graphical user interface for managing multiple HPE Primera 600 storage systems and HPE 3PAR StoreServ storage systems at a time. The software is available as a virtual appliance and can be downloaded from the HPE Software Depot. The software can be deployed in several supported virtual machine environments.

HPE 3PAR SSMC provides additional features and capabilities compared to the HPE Primera UI version 1.1.

HPE Application Tuner Express (HPE-ATX)

HPE Application Tuner Express is a utility for Linux® users to achieve maximum performance when running on multi-socket servers. Using this tool, you can align application execution with the data in-memory resulting in increased performance. HPE-ATX is designed to improve the

performance of NUMA unaware applications without requiring any changes to the application itself. It helps multi-process and multi-threaded applications running on multi-socket machines to achieve better NUMA placement of processes and threads that are related and share memory segments.

HPE-ATX offers the following launch policies to control the distribution of an application's processes and threads in a NUMA environment:

- **Round Robin**: Each time a process (or thread) is created, it will be launched on the next NUMA node in the list of available nodes. This ensures even distribution across all of the nodes.
- **Fill First**: Each time a process (or thread) is created, it will be launched on the same NUMA node until the number of processes (or threads) matches the number of CPUs in that node. Once that node is filled, future processes will be launched on the next NUMA node.
- **Pack**: All processes (or threads) will be launched on the same NUMA node.
- **None**: No launch policy is defined. Any child process or sibling thread that is created will inherit any NUMA affinity constraints from its creator.

For Oracle workloads, we strongly recommend launching the database listeners using HPE-ATX with "ff_flat" or "rr_flat" process launch policy while affinitizing to all the NUMA nodes where the database instance is running. HPE-ATX is fully supported by Hewlett Packard Enterprise and can be downloaded from the https://downloads.linux.hpe.com/SDR/project/hpe-atx/repo.html.

HPE OneView

The HPE OneView is a converged infrastructure management platform that provides a unified interface for the administration of systems in a data center. Through a single GUI, sometimes referred to as a single pane of glass, administrators can automate management and maintenance tasks that have traditionally been performed manually, and required several different tools. Within the data center, HPE OneView can manage physical systems, storage arrays, and network connectivity. HPE OneView can manage or monitor up to 80 HPE Superdome Flex 280 servers and HPE Superdome Flex systems. The HPE OneView Ansible library provides modules to manage HPE OneView using Ansible playbooks using HPE OneView REST APIs. HPE OneView Standard or Monitored mode enables:

- Server discovery
- A detailed inventory of physical and logical resources
- Comprehensive health monitoring, activities/alerts, and reporting
- Changing the BIOS settings, defining a server profile for efficiently standing up, and deploying servers in the future

Refer to the HPE Superdome Flex 280 Server Manageability for licensing and more details.

HPE Infosight for HPE Primera

The HPE Primera uses artificial intelligence (AI) and machine learning (ML), powered by HPE InfoSight, to predict and prevent disruptions across storage, servers, and virtual machines. Over the past decade, HPE InfoSight has analyzed application patterns across 1,2503 trillion data points, transforming how storage is managed and supported. Thousands of disruptions have been prevented saving our customers over 1.5 million work hours.[1]. This end-to-end app-aware approach for resiliency is why every HPE Primera is backed by a 100% Availability Guarantee.

Oracle Real Application Clusters

Oracle Real Application Clusters (RAC) is an option for Oracle Database that provides high availability (HA) and scalability to the Oracle Database without requiring any application changes. Oracle Database with the Oracle RAC option allows multiple database instances running on different servers to access the same physical database stored on shared storage. The database spans multiple systems but appears as a single unified database to the application. This provides a scalable computing environment, where capacity can be increased by adding more nodes to the cluster. While all servers in the cluster must run the same OS and the same version of Oracle, they need not have the same capacity, which allows adding servers with more processing power, and memory when more performance is required. This architecture also provides high availability, as RAC instances running on multiple nodes protect from a node failure.

Oracle Database 19c

Oracle Database 19c[2] builds upon the innovations of previous releases such as Multitenant, In-Memory, JSON support, Sharding, and many other features. This release introduces new functionality, providing customers with a multi-model enterprise-class database for all their typical use cases, including:

- Operational database use cases such as traditional transactions, real-time analytics, JSON document stores, and Internet of Things (IoT) applications.
- Analytical database use cases such as traditional and real-time data warehouses, data marts, big data lakes, and graph analytics.

[2] Oracle Database 19c Introduction and Overview

SOLUTION COMPONENTS

This solution included the HPE Superdome Flex 280 server with HPE Primera A650 all-flash storage, running on Oracle Database 19c. Figure 3 shows the solution diagram.

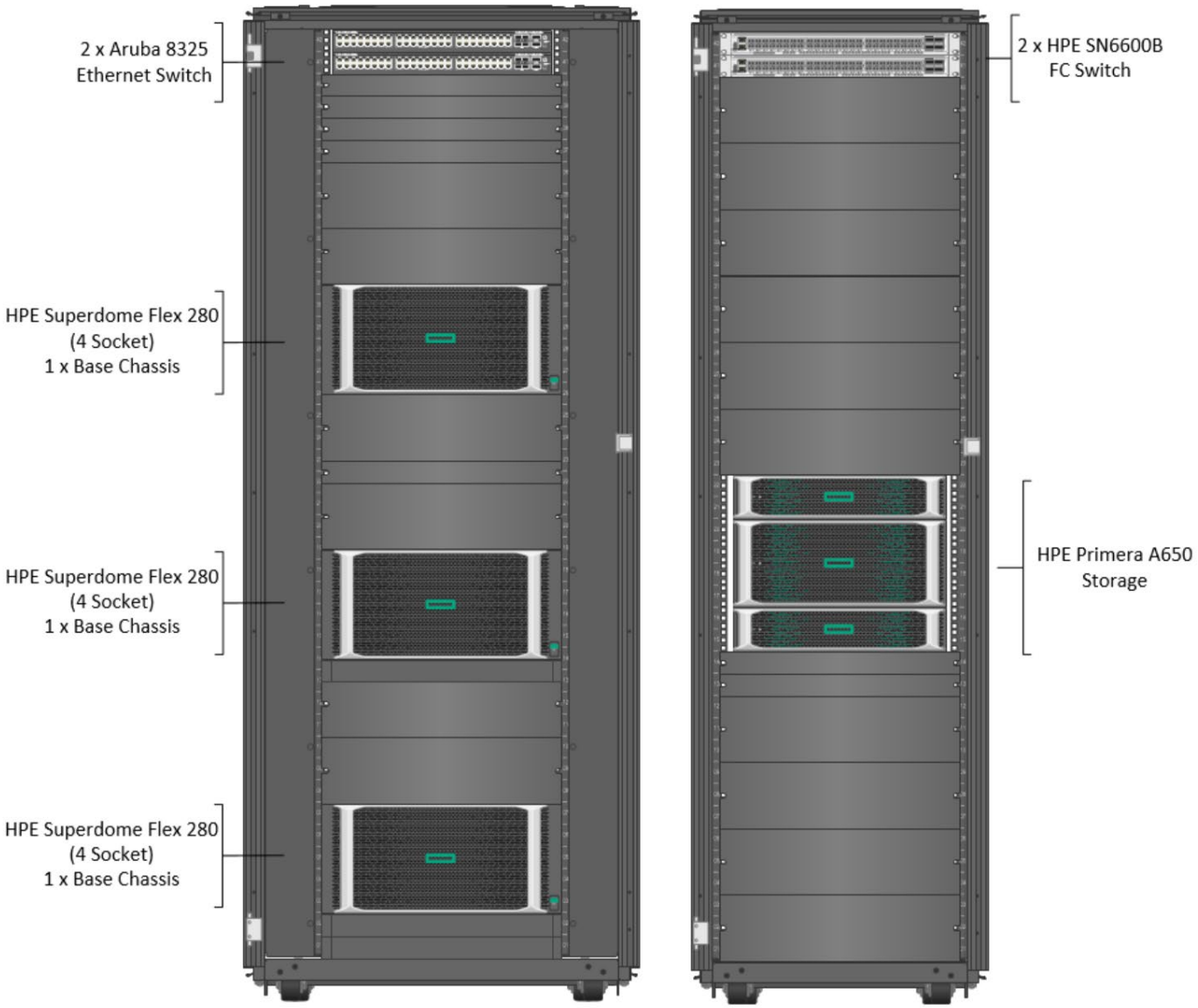

FIGURE 3. Solution hardware diagram

Hardware

Table 1 summarizes hardware components that were utilized in the design and construction of this Reference Architecture.

TABLE 1. Hardware components utilized in this solution

Component	Qty	Description
HPE Superdome Flex 280 server • 1x Base chassis consisting of the following components • 4 x Intel Xeon-Platinum 8380HL (2.9GHz/28-core/250W) Processor.[3] • 48 x HPE 64 GB DDR4 3200 MT/s RDIMM • 4 x HPE 800 GB SAS 12G MU SFF SSD • 4 x HPE SN1610Q 32Gb Fibre Channel Host Bus Adapter • 2 x HPE Ethernet 10Gb 2 Port 562T Adapter	3	4-socket server **NOTE:** These servers were also used for 2-socket testing. Two CPUs were disabled at the eRMC level using the deconfig command.
HPE Primera A650 all-flash storage • 4 Controller nodes • 48 x 1.92 TB SSD • 8 x 32Gb 4 Port FC HBAs	1	Storage for Oracle RAC database, redo logs, etc.
HPE B-series SN6600B Fibre Channel Switch	2	Fibre Channel Switch for providing connectivity between HPE Primera and Servers
Aruba 8325-48Y8C 48p 25G • 48 Ports x 1G/10G/25 GbE • 8 Ports x 40G/100 GbE	2	Layer 2 switch for 10Gb network connectivity

Software

Table 2 shows the software used in this solution configuration.

TABLE 2. Software used in this solution

Component	Version	Description
Red Hat Enterprise Linux	8.3	Operating System
HPE Application Tuner Express	1.0.3	Licensed tool sold by Hewlett Packard Enterprise. This tool optimizes application performance by loading data into memory close to the processor executing the computing instruction.
HPE Foundation Software for RHEL	2.4	This contains the required suite of technical support tools and utilities that enable HPE mission-critical systems to run at improved performance with enhanced technical support.

Application software

Table 3 shows the application software used in this solution configuration.

TABLE 3. Application software used in this solution

Component	Version	Description
Oracle Database	19c	Oracle Database 19.3.0.0.0 Enterprise Edition with the RAC option and the following patches • GRID Patch p32226239 (19.10) • DB SW Patch p32218454 (19.10)

[3] An Intel Xeon Platinum 8380HL processor supports up to 4.5 TB of memory per processor. In this Reference Architecture, only 768 GB memory was installed per processor. For a 768 GB memory per processor configuration, the Intel Xeon Platinum 8380H processor is sufficient and it supports up to 1.12 TB memory per processor.

BEST PRACTICES AND CONFIGURATION GUIDANCE FOR THE SOLUTION

Install HPE Foundation Software

The HPE Foundation Software should be installed on the server after installing the OS. This software consists of packages designed to ensure the smooth operation of the server. It includes Data Collection Daemon (DCD) for Linux, an agentless service that proactively monitors the health of hardware components in the server.

BIOS setting

The workload profile for the Oracle RAC database was set to custom profile, which consisted of Mission Critical workload profile settings, and the **Minimum Processor Idle Power Core C-State** option was set to **C1E** in the BIOS.

Setting the minimum C-state to C1E encourages Turbo Boost clock speed increases and also it requires low wake-up latency time from sleep as compared to higher C-states, such as C6.

Configure kernel boot options

The HPE Foundation Software sets some kernel boot options that are key to optimal performance. The boot options listed in Table 4 were used for optimal performance. Note that numa_balancing=disable and transparent_hugepage=never are both recommended by Oracle. When NUMA balancing is enabled, the kernel migrates the task's pages to the same NUMA node where the task is running. Due to the size of the Oracle SGA and the frequency of task migrations, the migration of pages can be expensive, and optimal performance can be achieved by disabling NUMA balancing.

Transparent huge pages are enabled by default in RHEL, and Oracle recommends disabling this setting to avoid memory allocation delays at runtime. Note that while dynamically-allocated transparent huge pages were disabled, statically-allocated huge pages were configured via the vm.nr_hugepages kernel parameter. Below are the kernel boot options that were used for the testing.

```
# cat /proc/cmdline
BOOT_IMAGE=(hd1,gpt2)/vmlinuz-4.18.0-240.el8.x86_64 root=/dev/mapper/rhel00-root ro resume=/dev/mapper/rhel00-
swap rd.lvm.lv=rhel00/root rd.lvm.lv=rhel00/swap rhgb transparent_hugepage=never console=ttyS0,115200
udev.children-max=32 nmi_watchdog=0 mce=2 uv_nmi.action=kdump bau=0 earlyprintk=ttyS0,115200 log_buf_len=8M
numa_balancing=disable pci=nobar crashkernel=512M,high
```

Table 4 shows the Kernel boot options for optimal performance.

TABLE 4. Kernel boot options for optimal performance

Kernel boot option	Description	How to set
transparent_hugepage=never	Disables transparent huge pages	`hpe-auto-config` command (part of HPE Foundation Software)
numa_balancing=disable	Disables automatic numa balancing	`hpe-auto-config` command (part of HPE Foundation Software)

RHEL OS settings

A complete list of the RHEL tuning parameters is shown in Appendix B. The shared memory parameter and the number of huge pages were set large enough to contain the Oracle SGA (buffer cache).

HPE Primera A650 all-flash array volumes

The HPE Primera A650 all-flash array was configured with the volumes listed in Table 5. Sixteen (16) volumes were configured in an Oracle ASM disk group named DATA, which contained the Oracle RAC database tablespaces, indexes, undo tablespace and temp tablespace. Sixteen (16) volumes were used for the ASM disk group REDOA, and sixteen more were consumed for REDOB, which contained the Oracle redo threads. In an Oracle RAC database, each instance requires its own set of redo log groups, which is known as a redo thread.

Table 5 shows the HPE Primera volumes.

TABLE 5. HPE Primera volumes

Quantity	RAID level, type	Description
1 x 200 GB	RAID6, thin	Oracle binaries
1 x 50 GB	RAID6, thin	Oracle ASM MGMT disk group
16 x 512 GB	RAID6, thin	Oracle ASM DATA disk group
16 x 128 GB	RAID6, thin	Oracle ASM REDOA disk group
16 x 128 GB	RAID6, thin	Oracle ASM REDOB disk group

Oracle configuration

A complete list of the Oracle parameters set for various tests is provided in Appendix D. The Oracle SGA was set large enough to minimize physical reads. For the OLTP test, the SGA was set to 990 GB.

Two redo log files of 650 GB each for every Oracle RAC node were configured to minimize the log file switching during the performance tests. Customer implementations should determine the log file size required to meet their business needs. An undo tablespace of 400 GB for each Oracle RAC instance was created to minimize overhead due to filling up the tablespace during a benchmark run. A temp tablespace of 600 GB was created.

HPE-ATX was used to evenly distribute the Oracle processes across all nodes in the server (see Appendix G for ATX configuration script).

CAPACITY AND SIZING

Workload description

Oracle performance tests were conducted using HammerDB, an open-source tool. For this Reference Architecture, HammerDB 3.3 was used to implement the online transaction processing (OLTP) workload. For the OLTP workload, HammerDB provides a real-world type scenario that consumes both CPU for the application logic and I/O. The HammerDB tool implements an OLTP-type workload with small I/O sizes of a random nature. The transaction results were normalized and used to compare test configurations. Other metrics collected during the testing came from the operating system and/or standard Oracle Automatic Workload Repository (AWR) statistics reports.

The OLTP test performed was highly CPU intensive and moderately I/O intensive. The environment was tuned for maximum user transactions. After the database was tuned, the transaction rates were recorded at various Oracle connection counts. Because customer workloads vary in characteristics, the measurement was made with a focus on maximum transactions.

HPE-ATX was used in all the test runs to achieve maximum performance when running on multi-socket servers. Refer to Appendix G for the ATX configuration script.

Analysis and recommendations for Oracle RAC scalability

Testing was conducted with one, two, and three Oracle RAC node configurations and all of these runs were on the HPE Superdome Flex 280 servers with two (2) socket and four (4) socket configurations separately.

2-socket configuration - Oracle RAC scalability

Figure 4 shows relative throughput (measured in transactions per minute) for Oracle RAC with one, two, and three nodes. All results are normalized to the single RAC node data for each Oracle connection count. When adding RAC nodes, throughput increased linearly, with two nodes providing up to 1.86 times and three nodes offering up to 2.62 times the number of transactions as the single RAC node instance.

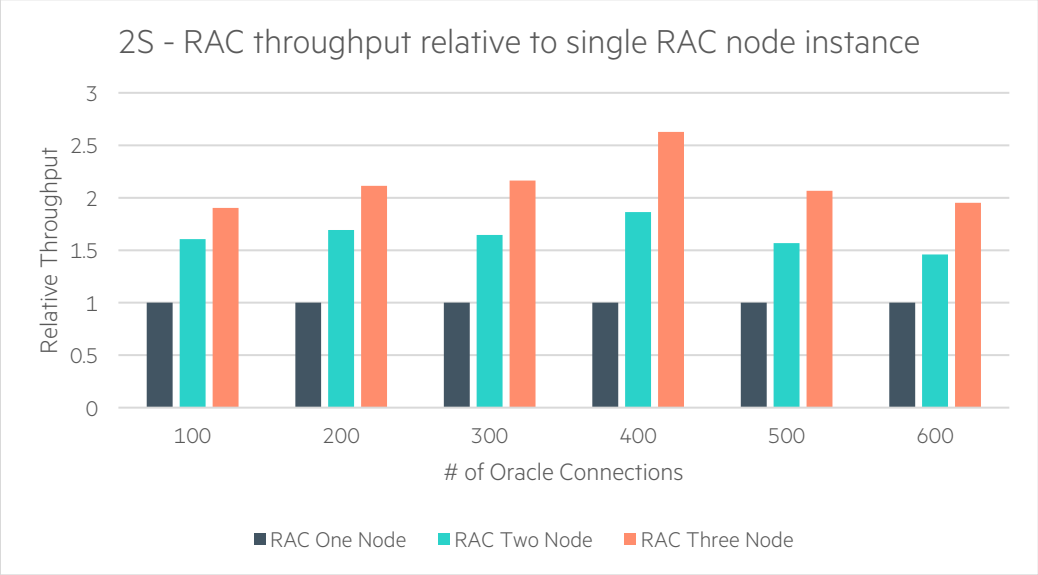

FIGURE 4. 2-socket – Oracle RAC scalability

Figure 5 shows the CPU utilization for each configuration as the number of Oracle connections is increased. With a single RAC node instance, utilization reached 99% CPU usage. As RAC nodes were added, the utilization decreased because the load is shared across Oracle RAC nodes and more time is spent on coordinating the activities between the nodes in the cluster.

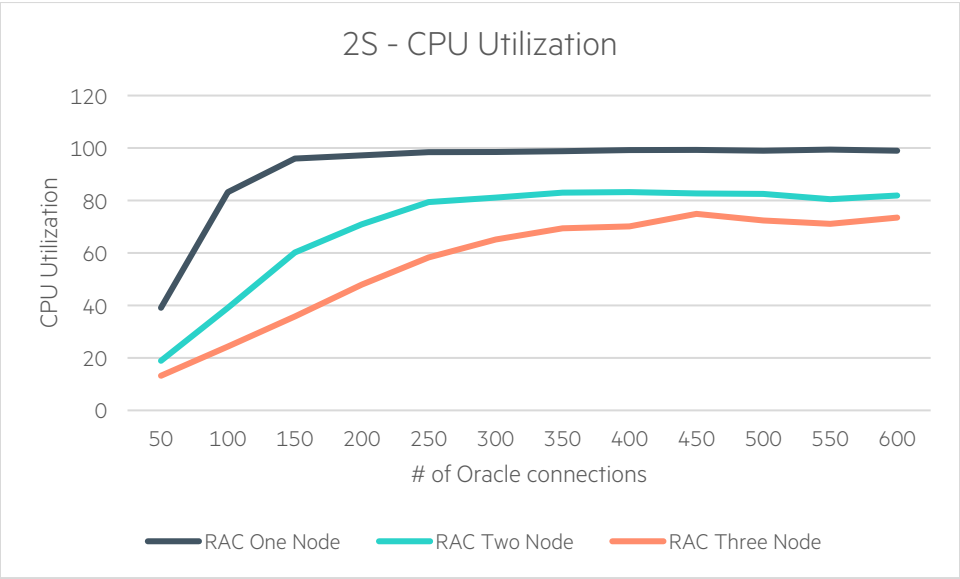

FIGURE 5. 2-socket – Oracle RAC CPU utilization

4-socket configuration - Oracle RAC scalability

Figure 6 shows the relative throughput (measured in transactions per minute) for Oracle RAC with one, two, and three nodes. As compared to the 2-socket configuration, each user was able to achieve more transactions per minute in the test run. All results are normalized to the single RAC node data for each Oracle connection count. When adding RAC nodes, throughput increased linearly, with two nodes providing up to 1.84 times and three nodes offering up to 2.44 times the number of transactions as the single RAC node instance.

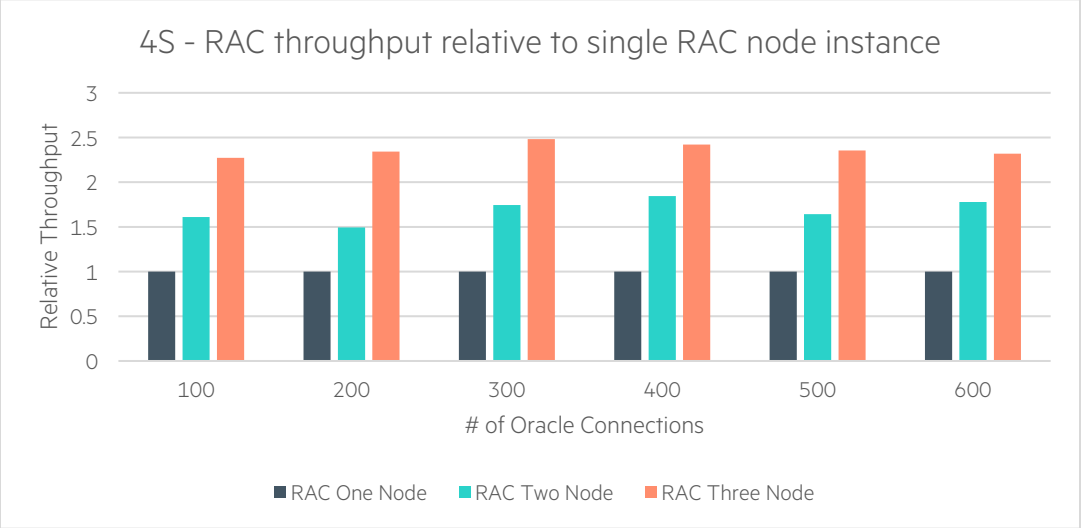

FIGURE 6. 4-socket – Oracle RAC scalability

Figure 7 shows that CPU utilization for each configuration as the number of Oracle connections increased. With a single RAC node instance, utilization reached 98%. As RAC nodes were added, the utilization decreased because the load is shared across Oracle RAC nodes and more time is spent on coordinating the activities between the nodes in the cluster. With the three-node configuration, CPU utilization reached around 65% for 600 users. It means, adding more RAC nodes helps to serve a higher number of users.

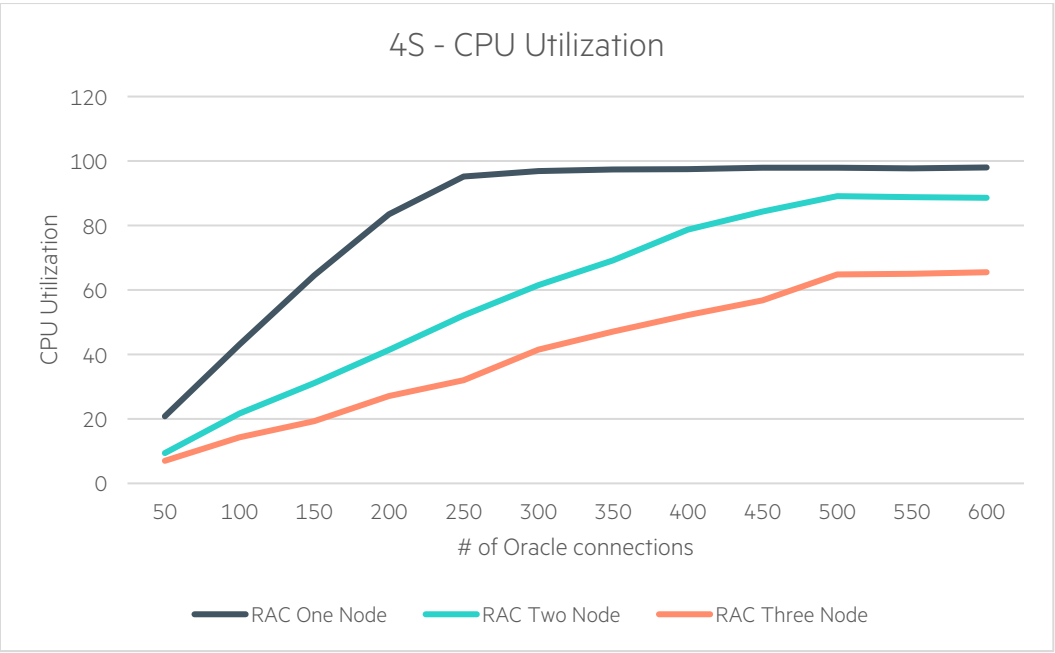

FIGURE 7. 4-socket – Oracle RAC CPU utilization

2-socket versus 4-socket

One of the benefits of the HPE Superdome Flex 280 solution is the ability to easily scale up to a server with more processing power when more capacity is required. The final test demonstrates, how Oracle throughput can be increased by moving from a 2-socket HPE Superdome Flex 280 server to a 4-socket HPE Superdome Flex 280 server when the CPUs are on the 2-socket system are fully utilized, and more capacity is required. For this test, the number of cores per processor in the HPE Superdome Flex 280 server was reduced to four, to demonstrate a scenario where processing power was the bottleneck. Figure 8 shows that Oracle throughput can be increased by a factor of 1.81 by moving the workload from the 2-socket to the 4-socket server.

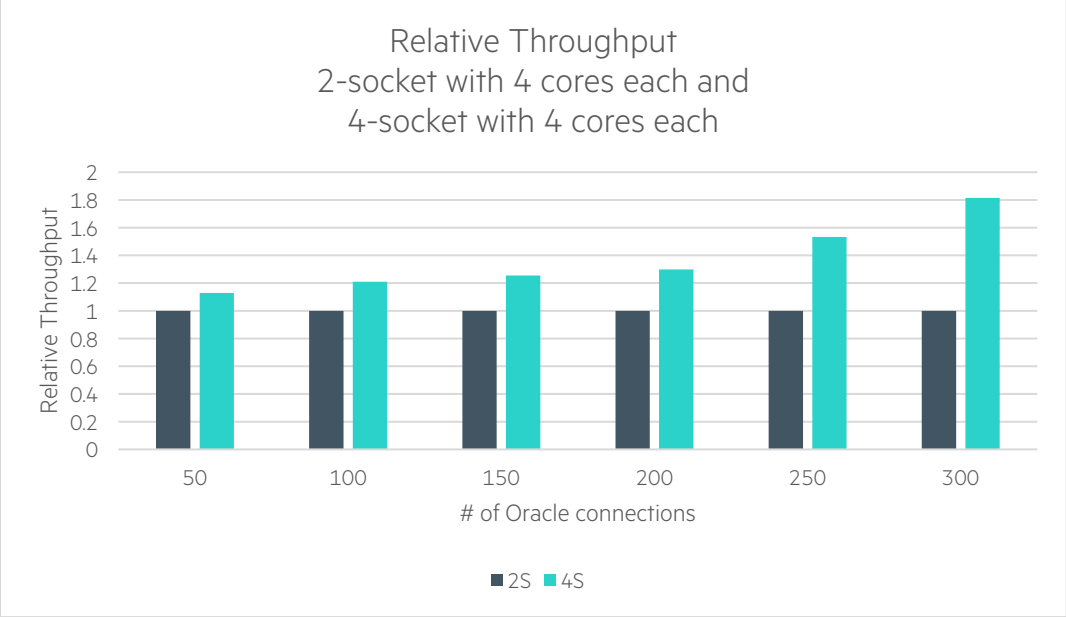

FIGURE 8. 2-socket vs 4-socket with 4 cores in each socket – Relative OLTP throughput

SUMMARY

The HPE Superdome Flex 280 server is designed to power data-intensive applications and provides support for 2 to 8 sockets with Intel Xeon Scalable processors in a single system. With the 3rd generation Intel Xeon Scalable processors, the HPE Superdome Flex 280 server provides significant OLTP performance improvement and the HPE Application Tuner Express allows customers to obtain optimal performance as workloads scale up, taking full advantage of HPE Superdome Flex 280 server processing and memory capacities.

IMPLEMENTING A PROOF-OF-CONCEPT

As a matter of best practice for all deployments, Hewlett Packard Enterprise recommends implementing a proof-of-concept using a test environment that matches as closely as possible the planned production environment. In this way, appropriate performance and scalability characterizations can be obtained. For help with a proof-of-concept, contact a Hewlett Packard Enterprise services representative (hpe.com/us/en/services/consulting.html) or your Hewlett Packard Enterprise partner.

APPENDIX A: BILL OF MATERIALS

NOTE

Part numbers are at the time of publication/testing and subject to change. The bill of materials does not include complete support options or other rack and power requirements. If you have questions regarding ordering, please consult with your Hewlett Packard Enterprise reseller or Hewlett Packard Enterprise sales representative for more details. hpe.com/us/en/services/consulting.html.

TABLE A1. Bill of materials – Server Infrastructure

Part number	Quantity	Description
		Rack and Switches
P9K16A	1	HPE 42U 800x1200mm Advanced Shock Rack
P9K16A 001	1	HPE Factory Express Base Racking Service
R9F64A	2	Aruba 8325-48Y8C 48p 25G SFP/+/28 8p 100G QSFP+/28 Front-to-Back 6 Fans and 2 PSU Bundle
R9F59A	2	Aruba X474 4-post Rack Kit
		HPE Superdome Flex 280 (Bill of material for One (1) 4-socket Server)
R4R03A	1	HPE Superdome Flex 280 4-socket Base Chassis
R4R14A.[4]	4	Intel Xeon-Platinum 8380HL (2.9GHz/28-core/250W) Processor Kit for HPE Superdome Flex 280
R4S37A	1	HPE Superdome Flex 280 PCIe Low Profile 16-slot Bulkhead with 1x 4-slot Riser Kit
R4S12A	1	HPE Superdome Flex 280 4-slot 2x8/2x16 PCIe Left Riser
R4S39A	2	HPE Superdome Flex 280 4-slot 2x8/2x16 PCIe Right Riser
R4S34A	1	HPE Superdome Flex 280 9.5mm SATA Internal DVD-RW Optical Drive
R4R08A	1	HPE Superdome Flex 280 2-4 Sockets UPI Internal Enablement Kit
R4S14A	1	HPE Superdome Flex 280 2x 1600W Platinum Hot Plug Power Supply
R4S15A	1	HPE Superdome Flex 280 2x 1600W Platinum Hot Plug Additional Power Supply
R4S22A	1	HPE Superdome Flex 280 8SFF Premium Storage Backplane Kit
R4R47A	1	HPE Superdome Flex 280 SmartRAID 3154-16i 16-port Internal 4GB Cache SAS 12G PCIe3 x8 Controller
R4R43A	2	HPE SLIM SAS 8i PLUG TO MINI SAS 4i PLUG CABLE ASSY (320mm)
R4R44A	2	HPE SLIM SAS 8i PLUG TO MINI SAS 4i PLUG CABLE ASSY (490mm)
R4S27A	48	HPE SD FLEX 280 64 GB 2RX4 DDR4-3200R
R4S18A	2	HPE Superdome Flex 280 2x C14 250V 10Amp Power Cord
R4R07A	1	HPE Superdome Flex 280 Chassis Bezel
R4S20A	1	HPE Superdome Flex 280 Chassis Intrusion Detection Kit
817738-B21	2	HPE Ethernet 10Gb 2-port BASE-T X550-AT2 Adapter
R2E09A	4	HPE SN1610Q 32Gb 2p FC HBA
R6A27A	4	HPE 800 GB SAS 12G MU SFF BC SSD

TABLE A2. Bill of materials – Storage Infrastructure

Part number	Quantity	Description
		Rack and Fibre Channel Switches
P9K38A	1	HPE 42U 600mmx1075mm G2 Enterprise Shock Rack
P9K38A-001	1	HPE Factory Express Base Racking Service
Q0U54B	2	HPE SN6600B 32Gb 48/24 Fibre Channel Switch
P9H32A	48	HPE B-series 32Gb SFP28 Short Wave 1-pack Transceiver
		HPE Primera A650 Storage
N9Z47A	1	HPE Primera 600 4-way Storage Base
N9Z61A	1	HPE Primera A650 4-node Controller
R0P95A	24	HPE Primera 600 1.92 TB SAS SFF (2.5in) SSD
N9Z39A	8	HPE Primera 600 32Gb 4-port Fibre Channel Host Bus Adapter
716195-B21	8	HPE External 1.0m (3ft) Mini-SAS HD 4x to Mini-SAS HD 4x Cable
N9Z50A	2	HPE Primera 600 2U 24-disk SFF Drive Enclosure
R0P95A	24	HPE Primera 600 1.92TB SAS SFF (2.5in) SSD
QK734A	16	HPE Premier Flex LC/LC Multi-mode OM4 2 fiber 5m Cable
QK735A	48	HPE Premier Flex LC/LC Multi-mode OM4 2 fiber 15m Cable

APPENDIX B: RHEL KERNEL SETTINGS

The following RHEL kernel parameters were set in the `/etc/sysctl.conf` file. The nr_hugepages were set large enough to accommodate the memory requirements of the OLTP workload.

```
kernel.sem = 250 64000 100 128
kernel.shmall = 4294967295
kernel.shmmax = 12094627905536
fs.file-max = 6815744
kernel.shmmni = 16384
fs.aio-max-nr = 3145728
net.ipv4.ip_local_port_range = 9000 65500
net.core.rmem_default = 1048576
net.core.wmem_default = 1048576
net.core.rmem_max=26214400
net.core.wmem_max=26214400
net.ipv4.tcp_rmem = 1048576 1048576 4194304
net.ipv4.tcp_wmem = 1048576 1048576 1048576
vm.nr_hugepages = 665536
vm.hugetlb_shm_group = 54321
net.ipv4.conf.ens2096f1.rp_filter = 2
net.ipv4.conf.ens2096f0.rp_filter = 2
net.ipv4.conf.bond0.rp_filter = 1
```

APPENDIX C: ORACLE USER ACCOUNT LIMITS

The following settings were included in the file `/etc/security/limits.d/oracle-limits.conf`.

```
oracle   soft   nofile    1024
oracle   hard   nofile    65536
oracle   soft   nproc    16384
oracle   hard   nproc    16384
oracle   soft   stack    10240
```

```
oracle   hard   stack    32768
oracle   hard   memlock    1158217728
oracle   soft   memlock    1158217728
```

APPENDIX D: ORACLE INITIALIZATION PARAMETERS

The following Oracle parameters were set in the init.ora initialization file.

```
racdb2.__data_transfer_cache_size=0
racdb3.__data_transfer_cache_size=0
racdb1.__data_transfer_cache_size=0
racdb2.__db_cache_size=943819063296
racdb3.__db_cache_size=943819063296
racdb1.__db_cache_size=943819063296
racdb2.__inmemory_ext_roarea=0
racdb3.__inmemory_ext_roarea=0
racdb1.__inmemory_ext_roarea=0
racdb2.__inmemory_ext_rwarea=0
racdb3.__inmemory_ext_rwarea=0
racdb1.__inmemory_ext_rwarea=0
racdb2.__java_pool_size=0
racdb3.__java_pool_size=0
racdb1.__java_pool_size=0
racdb2.__large_pool_size=2684354560
racdb3.__large_pool_size=2684354560
racdb1.__large_pool_size=2684354560
racdb1.__oracle_base='/u01/app/oracle'#ORACLE_BASE set from environment
racdb2.__oracle_base='/u01/app/oracle'#ORACLE_BASE set from environment
racdb3.__oracle_base='/u01/app/oracle'#ORACLE_BASE set from environment
racdb2.__pga_aggregate_target=354871672832
racdb3.__pga_aggregate_target=354871672832
racdb1.__pga_aggregate_target=354871672832
racdb2.__sga_target=1064615018496
racdb3.__sga_target=1064615018496
racdb1.__sga_target=1064615018496
racdb2.__shared_io_pool_size=536870912
racdb3.__shared_io_pool_size=536870912
racdb1.__shared_io_pool_size=536870912
racdb2.__shared_pool_size=117037858816
racdb3.__shared_pool_size=117037858816
racdb1.__shared_pool_size=117037858816
racdb2.__streams_pool_size=0
racdb3.__streams_pool_size=0
racdb1.__streams_pool_size=0
racdb2.__unified_pga_pool_size=0
racdb3.__unified_pga_pool_size=0
racdb1.__unified_pga_pool_size=0
*._enable_NUMA_support=TRUE
*._fast_cursor_reexecute=TRUE
*._high_priority_processes='VKTM*|LG*'
*._trace_pool_size=0
*._undo_autotune=FALSE
*.audit_file_dest='/u01/app/oracle/admin/racdb/adump'
*.audit_trail='NONE'
*.cluster_database=true
*.commit_logging='BATCH'
*.commit_wait='NOWAIT'
*.compatible='19.0.0'
*.control_files='+DATA/RACDB/CONTROLFILE/current.263.1070025357'
```

```
*.db_block_size=8192
*.db_create_file_dest='+DATA'
*.db_domain='orainfra.local'
*.db_name='racdb'
*.diagnostic_dest='/u01/app/oracle'
*.dispatchers='(PROTOCOL=TCP) (SERVICE=racdbXDB)'
family:dw_helper.instance_mode='read-only'
racdb2.instance_number=2
racdb1.instance_number=1
racdb3.instance_number=3
*.local_listener='-oraagent-dummy-'
*.lock_sga=TRUE
*.nls_language='AMERICAN'
*.nls_territory='AMERICA'
*.open_cursors=3000
*.pga_aggregate_target=338419m
*.pre_page_sga=FALSE
*.processes=8960
*.remote_login_passwordfile='exclusive'
*.sga_target=1015257m
racdb3.thread=3
racdb2.thread=2
racdb1.thread=1
*.trace_enabled=FALSE
racdb1.undo_tablespace='UNDOTBS1'
racdb2.undo_tablespace='UNDOTBS2'
racdb3.undo_tablespace='UNDOTBS3'
*.use_large_pages='ONLY'
```

APPENDIX E: MULTIPATH.CONF

The following entries were included in the `/etc/multipath.conf` file. Aliases were created for each HPE Primera volume, for usage in the udev rules file, dm-permission.rules, described in Appendix F. Only two of the alias entries are shown here, because of the large number of volumes that were used for the testing.

```
defaults {
        polling_interval 10
        max_fds 8192
        user_friendly_names yes
        }
devices {
        device {
        vendor "3PARdata"
        product "VV"
        path_grouping_policy "group_by_prio"
        path_selector "round-robin 0"
        path_checker tur
        features "0"
        hardware_handler "1 alua"
        prio "alua"
        failback immediate
        rr_weight "uniform"
        no_path_retry 18
        rr_min_io_rq 1
        detect_prio yes
        fast_io_fail_tmo 10
        dev_loss_tmo "infinity"
        }
}
```

```
multipaths {
multipath {
        wwid 360002ac0000000000000005c0007ea28
        alias data01
        }
multipath {
        wwid 360002ac0000000000000005d0007ea28
        alias data02
        }
# All the entries are not shown here for brevity
}
```

APPENDIX F: UDEV RULES

One udev rules file was created to set parameters for the HPE Primera volumes, and a second file was created to set the ownership and permissions for the Oracle volumes. The file `/etc/udev/rules.d/10-primera.rules` set the rotational latency, I/O scheduler, rq_affinity, nomerges and nr_requests parameters. For SSDs, the rotational latency is set to zero. The I/O scheduler was set to none. Setting rq_affinity to 2, forces block I/O completion requests to complete on the requesting CPU. The file contained the following settings:

```
# cat /etc/udev/rules.d/10-primera.rules
ACTION=="add|change", KERNEL=="dm-*", PROGRAM="/bin/bash -c 'cat /sys/block/$name/slaves/*/device/vendor | grep
3PARdata'", ATTR{queue/rotational}="0", ATTR{queue/scheduler}="none", ATTR{queue/rq_affinity}="2",
ATTR{queue/nomerges}="1", ATTR{queue/nr_requests}="128"
```

The `/etc/udev/rules.d/99-dm-permission.rules` file included the following settings for the HPE Primera volumes used for the Oracle ASM disk groups:

```
$ cat /etc/udev/rules.d/99-dm-permission.rules
ENV{DM_NAME}=="data-1", OWNER:="oracle", GROUP:="oinstall", MODE:="660"
ENV{DM_NAME}=="data-2", OWNER:="oracle", GROUP:="oinstall", MODE:="660"
ENV{DM_NAME}=="data-3", OWNER:="oracle", GROUP:="oinstall", MODE:="660"
ENV{DM_NAME}=="data-4", OWNER:="oracle", GROUP:="oinstall", MODE:="660"
ENV{DM_NAME}=="data-5", OWNER:="oracle", GROUP:="oinstall", MODE:="660"
ENV{DM_NAME}=="data-6", OWNER:="oracle", GROUP:="oinstall", MODE:="660"
ENV{DM_NAME}=="data-7", OWNER:="oracle", GROUP:="oinstall", MODE:="660"
ENV{DM_NAME}=="data-8", OWNER:="oracle", GROUP:="oinstall", MODE:="660"
ENV{DM_NAME}=="redoa-1", OWNER:="oracle", GROUP:="oinstall", MODE:="660"
ENV{DM_NAME}=="redoa-2", OWNER:="oracle", GROUP:="oinstall", MODE:="660"
ENV{DM_NAME}=="redoa-3", OWNER:="oracle", GROUP:="oinstall", MODE:="660"
ENV{DM_NAME}=="redoa-4", OWNER:="oracle", GROUP:="oinstall", MODE:="660"
ENV{DM_NAME}=="redoa-5", OWNER:="oracle", GROUP:="oinstall", MODE:="660"
ENV{DM_NAME}=="redoa-6", OWNER:="oracle", GROUP:="oinstall", MODE:="660"
ENV{DM_NAME}=="redoa-7", OWNER:="oracle", GROUP:="oinstall", MODE:="660"
ENV{DM_NAME}=="redoa-8", OWNER:="oracle", GROUP:="oinstall", MODE:="660"
ENV{DM_NAME}=="redoa-9", OWNER:="oracle", GROUP:="oinstall", MODE:="660"
ENV{DM_NAME}=="redoa-10", OWNER:="oracle", GROUP:="oinstall", MODE:="660"
ENV{DM_NAME}=="redoa-11", OWNER:="oracle", GROUP:="oinstall", MODE:="660"
ENV{DM_NAME}=="redoa-12", OWNER:="oracle", GROUP:="oinstall", MODE:="660"
ENV{DM_NAME}=="redoa-13", OWNER:="oracle", GROUP:="oinstall", MODE:="660"
ENV{DM_NAME}=="redoa-14", OWNER:="oracle", GROUP:="oinstall", MODE:="660"
ENV{DM_NAME}=="redoa-15", OWNER:="oracle", GROUP:="oinstall", MODE:="660"
ENV{DM_NAME}=="redoa-16", OWNER:="oracle", GROUP:="oinstall", MODE:="660"
ENV{DM_NAME}=="mgmt", OWNER:="oracle", GROUP:="oinstall", MODE:="660"
ENV{DM_NAME}=="data-9", OWNER:="oracle", GROUP:="oinstall", MODE:="660"
ENV{DM_NAME}=="data-10", OWNER:="oracle", GROUP:="oinstall", MODE:="660"
ENV{DM_NAME}=="data-11", OWNER:="oracle", GROUP:="oinstall", MODE:="660"
ENV{DM_NAME}=="data-12", OWNER:="oracle", GROUP:="oinstall", MODE:="660"
ENV{DM_NAME}=="data-13", OWNER:="oracle", GROUP:="oinstall", MODE:="660"
```

```
ENV{DM_NAME}=="data-14", OWNER:="oracle", GROUP:="oinstall", MODE:="660"
ENV{DM_NAME}=="data-15", OWNER:="oracle", GROUP:="oinstall", MODE:="660"
ENV{DM_NAME}=="data-16", OWNER:="oracle", GROUP:="oinstall", MODE:="660"
ENV{DM_NAME}=="redob-1", OWNER:="oracle", GROUP:="oinstall", MODE:="660"
ENV{DM_NAME}=="redob-2", OWNER:="oracle", GROUP:="oinstall", MODE:="660"
ENV{DM_NAME}=="redob-3", OWNER:="oracle", GROUP:="oinstall", MODE:="660"
ENV{DM_NAME}=="redob-4", OWNER:="oracle", GROUP:="oinstall", MODE:="660"
ENV{DM_NAME}=="redob-5", OWNER:="oracle", GROUP:="oinstall", MODE:="660"
ENV{DM_NAME}=="redob-6", OWNER:="oracle", GROUP:="oinstall", MODE:="660"
ENV{DM_NAME}=="redob-7", OWNER:="oracle", GROUP:="oinstall", MODE:="660"
ENV{DM_NAME}=="redob-8", OWNER:="oracle", GROUP:="oinstall", MODE:="660"
ENV{DM_NAME}=="redob-9", OWNER:="oracle", GROUP:="oinstall", MODE:="660"
ENV{DM_NAME}=="redob-10", OWNER:="oracle", GROUP:="oinstall", MODE:="660"
ENV{DM_NAME}=="redob-11", OWNER:="oracle", GROUP:="oinstall", MODE:="660"
ENV{DM_NAME}=="redob-12", OWNER:="oracle", GROUP:="oinstall", MODE:="660"
ENV{DM_NAME}=="redob-13", OWNER:="oracle", GROUP:="oinstall", MODE:="660"
ENV{DM_NAME}=="redob-14", OWNER:="oracle", GROUP:="oinstall", MODE:="660"
ENV{DM_NAME}=="redob-15", OWNER:="oracle", GROUP:="oinstall", MODE:="660"
ENV{DM_NAME}=="redob-16", OWNER:="oracle", GROUP:="oinstall", MODE:="660"
```

APPENDIX G: HPE-ATX CONFIGURATION SCRIPT

The following script was used to start the Oracle listener processes under ATX. The Round Robin policy was used to evenly distribute the processes across the NUMA nodes in the system. The listener should be started using the script below in each Oracle RAC node.

```
#!/bin/bash
lsnrctl stop
export ORACLE_HOME=/u01/app/19.0.0/grid;
/usr/bin/hpe-atx -p rr_flat -l rac_atx_listener.log /u01/app/19.0.0/grid/bin/lsnrctl start
```

RESOURCES AND ADDITIONAL LINKS

HPE Reference Architectures, hpe.com/info/ra

HPE Servers, hpe.com/servers

HPE Storage, hpe.com/storage

HPE Networking, hpe.com/networking

HPE Advisory and Professional Services, hpe.com/us/en/services/consulting.html

Oracle Solutions, https://hpe.com/info/oracle

HPE Superdome Flex 280 server QuickSpecs, https://www.hpe.com/psnow/doc/a00059763enw

HPE Superdome Flex 280 server - Product documentation, https://www.hpe.com/psnow/product-documentation?oid=1012865453&cc=us&lc=en

HPE Superdome Flex 280 server Architecture and RAS, https://www.hpe.com/psnow/doc/a50003250enw

HPE Primera, https://www.hpe.com/us/en/storage/hpe-primera.html

HPE Primera Red Hat Enterprise Linux Implementation Guide, https://support.hpe.com/hpesc/public/docDisplay?docLocale=en_US&docId=emr_na-a00088902en_us

Comparing Oracle IO workloads for OLTP and DSS with HPE Primera and 3PAR Storage, https://community.hpe.com/t5/around-the-storage-block/comparing-oracle-io-workloads-for-oltp-and-dss-with-hpe-primera/ba-p/7067671#.X3renuj7TE5

HPE Application Tuner Express, https://myenterpriselicense.hpe.com/cwp-ui/evaluation/HPE-ATX

HPE Serviceguard for Linux Quick Specs, https://support.hpe.com/hpesc/public/docDisplay?docLocale=en_US&docId=c04154488

HPE Serviceguard Toolkit for Oracle on Linux User Guide, https://support.hpe.com/hpsc/doc/public/display?docLocale=en_US&docId=emr_na-a00052274en_us&withFrame

HPE Serviceguard Toolkit for Oracle Data Guard User Guide, https://support.hpe.com/hpsc/doc/public/display?docLocale=en_US&docId=emr_na-a00052284en_us&withFrame

To help us improve our documents, please provide feedback at hpe.com/contact/feedback.

a50004699enw, July 2024

Hewlett Packard Enterprise