



Oracle 9i Real Application Clusters Using HyperFabric on Hewlett-Packard Platforms

Development Alliances Lab, Hewlett-Packard
June 2001

Introduction

In recent years, the requirement for highly available systems, able to scale on demand, has fostered the development of more and more robust cluster solutions. Prior to Oracle9i, HP and Oracle, with the combination of Oracle Parallel Server and HP ServiceGuard OPS edition, provided cluster solutions that lead the industry in functionality, high availability, management and services. Gartner Group's report on Cluster Solutions stated "HP has exhibited the strongest vision in high availability market with its combined high availability initiatives in servers, clustering management and services".

Now with the release of Oracle 9i Real Application Clusters (RAC) with the new Cache Fusion architecture based on an ultra-high bandwidth, low latency cluster interconnect technology, RAC cluster solutions have become more scalable without the need for data and application partitioning.

Hewlett-Packard Company and Oracle Corporation have been working closely on this combined cluster solution over the past three years. Their efforts have focused on RAC, ServiceGuard OPS edition and the cluster interconnect technology. This white paper describes the technical solutions provided by the two companies as well as the testing and implementation efforts leading to the release of Oracle 9i Real Application Clusters.

What is a cluster?

A cluster is a group of independent systems, which perform as a single system. A cluster consists of a number of servers or nodes, a cluster interconnect component and possibly a set of shared disks.

The server component in the cluster can have multiple processors, has its own memory, operating system, database instance as well as application software. The database instances in the cluster can share all data or nothing residing on the shared disks.

Clusters provide excellent solutions for high availability via redundancy of servers, system interconnects, disks, as well software components. This mechanism avoids single points-of-failure thus reducing both planned and unplanned downtime.

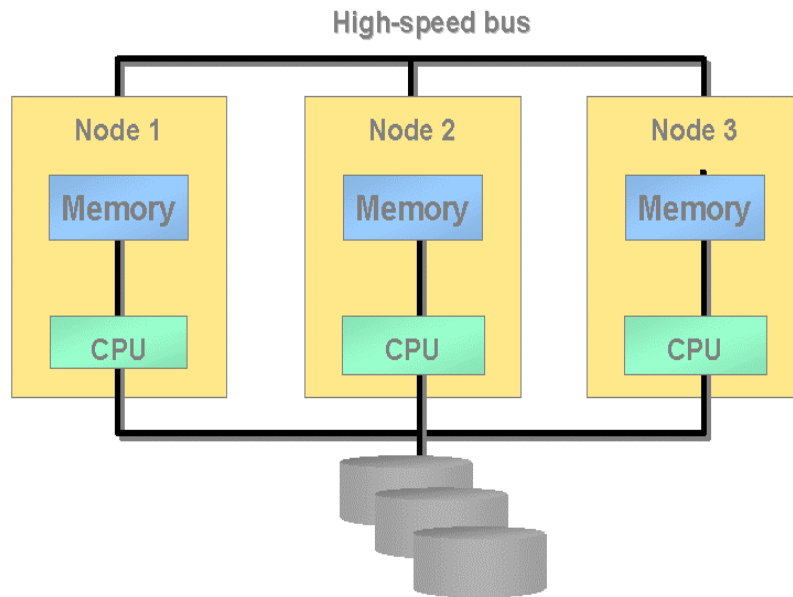
Current cluster solutions

Instances in the clustered database can either share all or nothing in the shared disks.

Shared data

All nodes in the cluster have the same access to the data on the shared disks. The nodes have the same ability to process the data – read, update or insert. This approach provides an excellent solution for highly available systems since the data is accessible even with only one available node. The efficiency of this approach depends greatly on the inter-node communication mechanism. Scalability in general is an issue with this type of cluster when compared with symmetric multiprocessing systems (SMP). When two or more nodes contend for the same data block the node that has the lock on the data block is forced to write it to disk before releasing the block to the other nodes. This operation of forcing a disk write is a slow process, which also requires message synchronization between nodes about the status of the data block. The combination of these activity results in limited scalability of shared disk cluster architectures.

Cluster scalability can be improved by data partitioning; however, this approach requires the client application to have good knowledge about the data. And over time, data analysis and data re-partitioning are necessary or the cluster performance will greatly degrade resulting in reduced system availability.

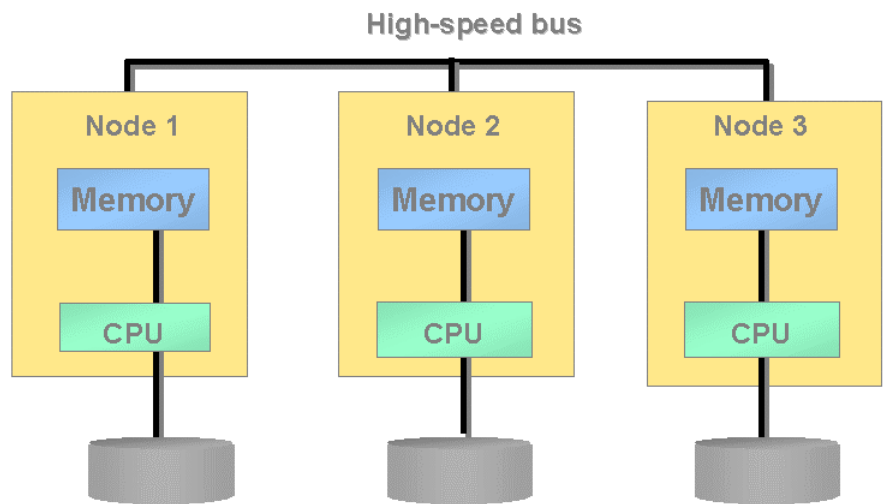


Clustered database with shared disk

Shared Nothing

In this setup, the database files are partitioned among the instances in the cluster. Each database instance owns or controls a separate set of data and has access to this data. In this architecture, data processing is done entirely by the instance owning the data. To support this environment, the data ownership is changed infrequently. Scalability of such clusters is excellent when the data partitioning is done correctly. High availability of this type cluster is constrained since each set of disks requires physical connectivity to two nodes. A single node failure will cause significant performance degradation since the surviving node will process twice the work previously assigned to it.

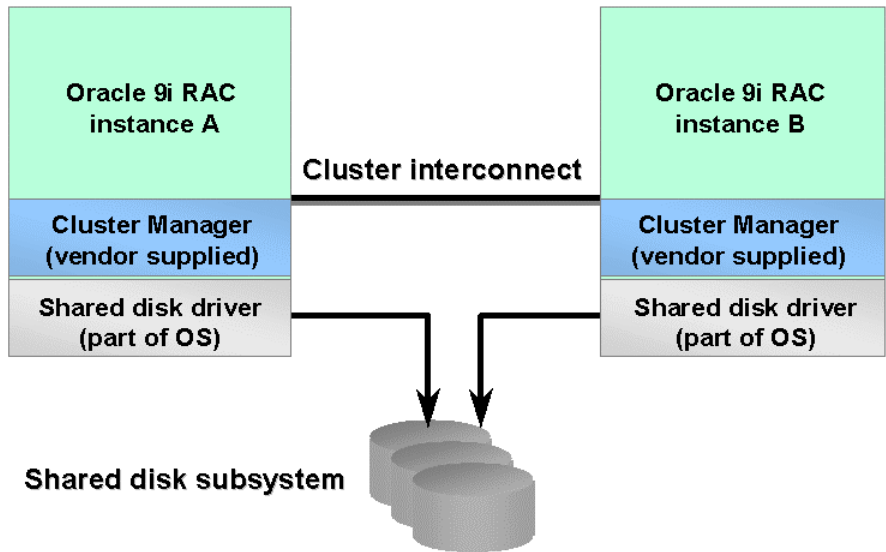
Shared nothing clustered database is suitable for Decision Support or Data Warehousing applications, which involve a large but mostly-static database with capability to process lots of complex and lengthy queries by a small number of users. Shared disk database architecture is commonly used for OLTP applications which require not only high availability but also ability to grow dynamically in term of its database size and number of users. Oracle Parallel Server uses the shared disk database architecture; its strength has been in the area of high availability as well as scalability. In Oracle 9i, Oracle Parallel Server renamed to Real Application Clusters continues to improve in these areas as well as functionality and usability.



Shared nothing clustered database

Oracle 9i Real Application Clusters – Cache Fusion technology

9i RAC Cache Fusion technology

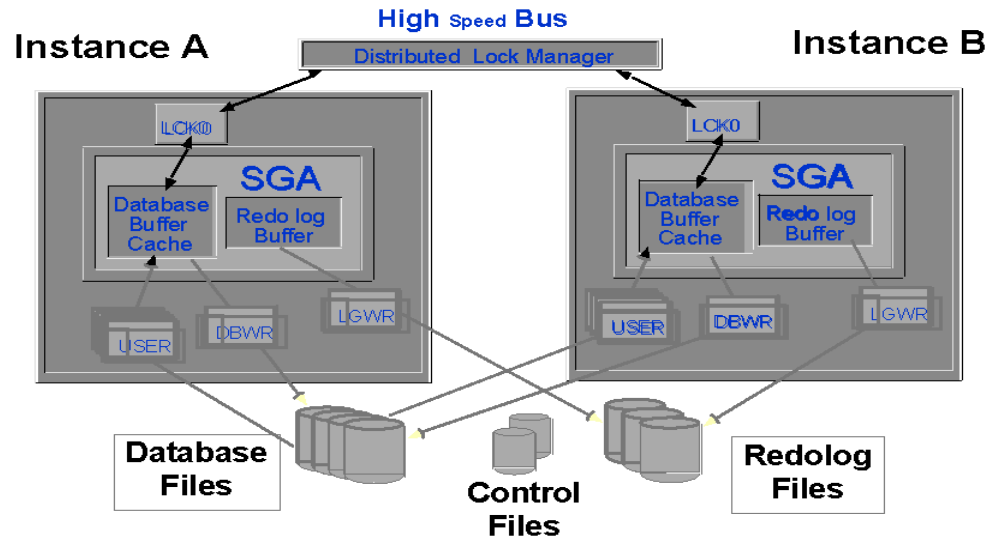


Oracle 9i cache fusion utilizes the collection of caches made available by all nodes in the cluster to satisfy database requests. Requests for a data block are satisfied first by a local cache, then by a remote cache before a disk read is needed. Similarly, update operations are performed first via the local node and then the remote node caches in the cluster, resulting in reduced disk I/O. Disk I/O operations are only done when the data block is not available in the collective caches or when an update transaction performs a commit operation.

Oracle 9i cache fusion thus provides Oracle users an expanded database cache for queries and updates with reduced disk I/O synchronization which overall speeds up database operations.

However, the improved performance depends greatly on the efficiency of the inter-node message passing mechanism, which handles the data block transfers between nodes.

9i RAC architecture



The efficiency of inter-node messaging depends on three primary factors:

- **The number of messages required for each synchronization sequence.** Oracle 9i's Distributed Lock Manager (DLM) coordinates the fast block transfer between nodes with two inter- node messages and one intra-node message. If the data is in a remote cache, an inter-node message is sent to the Lock Manager Daemon (LMD) on the remote node. The DLM and Cache Fusion processes then update the in-memory lock structure and send the block to the requesting process.
- **The frequency of synchronization (the less frequent the better).** The cache fusion architecture reduces the frequency of the inter-node communication by dynamically migrating locks to a node that shows a frequent access pattern for a particular data block. This dynamic lock allocation increases the likelihood of local cache access thus reducing the need for inter-node communication. At a node level, a cache fusion lock controls access to data blocks from other nodes in the cluster.
- **The latency of inter-node communication.** This is a critical component in Oracle 9i RAC as it determines the speed of data block transfer between nodes. An efficient transfer method must utilize minimal CPU resources, support high availability as well as highly scalable growth without bandwidth constraints.

HP Cluster Interconnect technology

HyperFabric

HyperFabric is a high-speed cluster interconnect fabric that supports both the industry standard TCP/UDP over IP and HP's proprietary Hyper Messaging Protocol (HMP). HyperFabric extends the scalability and reliability of TCP/UDP

by providing transparent load balancing of connection traffic across multiple network interface cards (NICs) and transparent failover of traffic from one card to another without invocation of MC/ServiceGuard. The HyperFabric NIC incorporates a network processor that implements HP's Hyper Messaging Protocol and provides lower latency and lower host CPU utilization for standard TCP/UDP benchmarks over HyperFabric when compared to gigabit Ethernet. Hewlett-Packard released HyperFabric in 1998 with a link rate of 2.56 Gbps over copper. In 2001, Hewlett-Packard released HyperFabric 2 with a link rate of 4.0 Gbps over fiber with support for compatibility with the copper HyperFabric interface. Both HyperFabric products support clusters up to 64-nodes.

HyperFabric Switches

Hewlett-Packard provides the fastest cluster interconnect via its proprietary HyperFabric switches, the latest product being HyperFabric 2, which is a new set of hardware components with fiber connectors to enable low-latency, high bandwidth system interconnect. With fiber interfaces, HyperFabric 2 provides faster speed – up to 4Gbps in full duplex over longer distance – up to 200 meters. HyperFabric 2 also provides excellent scalability by supporting up to 16 hosts via point-to-point connectivity and up to 64 hosts via fabric switches. It is backward compatible with previous versions of HyperFabric and available on IA-64, PA-RISC servers.

Hyper Messaging Protocol (HMP)

Hewlett-Packard, in cooperation with Oracle, has designed a cluster interconnect product specifically tailored to meet the needs of enterprise class parallel database applications. HP's Hyper Messaging Protocol significantly expands on the feature set provided by TCP/UDP by providing a true Reliable Datagram model for both remote direct memory access (RDMA) and traditional message semantics. Coupled with OS bypass capability and the hardware support for protocol offload provided by HyperFabric, HMP provides high bandwidth, low latency and extremely low CPU utilization with an interface and feature set optimized for business critical parallel applications such as Oracle 9i RAC.

Oracle 9i RAC tests/implementation on HP platforms

Oracle chose Hewlett-Packard as one of primary platforms to thoroughly test and deliver Oracle 9i RAC. The objectives of the testing have been:

- Ensuring high quality and robustness of Oracle 9i RAC by exhaustively testing the product on HP platforms before general release.
- Proving that large cluster configurations up to 16 nodes of Oracle 9i RAC are stable and resilient under stress.

- Proving the improved scalability of Oracle 9i RAC using industry standard benchmark kits such as Oracle Applications 11i standard benchmark and OLTP like workload.
- Certifying Oracle Applications 11i with Oracle 9i RAC.
- Proving Oracle and HP together provide an industry leading cluster solution with technology differentiators.

Test Methodology

- Simulate customer environments using the supported software stack and hardware configuration.
- Emulate customer installations using client, mid-tier servers as well as server variations.
- Tests are run against an increasing number of cluster nodes to ensure reliable results, starting with a small configuration and increasing to a larger configuration.
- Stress test all of Oracle 9i RAC functionality in an integrated, organized manner.

Test Details

The test framework includes a test driver, which is run against any cluster configuration. It is easy to setup, use and to vary workload as well as fault injection. It can handle both planned and unplanned cluster membership changes.

Individually, the tests include the following:

- **Stress tests** involve a number of orderly and random repetitions of Oracle start-ups of one or more database instances, addition and removal of workload, and shutdowns. The workload is a mixture of different types of OLTP transactions such as SELECT, INSERT, UPDATE, DELETE and VERIFY. Cluster configurations are tested via thousands of connections through both non-dedicated and dedicated MTS modes. Cache Fusion stress tests include testing of exclusive forced pinging, handling of past buffer images and forced state transition.
- **Recovery tests** include testing of block recovery caused by death of database instances, lost blocks, and cancellation by Control-C.
- **Destructive tests** include forced failures by software and hardware while the system is running with either minimal or high workload.
Oracle software - one or more of Oracle background processes is killed manually.
OS software – one or more of the cluster daemons is killed manually or the system is forced to reboot.
Hardware: Manual removal of network or disk connectivity or power supply.

Software Stack

- HP-UX 11.0
- ServiceGuard OPS Edition.
- Oracle 9i Enterprise Edition.
- User Datagram Protocol (UDP) and Hyper Messaging Protocol (HMP) for cluster interconnect.

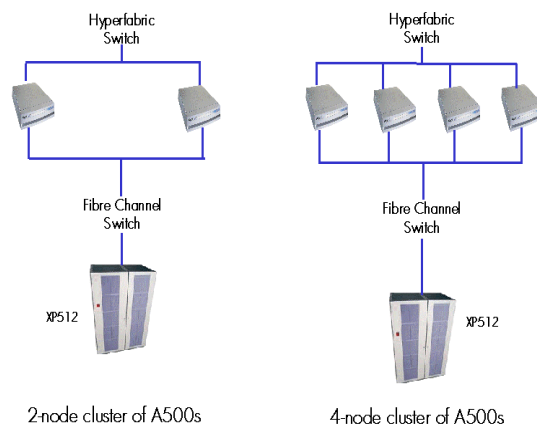
Hardware configurations

Oracle 9i RAC stress and destructive test environment

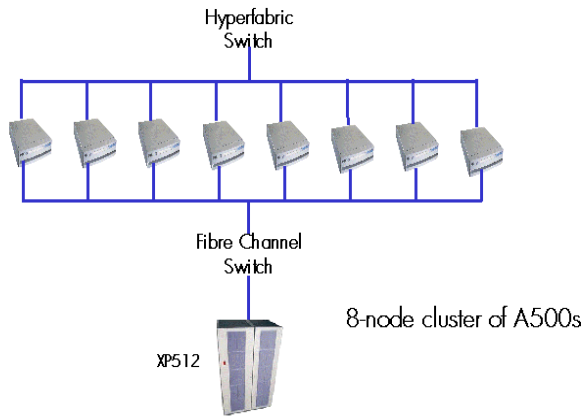
Thirty-four A500 systems were used to create a number of cluster configurations with sizes ranging from 2-nodes up to 16 nodes. A highlight of this environment was the use of HP HyperFabric switches for cluster interconnects. HP Surestore Disk Array XP512 provided the shared disks between the HP-UX servers as well as the HP Netervers (Windows 2000) used in the second test environment.

The following diagrams show the cluster configuration tested 2, 4, 8, 12 and 16 nodes. The HyperFabric switch can support up to 16 hosts, so in the case of smaller configurations such as 2 and 4-node clusters, these clusters share the same HyperFabric switch.

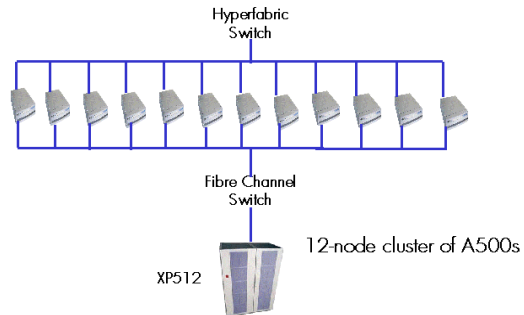
HP/Oracle 9i RAC test configuration



HP/Oracle 9i large cluster test configuration

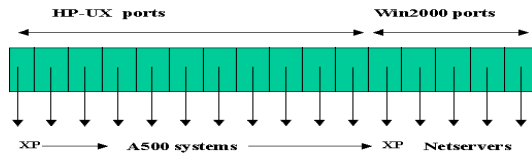


HP/Oracle 9i Large cluster test configuration



Role of HP SureStore E Disk Array XP512

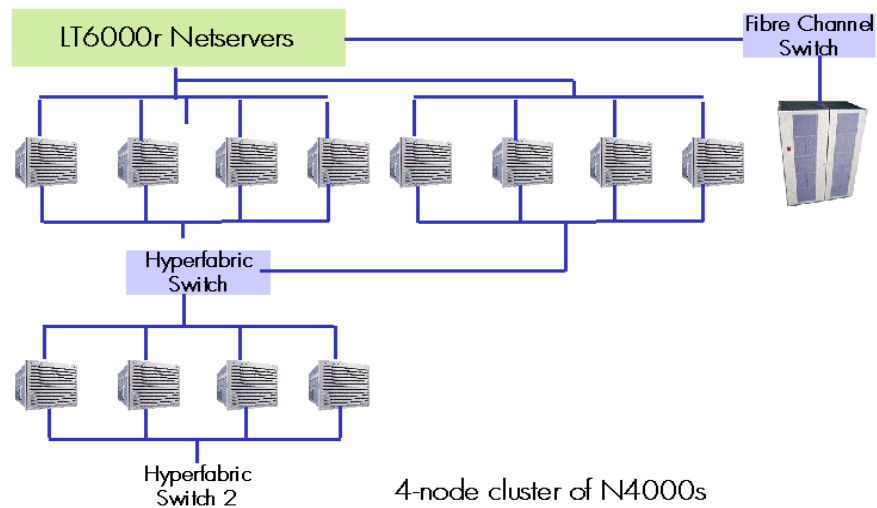
The following diagram shows the configuration of the 16-port SAN switch to accommodate the sharing of the XP512 between the HP-UX and the Netservers. The switch was zoned so that 11 ports were connected to the A500 systems and 5 ports were connected to the Netservers. Of the 11 A500 ports, 3 were connected to the XP512, the remaining 8 ports were connected to the A500 systems. For the Netservers, 1 port was used for the shared disks. A total of 8 16-port SAN switches were used in the test environment.



Oracle 9i scalability test environment.

- A cluster of 4 N4000 systems were connected via HP's fastest HyperFabric 2 switch.
- The database instances in this cluster shared 2 HP SureStore Disk Array FC60s via a SAN switch.
- The application servers layer uses 8 N4000 systems connecting to the database servers via HP HyperFabric 1 switch.
- 16 HP LT6000r (Windows 2000) were used by Load Runner to drive the Oracle Applications 11i benchmark. These Netserver's shared the same HP SureStore Disk Array, XP512, used by the A-class systems. The new Oracle Applications 11i benchmark kit was executed for this scalability test.

HP/Oracle 9i RAC test configuration



Product Specifications

This table below describes in details the specifications of HP products used in the Oracle 9i RAC test environment on HP Platforms.

A500 Servers	PA8500/440MHz 2 CPUs 4GB RAM 2x18GB internal disks 1.8GB/s memory bus	4 I/O slots 1.9GB/s I/O 64-bit HP-UX 11.0 +
N4000 database	PA8500/440MHz 8CPUs 24GB RAM 1MB memory cache	12 I/O slots 2x18GB Internal disks 64-bit HP-UX 11.0 +
N4000 Application	PA8600/550 MHz 8 CPUs 32GB RAM 1MB memory cache	12 I/O slots 2x32GB internal disks 64-bit HP-UX 11.0 +
Lt6000r Netservers	4 x 700 MHz CPU 4 GB RAM	4.8 GB disk Win 2000 Advanced Server
HyperFabric 1	PCI 4X Adapter Card 32/64-bit support 2.56 Gbps link rate Copper cables to 18m	Low CPU usage Support TCP/IP, SCN,HMP 64-bit HP-UX 11.0 Support HP MC/ServiceGuard 16 copper ports
HyperFabric 2	PCI 4X Adapter card with FC Fiber optic cables to 200m Interoperate with HF1 4 Gbps link rate	Management LAN port Support TCP/IP, HMP Hot swappable Adapter cards 16 fiber ports
XP512 SureStore	8 FC adapter cards, 4 ports each 18GB disks, over 2TB total 255 ldevs	8 ports for Netservers 24 ports for HP-UX servers 320 LUNs

Results/Findings

With the release of Oracle 9i Real Application Clusters, Oracle's and Hewlett-Packard's joint effort demonstrate that Oracle 9i RAC is a viable cluster solution which meets the scalability and high availability requirements of today's e-business environment.

- Tested on a wide range of HP Servers: A, L and N-classes.
- Support large cluster configurations from 2-nodes to 16-nodes.
- Support highly available, cost effective, multi-platform disk arrays.
- Successful concurrent usage of HP SureStore E Disk Array XP512 in the heterogeneous environment of HP-UX and Windows 2000 operating environments.
- Supports high bandwidth, low latency cluster interconnect.
- 200+ enhancements were made to Oracle 9i as a result of the testing effort.
- Stress tests were done on clusters up to 72 hours.
- Successful cluster reconfigurations during the destructive tests.
- Oracle Applications 11i certified on Oracle 9i RAC.
- 1.8 scalability with 2 N4000 database servers and 4 N4000 application servers using UDP cluster interconnect.
- 2.5 scalability with 3 N4000 database servers and 6 N4000 application servers using UDP cluster interconnect.
- 6.4 scalability with 8-node A500 cluster and 12.0 scalability with 16-node cluster using OLTP-like workload and UDP cluster interconnect.

At publication of this paper the scalability results of Oracle 9i RAC using HMP are not yet available. When these results are released, we expect even greater scalability of Oracle 9i RAC.

Summary

The result of the joint testing and implementation of Oracle 9i Real Application Clusters is a thoroughly tested and robust product. Using the standard benchmark kits to perform the scalability tests showed that the combination of Oracle 9i Cache Fusion and HP HyperFabric technologies give Oracle 9i Real Application Clusters dramatically improved scalability without the need for data and workload partitioning. Together, Hewlett-Packard and Oracle provide a high availability solution that has the power to scale and grow as the customer's business grows.

Acknowledgements

The author would like to acknowledge the following individuals for their contribution in the effort of development, testing, and support of Oracle 9i Real Application Clusters on Hewlett-Packard platforms:

Annie Chen, Kelly Huth, Hogan Flake, Alvin Tam, Alan Siu, Michael Izioumtchenko, Andy Yang, Stefan Pommerenk, Kotaro Ono, Quaid Hasta, Noel Rodriguez, Pete Arriaga, Larry Rogers, Veronique Mairesse, Juan-Garcia Rovetta, Sandy Gruver, Bill Cortright, Keith DeSilva, Yue-wen Chen and her team, Keshav Sharma and his team.

For more information about Hewlett-Packard MC/ServiceGuard OPS edition, check the website at

<http://www.hp.com/products1/unix/highavailability/>

Information about Hewlett-Packard servers can be found at

<http://www.hp.com/country/us/eng/prodserv.htm>

More information about Oracle 9i Real Application Clusters is available at

<http://technet.oracle.com/products/oracle9i/content.html>

HP-UX Release 11.00 and later (in both 32- and 64-bit configurations) on all HP 9000 computers are Open Group UNIX 95 branded products.

Oracle is a registered U.S. trademark and Oracle8i is a trademark of Oracle Corporation, Redwood City, California.