HITACHI
Inspire the Next

Deploying Oracle Real Application Clusters 11g on the Hitachi Adaptable Modular 2000 Family

For Oracle Enterprise Linux

By Dhiren J. Patel

January 2009

Executive Summary

Businesses juggling increasing demands for processing power from mission-critical applications like Oracle Real Application Clusters (RAC) 11g and from increasing numbers of users who need 24/7 access to data, require solutions that increase scalability, availability and performance. Database and storage administrators know that the scalability and performance of applications are dependent upon the scalability and performance of the underlying storage system.

The Hitachi Adaptable Modular Storage 2000 family provides a reliable, flexible, scalable and cost-effective modular storage system for Oracle RAC 11g. The 2000 family is ideal for demanding application requirements, delivers enterprise-class performance, capacity and functionality at a midrange price, and is the only midrange storage product with symmetric active-active controllers that provide integrated, automated hardware-based front-to-back-end I/O load balancing.

This document provides a reference configuration and guidance on deploying RAC 11g to organizations deploying applications on 2000 family storage systems. It is written for database administrators, system administrators and storage administrators with moderate levels of experience with Oracle RAC.

Contributors

The information included in this document represents the expertise, feedback and suggestions of a number of skilled practitioners. The author recognizes and sincerely thanks the following contributors and reviewers of this document:

- Patricia Brailey, SCSSI — Solutions Engineering
- Steve Burr, Global Solutions Services
- Eduardo Freitas, SCSSI — Solutions Engineering
- Francois Zimmermann, UK Global Solution Services

Table of Contents

Deploying Oracle Real Application Clusters 11g on the Hitachi Adaptable Modular 2000 Family

For Oracle Enterprise Linux

By Dhiren J. Patel

For businesses juggling increasing demands for processing power from mission-critical applications like Oracle Real Application Clusters (RAC) 11g and from increasing numbers of users who need 24/7 access to data, it's more important than ever to find solutions that increase scalability, availability and performance. Database and storage administrators know that the scalability and performance of applications are dependent upon the scalability and performance of the underlying storage system.

As the number of nodes or users in a RAC environment increase, the storage system must provide concurrent access to the existing physical disks without compromising performance. In addition, resource coordination within an Oracle RAC environment must remain optimal as the number of nodes increase within a cluster. This means that the global resource directory must coordinate resources across all nodes in a cluster.

The Hitachi Adaptable Modular Storage 2000 family offers all of these advantages to businesses that deploy Oracle Real Application Clusters 11g. Hitachi Dynamic Load Balancing Controllers and other features of the 2000 family models provide storage and database administrators the ability to meet increasing demand for capacity and to scale applications while avoiding performance bottlenecks.

This document provides a reference configuration and guidance on deploying RAC 11g to organizations deploying applications on 2000 family storage systems. It is written for database administrators, system administrators and storage administrators who have at least moderate levels of experience with Oracle RAC.

Hitachi Adaptable Modular Storage 2000 Family Features

The Adaptable Modular Storage 2000 family provides a reliable, flexible, scalable and cost-effective modular storage system for Oracle RAC 11g. The 2000 family is ideal for more demanding application requirements and delivers enterprise-class performance, capacity and functionality at a midrange price.

The 2000 family is the only midrange storage product with symmetric active-active controllers that provide integrated, automated hardware-based front-to-back-end I/O load balancing.

This ensures I/O traffic to back-end disk devices is dynamically managed, balanced and shared equally across both controllers, even if the I/O load to specific logical units (LUs) is skewed. Storage administrators are no longer required to manually define specific affinities between LUs and controllers, simplifying overall administration. In addition, this new controller design is fully integrated with standard host-based multipathing, thereby eliminating mandatory requirements to implement proprietary multipathing software.

No other midrange storage product that scales beyond 100TB has a serial-attached SCSI (SAS) drive interface. The new point-to-point back-end design virtually eliminates I/O transfer delays and contention associated with Fibre Channel arbitration and provides significantly higher bandwidth and I/O concurrency. It also isolates any component failures that might occur on back-end I/O paths.

Flexibility

- Choice of Fibre Channel and iSCSI server interfaces or both
- Resilient performance using LUs that can be configured to span multiple drive trays and back-end paths
- Choice of high-performance SAS and low-cost SATA disk drives
- Flexibility of using SAS or SATA drives that can be intermixed in the same tray
- Support for all major open systems operating systems, host bus adapters (HBAs) and switch models from major vendors

Scalability

- Ability to add capacity, connectivity and performance as needed
- Concurrent support of large heterogeneous open systems environments using up to 2048 virtual ports with host storage domains and 4096 LUs
- Ability to scale capacity to 472TB
- Ability to scale performance to more than 900K IOPS
- Seamless expansion due to data-in-place upgrades from Adaptable Modular Storage 2100 to Adaptable Modular Storage 2300 and to Adaptable Modular Storage 2500

Availability

- Outstanding performance and non-disruptive operations using Hitachi Dynamic Load Balancing Controller
- 99.999% data availability
- No single point of failure
- Hot swappable major components
- Dual-battery backup for cache
- Non-disruptive microcode updates
- Flexible drive sparing with no copy back required after a RAID rebuild
- Host multipathing capability
- Remote site replication
- RAID-5, RAID-1, RAID-1+0 and RAID-0 (SAS drives) support
- RAID-6 dual parity support for enhanced reliability when using large SATA and SAS drives

Performance

- No performance bottlenecks in highly utilized controllers due to Hitachi Dynamic Load Balancing Controller
- Point-to-point SAS backplane with a total bandwidth of 96Gbps and no overhead from loop arbitration
- Full duplex 3Gbps SAS drive interface that can simultaneously send and receive commands or data on the same link
- Up to 32 concurrent I/O paths provide up to 9600 megabytes per second of total system bandwidth
- 4Gbs host Fibre Channel connections
- Cache partitioning and cache residency to optimize or isolate unique application workloads

Simplicity

- Simplified RAID group placement using SAS backplane architecture
- Highly intuitive management software that includes easy-to-use configuration and management utilities
- Command line interface (CLI) and command control interface (CCI) that match GUI functionality
- Seamless integration with Hitachi storage systems, managed with a single set of tools using Hitachi Storage Command Suite software
- Consistency among most Hitachi software products whether run on Hitachi modular storage systems or Hitachi Universal Storage Platform™ systems

Hitachi Software

Several Hitachi software applications provide a foundation to deploy highly available business applications with Oracle Real Application Cluster. These management applications allow administrators to easily provision storage using automated wizards and robust command-line interfaces, and to provide optimum storage performance and availability.

Hitachi Dynamic Link Manager Advanced Software

Hitachi Dynamic Link Manager Advanced software consists of Hitachi Dynamic Link Manager multipathing software and Hitachi Global Link Manager software.

Hitachi Dynamic Link Manager Software

Hitachi Dynamic Link Manager software enhances the system's availability and reliability by ensuring failover of an access path to another path in the event of a fault on the path. It includes these features:

- **Load balancing** — When multiple access paths connect a host to a storage system, Dynamic Link Manager software distributes the load over all the access paths. This avoids data access congestion on a specific access path.
- **Fault tolerance** — If a fault occurs on an access path when multiple access paths connect a host to a storage system, Dynamic Link Manager software automatically switches to another access path. This allows the process to continue without being affected by the fault.
- **Automatic failback** — Dynamic Link Manager software automatically restores an access path online when it recovers from a fault. Usually the administrator restores a failed access path at a convenient time that minimizes the negative effect on business. This means that failback is usually performed manually. However, it is also possible to perform failback automatically with Dynamic Link Manager software. When automatic failback is activated, Dynamic Link Manager software automatically activates the path after the user solves the physical route problem.
- **Automatic path health check** — Dynamic Link Manager software detects faults by checking the status of access paths at user-defined intervals. This allows path errors to be identified and corrected.
- **I/O load-balancing algorithms** — Dynamic Link Manager software includes the following load-balancing algorithms that are especially suited for Hitachi storage systems:
  - Round robin
  - Extended round robin
  - Least I/O
  - Extended least I/O
  - Least blocks
  - Extended least blocks

Dynamic Link Manager software manages multiple access paths to storage subsystems, provides functionality for distributing the load across paths and switching to another path a failure occurs on a path being used, thus improving system availability and reliability.

Hitachi Global Link Manager Software

Hitachi Global Link Manager software controls access paths to storage systems. It consolidates, simplifies and enhances the management, configuration and reporting of multipath connections between servers and storage systems. Global Link Manager software manages all of the Dynamic Link Manager installations in an environment. Use Global Link Manager software to monitor all the connections to an Adaptable Modular Storage 2000 system and to report on those connections. Global Link Manager software also enables administrators to configure load balancing at a per-LU level. This functionality is important when the load-balancing algorithm must be set at the LU level to accommodate diverse I/O patterns.

Hitachi Storage Navigator Modular 2

Hitachi Storage Navigator Modular 2 is part of the Hitachi Basic Operating System software package. It manages and monitors Hitachi modular storage systems, including the Adaptable Modular Storage 2000 family, through either a GUI or a command-line interface. Use Storage Navigator to create RAID groups and logical units and to assign those logical units to hosts. Storage Navigator is also useful for monitoring events and status of the various components on a modular storage subsystem, including performance metrics such as IOPS, processor and cache utilization, among others.

Oracle Product Family

For deploying large, highly available business applications, Oracle provides a critical set of software to build a robust application infrastructure, including a Linux-based operating environment and a database with several features and options that provide storage management, clustering required for deployment of application infrastructure.

Oracle Enterprise Linux

Oracle Enterprise Linux is an open source operating system that is compatible with Red Hat Enterprise Linux. Oracle Enterprise Linux provides a low-cost, high-performance operating environment for deploying business-critical applications and Oracle RAC in enterprise environments.

Oracle Database 11g

Oracle Database 11g is a relational database management system released in 2007. It increases availability by protecting against human and other errors, and also offers security, scalability, reliability and performance. Oracle Database 11g has the following features and options:

Oracle Real Application Clusters

Oracle's Real Application Clusters (RAC) option for Oracle Database 11g facilitates deployment of one database across a cluster of servers. It provides for continued operation in situations like hardware failures and planned outages. RAC also allows administrators to scale applications without taking users offline.

Oracle Clusterware

Oracle Clusterware software groups multiple servers so that they function as one server, facilitating failure detection, recovery, messaging and locking. Use it to monitor, move and restart applications and to automatically manage all Oracle processes. Clusterware is part of Oracle Database 11g.

Oracle Automatic Storage Management

Oracle Database 11g's Oracle Automatic Storage Management (ASM) feature is a simple, consistent storage management interface, regardless of the server or storage platform on which it is implemented. ASM allows DBAs to manage database environments efficiently. ASMLib, a support library for Oracle ASM, allows DBAs to manage disk groups on Linux.

Reference Architecture

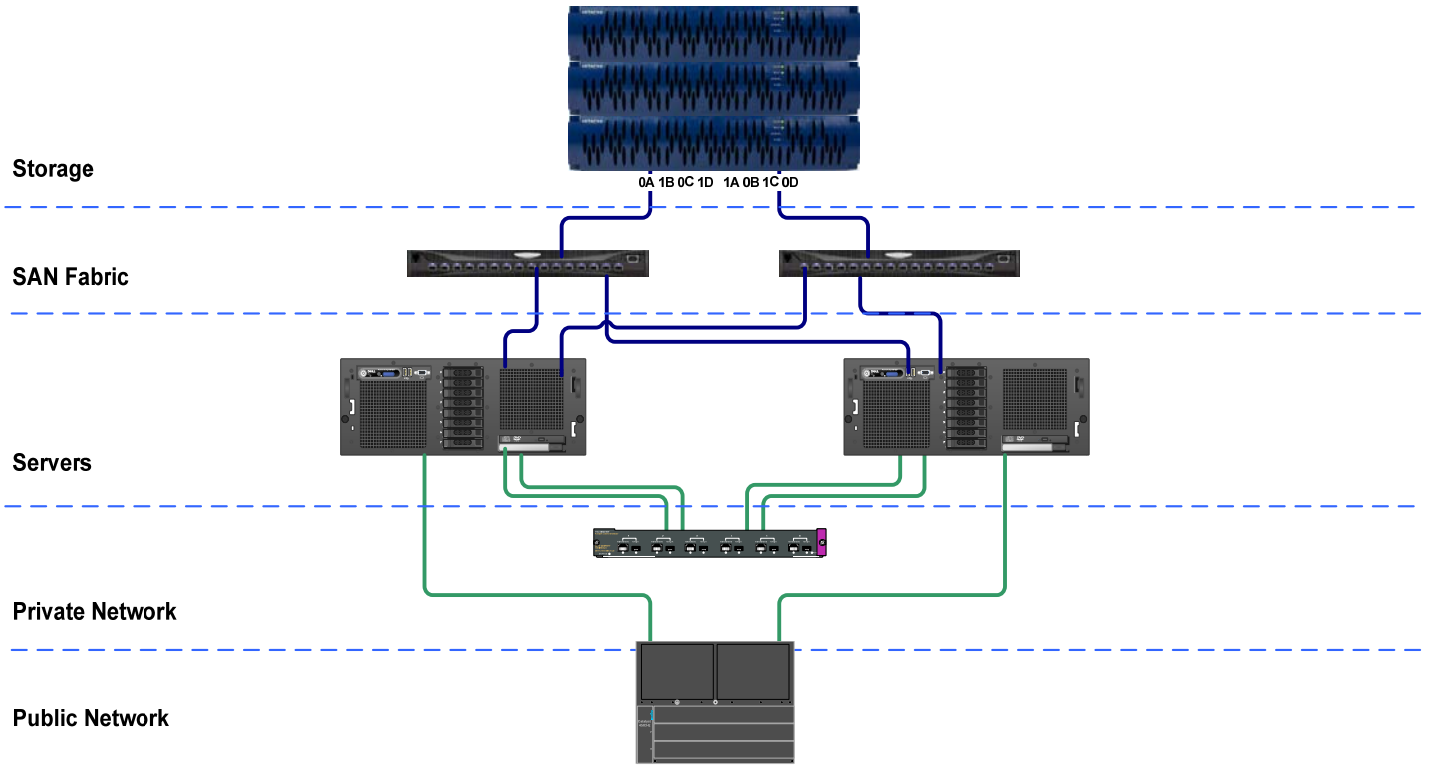
Figure 1 illustrates a reference configuration for a large application with an OLTP database that requires high availability, scalability and optimal performance for a heavy workload. This reference configuration also includes on-demand provisioning of additional servers and storage to make applications highly scalable without any code changes to the application or database.

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

- **Storage system** — Adaptable Modular Storage 2300 with two controllers, each with 4GB of cache and four front-end ports (host connectors). The storage system has five trays with total storage capacity of approximately 14TB.
- **SAN fabric** — Two 4GB/s Fibre Channel switches.
- **Servers** — Two servers each with four quad core processors of 1.9GHZ, 64GB memory, two 4/Gb/s Fibre Channel HBAs and two 1 Gb/s NICs.
- **Network switches** — Two network switches each with at least 1Gb/s transfer rate.

Figure 1 illustrates the reference architecture used in testing for this solution.

Figure 1. Oracle RAC on Adatable Modular Storage 2000 Reference Architecture



Recommended Architecture

When deploying Oracle RAC with Adaptable Modular Storage 2000, Hitachi Data Systems recommends that you use the following minimum hardware and software components. These recommendations apply to deployment and you can modify resources based on your application needs.

Hardware

A resilient and scalable hardware infrastructure is required for implementing large applications that include Oracle RAC. Hardware components for the recommended architecture include the following shared storage, servers and network components:

- **Shared storage** — Provide a robust shared storage infrastructure encompassing redundant switches and a mid-range I/O subsystem that can scale to handle growth and service multiple terabytes of database and application data. Although Oracle RAC is supported on SAN, NAS and iSCSI based devices, Hitachi Data Systems recommends Fibre Channel-SAN based storage infrastructure, as follows:
  - **Storage system** — The Adaptable Modular Storage 2000 family storage system provides the right level of scalability, availability and performance for applications and databases. The 2000 family can support raw devices, block devices, Oracle Cluster File System version 2 and Oracle Automated Storage Management for Oracle Real Application Cluster.
  - **SAN fabric** — In a shared Fibre Channel storage environment, a highly available Fibre Channel-SAN fabric pair made up of a 4Gb/s Fibre Channel switch or a Fibre Channel director provides adequate availability, performance and connectivity within the shared storage infrastructure required for business applications and Oracle RAC database.
- **Server** — Provide a cluster infrastructure that scales not only within a machine but also across several machines. To avoid single points of failure, the servers within a cluster must handle multiple CPUs, NICs and Fibre Channel HBAs, among other components. Use a server with the following characteristics:
  - At least two CPUs and enough speed to handle workload of the application
  - A minimum of 16GB of memory based on the workload of the application
  - Two qualified 4Gb/s Fibre Channel HBAs
  - Four 1Gb/s NICs
- **Network** — Two separate networks, a private network and a public network are required. The private network, also referred as *cluster interconnect*, must be scalable and must meet the low latency needs of the network traffic generated by cache synchronization of Oracle Real Application Clusters and inter-node communication among the nodes in the cluster. The public network is used for client connections to the applications and Oracle Real Application Clusters. Hitachi Data Systems recommends using a pair of 1Gb/s NICs for the cluster interconnect as well as for the public network.

Software

It is imperative to have adaptable and scalable software to implement large applications that include Oracle RAC. Various software applications for providing operating environment, I/O path management, database management, storage management and high availability are needed:

- **Operating system** — Oracle Enterprise Linux provides enterprise class low-cost, high performance operating environment for deploying business critical application and Oracle RAC. For maximum benefit, Hitachi Data Systems recommends using Oracle Enterprise Linux 5.1 64bit.
- **Hitachi Dynamic Link Manager** — To create a highly availably application infrastructure, you must deploy I/O multipathing as part of the storage and server infrastructure. Hitachi Dynamic Link Manager software provides high availability and performance between host and the I/O storage subsystem.
- **Hitachi Storage Navigator Modular 2** — Manages and monitors the 2000 family I/O subsystem using a GUI or CLI. Hitachi Storage Navigator Modular 2 comes with the 2000 family.

- **Oracle Database 11g** — Use the following versions of Oracle Database software and its associated options:
  - Oracle Database 11g Enterprise Edition Release 11.1.0.6.0 64bit
  - Oracle Real Application Cluster option
  - Oracle Clusterware 11.1.0.6
  - Oracle Automatic Storage Management with ASMLib

Deployment Considerations

The reference architecture introduced in this paper includes a wide variety of technologies, each of which requires special considerations individually as well as in conjunction with one or more technologies in the reference architecture. Deployment of the reference architecture involves configuring few key components, including storage, servers, networks and Oracle software and its associated features and options. These components require special consideration when they are used to build Oracle Real Application Clusters.

Storage Considerations

Various technologies in the storage stack are key for providing robust storage infrastructure for deployment of large applications with Oracle RAC. It is important to choose the right level of RAID for the database storage, the right type of disk driver and to manage I/O paths to achieve optimal performance and higher availability. Keep the following consideration in mind as you plan your storage infrastructure.

RAID

With the Adaptable Modular Storage 2000 family, physical disks are formatted into RAID groups. A slice created from one or more of these RAID group is presented to the host or hosts as a LU. Choose the RAID group configuration carefully based on the I/O characteristics of the application data that is to be stored in the database. Hitachi Data Systems recommends using RAID-1+0 for majority of the Oracle file types. Table 1 provides guidelines on implementing the RAID levels for various types of Oracle files managed by ASM disks.

Disk Drive

Selecting a drive type is as important as selecting a RAID type for a storage implementation. With the introduction of the 2000 family, users can choose between SAS and SATA drives to meet their applications' performance or capacity requirements. Due to performance and availability requirements, Hitachi Data Systems recommends the use of SAS drives for deployment of large applications with Oracle RAC.

I/O Path

The 2000 family's Dynamic Load Balancing Controller architecture allows host access to a LU from any of the front-end host ports on the two controller modules. This type of design is commonly referred to as symmetrical active-active. To achieve optimal performance from this symmetrical active-active design and subsequently achieve optimal database performance, use all front-end ports on each of the controller of the 2000 family of storage systems.

Hitachi Data Systems recommends the use of dual SAN fabrics, multiple HBAs and host-based multipathing software when deploying the reference architecture. At least two paths from the database hosts connecting to two independent SAN fabrics and two paths from each SAN fabric to two different controllers of the I/O subsystem is essential for ensuring the redundancy required for critical applications.

Hitachi Dynamic Link Manager is a critical component of the reference architecture. It enables hosts to see and access multiple paths to the same LU, enabling data to travel any available path for increased performance or continued access to data in the case of a failed path. For deployment of large applications with Oracle RAC, Hitachi Data Systems recommends using the round robin load-balancing algorithm to distribute load evenly over all available HBAs.

When using Hitachi Dynamic Link Manager software in an Oracle RAC environment with Oracle ASM and ASMLib, change the value of `ORACLEASM_SCANORDER` parameter to `ORACLEASM_SCANORDER="sddlm"` in the `/etc/sysconfig/oracleasm` file.

In addition, you must change the MISSCOUNT and DISKTIMEOUT (only when four or more paths are connected to voting disks) parameters of Clusterware based on following formulas.

To change the MISCOUNT parameter, use this formula:

MISCOUNT = number of paths connected to voting disk x 60 seconds

For example, to set MISCOUNT to 120, use the following command:

```
$CRS_HOME/bin/crsctl set css misscount 120
```

To change the DISKTIMEOUT parameter, use this formula:

DISKTIMEOUT = number of paths connected to voting disk x 60 seconds

For example, to set DISKTIMEOUT to 240, use the following command:

```
$CRS_HOME/bin/crsctl set css disktimeout 240
```

Database Storage Considerations

An Oracle database has several types of files that store a wide variety of data. For each type of file, performance characteristics can vary according to the database function and database workload. Various types of files include datafiles, redo log files, archive log files, temporary tablespace files, control files and so on. Use Oracle Automatic Storage Management feature to manage these different file types. Oracle ASM has its own logical structures to store and manage database data optimally. In an Oracle RAC and Clusterware environment, storage is required for voting disks as well as for Oracle Cluster Registry (OCR) disks. In addition, a wide variety of database tuning parameters are available. Incorporate the following considerations when implementing the reference architecture:

- Based on the I/O characteristics, segregate Oracle objects into separate tablespaces, ASM disk groups, or both, wherever possible. For example, place user data in a different ASM disk group than indexes and place online redo logs in a separate ASM disk group apart from the ASM disk group containing user data.
- For storage management of database data of the reference architecture, use Oracle ASM with ASMLib libraries. For optimal throughput, use ASM to spread datafiles over as many spindles as possible. Stripe each ASM disk group over a number of LUs. Also use EXTERNAL REDUNDANCY while configuring ASM disk groups.
- Create at least three online redo log files. Multiplex the online redo log files into a separate ASM disk group.
- Based on the reference architecture, follow the configuration layout of ASM disk groups shown in Table 1.

Table 1. ASM Disk Group Configuration Layout

ASM Disk Group	RAID Level	Application	Contents
DATADG1	RAID-1+0	System tablespaces SYSAUX tablespace USERS tablespace	System and user data
DATADG2	RAID-1+0	USERS tablespace	User data (preferably for indexes)
REDODG1	RAID-1+0*	Thread 1 Redo Log Group 1 Member 1 Thread 1 Redo Log Group 2 Member 1 Thread 1 Redo Log Group 3 Member 1	One member of each redo log group of one node of the RAC
REDODG2	RAID-1+0*	Thread 2 Redo Log Group 1 Member 1 Thread 2 Redo Log Group 2 Member 1 Thread 2 Redo Log Group 3 Member 1	One member of each redo log group of one node of the RAC
REDODG3	RAID-1+0*	Thread 1 Redo Log Group 1 Member 2 Thread 1 Redo Log Group 2 Member 2 Thread 1 Redo Log Group 3 Member 2	One member of each redo log group of one node of the RAC
REDODG4	RAID-1+0*	Thread 2 Redo Log Group 1 Member 2 Thread 2 Redo Log Group 2 Member 2 Thread 2 Redo Log Group 3 Member 2	One member of each redo log group of one node of the RAC
ARCHDG	RAID-5	Archive logs	Archived redo log files
TEMPDG	RAID-1+0	TEMP tablespace	TEMP tablespace

*Or RAID-5

- Hitachi Data Systems recommends using block devices for voting disks and OCR disks. Using block devices simplifies the solution and does not require any additional software. In addition, raw devices, an alternative to block devices, are deprecated (since Oracle Enterprise Linux 5).

Server Considerations

The primary consideration for the reference architecture is to provide scalability, availability and performance applications deployed on it. To meet this requirement, the hardware must be able to eliminate single points of failure and be able to scale up within an individual node of the cluster as well as scale out across a number of nodes in the cluster. Following are Hitachi Data System's recommendations:

- Each server of the cluster must be identical and must have the following characteristics:
  - At least two quad-core 64-bit servers
  - A minimum of 16GB of memory expandable up to 64GB

- To provide high availability and avoid single points of failure of the components, use redundant Fibre Channel HBAs, Fibre Channel switches or Fibre Channel directors and NICs (for public as well as private networks). Wherever possible, configure load balancing and failover among the redundant components.
- For optimal performance, ensure that each set of redundant components has identical characteristics. For example, ensure that all Fibre Channel HBAs have same transfer rates and other settings (for example, queue depth settings).
- For optimal performance, ensure that none of the components are bottlenecks either in the I/O path or in the data path, for example:
  - If a Fibre Channel HBA has a 4Gb/s transfer rate, the Fibre Channel switch and the I/O subsystem port must also have at least a 4Gb/s transfer rate.
  - The NICs for the private networks and the network switch must use the same transfer rate.

Network Considerations

You must have a robust, high-bandwidth private network and public network. The private network is used for cluster interconnect that is one of the key component to maintain synchronization among various nodes of the Oracle Clusterware and Oracle RAC. The public network has a critical role of enabling client connection to the application and the database. Keep the following points in mind when configuring private and public network in the your Oracle RAC environment:

- For each server in the Oracle RAC Clusterware configuration, use at least two identical, high bandwidth, low-latency NICs for the interconnect.
- Use NIC bonding to provide failover and load balancing of interconnects within a server.
- Set all NICs to full duplex mode.
- Use at least two public NICs for client connections to the application and database.

Summary

The Hitachi Adaptable Modular Storage 2000 family, with its newly introduced features like symmetrical active-active controllers and Hitachi Dynamic Link Manager software, enables customers to deploy highly available and scalable applications on Oracle Database 11g. The reference architecture using the 2000 family, Oracle Database 11g features or options such as Oracle Real Application Cluster, Oracle Automatic Storage Management along with reliable and scalable servers provide flexibility to provision and deprovision servers according to the performance and scalability requirements of business-critical applications.

Appendix A: References

For more information about related topics, see the following documents, which are available to registered users on the Hitachi Data Systems support portal:

- Hitachi Storage Command Suite Hitachi Dynamic Link Manager Software User's Guide for Linux
- Oracle Real Application Clusters Administration and Deployment Guide 11g Release 1 (11.1)
- Oracle Clusterware Installation Guide 11g Release 1(11.1) for Linux
- Best Practice Library – Deploying Oracle 11g RAC with Hitachi Enterprise-class Storage Systems

Training courses on the Hitachi Adaptable Modular Storage 2000 family are available from the Hitachi Data Systems Learning Center.

For additional information and contacts, see http://www.hds.com/.

Hitachi Data Systems Corporation

Corporate Headquarters 750 Central Expressway, Santa Clara, California 95050-2627 USA
Contact Information: + 1 408 970 1000 www.hds.com / info@hds.com

Asia Pacific and Americas 750 Central Expressway, Santa Clara, California 95050-2627 USA
Contact Information: + 1 408 970 1000 www.hds.com / info@hds.com

Europe Headquarters Sefton Park, Stoke Poges, Buckinghamshire SL2 4HD United Kingdom
Contact Information: + 44 (0) 1753 618000 www.hds.com / info.uk@hds.com

AS-003-00 January 2009