HP InfiniBand solution for Oracle RAC environments

Overview............................................................................................................................................ 2
Oracle RAC overview....................................................................................................................... 2
Protocols used by Oracle RAC in an InfiniBand environment.............................................................. 3
Internode synchronization in an Oracle RAC environment ................................................................. 4
Advantages of InfiniBand in an Oracle RAC environment.................................................................. 5
HP InfiniBand solution for Oracle RAC................................................................................................ 5
Solution overview ............................................................................................................................ 5
Sample 4-node configuration ......................................................................................................... 6
Deployment of the HP InfiniBand solution for Oracle RAC......................................................................... 7
HP PDC Installation Service ............................................................................................................... 7
Service benefits .................................................................................................................................. 8
HP PDC manual installation process .................................................................................................. 8
Hardware setup and configuration ...................................................................................................... 8
Installing HP PDC for Oracle10g RAC on Linux.................................................................................... 9
Installation of InfiniBand software components .................................................................................. 9
Oracle installation............................................................................................................................... 11
Oracle uDAPL support....................................................................................................................... 11
Verifying that the uDAPL protocol is being used for inter-node communication............................... 12
Conclusion........................................................................................................................................ 12
For more information.......................................................................................................................... 13
Overview

Oracle® Real Application Clusters (RAC) has a scale out, shared-disk clustering architecture. In a shared-disk architecture, all nodes in the cluster have access to, and can manipulate, the same dataset. Concurrent access to the data from multiple nodes is synchronized through a private cluster interconnect network. Depending on the amount of contention inherent in the application, the latency and efficiency of the cluster interconnect can directly affect the performance and scalability of the entire system. Most Oracle RAC deployments use standard Ethernet networks for the cluster interconnect, however applications that have high levels of contention will not scale beyond a certain point. For these applications, InfiniBand technologies can be deployed for the cluster interconnect network, enabling the system to achieve higher levels of performance than with standard Ethernet networks. This white paper explains how to implement an Oracle Real Applications Cluster (RAC) solution that leverages HP InfiniBand products for the cluster interconnect. It is intended for system administrators, system architects, and systems integrators who are considering the advantages of HP InfiniBand based solutions in an Oracle RAC environment.

Oracle RAC overview

Oracle RAC system is a shared-disk clustering architecture in which several servers share access to a common set of network storage devices. A database instance runs on each node in the cluster. The database data and log files are stored on the shared storage and accessed simultaneously by all of the nodes in the cluster. The database instances communicate through a cluster interconnect to maintain concurrency of the data in the database.

A typical Oracle RAC system has the following components:

- Two or more server nodes
- Public local area network (LAN)
- Cluster interconnect
- Fibre Channel storage area network (SAN) (usually redundant)
- Fibre Channel shared storage array
Protocols used by Oracle RAC in an InfiniBand environment

There are 3 major components in an Oracle RAC cluster that make use of the cluster interconnect:

- Oracle Cluster Ready Services (CRS)
- Database or ASM instances
- Oracle Cluster File System (OCFS)

On Linux, each of these components is designed to communicate across a standard Ethernet network using the UDP protocol.

When an InfiniBand fabric is used for the cluster interconnect, support for UDP is transparently provided via InfiniBand’s IP over InfiniBand (IPoIB) protocol. By default, all of the Oracle components can transparently use IPoIB for inter-cluster communication. In addition to IPoIB, Oracle has developed libraries that allow a database instance to use the User Direct Access Programming Library (uDAPL) protocol for communication between cluster database instances.

Figure 2 is a pictorial representation of the protocols used in an Oracle RAC environment.
Internode synchronization in an Oracle RAC environment

Because RAC is a shared-disk clustering architecture, the database instances on each node must cooperate in order to maintain the integrity of the data. Database users could potentially attempt to update the same records from different nodes. The database instances must coordinate user access to the data to ensure the database does not become corrupted.

There are four situations that warrant consideration when multiple Oracle instances access the same data. For reasons of simplicity, the examples refer to a two-node cluster with “node 1” and “node 2.”

- **Read/read** — User on node 1 wants to read a block that user on node 2 has recently read.
- **Read/write** — User on node 1 wants to read a block that user on node 2 has recently updated.
- **Write/read** — User on node 1 wants to update a block that user on node 2 has recently read.
- **Write/write** — User on node 1 wants to update a block that user on node 2 has recently updated.

The read/read case typically requires little or no coordination, depending the on the specific database implementation. In a traditional shared disk implementation, the request by the user on node 1 will be satisfied either via local cache access or by way of disk read operations. Oracle RAC allows the read request to be served by any of the caches in the cluster database where the order of access preference is local cache, remote cache, and finally disk I/O. If the query request is served by a remote cache, the block is transferred across the cluster interconnect from one node’s cache to another.

Both the read/write and write/write cases, in which a user on one or both nodes updates the block, coordination between the instances becomes necessary so that the block being read is a read consistent image (for read/write) and the block being updated preserves data integrity (for
write/write). In both cases, the node that holds the initially updated data block ships the block to the requesting node across the high speed cluster interconnect.

In the case of the write/read case scenario, a node wants to update a block that’s already read and cached by a remote instance. An update operation typically involves reading the relevant block into memory and then writing the updated block back to disk. In Oracle’s Parallel Server product (predecessor to Oracle RAC), once the update had been complete and the update block had been written to disk, the node waiting for the block would then read the new version of the block off of the disk. This “disk ping” created additional I/O, resulting in lower system performance. In an Oracle RAC environment, the node holding the updated block can transfer the updated block across the cluster interconnect. In this scenario, disk I/O for read is avoided and performance is increased as the block is shipped from the cache of the remote node into the cache of the requesting node.

Advantages of InfiniBand in an Oracle RAC environment

Because resolving contention for database blocks involves sending the blocks across the cluster interconnect, efficient inter-node messaging is the key to coordinating fast block transfers between nodes. The efficiency of inter-node messaging depends on three primary factors:

- The number of messages required for each synchronization sequence
- The frequency of synchronization – the less frequent, the better
- The latency, or speed, of inter-node communications

The first two factors depend mostly on the application being deployed on the RAC database. The performance of the cluster interconnect can be greatly enhanced through the use of InfiniBand technologies. InfiniBand is a low latency, high bandwidth interconnect that can be used to enhance the performance of the inter-node messaging. In addition to its high performance design, InfiniBand supports the uDAPL, a user mode API for memory-to-memory transfers between applications running on different nodes. uDAPL greatly reduces the latency and CPU overhead associated with inter-node communication, allowing the cluster to scale significantly better than standard Ethernet technologies.

HP InfiniBand solution for Oracle RAC

Solution overview

Many HP customers have successfully implemented Oracle RAC solutions using the award winning HP ProLiant line of servers and HP StorageWorks line of Fibre Channel connectivity products. With the addition of InfiniBand networking products, customers can implement an Oracle RAC configuration using InfiniBand as a high speed cluster interconnect.

A sample configuration could be composed of the following components:

- HP ProLiant DL380 G4 servers
- HP StorageWorks Fibre Channel SAN
  - HP StorageWorks Modular Smart Array (MSA) 1000
  - HP StorageWorks FCA2214DC Fibre Channel HBA
  - HP StorageWorks MSA SAN Switch 2/8
- HP InfiniBand cluster interconnect
  - HP NC570C Dual Port PCI-X InfiniBand HCA
  - HP 24 Port 4x Copper Fabric Switch
Sample 4-node configuration

Below is a sample Bill of Materials for a 4-node cluster based on the ProLiant DL380 G4 server using the MSA1000 for the shared storage and InfiniBand for the cluster interconnect:

<table>
<thead>
<tr>
<th>Quantity</th>
<th>Part Number</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Server Nodes</td>
</tr>
<tr>
<td>4</td>
<td>361011-001</td>
<td>DL380R04 X3.6-1MB/800, 2GB High Performance, 2 processors</td>
</tr>
<tr>
<td>8</td>
<td>343056-B21</td>
<td>2GB of Advanced PC2 PC3200 DDR2 SDRAM DIMM Memory Kit (2 x 1024 MB)</td>
</tr>
<tr>
<td>8</td>
<td>286776-B22</td>
<td>36.4GB 15,000 rpm, U320 Universal Hard Drive, 1&quot; (internal server storage)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cluster Interconnect</td>
</tr>
<tr>
<td>4</td>
<td>376158-B21</td>
<td>HP NC570C PCI-X Dual-port 4x Fabric Adapter</td>
</tr>
<tr>
<td>4</td>
<td>376232-B21</td>
<td>HP 1m 4x Fabric Copper Cable</td>
</tr>
<tr>
<td>1</td>
<td>376227-B21</td>
<td>HP 24-Port 4x Fabric Copper Switch</td>
</tr>
<tr>
<td>Quantity</td>
<td>Part Number</td>
<td>Description</td>
</tr>
<tr>
<td>----------</td>
<td>---------------</td>
<td>---------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td></td>
<td>321835-B21</td>
<td>StorageWorks 2 Gb, Dual Port, 64-Bit/133 MHz PCI-X-to-Fibre Channel Host Bus Adapter for W2K, Microsoft® Windows® Server 2003 and Linux</td>
</tr>
<tr>
<td>8</td>
<td>221692-B22</td>
<td>5 Meter FC LC-LC Multi-Mode cable</td>
</tr>
<tr>
<td>2</td>
<td>288247-B21</td>
<td>MSA SAN Switch 2/8 (embedded 8 port FC Switch)</td>
</tr>
<tr>
<td>1</td>
<td>201723-B22</td>
<td>Modular Smart Array 1000 with 256 MB cache</td>
</tr>
<tr>
<td>1</td>
<td>218231-B22</td>
<td>MSA1000 Controller with 256 MB cache</td>
</tr>
<tr>
<td>1</td>
<td>302970-B21</td>
<td>HP StorageWorks Modular Smart Array 30 Dual Bus Enclosure</td>
</tr>
<tr>
<td>1</td>
<td>400982-002</td>
<td>SCSI cable to connect MSA30 enclosure to MSA1000</td>
</tr>
<tr>
<td>28</td>
<td>286776-B22</td>
<td>36.4GB 15,000 rpm, U320 Universal Hard Drive, 1&quot; (shared storage)</td>
</tr>
<tr>
<td></td>
<td>308225-B23</td>
<td>HP PDC Cluster Kit for Linux</td>
</tr>
</tbody>
</table>

### Deployment of the HP InfiniBand solution for Oracle RAC

Due to their cluster architecture, Oracle RAC systems can be difficult to configure and deploy, and assembling a fully supported software/hardware stack can be challenging. To simplify deployment of Oracle RAC systems, HP offers the Parallel Database Cluster Kit for Oracle RAC on Linux (PDC for Linux). The HP PDC for Linux simplifies the deployment of Oracle RAC environments by providing a set of utilities that assist in the manual deployment, as well as a fixed price service where HP service personnel deploy the environment for the customer. Customers seeking to deploy the HP InfiniBand Solution for Oracle RAC can leverage the HP PDC for Linux to simplify the deployment of the solution. For more information on the HP PDC for Linux product go to [http://h18004.www1.hp.com/solutions/enterprise/highavailability/oracle/index.html](http://h18004.www1.hp.com/solutions/enterprise/highavailability/oracle/index.html).

### HP PDC Installation Service

The HP Installation and Startup Service for Parallel Database Clusters on Linux is designed to help you manage the requirements of your constantly changing business environment. The service utilizes an HP-developed cluster kit and processes to provide the help you need to deploy specifically designed and certified HP server and storage configurations that are tailored for Oracle10g Real Application Cluster (RAC) technology. HP service specialists will perform all of the activities required to implement a database on your fully integrated HP hardware solutions. The HP Installation and Startup Service for Parallel Database Clusters is a fixed-deliverable, fixed-price service for customers implementing one of
the HP predefined cluster solutions. For cluster solutions built from other hardware configurations or for those that scale beyond the boundaries of this service, HP offers other cluster configuration services to meet your requirements.

**Service benefits**

- Reliable, repeatable installation processes
- Proven process to quickly bring your Oracle9i RAC Parallel Database Cluster solution on HP platforms up to application readiness
- Delivery of the service at a mutually scheduled time
- Availability of an HP service specialist to answer basic questions related to this service during the customer orientation session
- Configuration services from HP so your staff can remain focused on operating your business


**HP PDC manual installation process**

In addition to the PDC installation service, customers may order the kit components and perform the installation themselves. The document *Installing HP Parallel Database Cluster (PDC) for Oracle 10g Real Application Clusters (RAC) on Linux IA32* is included with the PDC and contains detailed information for using the PDC scripts to configure and install an Oracle RAC solution. For detailed installation instructions, refer to that document. The steps that follow are a high level overview of the PDC installation process.

---

**Note:** The latest set of scripts, supported configurations and installation instructions can be obtained by e-mailing RAC_Contact@hp.com.

---

**Hardware setup and configuration**

Follow the appropriate installation guides to prepare your site and rack mount the MSA1000, ProLiant servers, and HP 24 port 4x InfiniBand switch. Using Figure 3 as a guide, connect the HBA and HCA in each server to the SAN switches and the InfiniBand switch.

When configuring the system make sure to:

- Cable one port from the HBA in each node to each redundant MSA SAN Switch 2/8. It is recommended for troubleshooting and support purposes that the same HBA port on each server be connect to the same SAN switch. i.e., connect HBA port 1 on each node to the MSA SAN Switch 2/8 in the left-hand bay (looking at the MSA from behind), and HBA port 2 to the SAN switch in the right-hand bay.
- Connect the NC570C HCA on each node to the HP 24 port 4x Fabric Switch.
- Temporarily disconnect or power down the MSA1000 while installing Red Hat Enterprise Linux to avoid having the installer put the master boot record on the shared storage rather than the local storage.

When configuring the storage make sure to do the following:
• Set the host_mode for each connection to Linux. In a redundant configuration, each node in the cluster should have 2 connections, one for each port on the HBA.
• If Selective Storage Presentation (SSP) is used to control access to the LUNs, make sure to present the LUNs used for storage by the database to ALL of the nodes in the cluster.

Installing HP PDC for Oracle10g RAC on Linux

Follow the procedure described in the *Installing HP PDC for Oracle 10g RAC on Linux IA32* (PDC Installation) document. The PDC installation procedure outlines all of the steps necessary to install and configure the operating system, servers, and storage to run in an Oracle RAC environment. The PDC kit provides a set of scripts that automate the installation of HP and Oracle software pieces. Once the procedure has been completed, the system should be ready to accept an Oracle10g installation. Prior to proceeding with the Oracle installation, the InfiniBand components should be configured.

Installation of InfiniBand software components

Follow the instructions in the *Installing the HCA Drivers* section of the Dual-Port 4x Adapter User Guide to install the HCA drivers. The drivers will automatically be installed with support for all of the protocols used by Oracle10g RAC including uDAPL and IPoIB.

Once the HCA drivers have been installed, follow the instructions in the *Configuring IPoIB Drivers* section of the Dual-Port 4x Adapter User Guide to configure an IB interface for use as the cluster interconnect. In most cases, the PDC installation scripts will have configured eth1 as the cluster interconnect. The interface should be disabled and the ib interface should be configured as a replacement. In order to disable eth1 and configure the virtual InfiniBand NIC, use the following procedure on each node:

1. Unmount any OCFS volumes that are currently mounted.
2. Shutdown the eth interface using the following command:
   ```bash
   ifdown ethX
   ```
3. Edit the `/etc/sysconfig/network-scripts/ifcfg-ethX` file, setting the `ONBOOT` parameter to the value `NO` in order to prevent the eth device from starting up automatically on boot. Below are the contents of an example configuration file:
   ```
   # Broadcom Corporation|NetXtreme BCM5703 Gigabit Ethernet
   DEVICE=eth1
   BOOTPROTO=static
   HWADDR=00:02:A5:EF:5F:8E
   IPADDR=192.168.0.1
   NETMASK=255.255.255.0
   ONBOOT=no
   TYPE=Ethernet
   ```
4. Create the `/etc/sysconfig/network-scripts/ifcfg-ib0` file for the virtual InfiniBand NIC. Make sure to use the same networking configuration for the ib device as the eth device. Below are the contents of an example configuration file:
   ```
   DEVICE=ib0
   BOOTPROTO=static
   IPADDR=192.168.0.1
   NETMASK=255.255.255.0
   ONBOOT=yes
   TYPE=Ethernet
   ```
5. Startup the ib interface using the following command:
   ```bash
   ifup ib0
   ```
6. Verify that the ib device started successfully by examining the output of the command `ifconfig`. Output from the command should contain something similar to the following:

```
i0   Link encap:Ethernet  HWaddr DD:86:86:86:86:86
    inet addr:192.168.0.1  Bcast:192.168.0.255  Mask:255.255.255.0
    UP BROADCAST RUNNING MULTICAST  MTU:2044  Metric:1
    RX packets:967 errors:0 dropped:0 overruns:0 frame:0
    TX packets:1264 errors:0 dropped:0 overruns:0 carrier:0
    collisions:0 txqueuelen:128
    RX bytes:65612 (64.0 Kb)  TX bytes:1052848 (1.0 Mb)
```

7. Once the virtual InfiniBand NIC has been configured on each node, ensure that the nodes can communicate through the ib devices.

8. If the Cluster Verification Utility (CVU) packaged with the HP PDC scripts is run after the IB device has been configured, some versions of the CVU will report a failure during the Private Network test of the Network Configuration section:

```
#   Network Configuration
+  Number of network interfaces  2  >= 2  pass
+  Number of network interfaces  2  >= 2  pass
E  Private Network              disabled enabled fail
+  Public Network               enabled enabled pass
+  Local Host Loop Back         OK    ==    pass
+  Local Host Loop Back         OK    ==    pass
+  topspin1san Host Ping        topspin1san == ping OK  pass
+  topspin1san IP Ping          192.168.0.1 == ping OK  pass
+  topspin2san Host Ping        topspin2san == ping OK  pass
+  topspin2san IP Ping          192.168.0.2 == ping OK  pass
+  topspin1 Host Ping           topspin1    == ping OK  pass
+  topspin1 Host Ping           topspin1    == ping OK  pass
+  topspin1 IP Ping             15.13.183.67 == ping OK  pass
+  topspin2 Host Ping           topspin2    == ping OK  pass
+  topspin2 IP Ping             15.13.183.68 == ping OK  pass
```

9. The CVU will report this failure because it is unable to detect the IB device. This failure can safely be ignored. All of the other tests in the Network configuration section should pass, and any other failures should be investigated before continuing.
Oracle installation

Once the ib interface has been configured on all of the nodes, Oracle10g can be installed using the procedure outlined in the Oracle® Real Application Clusters Installation and Configuration Guide 10g Release 1 (10.1) for AIX-Based Systems, Apple Mac OS X, hp HP-UX, hp Tru64 UNIX®, Linux, Solaris Operating System, and Windows Platforms. When installing Oracle’s Cluster Ready Services (CRS), the ib device should be configured for use as the private cluster interconnect.

When installing CRS, the Oracle Universal Installer (OUI) will display the Private Interconnect Enforcement page. This page allows specific NICs to be designated for use by Oracle as the public network, private network, or unused. On this page, be sure to set the IB device configured earlier for use as the private network as shown in the screen shot below:

Figure 4. Oracle Universal Installer: Private Interconnect Enforcement page

<table>
<thead>
<tr>
<th>Interface Name</th>
<th>Subnet</th>
<th>Interface Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>eth0</td>
<td>15.13.183.0</td>
<td>Public</td>
</tr>
<tr>
<td>ib0</td>
<td>192.168.0.0</td>
<td>Private</td>
</tr>
</tbody>
</table>

All other steps in the installation process can be completed using the procedure described in the Oracle10g installation document.

Oracle uDAPL support

In order to support Oracle Cluster File System (OCFS) and the uDAPL protocol, you will need to install additional patches or perform additional configuration steps. Contact your Oracle sales or support
representative in order to gather the needed patches and information necessary to enable uDAPL and OCFS support.

**Verifying that the uDAPL protocol is being used for inter-node communication**

Once all of the Oracle software components have been successfully installed and a database has been created, the alert log can be examined to determine if the instance is using uDAPL for inter-instance communication. The alert log should contain output similar to the following:

```
db_name                    = orcl
open_cursors              = 300
pga_aggregate_target      = 25165824
Mon Dec 6 11:58:42 2004
cluster interconnect IPC version: Oracle UDAPL Jul 1 2004 06:25:39
IPC Vendor 1 proto 1 Version 1.0
PMON started with pid=2, OS id=28186
DIAG started with pid=3, OS id=28191
LMON started with pid=4, OS id=28197
```

The `cluster interconnect IPC version` field indicates which protocol is being used across the interconnect.

Alternatively, the following procedure can be used to determine if the database instance is using uDAPL:

- Open a `sqlplus` session and connect as `sysdba`.
- Enter the following commands at the `sqlplus` prompt
  - `Oradebug setmypid`
  - `Oradebug ipc`
- Check the location specified in the `user_dump_directory` initialization parameter for the latest `.trc` file.

Once the system has successfully been configured to use the uDAPL protocol, the installation is complete.

**Conclusion**

Due to its architecture, an Oracle RAC requires a high performance interconnect to achieve high levels of scalability. The HP InfiniBand products can be deployed to provide an Oracle Real Application Cluster with a high bandwidth, low-latency interconnect.

Oracle Real Application Cluster environments have a reputation of being difficult to integrate and deploy successfully. The HP Parallel Database Cluster kit and accompanying fixed price service can greatly reduce the complexity involved in deploying an Oracle RAC system.
For more information

For additional information, refer to the resources detailed below.

<table>
<thead>
<tr>
<th>Resource description</th>
<th>Web address</th>
</tr>
</thead>
<tbody>
<tr>
<td>HP StorageWorks Modular Smart Arrays (MSA)</td>
<td><a href="http://h18006.www1.hp.com/storage/arraysystems.html">http://h18006.www1.hp.com/storage/arraysystems.html</a></td>
</tr>
<tr>
<td>HP Parallel Database Clusters (PDC) for Oracle RAC</td>
<td><a href="http://h18004.www1.hp.com/solutions/enterprise/highavailability/oracle/index.html">http://h18004.www1.hp.com/solutions/enterprise/highavailability/oracle/index.html</a></td>
</tr>
<tr>
<td>Installing HP Parallel Database Cluster (PDC) for Oracle 10g Real Application Clusters (RAC) on Linux IA32</td>
<td>Included with the PDC kit.</td>
</tr>
</tbody>
</table>

© 2005 Hewlett-Packard Development Company, L.P. The information contained herein is subject to change without notice. The only warranties for HP products and services are set forth in the express warranty statements accompanying such products and services. Nothing herein should be construed as constituting an additional warranty. HP shall not be liable for technical or editorial errors or omissions contained herein.

UNIX is a registered trademark of The Open Group. Microsoft and Windows are U.S. registered trademarks of Microsoft Corporation. Oracle is a registered US trademark of Oracle Corporation, Redwood City, California. Linux is a U.S. registered trademark of Linus Torvalds.

[02/2005]-1