

## Oracle RAC over InfiniBand

### Introduction to InfiniBand

InfiniBand Architecture (IBA) is an industry-standard, channel-based, switched-fabric, high-speed interconnect architecture with low latency and high throughput.

With Oracle support of IP over InfiniBand (IPoIB), Oracle Real Application Clusters (RAC) customers can now take advantage of greatly improved performance and scalability while using a scalable number of inexpensive high-performance servers. In addition, the Cisco® Small Computer System Interface (SCSI) Remote Direct Memory Access (RDMA) Protocol (SRP) stack has been deployed in numerous Oracle 10g environments to take advantage of the fast 10-gigabit transport to the database or application server.

This document discusses the benefits of Oracle over InfiniBand and provides details about the current implementation of this solution.

### Benefits of Oracle over InfiniBand

The performance and scalability of applications in today's typical data center depend on an efficient communication facility.

#### High Bandwidth

- Increases throughput—Standard servers connect to the InfiniBand network from a PCI-X or PCI-E host channel adapter (HCA). The HCAs have two 10-Gbps ports, and each of those ports supports an aggregate throughput of 20 Gbps—a bandwidth improvement of 20x compared to a typical Gigabit Ethernet card.
- Eliminates CPU load—InfiniBand uses RDMA, which allows send and receive buffers to be passed directly to the application, bypassing the operating system kernel, eliminating CPU-intensive memory copying operations, and leaving cycles free for other work. As a result, application performance is improved on existing database servers and application servers.
- Uses efficient protocols—By using the IPoIB stack on an InfiniBand attached server, a customer can achieve throughput of 200 Mbps or more between servers. In a clustered Oracle grid environment, this makes InfiniBand a compelling technology because it provides an easy-to-implement means of improving Cache Fusion performance.

An Oracle architect will be able to identify the expected performance improvements with benchmarks such as *netperf*, *netio*, and *IOMeter*. However, because most applications are customer specific, an evaluation of the specific environment to be deployed will likely be undertaken before taking the Oracle grids into production, to better understand the actual application environment scalability.

### Reduced Complexity and Cost

The expansive bandwidth and low latency allows sharing of the same physical link for the following purposes:

- Communication between the nodes of a cluster
- Communication with back-end database servers
- Network area storage (NAS)

Thus, aggregating network pipe technology in the data center helps reduce complexity and cost.

Figure 1 shows a sample environment. This example uses Cache Fusion to show the importance of adequate bandwidth when running Oracle RAC. In this example, the fabric is a simple one, made up of two Oracle databases (nodes A and B) and two clients (users 1 and 2). The two Oracle database nodes access shared storage over a shared Fibre Channel or InfiniBand network.

User A on node A requests record X on table EMP for modification. Database node A acquires the lock and updates the data.

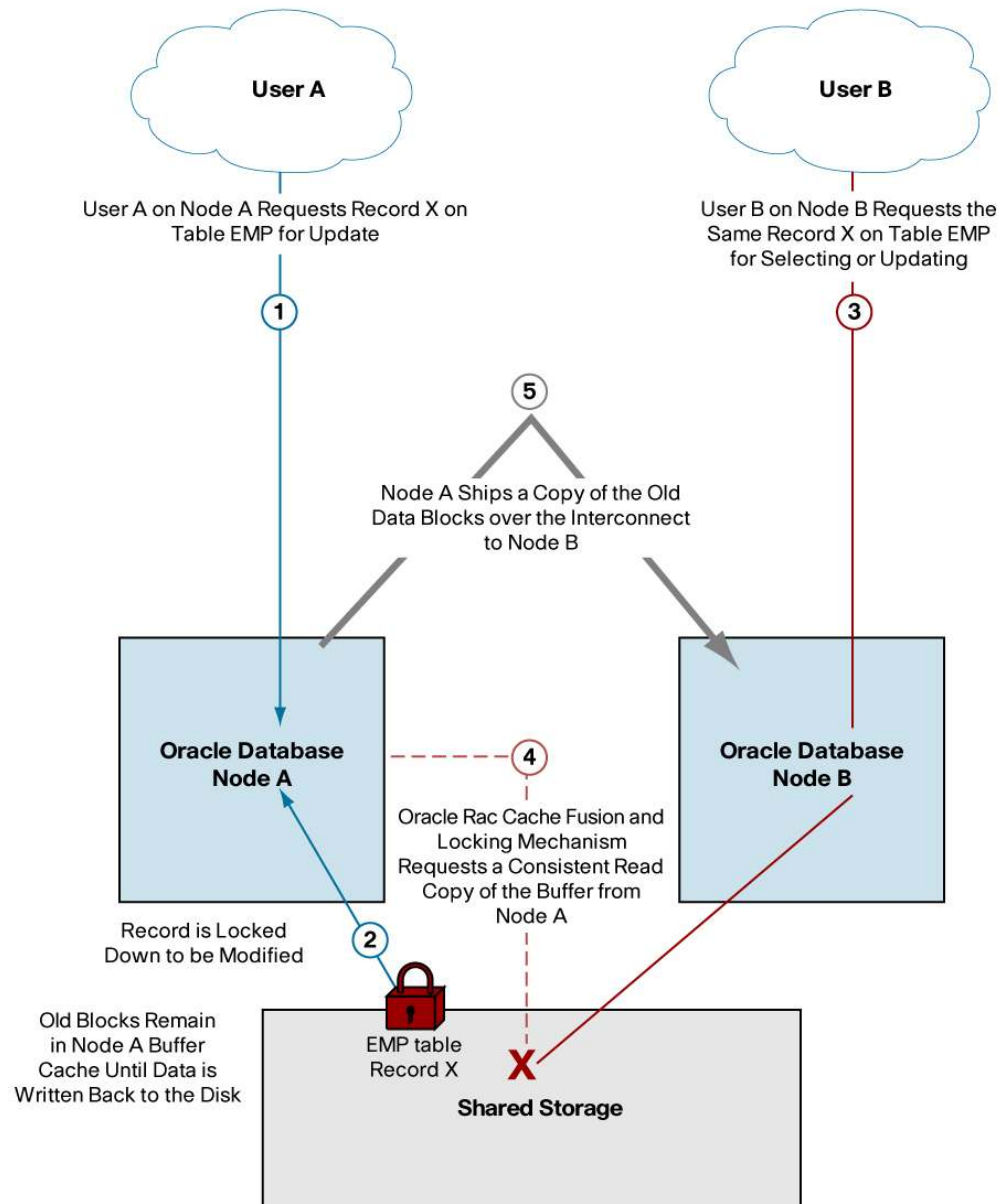
**Note:** During this process, node A has the old data blocks in its buffer cache because the new data is not yet been written back to the disk.

At this time, user 2 on node B requests the same record X on table EMP for selection or modification.

Because node A is the owner of the lock and is in the process of modifying that record, database node B requests a consistent read copy of the buffer from node A.

Node A ships a copy of the old buffer from its buffer cache over the interconnect to node B; therefore, it is not just the record that is being shipped but also the old data blocks.

**Figure 1.** Example of Importance of Adequate Bandwidth When Running Oracle RAC



Now scale this scenario to tens of clients accessing, modifying, and updating thousands of records. In the real world, the numbers are higher, but even in a small environment it is easy to see the importance of high bandwidth in scaling performance. The less time databases spend waiting for work, and the less time application servers spend waiting for responses, the better the ability to scale without having to sacrifice performance.

### Oracle RAC Components That Work with InfiniBand

The following Oracle RAC components have been shown to run successfully over InfiniBand without modification:

- Oracle Clusterware (Oracle Cluster Ready Services [CRS] in Oracle 10g Release 1)
- Oracle RAC Database
- Oracle Automated Storage Management

Additionally, Oracle Cluster File System (OCFS) works over InfiniBand with some modification; however, this is not formally supported by Oracle.

#### **Limitations and Modifications**

- OCFS should not be used in a Cisco VFrame Server Fabric Virtualization Software environment.
- Replace the load\_ocfs script with the Cisco script load\_ocfs. The script is located found in /usr/sbin or /sbin.

#### **Oracle and InfiniBand Drivers**

Oracle uses IPoIB for cluster membership voting between cluster nodes Cache Fusion traffic, Internet Protocol Control (IPC) for parallel queries, and cluster membership in Automatic Storage Management (ASM) and OCFS.

**Note:** OCFS has a known problem with MAC addresses. OCFS depends on the host MAC address to resolve cluster nodes, and the InfiniBand port MAC address can change across boots. For this reason, the load\_ocfs script must be modified.

#### **Oracle RAC and Cisco VFrame**

Cisco VFrame has been successfully tested with Oracle 10g R1. Oracle 10g R2 is being tested with Cisco VFrame at the time of this writing.

OCFS should not be used in a Cisco VFrame environment because of specific OCFS requirements.

#### **Oracle over InfiniBand Certifications**

Oracle RAC 10g over IPoIB on Red Hat Enterprise Linux Version 3.0 (RHEL 3.0) is certified. Oracle RAC 10g R1 has been certified.

Oracle has run the complete suite of regression and stress test with the Cisco driver and Cisco Server Fabric Switch (SFS). Oracle used filers (NetApp) for storage.

#### **Informal Testing Results**

Internal testing has been performed for Oracle RAC over IPoIB. The largest cluster built to date is 12 nodes and 100 GB, which is similar to a TPC-H benchmark cluster. This cluster performs well with no scaling problems.

Though Oracle has not formally tested Oracle RAC on storage over the SRP, extensive testing in Cisco labs has shown no problem with current host drivers and the Cisco SFS operating system.

Additionally, successful Oracle I/O tests have been performed in EMC and IBM labs using their equipment.

Many clients currently use Oracle over InfiniBand successfully in their data centers.

### Specific Performance Tests

Oracle Certification testing is used on an ongoing basis with Cisco releases.

The following tests have also been used:

- Oracle Automated Stress Test (OAST)
- TPC-H
- OraSim and IOTest, which simulate Oracle I/O without the need for a full Oracle database installation to test Oracle I/O

### Benchmark Results

Oracle has published two benchmarks for InfiniBand:

- Cisco SFS InfiniBand switch and Hewlett-Packard (HP) superdomes with HP drivers (using the HP Advanced Technology Attachment Packet Interface [ITAPI] acceleration API)
- HP servers with 12 nodes and InfiniCon IPoIB drivers

Oracle benchmarks are available at <http://www.tpc.org>.

An additional white paper is available from Oracle at

[http://www.oracle.com/technology/products/bi/db/10g/pdf/twp\\_bi\\_dw\\_build\\_multi\\_tb\\_dw\\_using\\_rac\\_linux\\_0406.pdf](http://www.oracle.com/technology/products/bi/db/10g/pdf/twp_bi_dw_build_multi_tb_dw_using_rac_linux_0406.pdf).

### High-Availability Support

High-availability support is available for InfiniBand products and is supported for Oracle RAC.

The IPoIB and SRP drivers can be configured for high availability, as described in the *Linux Host Drivers Guide*.

The Oracle configuration does not require additional steps to activate high availability.

### Configuring Oracle RAC over InfiniBand

Detailed steps and best practices for Oracle RAC configuration over InfiniBand are provided in the whitepaper “Configuring Oracle RAC over InfiniBand.” Here is an overview of the steps:

1. Set up the Cisco SFS as described in the appropriate hardware user guide.
2. Install InfiniBand drivers, as described in the Host-Side Driver User Guide. Use `tsinstall` to perform the driver installation rather than performing an RPM installation. An RPM installation will not upgrade the firmware, and using old firmware on the HCA can cause severe problems.
3. Configure the IP address for the InfiniBand (ib) interface. This step is the same as configuring an Oracle RAC installation with Gigabit Ethernet (GigE); you are simply configuring the InfiniBand interface instead of the GigE interface.
4. Configure IPoIB high availability, as described in the Host-Side Driver User Guide. With InfiniBand, you will not have to configure any bonding as you do with Ethernet interfaces.
5. Install Oracle Clusterware and Oracle RAC, as described in the Oracle Installation Guide. You will need to specify the IP address of the InfiniBand interface as the `CLUSTER_INTERCONNECTS` parameter in the init file; Oracle 10g R1 requires this.

## Verifying Oracle RAC over InfiniBand Configuration

Detailed steps and best practices for verifying Oracle RAC over InfiniBand are provided in the whitepaper “Configuring Oracle RAC over InfiniBand.” An overview of the steps follows :

1. Verify that the Oracle Clusterware is communicating over InfiniBand.  
Check `ocssd<node_number>.log`
2. Verify that Oracle Cache Fusion and IPC are communicating over InfiniBand.  
Login to `sqlplus` as `sysdba`.  
`oradebug setmypid`  
`oradebug ipc`
3. Check the latest trace file in the `udump` directory. Find the IP address and verify that it matches the InfiniBand interface IP address.



**Americas Headquarters**  
Cisco Systems, Inc.  
170 West Tasman Drive  
San Jose, CA 95134-1706  
USA  
[www.cisco.com](http://www.cisco.com)  
Tel: 408 526-4000  
800 553-NETS (6387)  
Fax: 408 527-0883

**Asia Pacific Headquarters**  
Cisco Systems, Inc.  
168 Robinson Road  
#28-01 Capital Tower  
Singapore 068912  
[www.cisco.com](http://www.cisco.com)  
Tel: +65 6317 7777  
Fax: +65 6317 7799

**Europe Headquarters**  
Cisco Systems International BV  
Haarlerbergpark  
Haarlerbergweg 13-19  
1101 CH Amsterdam  
The Netherlands  
[www-europe.cisco.com](http://www-europe.cisco.com)  
Tel: +31 0 800 020 0791  
Fax: +31 0 20 357 1100

Cisco has more than 200 offices worldwide. Addresses, phone numbers, and fax numbers are listed on the Cisco Website at [www.cisco.com/go/offices](http://www.cisco.com/go/offices).

©2006 Cisco Systems, Inc. All rights reserved. CCVP, the Cisco logo, and the Cisco Square Bridge logo are trademarks of Cisco Systems, Inc.; Changing the Way We Work, Live, Play, and Learn is a service mark of Cisco Systems, Inc.; and Access Registrar, Aironet, BPX, Catalyst, CCDA, CCDP, CCIE, CCIP, CCNA, CCNP, CCSP, Cisco, the Cisco Certified Internetwork Expert logo, Cisco IOS, Cisco Press, Cisco Systems, Cisco Systems Capital, the Cisco Systems logo, Cisco Unity, Enterprise/Solver, EtherChannel, EtherFast, EtherSwitch, Fast Step, Follow Me Browsing, FormShare, GigaDrive, GigaStack, HomeLink, Internet Quotient, IOS, IP/TV, iQ Expertise, the iQ logo, iQ Net Readiness Scorecard, iQuick Study, LightStream, Linksys, MeetingPlace, MGX, Networking Academy, Network Registrar, Packet, PIX, ProConnect, RateMUX, ScriptShare, SlideCast, SMARTnet, StackWise, The Fastest Way to Increase Your Internet Quotient, and TransPath are registered trademarks of Cisco Systems, Inc. and/or its affiliates in the United States and certain other countries.

All other trademarks mentioned in this document or Website are the property of their respective owners. The use of the word partner does not imply a partnership relationship between Cisco and any other company. (0609R)