

Deploying Oracle 11gR2 Enterprise Edition Real Application Cluster with Oracle Enterprise Linux 5.4 on Hitachi Compute Blade 2000 and Hitachi Adaptable Modular Storage 2500

Reference Architecture Guide

By Anantha Adiga

September 2011

Feedback

Hitachi Data Systems welcomes your feedback. Please share your thoughts by sending an email message to SolutionLab@hds.com. Be sure to include the title of this white paper in your email message.

Table of Contents

Solution Overview	4
Key Solution Components.....	5
Hitachi Compute Blade 2000.....	6
Hitachi Adaptable Modular Storage 2500	6
Hitachi Dynamic Provisioning.....	7
Hitachi Dynamic Link Manager Advanced	8
Hitachi Storage Navigator Modular 2	8
Oracle Enterprise Linux OS.....	8
Oracle Database11g Release R2	9
Fusion-io Card PCIe SSD Card	10
Brocade Storage Area Network Switches	10
Solution Design	10
High-level Infrastructure	10
Storage Architecture.....	11
Server and Application Architecture	14
SAN Architecture	15
Physical and Virtual Network Cluster Interconnect Architecture.....	17
Engineering Validation.....	19
Workload	19
Methodology	19
Data Gathering	20
Results Summay	20
Conclusion	27

Deploying Oracle 11gR2 Enterprise Edition Real Application Cluster with Oracle Enterprise Linux 5.4 on Hitachi Compute Blade 2000 and Hitachi Adaptable Modular Storage 2500

Reference Architecture Guide

Customers need an integrated solution so their Oracle infrastructure does the following:

- Accelerate the time to market and return on investment
- Reduce the need for expensive system integrators
- Simplify support
- Reduce the overall cost of the solution

To meet this need, this reference architecture from Hitachi Data Systems includes servers, storage devices, and storage software validated for a single instance configuration and a real application cluster (RAC) configuration. This integrated solution combines Hitachi hardware and software with third-party products from software, hardware, and networking vendors for reliability, high availability, scalability, and performance. Use this reference as one resource to build a flexible infrastructure that meets your infrastructure requirements and budget.

This reference architecture is for database administrators, storage administrators, and IT personnel responsible for planning and deploying Oracle 11g RAC solutions. It assumes familiarity with Hitachi Adaptable Modular Storage 2500, storage area networks, Oracle 11gR2 RAC Database, Oracle ASM, and Smart Flash Cache.

Solution Overview

While the test configuration for the reference architecture for Oracle 11gR2 shows this converged solution's suitability for OLTP workload database servers, this delivers a cost-effective, high-performance, and extensible platform that can be tailored to meet your specific needs.

The Hitachi Data Systems test configuration consisted of the following infrastructure:

- The compute infrastructure used two Hitachi Compute Blade 2000 server blades.
 - Blade 1 and Blade 2 were Hitachi Compute Blade 2000 E57A1 server blades. Each server blade had 16 processor cores and 256GB RAM.
 - PCI SSD cache had over 300GB per card. The SSD on the PCI card, 320MLC, serves as a second level cache for the Oracle buffer for each instance by enabling Smart Flash Cache in Oracle database server.
- The storage infrastructure used a Hitachi Adaptable Modular Storage 2500 to provide the following:
 - A robust, scalable architecture hosted in Oracle Server environments to store, manage, and access large and growing amounts of information.
 - Wide striping and thin provisioning using Hitachi Dynamic Provisioning to minimize storage footprint and to maximize database performance.
 - Storage Navigator Module 2 for storage administration.
 - Storage software plug-ins that have been integrated into Oracle Enterprise Manager to monitor Hitachi storage. Hitachi Data Systems developed these software plug-ins for the Oracle Enterprise Manager.
- The SAN infrastructure consisted of redundant paths from storage to the servers.
- The network infrastructure consisted of multiple private cluster interconnects to provide a robust and highly available cluster infrastructure for the two node Oracle RAC configuration.

Figure 1 shows the infrastructure used to host the Oracle 11gR2 RAC database server environment.

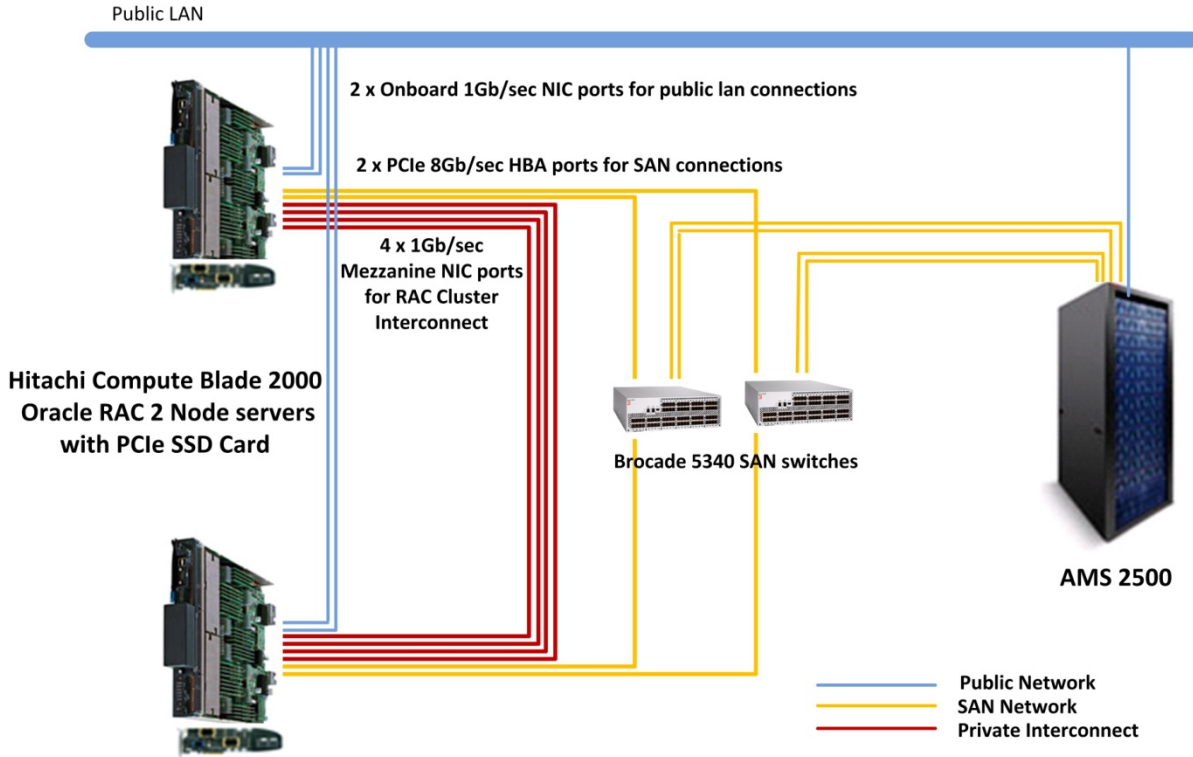


Figure 1

Key Solution Components

Table 1 and Table 2 lists the components used in this reference architecture.

Table 1. Reference Architecture Hardware Components

Component	Description	Version	Qty
Database Server	Hitachi Compute Blade 2000 E57 <ul style="list-style-type: none"> Two Intel Xeon X7560 at 2.27GHz, 16 Cores, 256GB RAM (8x32GB) One Emulex 8Gb/sec dual port Mezzanine HBA 	EFI BIOS Version 4.6.3 8.2.0.86	2
Management Servers	Oracle Grid Control Management Server	N/A	1
	Hitachi Command Suite Server	N/A	1
Cache Memory	PCIe SSD cache <ul style="list-style-type: none"> 320GB per card 	Version 5.0.7 Rev. 101971	2

<i>Component</i>	<i>Description</i>	<i>Version</i>	<i>Qty</i>
Storage Array	Hitachi Adaptable Modular Storage 2500 <ul style="list-style-type: none"> • 300GB SAS 15k • 600GB SAS(SFF) 10k • 32GB cache • 2x8 Host Port Fibre Channel Controllers 	08B5TA1H	1
SAN Switches	Brocade 5340 8Gb/sec <ul style="list-style-type: none"> • 48 to 80 ports 	V6.04b	2

Table 2. Reference Architecture Software Components

<i>Component</i>	<i>Description</i>	<i>Version</i>
Operating System	Oracle Enterprise Linux	5 Update 4
Database Software	Oracle	11gR2 11.2.0.2
Database Management Software	Oracle Enterprise Grid Control Manager	11gR2 11.2.0.2
Storage Management Software	Hitachi Command Suite	7
	Hitachi Storage Navigator Module 2	11.5

Hitachi Compute Blade 2000

Hitachi Compute Blade 2000 is an enterprise-class blade server platform. It features the following:

- Balanced system architecture that eliminates bottlenecks in performance and throughput
- Configuration flexibility
- Eco-friendly power-saving capabilities
- Fast server failure recovery using a N+1 cold standby design that allows replacing failed servers within minutes

Use virtualization on Hitachi Compute Blade 2000 to consolidate application and database servers for backbone systems. Removing performance and I/O bottlenecks opens opportunities for increasing efficiency and utilization rates. Also, it reduces the administrative burden in your data center.

Hitachi Compute Blade 2000 [supports Oracle 11gR2 RAC](#).

For more information, see [Hitachi Compute Blade Family](#) on the Hitachi Data Systems website.

Hitachi Adaptable Modular Storage 2500

Hitachi Adaptable Modular Storage 2500, a member of the Hitachi Adaptable Modular Storage 2000 family, provides a reliable, flexible, scalable, and cost-effective modular storage system. Its symmetric active-active controllers provide integrated, automated, hardware-based, front-to-back-end I/O load balancing.

Both controllers in the storage system dynamically and automatically assign the access paths from the controller to a logical unit (LU). All LUs are accessible, regardless of the physical port or the server from which the access is requested.

Controller utilization rates are monitored to maintain a more even distribution of workload between the two controllers. Storage administrators are not required to manually define specific affinities between LUs and controllers, simplifying overall administration.

This controller design is fully integrated with standard host-based multipathing. This eliminates mandatory requirements to implement proprietary multipathing software.

The point-to-point back-end design nearly eliminates I/O transfer delays and contention associated with Fibre Channel arbitration. The design provides significantly higher bandwidth and I/O concurrency. It isolates any component failure that might occur on back-end I/O paths.

Use the Oracle EM Grid Control System Monitoring Plug-in for Hitachi Storage to monitor and manage your storage infrastructure from within Oracle Enterprise Manager Grid Control. This provides real-time visibility to utilization, availability, and performance metrics. Optimize your hardware infrastructure for Oracle database applications using the following:

- Detailed reports on storage subsystem configuration
- LUN configuration
- Database usage summary
- ASM diskgroup to storage devices
- Storage devices to OS mapping
- Database files to storage mapping
- Storage subsystem statistics
- Database performance,
- Storage pool performance,
- Port performance and raid group performance

For more information, see [Hitachi Adaptable Modular Storage 2000 Family](#) on the Hitachi Data Systems website.

Hitachi Dynamic Provisioning

On Hitachi Adaptable Modular Storage 2500, Hitachi Dynamic Provisioning provides wide striping and thin provisioning functionalities.

Using Hitachi Dynamic Provisioning is like using a host-based logical volume manager (LVM), but without incurring host processing overhead. It provides one or more wide-striping pools across many RAID groups. Each pool has one or more dynamic provisioning virtual volumes (DP-VOLs) of a logical size you specify of up to 60TB created against it without allocating any physical space initially.

Deploying Hitachi Dynamic Provisioning avoids the routine issue of hot spots that occur on logical devices (LDEVs). These occur within individual RAID groups when the host workload exceeds the IOPS or throughput capacity of that RAID group. This distributes the host workload across many RAID groups, which provides a smoothing effect that dramatically reduces hot spots.

Hitachi Dynamic Provisioning has the benefit of thin provisioning. Physical space assignment from the pool to the DP-VOL happens as needed using 1GB pages, up to the logical size specified for each DP-VOL. There can be a dynamic expansion or reduction of pool capacity without disruption or downtime. An expanded pool can be rebalanced across the current and newly added RAID groups for an even striping of the data and the workload.

For more information, see the [Hitachi Dynamic Provisioning datasheet](#) and [Hitachi Dynamic Provisioning](#) on the Hitachi Data Systems website.

Hitachi Dynamic Link Manager Advanced

Hitachi Dynamic Link Manager Advanced is a software package that combines all the capabilities of Hitachi Dynamic Link Manager and Hitachi Global Link Manager into a comprehensive multipathing solution. It includes capabilities such as the following:

- Path failover and failback
- Automatic load balancing to provide higher data availability and accessibility.

Used for storage area network multipathing, the Hitachi Dynamic Link Manager configuration for this solution used the extended round-robin multipathing policy. This policy automatically selects a path by rotating through all available paths. This balances the load across all available paths and optimizing IOPS and response time.

For more information, see [Hitachi Dynamic Link Manager](#) on the Hitachi Data Systems website.

Hitachi Storage Navigator Modular 2

Hitachi Storage Navigator Modular 2 enables essential management and optimization functions for the Hitachi Adaptable Modular Storage 2000 family. It provides a web-accessible graphical management interface and a command line interface (CLI) to allow ease of storage management.

You need Storage Navigator Modular 2 to take advantage of the full features in Adaptable Modular Storage 2500.

Use Storage Navigator Modular 2 for:

- RAID-level configurations
- LUN creation and expansion
- Online microcode updates and other system maintenance functions
- Performance metrics

For more information, see [Hitachi Storage Navigator Modular 2](#) on the Hitachi Data Systems website.

Oracle Enterprise Linux OS

Oracle Enterprise Linux is an enterprise-class operating system that is fully compatible with the Red Hat Enterprise Linux kernel. Derived from the stable 2.6.32 mainline Linux kernel, it was built and tested to run Oracle hardware, databases, and middleware.

For more information, see the [Oracle Enterprise Linux](#) website.

Oracle Database 11g Release R2

The following are features of [Oracle Database 11g Release 2](#).

Oracle Real Application Clusters

The Oracle Real Application Clusters (RAC) option for Oracle Database 11gR2 facilitates deployment of one database across a cluster of servers. It provides for continuous operation in situations such as hardware failure and planned outages. RAC allows administrators to scale applications without taking users offline.

Oracle Clusterware

Oracle Clusterware groups multiple servers so that they function as one server. This facilitates failure detection, recovery, messaging, and locking. Use it to monitor, move, and restart applications and to manage all Oracle processes automatically. Clusterware is part of the grid infrastructure component in Oracle Database 11gR2.

Oracle Automatic Storage Management

Oracle Database Automatic Storage Management (ASM) system combines the features of a volume manager and an application-optimized general purpose file system. It is optimized for use with Oracle products. ASM is part of the grid infrastructure component in Oracle Database 11gR2. You can use storage level striping and mirroring. Hitachi Data Systems recommended practice for ASM is to choose **External Redundancy** for mirroring.

Oracle Smart Flash Cache

The Oracle Smart Flash Cache provides a way to increase the effective size of the Oracle buffer cache without adding more RAM to the system.

Oracle database blocks stored on the disk are loaded into memory when a client session requires it. This memory area is called the system global area (SGA). When SGA needs more space, Flash Cache moves lesser used database blocks to the second level cache on SSD flash memory instead of evicting the blocks. If the SGA needs one of those lesser used blocks later, Flash Cache fetches it from flash memory instead of the disk. Since Flash Cache fetches data from memory instead of a physical drive, this can benefit transaction throughput and application response times.

This release supports Flash Cache on Solaris and Linux platforms.

Oracle Enterprise Manager Grid Control

Oracle Enterprise Manager 11g Database Control Release 11.2.0.2.0 is the system management and monitoring software for heterogeneous environments used in this reference architecture. It is a centralized management console. The features in this release extend resource administration to manage other components in the solution stack. In addition to the databases, including RAC, resource administration includes servers, storage, operating systems, and middleware.

Hitachi Data Systems has developed storage plug-ins and adaptors to extend the centralized management functionality of Hitachi storage. This solution uses two of the monitoring and analysis features, Automatic Workload Report (AWR) and Automatic Database Diagnostic Monitor (ADDM). They have been used to detect and tune database transactions for performance and scalability.

Fusion-io Card PCIe SSD Card

This reference architecture uses a 320GB Fusion-io PCI Express x4 PCIe solid state storage device (SSD) card as an extended cache to extend memory capacity and improve database performance. The SSD cache has very low access latency and provides very high IOPS. Oracle applications can use this card as an extended Oracle buffer cache by utilizing Oracle's Flash Cache and as a fast database storage device by utilizing Oracle ASM with the preferred read option.

For more information on the card used in this solution, see the [Fusion-io ioDrive](#) website.

Brocade Storage Area Network Switches

This environment uses two Brocade 5340 switches. These switches provide 8Gb/sec connectivity between the physical servers and the Hitachi Adaptable Modular Storage 2500, and 48 to 80 ports per switch.

For more information, see the [Brocade and Hitachi Data Systems Products and Solutions](#) website.

Solution Design

This is a detailed description of the test setup used for the Hitachi Compute Blade 2000 reference architecture for small to medium scale OLTP workload on a two node Oracle 11gR2 RAC environment. The purpose of the test was to illustrate the performance and scalability for two variations of OLTP workload profiles.

High-level Infrastructure

The architecture that Hitachi Data Systems used for testing contains the following:

- **Servers** — Two Hitachi Compute Blade 2000 E57A1 server blades were used. Each server blade had two quad core 2.27GHz processors, 256GB RAM (8x32GB), one dual port 8Gb/sec Fibre Channel PCI Emulex HBA on the PCIe slot, and six 1Gb/sec NICs (two on-board and four on mezzanine slot 0).
- **Storage System** — Hitachi Adaptable Modular Storage 2500 with two controllers, each with 4GB cache and sixteen front-end ports (host connectors). The storage system has nine trays with a total raw storage capacity of approximately 125TB. Two front-end ports on each controller were used to map the boot and database storage LUNs.
- **Local Storage** — Two PCIe Fusion-io Multi Level Cell MLC 320GB cards were used. One card was installed on PCIe slot 0 on each E57A1 server blade. Each flash memory card extends the Oracle buffer cache of one Oracle instance using Oracle Flash Cache.
- **Cluster Interconnect** — Four 1Gb NIC ports on mezzanine slot 0 of each server blade is configured for multiple private Oracle RAC cluster interconnects
- **SAN Fabric** — Two 8GB/sec Fibre Channel switches were installed. Two zones were created on each switch to zone the two PCIe HBA ports on each server blade and the four storage host ports.

Figure 2 illustrates the reference architecture used in testing the Oracle 11gR2 RAC solution.

AMS 2500 Storage:
300GB 15K SAS, 600GB 10K SAS(SFF) disks

Hitachi Compute Blade 2000: Oracle 11gR2 RAC Nodes

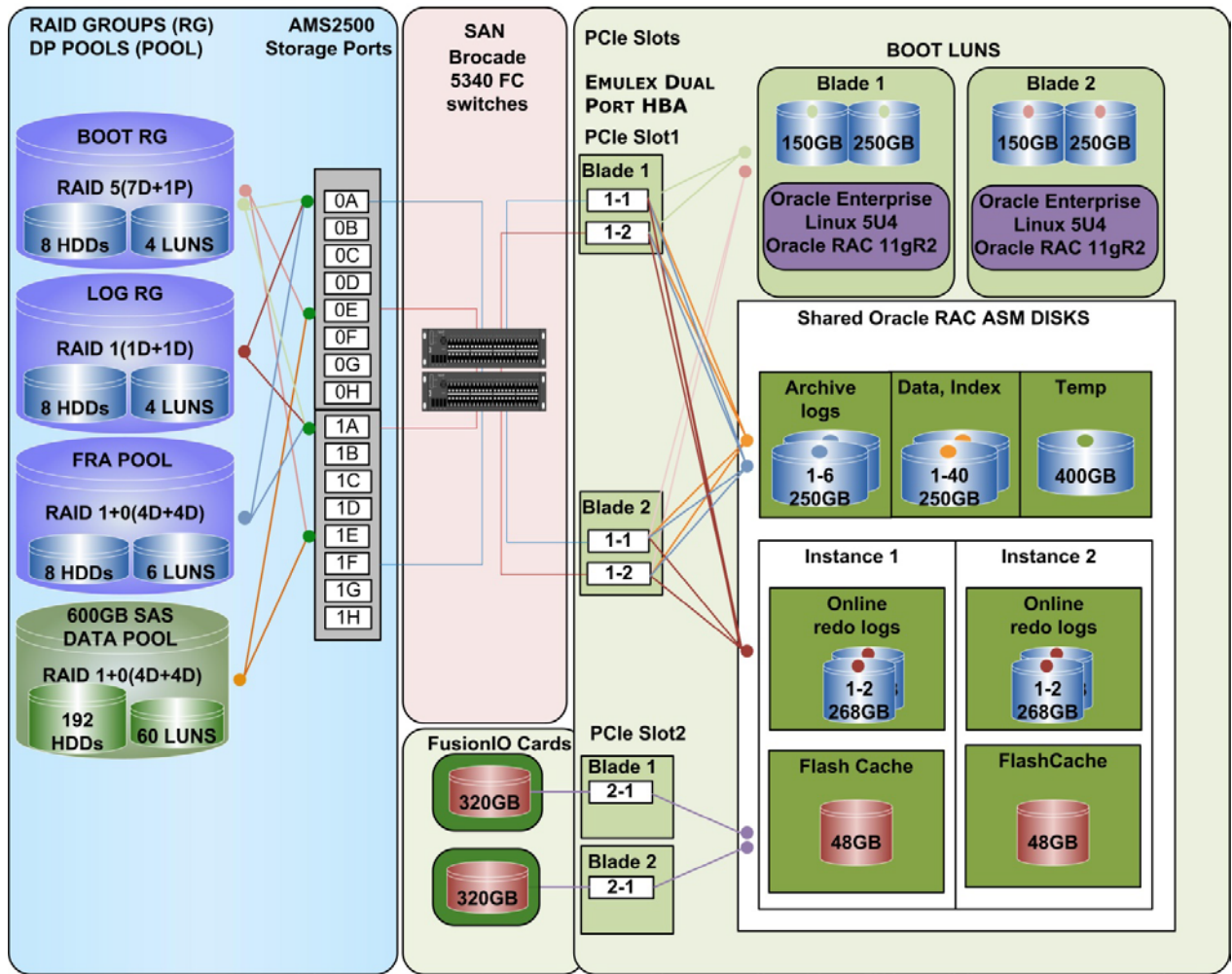


Figure 2

Storage Architecture

This describes the storage architecture for the Oracle database environment deployed for this reference architecture. It takes into consideration Hitachi Data Systems and Oracle recommended practices for the deployment of database storage design.

Storage Configuration

This reference architecture uses RAID groups and dynamic provisioning storage pools on Hitachi Adaptable Modular Storage 2500.

Oracle RAC database requires shared storage for cluster nodes to access the database on the persistent storage. Shared storage is configured through host groups and zoning on the SAN, while Oracle Clusterware and Oracle ASM manages the shared storage on the server for the RAC database.

Figure 3 describes the RAID groups, dynamic provisioning pools, and host groups on Adaptable Modular Storage 2500 for the operating system and database disks.

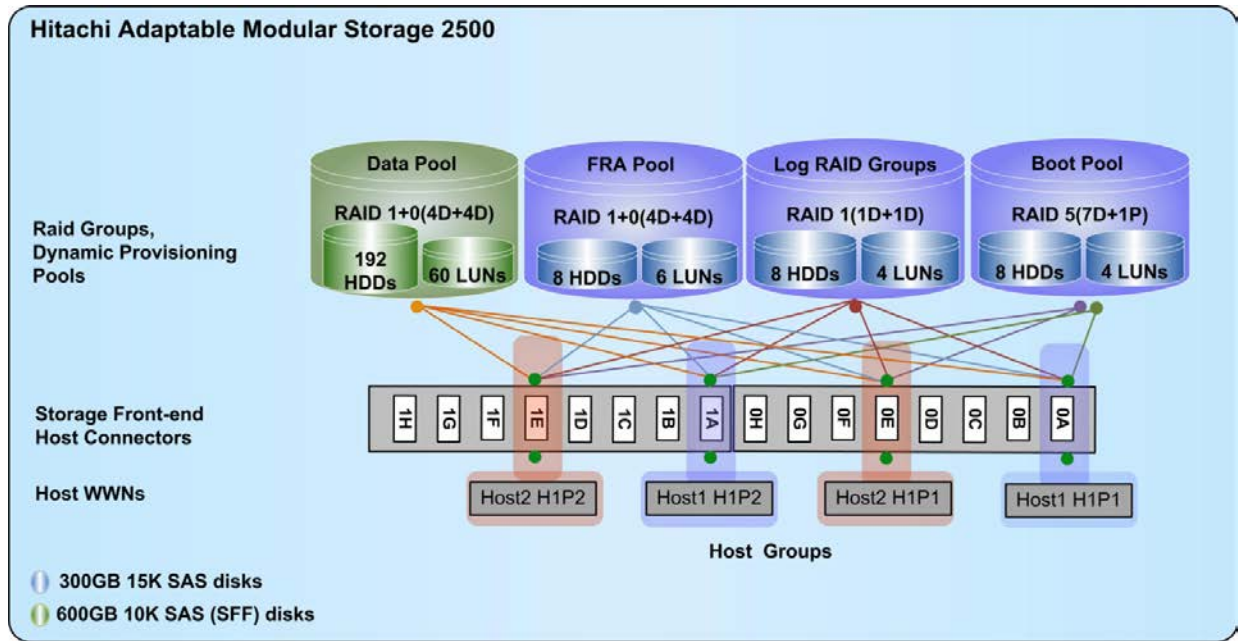


Figure 3

Details of the RAID groups and dynamic provisioning pools are given below. Table 3 and Table 4 show the RAID groups and LUNs. Table 5 and Table 6 show the dynamic provisioning pools and LUNs used for the operating system and database storage.

Table 3. RAID Groups Used for the Oracle Online Redo Logs

RAID Group	RAID Level	Drive Type	No of Drives	Capacity (GB)
001	RAID1(1D+1D)	300GB 15k RPM SAS	2	267.8
002	RAID1(1D+1D)		2	267.8
004	RAID1(1D+1D)		2	267.8
005	RAID1(1D+1D)		2	267.8

Table 4. RAID LUNs

RAID Group	LUNs	LUN Size (GB)	Purpose	Storage Port
001	011	267	Oracle RAC Node 1 online redo log files	0A, 1A
002	012	267		0A, 1A
004	013	267	Oracle RAC Node 2 online redo log files	0A, 1A
005	014	267		0A, 1A

Table 5. Dynamic Provisioning Pools

<i>Dynamic Provisioning Pool ID</i>	<i>DP RAID Group</i>	<i>RAID Level</i>	<i>Drive Type</i>	<i>No of Drives</i>	<i>Pool Capacity</i>
000	099	RAID5(7D+1P)	300GB 15k RPM SAS	8	1.8TB
002	088	RAID5(8D+1P)		9	2TB
003	064-087	RAID1+0(4D+4D)	600GB 10k RPM SAS(SFF)	192	50.2TB

Table 6. Dynamic Provisioning Volumes LUNs

<i>Dynamic Provisioning Pool ID</i>	<i>LUNs</i>	<i>LUN Size (GB)</i>	<i>Purpose</i>	<i>Storage Port</i>
000	000	150	Oracle RAC Node1 Boot Volumes	0A, 1A
	001	250		
002	002	150	Oracle RAC Node2 Boot Volumes	0A, 1A
	003	250		
002	055 056 057 058 059	250	Archive Logs and Flash Recovery Area	0A, 1A
003	061-100	250	CRS Voting Disks Oracle System Undo Temp OLTP Application Tablespaces	0E, 1E

Database Layout

The database layout design uses recommended practices from Hitachi Data Systems for Hitachi Adaptable Modular Storage 2500 for short random I/Os, such as the ones in OLTP transactions. It also takes into account the Oracle ASM best practices when using Hitachi storage.

The storage design for database layout needs to be based on the requirements of a specific application implementation. The design can greatly vary from implementation to another. The components in this solution set have the flexibility for use in various deployment options to provide the right balance between performance and ease of management for a given scenario.

- **Data and Indexes Table Space**—One Hitachi Dynamic Provisioning pool is assigned for the application data and indexes. The allocated capacity is 50TB. The small file table space consists of several 32GB data files. The table space is set to small initial size with auto extend enabled to maximize storage utilization from thin provisioning.
- **Temp Table Space**—Temp table space in this test configuration is in the Data and Indexes ASM diskgroup. A separate dynamic provisioning pool can be assigned to Temp table space considering the backup and recovery requirements and the type of workload. Temp table space was created with BIGFILE table space option.

- **Online Redo Logs**—One raid group is assigned to each Online Redo Log file of the two instances in the RAC cluster.
- **Size Settings**—The database block size is set to 8KB. ASM allocation unit is set to 8MB.

Table 7 below lists the disk map from the Storage LUNS to the OS devices and to the ASM Disk groups.

Table 7. Storage LUNS, OS Disks and Oracle ASM Disk Map

<i>LUN</i>	<i>OS device /dev/...</i>	<i>ASM disk</i>	<i>ASMDG</i>	<i>Purpose</i>
08-010	sdc1-sde1	OCR1-OCR3	SUTD	OCR and Voting Disks
061-100	sdbg1-sdcq1	SUTD1-SUTD40	SUTD	Sys, Undo, Temp, Data
058-060	sdaa1-sdaa2	SUTD45-SUTD46	SUTD	Temp
011-012	sdf1- sdg1	REDO11-REDO12	REDO1	Online REDO logs groups node 1
013-014	sdh1-sdi1	REDO21-REDO22	REDO2	Online REDO logs groups node 2
055-059	sdax1-sdbb1	ARCH1-ARCH5	ARCFR	Archive logs and Flash recovery area.
N/A	fioa	OFCFUIO1	OFC1	Flash Cache device node1
N/A	fioa	OFCFUIO2	OFC2	Flash Cache device node2

Server and Application Architecture

The reference architecture used two Hitachi Compute Blade 2000 server blades for the database servers. The test scenario includes a single RAC instance on each blade. Each blade has 16 CPU cores and 256GB RAM. The number of CPU cores provides the compute power for the Oracle RAC database to handle complex database queries and large volume of transaction processing in parallel.

At peak loads, the CPUs showed as high as 85% utilization. In the test scenario, only a single database instance occupied various SGA sizes. Depending on the nature of the workload, there is sufficient capacity CPU and memory to accommodate another database instance.

The design principle is to size the server CPU to meet the performance requirement at 85% busy, leaving 15% headroom for transient burst loads.

There is no additional third party volume manager software and cluster management software needed in the reference architecture. Oracle 11gR2 Clusterware provides the necessary functionality for cluster management functionality including a comprehensive framework to setup automatic failover configurations.

The management server for Oracle 11gR2 grid control and Hitachi Storage Navigator Modular 2 meets the need of the database and storage administrators in monitoring and managing the database, the database servers, and the storage.

The Oracle Grid Control Management Server, Hitachi Storage Navigator Modular 2, and Hitachi Command Center can be hosted on Hitachi Compute Blade 2000 server blade by carving out logical partitions.

Figure 4 represents the logical structure of the software solution stack.

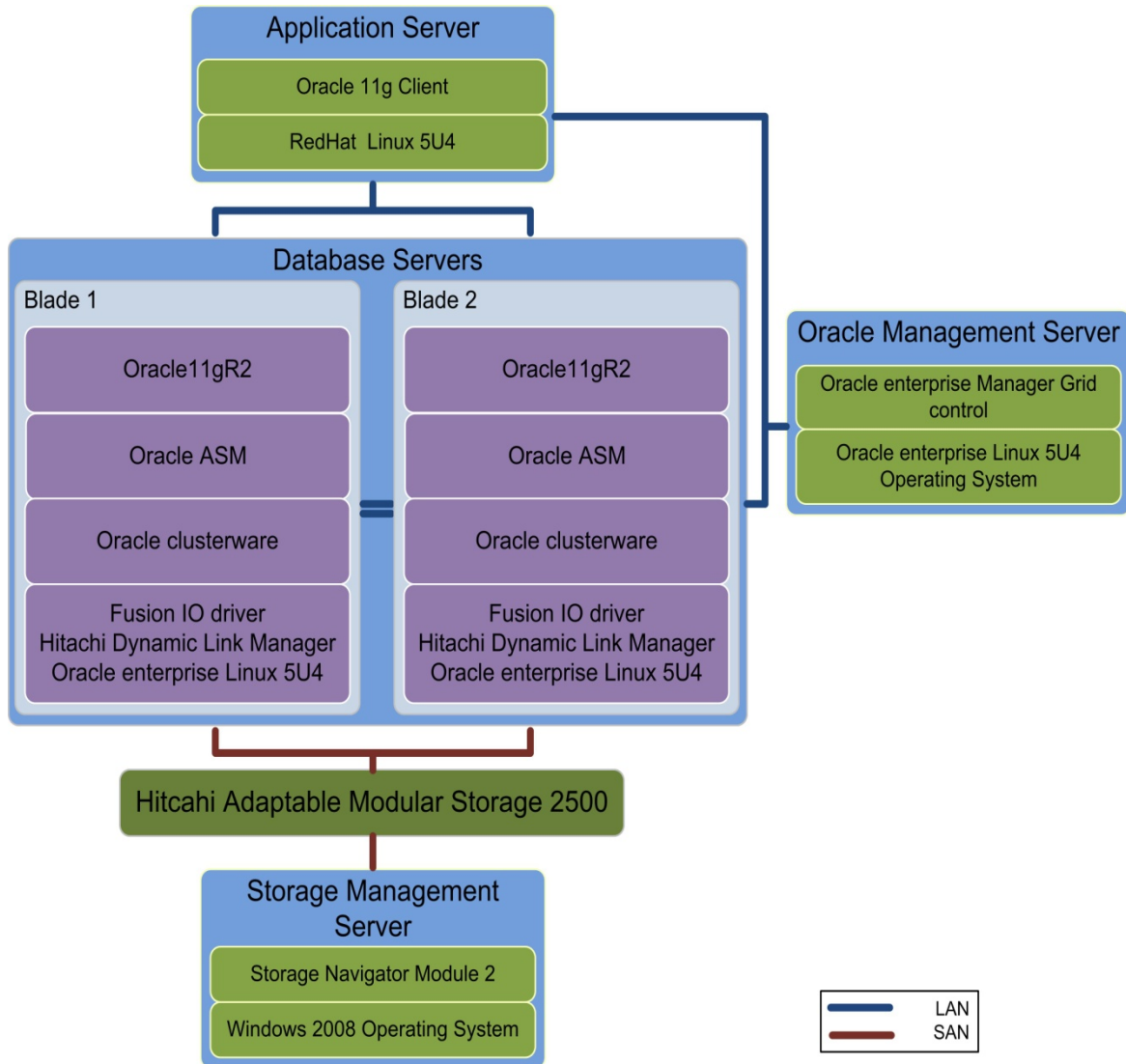


Figure 4

SAN Architecture

The provisioned LDEVs were mapped to multiple ports on a Hitachi Adaptable Modular Storage 2500. LDEV port assignments ensure that the host had two paths to the storage system for high availability.

The environment used two Brocade 5340 switches to provide scalability and high availability to the environment. See “Storage Configuration” for the host configuration details.

- The RAC database servers used two HBA ports of the Emulex PCIe HBA card. This enables a two-path connection for all LUNs mapped to that server. The following is how host ports were used on the Adaptable Modular Storage 2500:
 - Host ports 0A and 1A are used by RAC host 1 for its SAN boot LUNs
 - Host ports 0E and 1E are used by RAC host 2 its SAN boot LUNs
 - Host ports 0A, 1A, 0E, and 1E are used by RAC host 1 and 2 for the database LUNs.

- Each server has two SAN boot LUNs. The LUNs are mapped as following:
 - Database server host 1 LUNs are mapped to host ports 0A and 1A.
 - Database server host 2 LUNs are mapped to host ports 0E and 1E.
- The database disk LUNs have shared access by both database server hosts. It is a requirement for Oracle RAC database.

The environment used two Brocade 5340 switches to provide scalability and high availability to the environment. Table 8 lists the zoning details for the SAN.

Table 8. SAN Switch Architecture

<i>Server</i>	<i>HBA Ports</i>	<i>Switch Zone</i>	<i>Storage Port</i>	<i>Switch</i>	<i>Use</i>
RAC Nd-1	HBA1-1	BS2K_06_B1_HBA1_1_ASE45_26_0A	0A	5300-05	Boot Disks
RAC Nd-1	HBA1-2	BS2K_06_B1_HBA1_2_ASE45_26_1A	1A	5300-06	
RAC-Nd-2	HBA1-1	BS2K_06_B2_HBA1_1_ASE45_26_0E	0E	5300-05	Boot Disks
RAC-Nd-2	HBA1-2	BS2K_06_B2_HBA1_2_ASE45_26_1E	1E	5300-06	
RAC Nd-1	HBA1-1	BS2K_06_B2_HBA1_1_ASE45_26_0A	0A	5300-05	Database Disks
RAC-Nd-1	HBA1-2	BS2K_06_B2_HBA1_2_ASE45_26_1A	1A	5300-06	
RAC Nd-2	HBA1-1	BS2K_06_B2_HBA1_1_ASE45_26_0E	0E	5300-05	
RAC-Nd-2	HBA1-2	BS2K_06_B2_HBA1_2_ASE45_26_1E	1E	5300-06	

Hitachi Data Systems recommends the use of dual SAN fabrics, multiple HBAs, and host-based multipathing software when deploying this reference architecture. You need at least two paths to ensure the redundancy required for critical applications from the following:

- Database hosts connect to two independent SAN fabrics
- SAN fabric to two different controllers of the I/O subsystem

When designing your SAN architecture, follow these recommended practices to ensure a secure, high-performance, and scalable database deployment:

- Use at least two HBAs and place them on different I/O buses within the server. This distributes the workload over the server's PCIe bus architecture. Dual-port HBAs accomplish this by dividing the bandwidth of each port between PCIe lanes on the expansion bus.
- Use dual SAN fabrics, multiple HBAs, and host-based multipathing software in a business-critical deployment. Connecting two or more paths from the database servers to two independent SAN fabrics are essential to ensure the redundancy required for critical applications.
- Zone your fabric appropriately for multiple, unique paths from HBAs to storage ports. Use single initiator zoning. Use at least two Fibre Channel switch fabrics to provide multiple independent paths to Hitachi Adaptable Modular Storage 2500 to prevent configuration errors from disrupting the entire SAN infrastructure.
- For large bandwidth requirements that surpass a single HBA's port capability, use additional HBAs and use the round robin load-balancing setting for Hitachi Dynamic Link Manager.

Physical and Virtual Network Cluster Interconnect Architecture

Hosts in this reference architecture have separate 1Gb/sec physical network adapters for independent use by different types of data traffic.

Oracle 11gR2 introduces Single Client Access Name (SCAN). Prior to Oracle 11gR2, adding or removing nodes from the cluster required changing `tnsnames.ora` for the node VIP addresses. With SCAN, Oracle eliminated the need to change `tnsnames.ora` entries.

Oracle 11gR2 Grid Infrastructure installations require using SCAN, since it is essential during the creation of the Oracle RAC 11gR2 database.

Setup SCAN using one of the following options.

- **Using the DNS**—Create a single name that resolves to three IP addresses in the DNS. Do not assign the IP address to a NIC, as Oracle Grid Infrastructure will make that assignment.
- **Using GNS**— Instead of listing SCAN static addresses in the DNS, create a sub-domain with a static virtual IP address in the DNS for the GNS to run. Obtain the node VIP and the SCAN VIP from the DHCP server when using GNS.

When validating this test configuration, SCAN was configured using the DNS.

Use of Oracle Grid Infrastructure Release 2 Patch Set 1 (11.2.0.2) and later requires multicasting on the private interconnect. For this reason, you must enable the cluster to multicast at least for the following:

- Across the broadcast domain, as defined for the private interconnect
- On IP subnet ranges 224.0.0.0 through 224.0.0.0.24 and 230.0.1.0 through 230.0.1.0.24

You do not need to enable multicast communications across routers.

For the private vLAN, the interface must support the following:

- User datagram protocol (UDP) using high-speed network adapters.
 - UDP is the default interface protocol for Oracle RAC.
- Switches that support TCP/IP, with a minimum 1 gigabit Ethernet connection required.
 - TCP is the interconnect protocol for Oracle Clusterware.
 - A dedicated switch is recommended for the interconnect. Token-rings or crossover cables for the interconnect are not supported.
 - For the private network, the endpoints of all designated interconnect interfaces must be completely reachable on the network.

Figure 5 shows the public and private network configuration for the test environment.

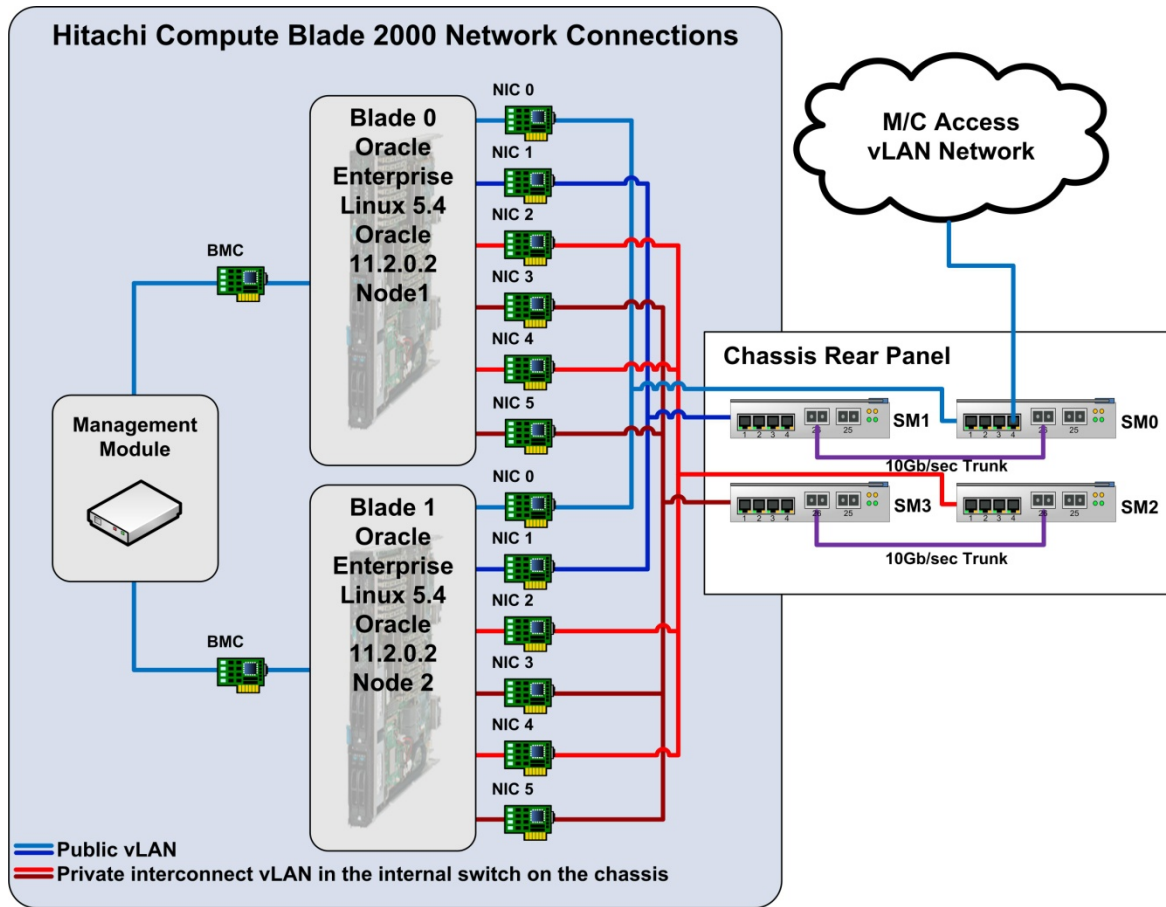


Figure 5

This configuration uses 1Gb/sec physical network adapters:

- Two adapters for public network
- Four adapters for Oracle cluster private interconnect

The ports on each blade connecting to the switches in the chassis are 1 gigabit Ethernet.

- NIC0 and NIC1 are the two onboard NICs.
- NIC2 to NIC 5 are the four ports from the mezzanine NIC card.

The uplink from the internal switches is 10 gigabit Ethernet Fibre Channel.

In the Oracle RAC configuration, the private interconnect is a critical component for improved performance. Use Linux networking parameters to improve Oracle RAC interconnect performance. These parameters tell the Linux operating system how much buffer space to reserve for each socket. In this reference architecture, the network latency of the cluster interconnect was within an acceptable limit of 1 millisecond.

- `net.core.rmem_default=262144`
- `net.core.rmem_max=4194304`
- `net.core.wmem_default= 262144`
- `net.core.wmem_max = 1048576`

Engineering Validation

This gives the validation of this reference architecture for functionality and performance for Oracle RAC scalability.

Workload

A Swingbench OLTP workload was used in the testing. It simulates an order entry system that consists of five transactions.

Two workload profiles were used in to validating the reference architecture, as follows:

- **P1 (Profile 1)**—48% read transactions, 52% write transactions
- **P2 (Profile 2)**—82% read transactions, 17% write transactions

The read-intensive profile was tested to show the benefit of using Flash Cache and PCIe SSD cards to improve performance of RAC scalability. Database layout for the OLTP workload has been built on partitioned tables and indexes for improved performance gains.

Methodology

The Oracle RAC scalability tests stressed both nodes in the cluster with concurrent user connections.

1. The workload was applied to first node.
 - Once the workload reached stability, a snap of the TPM counter was taken.
2. The workload was applied to the second node.
 - Once the workload reached stability, a snap of the TPM counter was taken.

Database IOPS was calculated by summing up the average reads and average writes across all of the tablespaces. Cluster scalability was calculated by the ratio of the stable load TPM count load on the first node to the stable load TPM count on the second node.

Two test scenarios were designed to quantify the performance gain achieved with the Flash Cache and PCIe SSD cards as follows:

- Large size SGA without Flash Cache. The SGA size was 10GB.
- Large size SGA with Flash Cache. The Flash Cache was set to 48GB in all test iterations.

Tests stressed each node in the cluster using both workload profiles by changing the number of connections to identify the performance at various CPU and I/O usage levels.

Data Gathering

Performance statistics were collected at these levels:

- **Storage**
 - Hitachi Storage Navigator Modular 2 performance data collected storage performance.
- **Operating System**
- **Database**
 - Oracle Automatic Workload Repository report collected database performance.
 - Swingbench collected application-level statistics for transactions executed and response times, including relevant Oracle level wait events and statistics.

The performance data was used for the following:

- Tune the architecture for performance and scalability
- Study the PCIe SSD cache

Results Summary

This summarizes the key observations from the test results.

RAC Performance IOPS and Response Times

The RAC nodes were stressed with read and write intensive OLTP workload profiles. The results were analyzed and system was tuned for the following:

- Disk bandwidth
- System global area (SGA)
- Log buffer sizes

The performance gain was seen after database tuning and addressing resource bottleneck.

- The RAC configuration is a two node Oracle RAC server.
- The SN configuration is a single node RAC server.

Table 9 summarizes the test results for this test.

Table 9. RAC and SN Performance by IOPS and Response Time

<i>Profile</i>	<i>Flash Cache</i>	<i>RAC or SN</i>	<i>IOPS</i>	<i>Response Time (ms)</i>
P1	With	RAC	24205	5.30
		SN	15836	2.90
	Without	RAC	25859	9.20
		SN	17375	6.80
P2	With	RAC	35874	3.14
		SN	24721	2.16
	Without	RAC	28574	6.89
		SN	19151	4.24

Figure 6 shows that database level IOPS when using and when not using Flash Cache. A single node RAC configuration and a two node RAC configuration were tested. During this testing, cluster interconnects traffic and the Oracle RAC global cache parameters were monitored.

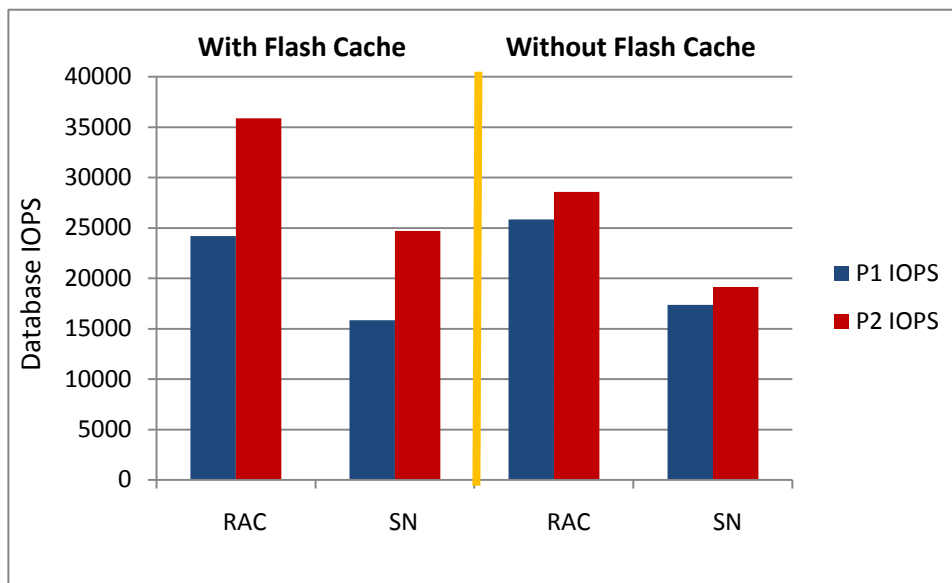


Figure 6

Figure 7 shows the response time when using and not using Flash Cache. The average response time of the database calculated across all of the table spaces.

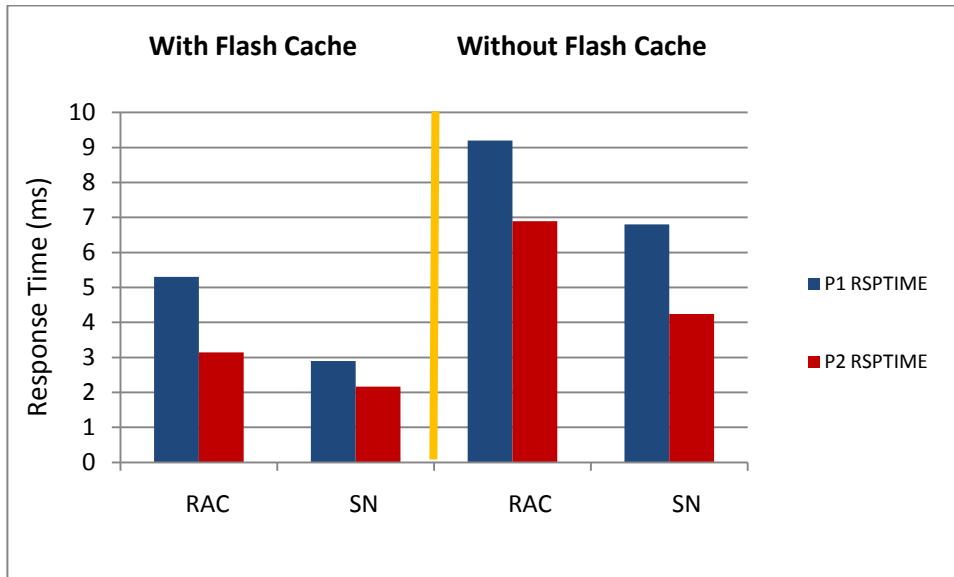


Figure 7

The testing showed that the read intensive workload in workload profile P2 benefited from using Flash Cache. In Figure 6, profile P2 delivered higher IOPS with Flash Cache in single node and RAC environments. Figure 7 shows the reduced response times with Flash Cache in the single node and RAC configurations with profile P2.

TPM and Scalability on 2 Node Oracle RAC Configuration.

These tests determined the number of transactions per minute achieved on a two node RAC cluster.

Workload profile P1 was exercised on Node 1 of the RAC cluster without using Flash Cache. After the load stabilized on Node 1, the workload was exercised on Node 2. Then, the test was repeated using Flash Cache.

Figure 8 and Figure 9 show the transactions per minute achieved with workload profile P1 and 80 300GB 15k RPM SAS disks with and without using Flash Cache setting on Node 1 and Node 2 of the RAC cluster. The test was rerun with 192 600GB 10k RPM SAS disks.

The exercise was repeated with workload profile P2. Figure 8 and Figure 9 show the test results of workload profile P1 and 2 with and without the Flash Cache setting on each node of the RAC cluster.

Table 10 summarizes the results of the OLTP workload transactions per minute tests on RAC Node1.

Table 10. RAC Node 1 OLTP Workload Transactions per Minute Test Results

Disk Configuration	Flash Cache	Profile	Node	Transactions/min.
80 300GB 15k RPM SAS disks	With	Profile 1	1	55247
			2	57136
	Without	Profile 1	1	31680
			2	30477
192 600GB 10k RPM SAS disks	With	Profile 1	1	82780
			2	84055
		Profile 2	1	130745
			2	168953
	Without	Profile 1	1	95538
			2	90160
		Profile 2	1	132518
			2	138280

Figure 8 and Figure 9 show the transactions per minute on Node 1 and Node 2 of the RAC cluster.

There was a significant improvement in the transactions per minute after removing a disk bandwidth bottleneck. Profile1 (P1), the write intensive profile, did better without Flash Cache. Profile 2, the read intensive profile, benefitted from the Flash Cache configuration.

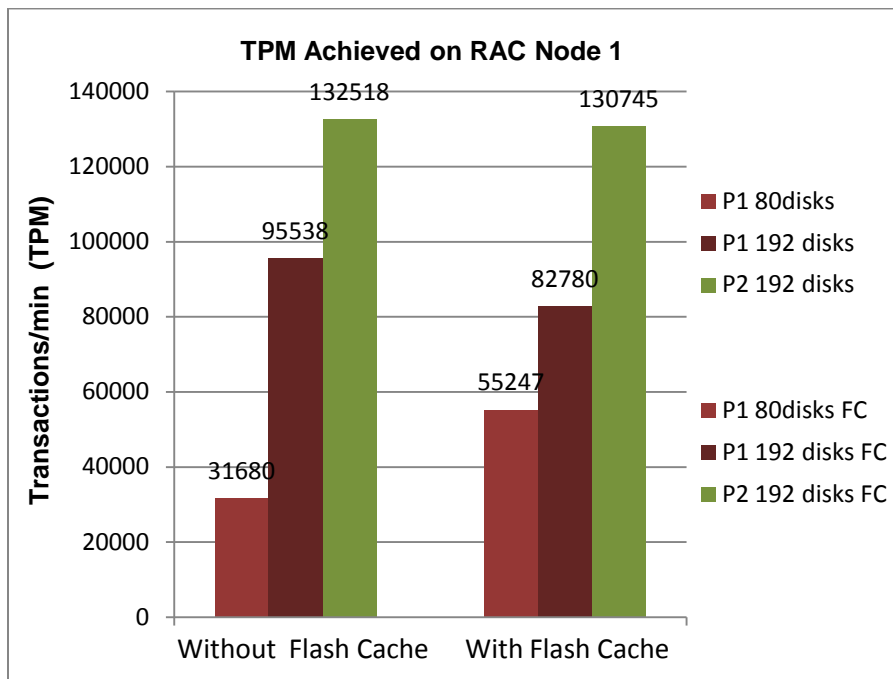


Figure 8

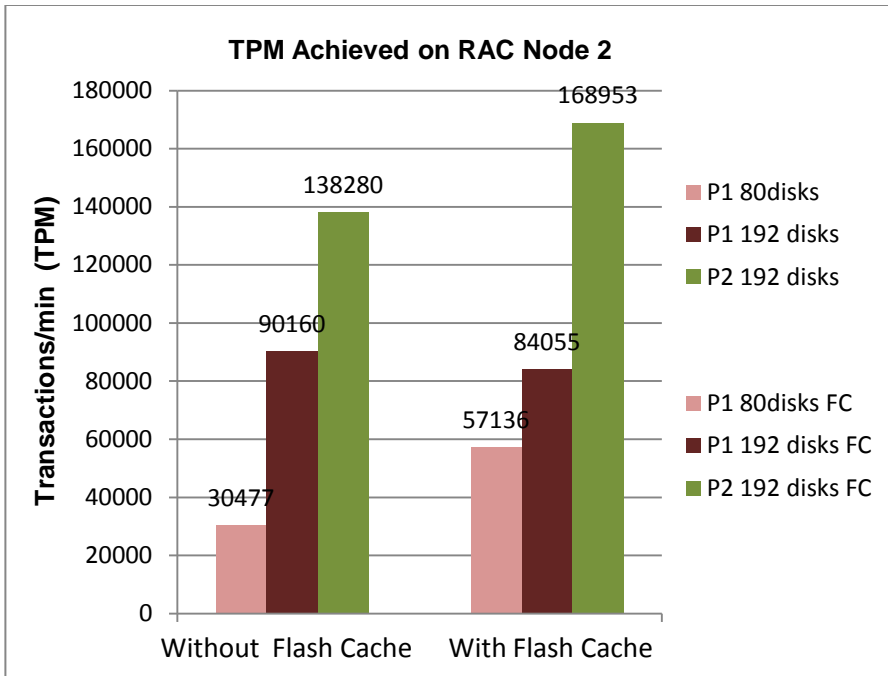


Figure 9

Figure 10 shows the combined transactions per minute achieved from a two node Oracle RAC cluster after doubling the user load through the second node. Once the initial load stabilized and reached its peak transactions per minute, the load on the second node was started.

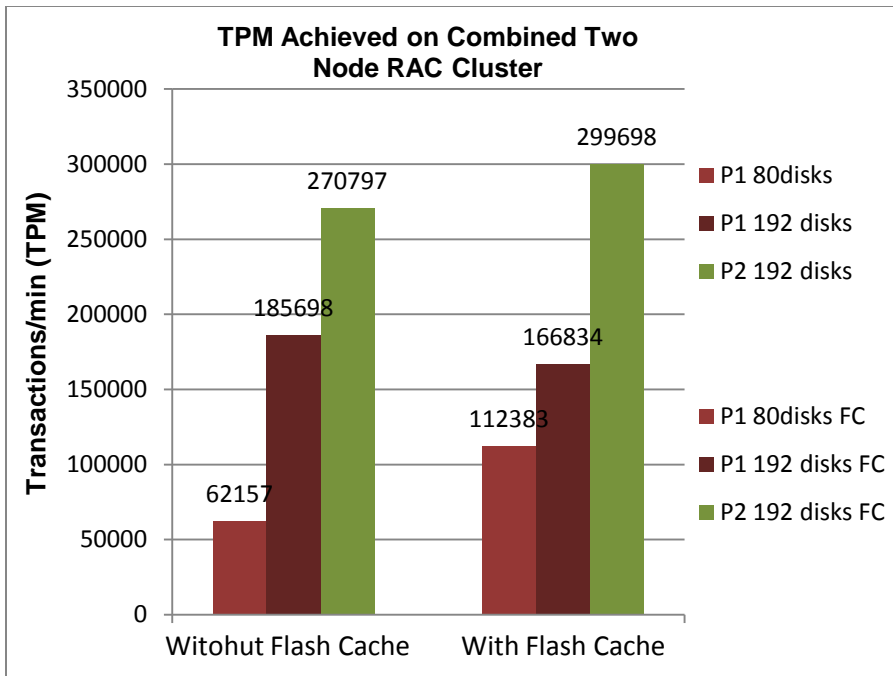


Figure 10

Figure 11 shows the scalability factor on a two node Oracle RAC for the two workload profiles. The scalability factor is calculated when the load stabilized and reached peak on the second node. It is observed that read intensive profile, P2, benefitted when using Flash Cache.

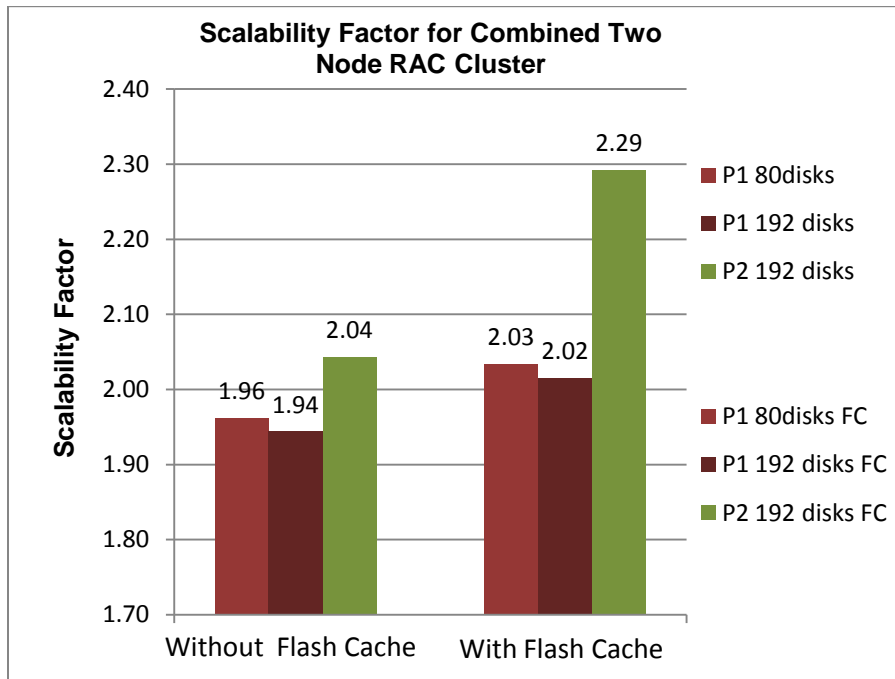


Figure 11

Flash Cache Comparison in Read-Heavy OLTP Workloads.

Table 11 shows Swingbench a query response time results with a highly stressed database. When Oracle system global area is at a premium, it shows that Flash Cache helps improve performance.

Table 11. Comparing Average Query Response Time With and Without Flash Cache

Flash Cache	Configuration	Average Query Response Time (ms)
With	Single Node	29
	Two Node	45
Without	Single Node	50
	Two Node	73

The results in Table 11 only compare two specific configurations in a lab environment. These results do not portray the maximum capabilities of any system, database software program, or storage device. The RAC nodes were stressed with a high user load to make the shared storage tier a bottleneck that increased the response time. This configuration only covered a single Flash Cache. Extending the use of Flash Cache should provide improved response times.

Figure 12 graphs the query response time for Profile 2 with and without using Flash Cache. Results are shown in the single node (SN) and two node (RAC) configurations.

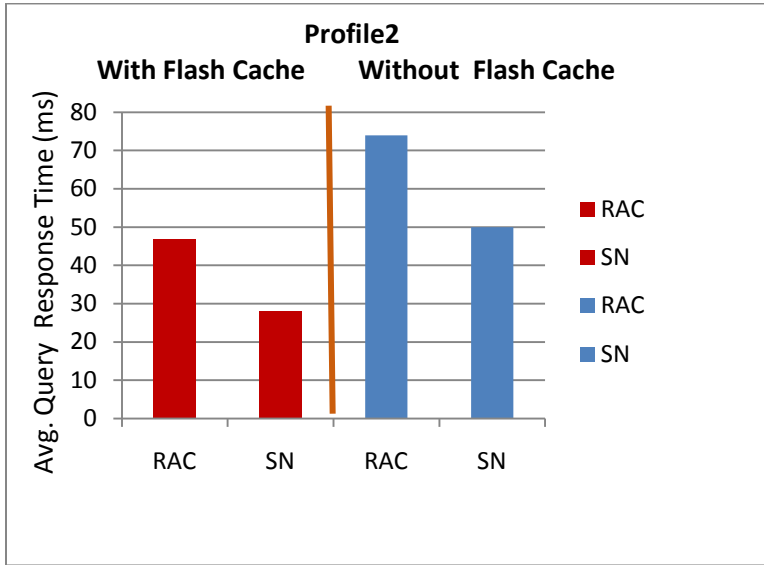


Figure 12

System Load

The server utilization data was collected during all the validation tests.

Figure 13 shows the percentage utilization of the CPU cores with Flash Cache. Note the reduced I/O waits when compared to Figure 14, the results without Flash Cache.

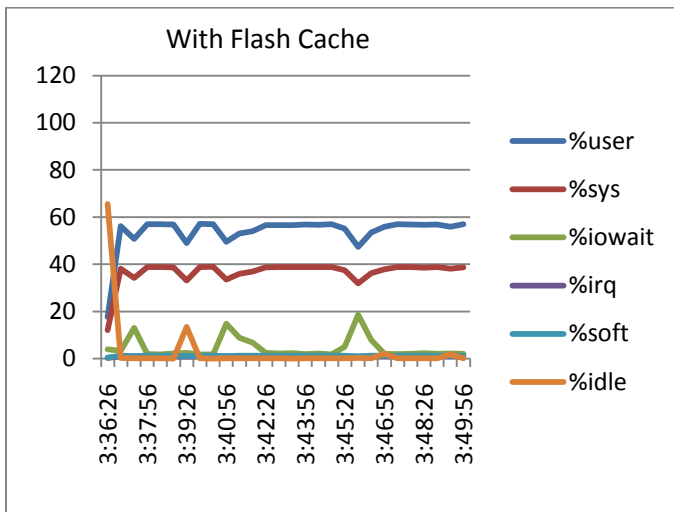


Figure 13

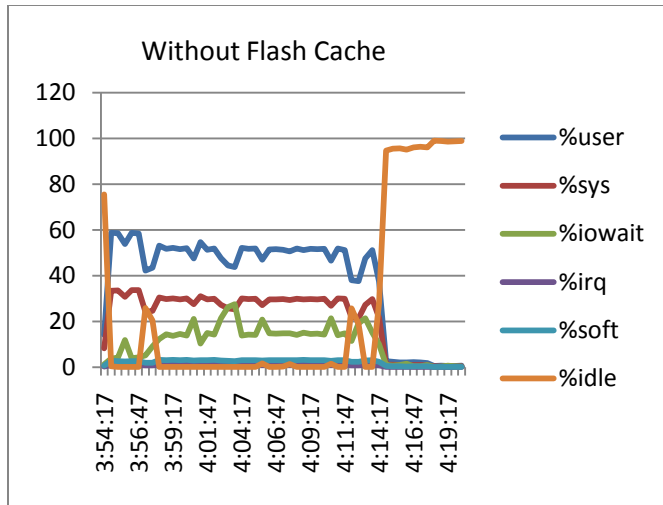


Figure 14

The read intensive workload benefited from using Flash Cache. The write intensive workload gained good performance after providing more disk bandwidth. There is room for further performance improvements through segment tuning of certain database table and index objects.

This configuration uses Flash Cache as an ASM device. There are Flash Cache configurations and size coverage that may provide better performance for read intensive workloads. There is a recommendation in Oracle documentation on how to size Flash Cache based on the size of the Oracle system global area.

Conclusion

A configuration using a two node Hitachi Compute Blade 2000 with a Hitachi Adaptable Modular Storage 2500 is an ideal server and storage solution for OLTP applications using Oracle 11gR2 RAC. Oracle supports Hitachi Compute Blade 2000 for Oracle 11gR2 RAC on Linux platform.

The solution works seamlessly with PCIe SSD cache for improving performance for certain types of workloads.

Using this reference architecture as a guide, you can design a larger infrastructure that meets your needs by adding the following:

- Additional blades in the Hitachi Compute Blade 2000 chassis
- Additional PCIe SSD cache
- Utilizing mezzanine HBA cards and Hitachi Compute Blade 2000 internal switches
- Additional disks and host ports on the Adaptable Modular Storage 2500
- Storage monitoring and database management through Hitachi storage plug-ins for Oracle Enterprise Manager

Hitachi Data Systems Global Services offers experienced storage consultants, proven methodologies and a comprehensive services portfolio to assist you in implementing Hitachi products and solutions in your environment. For more information, see the [Hitachi Data Systems Global Services](#) website.

Live and recorded product demonstrations are available for many Hitachi products. To schedule a live demonstration, contact a sales representative. To view a recorded demonstration, see the [Hitachi Data Systems Corporate Resources](#) website. Click the **Product Demos** tab for a list of available recorded demonstrations.

Hitachi Data Systems Academy provides best-in-class training on Hitachi products, technology, solutions and certifications. Hitachi Data Systems Academy delivers on-demand web-based training (WBT), classroom-based instructor-led training (ILT) and virtual instructor-led training (vILT) courses. For more information, see the [Hitachi Data Systems Academy](#) website.

For more information about Hitachi products and services, contact your sales representative or channel partner or visit the [Hitachi Data Systems](#) website.

 **Hitachi Data Systems Corporation**

Hitachi is a registered trademark of Hitachi, Ltd., in the United States and other countries. Hitachi Data Systems is a registered trademark and service mark of Hitachi, Ltd., in the United States and other countries. All other trademarks, service marks and company names mentioned in this document are properties of their respective owners.

Notice: This document is for informational purposes only, and does not set forth any warranty, expressed or implied, concerning any equipment or service offered or to be offered by Hitachi Data Systems Corporation

© Hitachi Data Systems Corporation 2011. All Rights Reserved. AS-106-00 September 2011

Corporate Headquarters

750 Central Expressway,
Santa Clara, California 95050-2627 USA
www.HDS.com

Regional Contact Information

Americas: +1 408 970 1000 or info@hds.com
Europe, Middle East and Africa: +44 (0) 1753 618000 or info.emea@hds.com
Asia Pacific: +852 3189 7900 or hds.marketing.apac@hds.com