

Achieving Mainframe-Class Performance on Intel Servers Using InfiniBand Building Blocks

*An Oracle White Paper
April 2003*

Achieving Mainframe-Class Performance on Intel Servers Using InfiniBand Building Blocks

Executive Overview	1
Optimizing RAC Performance with InfiniBand	2
Latency	2
Bandwidth.....	3
CPU Utilization.....	3
InfiniBand's Advantage: RDMA.....	4
RDMA in Oracle 9i Real Application Clusters.....	4
Deploying Oracle Clusters over InfiniBand	5
Additional Benefits from I/O Consolidation	5
InfiniBand System Availability	6
Summary	6
About Oracle Real Application Clusters (RAC)	7
Oracle Real Application Clusters on Linux	7

Achieving Mainframe-Class Performance on Intel Servers Using InfiniBand Building Blocks

EXECUTIVE OVERVIEW

Information technology managers are struggling to balance the need for application performance, scalability, and availability with the reality of flat or shrinking budgets. To address this challenge, more and more of them are turning to open operating systems, standard high volume low-cost computing hardware, and clustering technologies.

These technologies make it easier to implement clustered database solutions like Oracle9i Real Application Clusters. Oracle9i Real Application Clusters enables IT managers to create robust, scalable, and highly available databases running on low cost Intel based servers using the Linux operating system. By leveraging low-cost hardware, Oracle9i Real Application Clusters also significantly improves return on investment and reduces total cost of ownership.

And now, with a new class of interconnect standard called InfiniBand, Oracle9i Real Application Clusters builds upon these economic benefits with a dramatic performance and scalability boost that puts it in a class previously reserved for mainframe hardware.

Recent tests by Oracle and Dell, in association with Topspin and Mellanox, demonstrate that InfiniBand provides a two to four times performance improvement for Oracle 9i Real Application Clusters. Even better performance is expected as the number of nodes is scaled up.

With InfiniBand, Oracle9i database clusters benefit from network performance increases relative to Gigabit Ethernet interconnects including: (Table 1)

- Ten Times higher bandwidth
- Ten Times better interprocess (IPC) latency
- Ten Times better server CPU utilization

This combination enables major benefits at the database level within the cluster, including much higher block transfer rates and more efficient CPU utilization, and between the cluster and the application tier using Oracle Net.

Leveraging InfiniBand, customers deploying Oracle9i Real Application Clusters for Linux, now have an attractive alternative to Gigabit Ethernet that allows them to actually achieve the performance and scalability of expensive proprietary servers with the ease and economics of industry-standard components.

Network Performance Comparison

	InfiniBand	Gigabit Ethernet
Bandwidth	10 GBps	1 GBps
Latency	< 10 us	100 + us
CPU Utilization	1-3%	50%

Source: Topspin DAPL tests using dual-Xeon 2.4GHz PCI-X servers running RedHat Linux; Ethernet tests use Intel GigE NIC with checksum offload

Table 1

OPTIMIZING DATABASE CLUSTER PERFORMANCE WITH INFINIBAND

Oracle 9i RAC
Low Latency – provide fast (microsecond) passage of small (<512 bytes) lock messages between nodes in the cluster
High Bandwidth – enable high throughput (tens of Gigabytes per second) for larger cache to cache transfers (>4Kbytes) between nodes and between nodes and back end storage
Low CPU Utilization – minimize the fraction of CPU and memory bus cycles devoted to communications to free up cycles for additional database requests

Table 2

Oracle 9i Real Application Clusters enables high availability and scalability by running a single instance of the database across multiple individual server nodes with shared storage. This “shared everything” model allows the size and power of the database to scale beyond the boundaries associated with single server implementations.

The ideal deployment for Oracle9i Real Application Clusters involves running the database over a clustered infrastructure made up of standard high-volume servers and a high-speed interconnect. The interconnect must deliver the following (Table 2).

- Low latency
- High bandwidth
- Minimal CPU utilization

LATENCY

Latency directly affects cluster scalability. Clustered databases require synchronization between nodes, and the slower the synchronization, the less they can scale. In Oracle9i Real Application Clusters, cluster scalability relates directly to the time it takes to access physical storage. For database instances to access physical storage, each instance must have exclusive access (or lock) to data. The longer it takes to acquire locks from other nodes, the lower the overall performance of the database. In On-Line Transaction Processing (OLTP) environments this means that the number of transactions that can be performed is directly proportional to the latency of the cluster interconnect.

InfiniBand Advantage: At the network level, InfiniBand provides ten times better latency (<10microseconds) versus Gigabit Ethernet (up to 100microseconds) for low- level data transfers. At the Oracle9i Real Application Clusters application level, this translates to 300-600 percent better latency (figure 1).

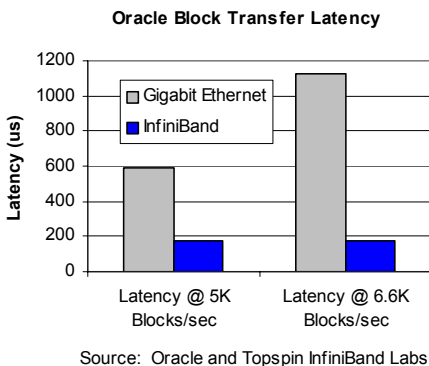
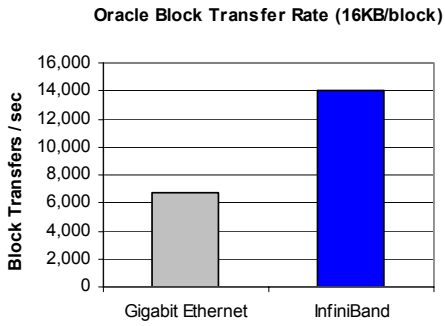


Figure 1

BANDWIDTH

Bandwidth also affects scalability due to the large amount of data transferred between database nodes. Oracle9i Real Application Clusters introduces more traffic between nodes via CacheFusion, which sends cached database blocks directly from one instance to another within a cluster. This eliminates the need to go directly to physical storage, decreasing latency and increasing scalability.



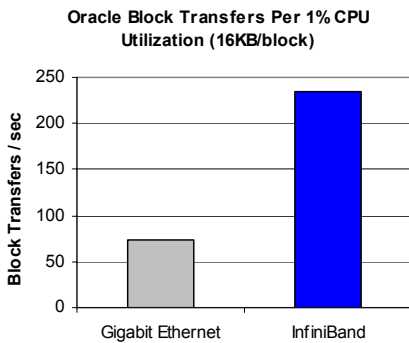
Source: Oracle and Topspin InfiniBand Labs

Figure 2

InfiniBand Advantage: At the network level, InfiniBand provides ten times higher bandwidth (10 Gbps) versus Gigabit Ethernet (1 Gbps). At the Oracle9i Real Application Cluster application level, this translates to a 210 percent better throughput (Figure 2). This advantage increases as more server nodes are added to the cluster.

CPU UTILIZATION

CPU and memory bus utilization are also key factors in database scalability in a clustered environment. As it scales, Oracle 9i Real Application Cluster can become 'CPU bound' with a large fraction of the CPU's processing and memory cycles consumed by TCP communication stack processing and redundant memory copies.



Source: Oracle and Topspin InfiniBand Labs

Figure 3

An inherently unreliable network, Ethernet runs the TCP protocol to ensure reliable communication. TCP adds a large amount of overhead to the CPU and network, including management payloads and processing overhead. This overhead severely limits how bandwidth can be used, and burdens hosts with packaging and processing network traffic, stealing valuable CPU cycles and adding a tremendous amount of latency to individual messages. This makes Ethernet an impediment to building large scale Oracle RAC clusters.

InfiniBand Advantage: At the network level, InfiniBand provides at least ten times better CPU utilization vs. Gigabit Ethernet clusters by implementing the communications stack in hardware and taking advantage of RDMA capabilities. At the Oracle 9i Real Application Cluster application level, this translates into over 300 percent better CPU utilization (Figure 3).

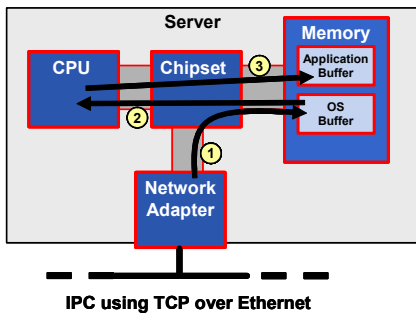


Figure 4

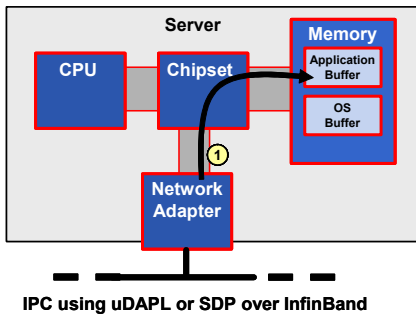


Figure 5

INFINIBAND'S ADVANTAGE: RDMA

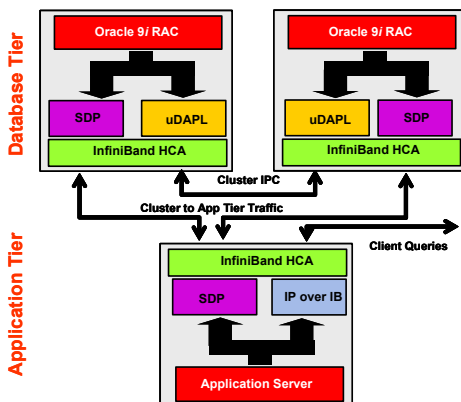
InfiniBand uses Remote Direct Memory Access (RDMA) to dramatically improve how servers communicate with other servers (and I/O) without the need for heavy protocols like TCP/IP. In addition to its high bandwidth capabilities, InfiniBand uses RDMA to dramatically free up CPU and memory bus cycles by slashing data copy overhead.

Using RDMA, server adapters (Host Channel Adapters or HCAs) read and write data directly into the memory subsystem of a server or I/O device. Because RDMA is built into the lowest levels of network interfaces, there is no need for a high overhead protocol driver to verify integrity and de-multiplex messages to applications. Instead, messages are moved directly into or out of application memory space from the network.

As shown in the diagram, it takes three copies across the server's internal memory bus to move data from the network to the application in a TCP model (Figure 4). In the InfiniBand RDMA model, this is reduced to a single direct data placement by the network adapter (HCA) from the network into the application memory (Figure 5).

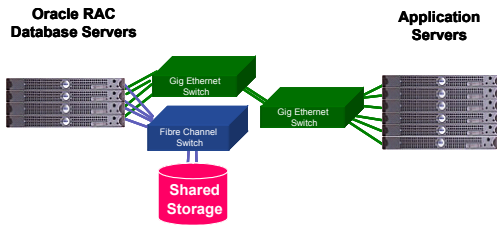
RDMA IN ORACLE9i REAL APPLICATION CLUSTERS

Oracle9i RAC takes advantage of two RDMA protocols provided by InfiniBand to accelerate messaging within the cluster and between the application tier and the cluster (Figure 6). uDAPL (direct access programming library) is used to communicate between nodes in the cluster and provides the extremely low latency required to scale the cluster to large numbers of nodes. A second protocol, SDP (sockets direct protocol), is used to connect the application tier to the cluster.



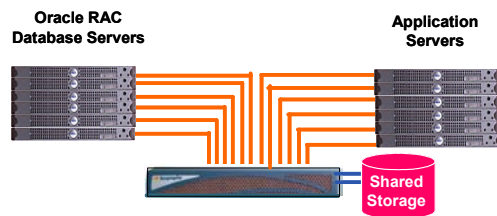
RDMA Protocols for Oracle 9i RAC and Application Server

Figure 6



RAC Cluster Using Gigabit Ethernet

Figure 7



**InfiniBand Enables Larger RAC Cluster
And Storage I/O Consolidation**

Figure 8

DEPLOYING ORACLE CLUSTERS OVER INFINIBAND

Moving from an Ethernet-attached database cluster to InfiniBand involves using an InfiniBand switch for all server-facing connections inside the cluster. As shown in the diagram, the traditional Ethernet-attached cluster involves running both Gigabit Ethernet and Fibre Channel connections into the database server nodes (Figure 7).

In the InfiniBand-attached case, these connections are terminated at the InfiniBand switch, through Fibre Channel and Ethernet to InfiniBand gateways which convert between the three protocols at full line speed, and all data and storage traffic travels across InfiniBand inside the database cluster (Figure 8).

The same holds true for the application servers, which can also be connected via InfiniBand. In the Oracle space, the application tier is connected to the database using SDP as the protocol. The Topspin SDP solution contains algorithms for determining when to use RDMA and when to actually move and copy data. This provides for faster throughput due to better CPU utilization, reduced latency, and bigger bandwidth.

ADDITIONAL BENEFITS FROM I/O CONSOLIDATION

In addition to the Oracle9i Real Application Cluster performance benefits described above, the InfiniBand connected cluster receives the following additional benefits through I/O consolidation.

- **Physical simplicity:** All server-facing connections are over a single standard Interconnect which can be single- or dual-connected to each server for reliability
- **Interconnect cost:** One or two InfiniBand connectors can replace six or more LAN (Ethernet), storage (Fibre Channel) and dedicated cluster interconnects. Cutting down on the number of server adapters, cables and switch ports required directly impacts system cost.
- **I/O Scalability:** InfiniBand provides enough bandwidth to run all clustering, storage, and communications traffic over a single “fat” pipe. By trunking and load balancing multiple storage connections at the Fibre Channel and Ethernet gateways, and sharing them across all servers, the system provides a smooth way to provide additional I/O bandwidth into the cluster as performance requirements increase.

INFINIBAND SYSTEM AVAILABILITY

Oracle and Dell are actively pursuing a strategy to bring this solution to market as quickly as possible. Oracle, Dell, Topspin, and Mellanox engineering teams are working together to assess the benefits of next-generation technologies such as Infiniband to customers deploying low-cost Intel architecture clusters. Oracle's R&D labs are running Infiniband performance tests on Dell systems that show more than a doubling of interconnect performance, which translates to higher reliability and better response times. Future tests and tuning are expected to show even more dramatic performance increases in the next several months. As a result, Oracle is planning to include full Infiniband support in the next major release of the flagship database product.

A complete set of InfiniBand switches and gateways, server adapters, and upper level RDMA protocols for building highly scalable database clusters are available today from Oracle's partner, Topspin Communications. These switches include integrated management, which enables them to drop seamlessly into existing data center Fibre Channel and Ethernet management environments. Oracle has worked closely with Topspin and Mellanox to ensure interoperability between RAC and InfiniBand.

SUMMARY

InfiniBand provides an unbeatable interconnect for building highly scalable, standards-based, Oracle solutions that includes databases in a RAC configuration, and application tiers connected to the database servers via an Oracle Net configuration. Complete hardware solutions for creating InfiniBand-connected clusters are now available and server vendors have recently highlighted the importance of InfiniBand on their roadmaps. Customers will experience dramatic ROI and other benefits, including much higher performance for a given cluster size as well as scalability to larger cluster sizes as business demands increase.

ABOUT ORACLE9i REAL APPLICATION CLUSTERS



Oracle9i Real Application Clusters provides unlimited scalability and high availability for packaged and custom applications. By running Oracle Database in a clustered hardware configuration, multiple nodes can be simply managed as a single system.

Oracle9i Real Application Clusters enables all applications to exploit cluster database availability, scalability and performance. Because applications can treat the cluster as a single system, they do not need to be modified or partitioned to achieve near-linear cluster database scalability. This means you can horizontally scale the database tier as demand continues to grow, without changing the application.

Oracle9i Real Application Clusters is self-tuning and adapts to the changing nature of the database workload. It dynamically shifts database resources across the cluster servers for optimal performance.

ORACLE DATABASE CLUSTERS ON LINUX

Linux is gaining tremendous momentum in small, medium and large size enterprises around the world.

Oracle is fully committed to supporting the Linux operating system. Oracle believes that Linux is one of the most attractive and cost effective environments currently available. In fact, Oracle was the first commercial database available on Linux. By supporting Linux with Oracle's industry leading products, Oracle is enabling customers to deploy enterprise-class solutions on the lowest-cost hardware and operating system infrastructure.

Over the past few years Oracle and its customers have gained a wealth of knowledge about running Oracle on Linux for enterprise-class deployments. Combining this knowledge with the opportunity to drastically improve performance and lower IT infrastructure costs has provided the catalyst for Oracle to take another step forward in improving the performance and extending the capabilities of Linux even further. Oracle has made this possible by dramatically increasing the performance of communications between clustered nodes via InfiniBand.



Oracle9i Real Application Clusters on InfiniBand

April 2003

Author: Raja Srinivasan

Contributing Authors:

Oracle Corporation

World Headquarters

500 Oracle Parkway

Redwood Shores, CA 94065

U.S.A.

Worldwide Inquiries:

Phone: +1.650.506.7000

Fax: +1.650.506.7200

www.oracle.com

Oracle is a registered trademark of Oracle Corporation and/or its affiliates. Various product and service names referenced herein may be trademarks of Oracle Corporation. All other product and service names mentioned may be trademarks of their respective owners.

Copyright © 2003 Oracle Corporation

All rights reserved.