Oracle Real Application Clusters 10g Release 2: Installation and Configuration of Linux Clusters Using RDS over InfiniBand Interconnect

An Oracle White Paper
December 2006
Version 1.9

Oracle RAC 10g Release 2 on Linux Cluster using RDS over IB Interconnect Installation and Configuration

INTRODUCTION

In Oracle Real Application Cluster (RAC) Technology, Cluster Interconnect is the key to maximizing its performance. RDS, Reliable Datagram Sockets protocol provides reliable datagram services multiplexing UDP packets over InfiniBand connection improving performance to Oracle RAC. It provides high performance cluster interconnect for Oracle RAC 10g Release 2, utilizing InfiniBand which has 10X bandwidth advantage and 10X latency reduction vs. Gigabit Ethernet.

The reliable delivery capabilities inherent in InfiniBand that offload end-to-end error checking to the InfiniBand fabric, freeing CPU cycles for application processing, thereby enabling processor scaling far greater than is possible with Ethernet.

The InfiniBand architecture delivers three levels of full duplex performance; the 1X link (2.5 gbit/s), the 4X link (10 gbit/s), and the 12X link (30 gbits/s)

IPoIB

Internet Protocol over InfiniBand **(IPoIB**) defines how Internet Protocol utilizes InfiniBand as a Link Layer protocol such as Ethernet. IPoIB provides significantly improved bandwidth, latency, and reliability characteristics over Ethernet. The uses of IPoIB are entirely transparent to TCP/IP based applications, thereby providing system wide improvements.

Oracle uses IPOIB for CSS (node monitor) communication.

RDS

Reliable Datagram Sockets (**RDS**) is a reliable-socket off-load driver and inter-processor communication (IPC) protocol with low overhead, low-latency, high-bandwidth. **RDS** enables enhanced application performance and cluster scalability. RDS over InfiniBand uses approximately 50% less CPU per operation than IPoIB and operates with approximately half the latency of User Datagram Protocol (UDP) over Ethernet.

Oracle uses RDS for cross instance database communication.

Advantages of IB/RDS as Interconnect Protocol in Oracle RAC 10g Release 2.

- High throughput gain over UDP Gigabit Ethernet (GIGE)
- Less latency of UDP GIGE and IPoIB
- Low CPU utilization
- Easy to install and configure
- Supports fail-over across Host Channel Adapter (HCA) ports and cards
- Stable deterministic performance under heavy CPU load

NETWORK CONFIGURATION FOR ORACLE CLUSTER

Component	Specification
Network	• eth0: Public network (GIGE may be used) • ib1: Infiniband network for Cluster Interconnect • Switch for Cluster Interconnect o Configuration 1 One dual port IB Card with cable connected to port 1 and port 2 of HCA1 o Configuration 2 Two dual port IB Cards with cable connected to port 1 of each HCA1 and HCA2
Clusterware	• Oracle Clusterware 10.2.0.3 (This is the first Oracle version to support this new technology. At present this is only supported on Linux platform)
Software and patch level (Tested)	• Linux Kernel 2.6.9-34.Elsmp • Oracle Clusterware and RAC 10.2.0.3 • SilverStorm RDS 3.3.0.10.1
Additional HA Software	• SilverStorm RDS 3.3.0.10.1

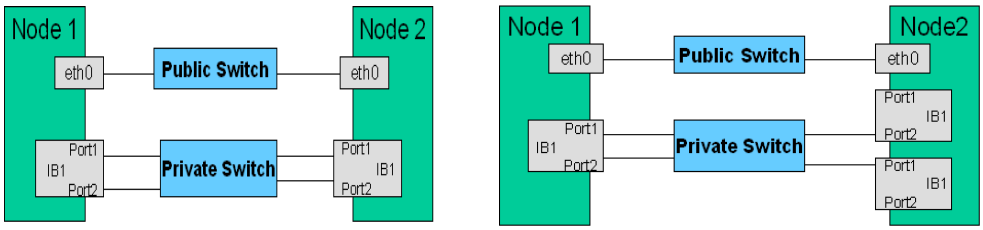

Figure 1.1 Network Configuration for Cluster

- Following diagram displays the configuration for cluster with two private switches connected to DUAL port HCA to avoid single point of failure (SPOF)
- Both inter link switches (ISL) should be in one subnet to enable RDS failover transparent to RAC

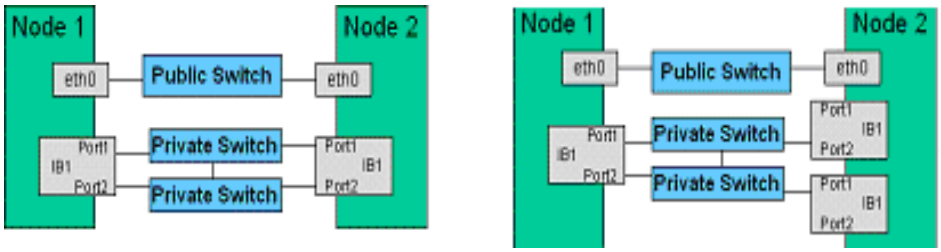

Figure 1.2 Network Configuration for cluster (two private switches)

- Primarily we will explain about single switch configuration in this paper.

FABRIC HARDWARE INSTALLATION

- Install the High performance10Gbit/s, Full Bisectional Bandwidth (FBB) InfiniBand switching fabric such as SilverStorm 9024 Switch that supports Ethernet and Fiber Channel.
- Connect one end of a Category 5 or 6 Ethernet cable to the RJ-45 connector on the switch and the other end to the OOB LAN workstation.
- Install Host Channel Adapters (HCA) such as SilverStorm HCA 7000 in each server in a 133MHZ PCI-X slot

- Connect the switch to IB-enabled hosts using 4X-to-4X IB or IB/Fiber Optic cables. Recommended distance limits for the IB/Fiber Optics is 100 meters
- Power up the switch and monitor its boot process.
- The following are ways to determine that the system has started successfully:
    - The IB link status indicator LEDs are lit up on the switch ports that are connected to an IB host.
    - The user is able to bring up Chassis Viewer through a web browser on the OOB LAN.
    - The homepage displays the 9024 switch ports

- There are three ways to view the boot process and configure the switch settings:
    - From a terminal connect to the switch using 'ssh' as user admin
    - Using the switch RS-232 port that is connected to a terminal, view or configure the Switch settings with the following command line interface (CLI) from the terminal:

        - Verify the system IP address:
            - **ShowChassisIpAddr**

        - Change the default IP address:
            - **setChassisIpAddr -h <new ipaddress> -m <new netmask>**

        - Change the default gateway IP address:
            - **setDefaultRoute -h <new ipaddress>**

        - Exit the CLI:
            - **logout**

        Example:

        # ssh open <switch ip address>

        username-> admin
        password->
        Welcome to the SilverStorm 9024 CLI. Type 'list' for the list of commands.
        -> showChassisIpAddr
        Chassis IP Address: 10.35.58.21 Net mask: 255.255.252.0
        ->
        -> setChassisIpAddr -h 10.35.58.21 -m 255.255.255.0
        You may need to reconnect if you have connected via: 10.35.58.21
        OOB IP Address/netmask successfully updated
        ->
        -> setDefaultRoute -h 10.35.58.21
        You will have to reboot in order for the setting to take effect

    - Using the **Quicksilver Chassis Viewer** GUI via web browser, using http://<Switch_Name> or http://<Switch_IP_Address>
      Chassis Viewer is browser based management software with the following management, configuration, monitoring and diagnostics functionality.

        - Manage and view user-defined data
        - Manage and monitor log files

- Manage firmware updates
- Monitor component status and switch-level detailed information
- Configure the InfiniBand, Ethernet, and Fiber Channel features

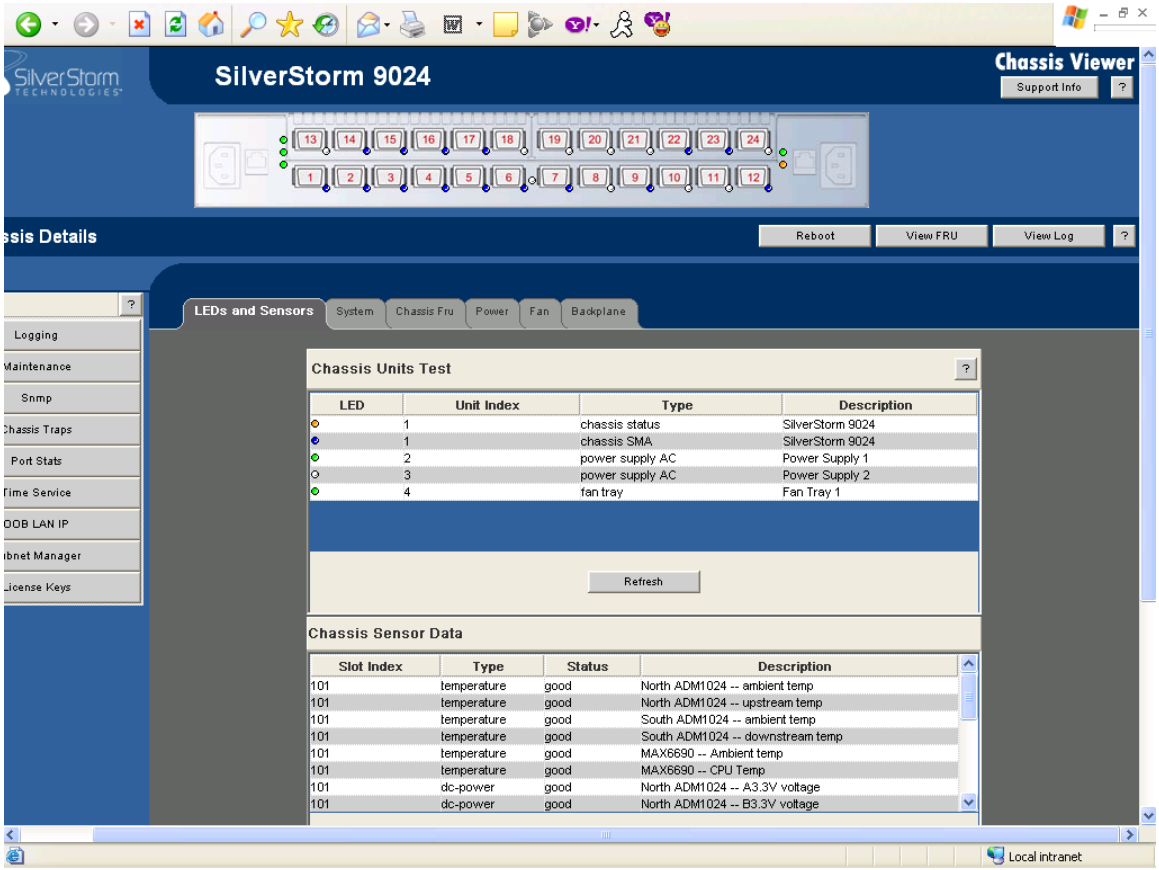

Figure 2. Chassis Viewer

RDS SOFTWARE INSTALLATION

INSTALL OS PACKAGES

- Install the same recommended OS version for Oracle 10.2.0.3 on all hosts
- Master host i.e. the local host used for installing the software, should have full OS installed including tcl and expect packages

  # rpm –q tcl expect

- Configure each host to enable the master host to connect through ‘ssh’ or ‘rsh’.

- For security reasons, it is highly recommended to configure 'ssh' in a production environment. In case you use 'rsh' for installation, ensure that rsh, and rlogin are disabled after the installation is done

- On the master host, edit the **"/etc/hosts"** file and add all ethernet and IPoIB addresses for all hosts and chassis in the fabric into it. Alternatively these addresses can be entered into DNS server and configure the hosts to access DNS server for name resolution.

  Example:
  # cat /etc/hosts

```
10.35.58.73      node1.example.com        node1
10.35.58.74      node2.example.com        node2
........
10.35.58.114     node1-vip.example.com    node1-vip
10.35.58.115     node2-vip.example.com    node2-vip
........
192.128.96.45    node1-ib.example.com     node1-ib
192.128.96.46    node2-ib.example.com     node2-ib
........
```

INSTALL THE FABRIC ACCESS SOFTWARE ON THE MASTER HOST

- Download the latest release of InfiniServ 3.3.0.10.1 host access software from the vendor website http://www.silverstorm.com/products/software.asp. (It requires a license)
- Copy and extract the tar file 'InfiniServ.3.3.0.10.1.tgz' to the **/root** directory as root user.

  # cd /root ; tar -xvfz InfiniServ.3.3.0.10.1.tgz

- During installation, you will be asked for the following input:
  - Number of IP over IB interfaces to configure
  - Interface names starting with ib1
  - Base IPV4 address and netmask in dot notation for ib1
  - Option for selecting between automatic or manual IPoIB configuration
  - Enable ib1 to autostart?
  - Select HCAs to update firmware version

  NOTE: 2-port redundant configuration is default for automatic option.

- Start the installation:

  # cd /root/InfiniServ.3.3.0.10.1 ;

  # ./INSTALL ? -- for help and install options

  # ./INSTALL

First select the “Install/Uninstall Software” option and then select the “Perform the selected actions” option to install all the drivers.

- Upon completion of the installation, remove the stage directory
- Create the following files:
  /etc/sysconfig/iba/hosts -- listing the Ethernet hostname of all hosts in the cluster except the master node
  /etc/sysconfig/iba/allhosts -listing the Ethernet hostnames including the master node
  /etc/sysconfig/iba/chassis -- edit it to add all the chassis Ethernet names (ip_addr.)

Examples:

```
# cat /etc/sysconfig/iba/hosts

  node1-ib
  node2-ib
  node3-ib
……

# cat /etc/sysconfig/iba/allhosts

  node1
  node2
  node3
  ……

# cat /etc/sysconfig/iba/chassis

  I9024
  10.35.58.21
```

- Review the Fast Fabric configuration file “/etc/sysconfig/fastfabric.conf”

Example:

```
# cat /etc/sysconfig/fastfabric.conf

if [ "$CONFIG_DIR" = "" ]
then
        if [ -d /etc/sysconfig ]
        then
                CONFIG_DIR=/etc/sysconfig
        else
                CONFIG_DIR=/etc
        fi
        export CONFIG_DIR
fi
export HOSTS_FILE="${HOSTS_FILE:-$CONFIG_DIR/iba/hosts}"
export CHASSIS_FILE="${CHASSIS_FILE:-$CONFIG_DIR/iba/chassis}"
export FF_IPOIB_SUFFIX="${FF_IPOIB_SUFFIX:--ib}"
export MGMT_HOST="${MGMT_HOST:-localhost}"
…….
…….
```

- Reboot the master host
- Run the Fast Fabric ToolSet User interface (TUI) 'iba_config' menu system to update the firmware and to verify connectivity with the switch

  # /sbin/iba_config

  Select the "Chassis Admin via Fast Fabric" option, then select the following options:

  Verify Chassis via Ethernet ping *(# pingall –C –p)
  Update Chassis Firmware (# ibtest –C -a run –P upgrade)
  Reboot Chassis (# ibtest – C reboot)

  NOTE: The Chassis Viewer GUI can be used to update firmware individually

  *Equivalent command line interface to carry out the specific task

INSTALL THE FABRIC ACCESS SOFTWARE ON REMAINING HOSTS IN THE CLUSTER

- Run the Fast Fabric ToolSet User interface (TUI) '/sbin/iba_config' menu system for installing the software in other hosts. After IPoIB is installed, configure the ifcfg-ib1 interface on each host to use the IPoIB address found in /etc/hosts (or via DNS) on each given host. Reboot all servers. Refresh the known hosts and ssh keys so the IPoIB addresses are included.

  # /sbin/iba_config

  Select the "Host Setup via Fast Fabric" option and then select the following options:

  Verify the hosts are accessible via the Ethernet network *(# pingall –p)
  Verify rsh setup (# check_rsh –l)
  Configure ssh (# setup_ssh –l)
  Replicate /etc/hosts file in all hosts (# scpall –p /etc/hosts)
  Review host OS versions (# uname –a)
  Select Install/Upgrade InfiniServ Software (# ibtest load)
  Select Configure IPoIB IP Address (# ibtest configipoib)
  Reboot all hosts (# ibtest reboot)
  Refresh ssh Known Hosts (# setup_ssh –C)

  * Equivalent command line interface to carry out the specific task

VERIFY FABRIC HARDWARE CONNECTIVITY

- Run the Fast Fabric ToolSet User interface (TUI) 'iba_config' menu system to verify connectivity:

  # /sbin/iba_config

- Select the “Chassis Admin via Fast Fabric” menu item and then select the following option to verify integrity of all ports

  Show Status of Chassis IB ports (# showallports –C)

- Select the “Host Admin via Fast Fabric” menu item and then select the following options to verify the fabric configuration, integrity of all ports, that all hosts can see each other via the subnet agent (SA) and have fully populated their SA cache and the master host can ping other hosts via IPoIB

  See Summary of Fabric Components (# fabric_info)
  Show Status of Host IB ports (# showallports)
  Verify Hosts see each other (# ibtest sacache)
  Verify Hosts ping via IPoIB (# ibtest ipoibping)

  NOTE: Review the output of ‘showallports’ command to ensure no links have excessive symbol errors (i.e., more than 200). For excessive symbol errors look for any configuration or network hardware issues.

ORACLE CLUSTERWARE AND RAC10gR2 INSTALLATION

- Using Cluster Verification Utility, verify the prerequisites for the CRS installation:

  # cluvfy stage -pre crsinst -n <node_list>

  NOTE: You can find the utility on the Oracle Clusterware installation media.

- Apply all OS patches for the Linux 2.6.9-34.Elsmp Kernel as recommended in the Oracle 10gR2 Installation manual

  # rpm –ivh <*package_name*>.rpm

- Edit the file **“/etc/sysctl.conf**” and set the value for shared memory, semaphore, message, network buffer and data buffer the kernel parameters as recommended in the Oracle 10gR2 Installation manual.

- Confirm IPoIB network is configured properly (Refer to the “Verify Fabric Hardware connectivity” Section above)

- Verify current IPoIB /RDS failover configuration. The default configuration is for Dual port HCA. For Multiple HCA, edit the “**ipoib.cfg**” file as follows:

  - For RDS fail-over between ports of a single HCA, cable both HCA ports to the InfiniBand fabric and configure IPoIB editing **“/etc/sysconfig/ipoib.cfg”** file with primary and secondary ports, either by referencing port numbers or GUIDs as follows:

    {

CREATE; NAME="ib1";
PRIMARY ={PORT=1 | PORTGUID=0x66A00b1000101;}
SECONDARY={PORT=2 | PORTGUID=0x66A01b1000101;}}

- For RDS fail-over between two HCAs in a server, configure IPoIB editing **"/etc/sysconfig/ipoib.cfg"** file, referencing port GUIDs for port 1 of each HCA as follows:

{
CREATE; NAME="ib1";
PRIMARY ={PORTGUID=0x66A00b1000101;}
SECONDARY={PORTGUID=0x67200b1000101;}}

- Install Oracle Clusterware software 10.2.0.1 in either local or shared CRS_HOME, depending upon your current architecture

  NOTE: Upgrade the software to 10.2.0.3 minimum before creating the database.

- Configure OCR Disks and mirrored Voting disks
  - OCR Disks with two partitions /dev/raw/raw1 and /dev/raw/raw2, each of size 200MB.
  - Voting Disks with three partitions /dev/raw/raw3, /dev/raw/raw4 and /dev/raw/raw5, each of size 200MB with normal redundancy

- Configure ASM with available block device.

- Configure network with three IP addresses for:
  - Public network interface (GIGE may be used)
  - IB for Cluster Interconnect
  - Virtual IP address

Examples:

 # cat /etc/hosts

10.35.58.73	**node1.example.com**	**node1**
10.35.58.114	**node1-vip.example.com**	**node1-vip**
192.128.96.45	**node1-ib.example.com**	**node1-ib**

……
…….

- Install Oracle 10.2.0.1 RDBMS software and upgrade the software to 10.2.0.3 before creating the database. Clusterware and RDBMS version has to be 10.2.0.3 to relink Oracle successfully with RDS.

- Create the database using ASM storage

- Configure RAC to utilize Infiniband and RDS after creating the database

  - To configure RAC over RDS,
    - First validate the Oracle RAC operation over IPoIB

- Shut down all Oracle instances
- Build the RAC IPC library for RDS (i.e. relink Oracle binary with RDS) by performing the following as oracle user

  $ cd $ORACLE_HOME/rdbms/lib
  $ make -f ins_rdbms.mk ipc_rds ioracle

- Restart Oracle instances

  NOTE: For non-shared ORACLE_HOME, the above commands must be run on each node of the RAC cluster.

- For existing Oracle installations, to configure the cluster interconnect with RDS, change the cluster interconnect interface on each node using the Oracle Interface Configuration Tool (**oifcfg**), as follows:

  - Shutdown all Oracle instances
  - Using "**oifcfg**", change the cluster interconnect to InfiniBand IP address

    $ oifcfg getif –global
    $ oifcfg delif –global <if_name, ex: eth1 or ib1>
    $ oifcfg setif –global ib1/192.128.96.0:cluster_interconnect

    NOTE: oifcfg requires CRS to be running.

  - Stop the CRS stack as root with "**crsctl stop crs**" on all nodes
  - Modify the **/etc/hosts** file on each node to map the new IP address with the existing private hostname.
  - Start the CRS stack as root with "**crsctl start crs**" on all nodes
  - Restart all Oracle instances.
  - RAC will now utilize InfiniBand with UDP IPC traffic passed using IPoIB for CSS communication
    Confirm IPoIB use via IPoIB interface statistics or by checking SilverStorm switch port statistics via switch GUI.
  - Relink the Oracle binary to use the RDS as stated above. Verify this by searching for the "**cluster interconnect IPC version:Oracle RDS/IP (generic)** " line in the alert log file

  CAUTION: Shutting down the previously used network interface or by removing the old Network Interface Card used for cluster interconnect may cause the node to reboot. To prevent this to occur, ensure that the new IP address is being used for cluster interconnect at both the Database and CSS level. Use the following commands to verify:

  $ egrep `olsnodes -p -l|awk ' { print $2 } '` /etc/hosts; $oifcfg getif -global|egrep interconnect|awk ' { print $2 }'

  Both commands should return the same subnet and make sure the displayed hostname is the correct private hostname.

- To revert back RAC to UDP, shut down all Oracle instances, log in as Oracle user and execute the following on each node of the cluster:

  $ cd $ORACLE_HOME/rdbms/lib
  $ make -f ins_rdbms.mk ipc_g ioracle

  Also, update **/etc/hosts** and run oifcfg as required to restore the desired cluster interconnect, per instructions above.

RDS over IB MANUAL CONFIGURATION

MODIFY IPoIB CONFIGURATION

- IP over IB requires the configuration file **"/etc/sysconfig/ipoib.cfg"** to specify parameters for each IP over IB device. The default configuration file provides for a 2 port redundant configuration. If you desire a different configuration for IP over IB, manually edit the file as mentioned below:

  - Edit the configuration file, "/etc/sysconfig/ipoib.cfg", add a CREATE block to the file in the following format:

    Example:

    #cat /etc/sysconfig/ipoib.cfg

    {CREATE; NAME="ib1"; PORTGUID="0x00066a00b1000101;}

    For each CREATE block (IP Link Layer interface) defined in the **"/etc/sysconfig/ipoib.cfg"** file, create an interface configuration file, **"/etc/sysconfig/network-scripts/ifcfg-<BLOCK_NAME>.".**

    Example:

    #cat /etc/sysconfig/network-scripts/ifcfg-ib1

    DEVICE=ib1
    BOOTPROTO=static
    IPADDR=192.128.96.45
    NETMASK=255.255.252.0
    ONBOOT=yes

  - (Re)start the IpoIB driver **# /etc/init.d/ipoib restart**
    or
    Start IpoIB **#iba_start ipob**

  - To reconfigure IPoIB, you can use 'iba_config' text user interface command as stated in the above section.

MODIFY RDS CONFIGURATION

- Readonly parameters can only be set at driver load, in order to do so, edit the file /etc/modprobe.conf, add parameter=value in the options RDS line.

  Example:

  # cat /etc/modprobe.conf

  options rds MaxDataSendBuffers=200

  and then restart RDS or reboot.

- To change a runtime configurable parameter, write the parameter to the **'/poc/driver/rds/config'** file

  Example:

  # echo " Heartbeat =0" > /proc/driver/rds/config

CONFIGURE "rdsping" SERVER MANUALLY

- Start 'rdsping' server to verify RDS connectivity between nodes in a RAC cluster.
- Start the server side of rdsping as follows:

  # /sbin/rdsping –d

- Once server side rdsping process is running, ping via RDS as follows:

  # /sbin/rdsping <host> [-n num]

  -- where <host> resolves to the IPoIB IP address

- For "rdsping" server to start automatically upon node reboot:

  # echo /sbin/rdsping –d >/etc/rc.d/rc3.d/S97rdsping
  # chmod 755 /etc/rc.d/rc3.d/S97rdsping

RDS SETUP VERIFICATION

Verify RDS configuration/setup and Cluster Interconnect network and protocol settings by following UNIX commands and SQL statements:

- Verify the PCI IB card model

  # lspci –vv --look for “InfiniBand” section

  Example:

  InfiniBand: Mellanox Technologies MT23108 InfiniHost (rev a1)
  Subsystem: Mellanox Technologies MT23108 InfiniHost
  Control: I/O+ Mem+ BusMaster+ SpecCycle- MemWINV+ VGASnoop- ParErr- Stepping- SERR+ FastB2B
  Status: Cap+ 66Mhz+ UDF- FastB2B- ParErr- DEVSEL=medium >TAbort- <TAbort- <MAbort- >SERR- <PERR-
  Latency: 64, Cache Line Size 10
  Interrupt: pin A routed to IRQ 66
  Region 0: Memory at df300000 (64-bit, non-prefetchable) [size=1M]
  Region 2: Memory at d7800000 (64-bit, prefetchable) [size=8M]
  Region 4: Memory at c8000000 (64-bit, prefetchable) [size=128M]
  Capabilities: <available only to root>

- Verify if IB Driver is installed

  # modinfo –d ics_dsc

  Example:

  SilverStorm Technologies Inc. IB Discovery Driver, version 3.3.0.10.1

- Verify if RDS is configured

  # /sbin/chkconfig –list|grep rds

  Example:

  rds 0:off 1:off 2:off 3:on 4:off 5:on 6:off

- Verify HCA driver modules (ipoib and rds)

  # ls /lib/modules/2.6.9-34.ELsmp/iba ipoib.ko rds.ko

- Verify that the HCA drivers are running:

  # /sbin/lsmod|egrep rds
  # /sbin/lsmod|egrep ib

  Example:

  # /sbin/lsmod|egrep rds ; /sbin/lsmod|egrep ib

  rds 94796 96
  ics_offload 11040 2 rds
  ipoib 122588 1 rds
  ics_dsc 74092 3 rds,ipoib,ics_srp
  ibt 714628 8 rds,ics_offload,ipoib,ics_srp,ics_dsc,mt23108vpd
  libcrc32c 6721 1 crc32c

```
ipoib              122588  1       rds
ics_dsc             74092  3       rds,ipoib,ics_srp
ibt                714628  8       rds,ics_offload,ipoib,ics_srp,ics_dsc,mt23108vpd
```

- Findout the IP address of HCAs in a node

  # /sbin/ipoib_path <node>

  Example:

```
Name: node1
Addr: 10.35.58.73
1 Paths:
Path: 0
   DGID: 0xfe70000000000000:0002c9020021ecb9
   DLID: 8
   SGID: 0xfe70000000000000:0002c9020021ecb9
   SLID: 8
   SL:   0
   PKey: 0xffff
   Mtu:  2048
   Rate: 10g
   Life:  134 ms
```

- Verify status of IB port1 and port2

  # /sbin/p1info
  # /sbin/p2info

  Example:

```
Port 1 Info
  PortState: Active         PhysState: LinkUp    DownDefault: Polling
  LID:    0x0008           LMC: 0
  Subnet: 0xfe80000000000000  GUID: 0x0002c9020021ecb9
  SMLID:  0x0001   SMSL:  0   RespTimeout :  33 ms  SubnetTimeout:  536 ms
  M_KEY:  0x0000000000000000  Lease:     0 s       Protect: Readonly
  MTU:       Active:    2048  Supported:    2048  VL Stall: 0
  LinkWidth: Active:     4x  Supported:    1-4x  Enabled:    1-4x
  LinkSpeed: Active:   2.5Gb  Supported:   2.5Gb  Enabled:   2.5Gb
  VLs:       Active:     4+1  Supported:     4+1  HOQLife: 4096 ns
  Capability 0x02010048: CR CM SL Trap
  Violations: M_Key:     0 P_Key:     0 Q_Key:     0
  ErrorLimits: Overrun: 15 LocalPhys: 15  DiagCode: 0x0000
  P_Key Enforcement: In: Off Out: Off  FilterRaw: In: Off Out: Off
```

- Verify RDS Version and value of tunable parameters

  # cat /proc/driver/rds/config

  Example:

```
rds version 3.3.0.10.1
for SilverStorm Technologies Inc. Infiniband(tm) Rds , version 3.3.0.10.1
Built for Linux Kernel 2.6.9-34.ELsmp

RdsDbgLvl - Logging for Rds
```

Bit masks are as follows:
0x80000000 - Serious Errors
0x40000000 - Errors
0x20000000 - Warnings
0x10000000 - Informational messages

RdsTraceLvl - Time tracing for Rds

Bit masks are as follows:
0x00000000 - Trace All
0x00001000 - Trace Sends
0x00002000 - Trace Recvs
0x00004000 - Trace Poll
0x00008000 - Trace Ctrl

Parameter Values:

RdsDbgLvl=0xc0000000
RdsTraceLvl=0x00000000
MinRnrTimer=10
PerfCounters=1
PendingRxPktsHWM=75 **-- Pending Rx buffers High Water Mark (percentage)**
Heartbeat=0 **- Heartbeat ON/OFF**
SessionCloseTimeWait=5000 **- SessionCloseTimeWait (milliseconds)**

Read Only Parameters (set at module load time):

UserBufferSize=4096 **- User buffer size**
MaxDataRecvBuffers=500 **- Max data recv buffers**
MaxDataSendBuffers=100 **- Max data send buffers**
MaxCtrlRecvBuffers=100 **- Max ctrl recv buffers**
MaxCtrlSendBuffers=50 **- Max ctrl send buffers**
DataRecvCoalesceFactor=10 **- Max Recv buffer coalescing (percentage)**
DataRecvBufferLVM=50 **- Recv buffer Low Water Mark (percentage)**
MaxRecvMemory=32000 **- Performance Counters ON/OFF**

- Verify IB Interface Configuration

  # ifconfig –ib1

  Example:

  ib1 Link encap:Ethernet HWaddr 26:02:C9:21:CC:B9
  inet addr:192.128.96.45 Bcast:192.128.96.255 Mask:255.255.255.0
  inet6 addr: fe80::2402:c9ff:fe21:ccb9/64 Scope:Link
  UP BROADCAST RUNNING MULTICAST MTU:2044 Metric:1
  RX packets:8143830 errors:0 dropped:0 overruns:0 frame:0
  TX packets:8061004 errors:0 dropped:2 overruns:0 carrier:0
  collisions:0 txqueuelen:1000
  RX bytes:1524279209 (1.4 GiB) TX bytes:993291326 (947.2 MiB)

- List Network Interface configuration

  # oifcfg iflist

  Example:

eth0 10.35.58.0
eth1 10.35.44.0
ib1 192.128.96.0

- Get the subnet and interface for public / private network

  # oifcfg getif

  Example:

  eth0 10.35.58.0 global public
  ib1 192.128.96.0 global cluster_interconnect

- View active cluster nodes from master node

  # cat /etc/sysconfig/iba/hosts

  Example:

  node1-ib
  node2-ib
  node3-ib
  node4-ib
  node5-ib
  .....
  .....

- Verify if the node is accessible via RDS

  # /sbin/rdsping <host>

  Example:

  rdsping node5-ib
  RDS-PING (node5-ib) 64 bytes of data.
  recvd 64 bytes from node5-ib: time = 24.050000 usec
  recvd 64 bytes from node5-ib: time = 23.950000 usec
  recvd 64 bytes from node5-ib: time = 26.100000 usec

- Verify private hostname used for cluster_interconnect

  # olsnodes –p –l

  Example:

  node5 node5-ib

- Verify which interface being used for cluster_interconnect

  # ocrdump ;egrep ib1 OCRDUMPFILE

Example:

[SYSTEM.css.interfaces.global.ib1]
[SYSTEM.css.interfaces.global.ib1.192|d128|d96|d0]
[SYSTEM.css.interfaces.global.ib1.192|d128|d96|d0.1]

- Verify that the IB and RDS Drivers started up successfully during node startup

 # cat /var/log/messages
 # cat /var/log/boot.log
 # dmesg

Example:

SilverStorm Technologies Inc. InfiniBand(tm) Transport Driver, version 3.3.0.10.1
Copyright (C) 2000-2002 InfiniCon Systems(r)
Copyright (C) 2005 SilverStorm Technologies Inc.
Copyright (C) 2000-2001 Intel Corporation
Built for Linux Kernel 2.6.9-34.ELsmp
mt23108vpd: no version for "SpinRwLockInit" found: kernel tainted.
Initializing SilverStorm Technologies Inc. MT23108/MT25208 Verbs Provider Driver, version 3.3.0.10.1
for SilverStorm Technologies Inc. Infiniband(tm) Transport Driver, version 3.3.0.10.1
Built for Linux Kernel 2.6.9-34.ELsmp
Adding CA (vendor=0x15b3, device=0x5a44)
CR space at 0xdf300000 1 MB (0x100000 Bytes)
UAR space at 0xd7800000 8 MB (0x800000 Bytes)
HCA DDR memory at 0xc8000000 128 MB (0x8000000 Bytes)
ACPI: PCI interrupt 0000:0b:00.0[A] -> GSI 101 (level, low) -> IRQ 66
Interrupt pin 1 routed to IRQ 66
Set PCI Max Read Byte Count (at 0x70) to 4096 bytes
Set PCI Max Outstanding Split Transactions (at 0x70) to 2
InfiniServ HCA 1, Firmware version: 3.3.5
FW: 6291456 bytes
QPC/EQPC: 65536 QPs 18874368 bytes
RDB: 4 RDMA Resp 8388608 bytes
CQ: 16384 CQs 1048576 bytes
MPT: 524288 MPTs (131072 MRs, 262144 MWs) 33554432 bytes
MTT: 4194304 entries (8 per seg) 33554432 bytes
UAR Scratch: 2048 UARs 65536 bytes
EQC: 64 EQs 4096 bytes
MCG: 8192 MCGs 524288 bytes
AV: 995200 AVs 31846400 bytes
Cmds: DB: 1 Max Outstanding: 64, Max In Mbx: 1024 bytes, Max Out Mbx: 1024 bytes
Cmd Timeout: A: 5000000 usec, B: 10000000 usec, C: 20000000 usec, D: 50000000 usec
Node Guid: 0x0002d9020421ceb5
Port 1 Guid: 0x0002d9020421ceb96
Port 2 Guid: 0x0002d9020421cebc
ICS DSC:Initializing SilverStorm Technologies Inc. IB Discovery Driver, version 3.3.0.10.1
ICS DSC:Built for Linux Kernel 2.6.9-34.ELsmp
ICS DSC:Found 1 HCAs
ICS SRP:Initializing SilverStorm Technologies Inc. Virtual HBA (SRP) SCSI Driver, version 3.3.0.10.1
ICS SRP:Built for Linux Kernel 2.6.9-34.ELsmp
ICS SRP:Using Physical memory model.
ICS SRP:Found 1 HCAs
IPOIB: Initializing SilverStorm Technologies Inc. IP over IB Driver, version 3.3.0.10.1
IPOIB: Built for Linux Kernel 2.6.9-34.ELsmp
divert: allocating divert_blk for ib1

```
Initializing ics_offload version 3.3.0.10.1
for SilverStorm Technologies Inc. Infiniband(tm) Transport Driver, version 3.3.0.10.1
Built for Linux Kernel 2.6.9-34.ELsmp
NET: Registered protocol family 30
Initializing rds version 3.3.0.10.1
for SilverStorm Technologies Inc. Infiniband(tm) Rds , version 3.3.0.10.1
Built for Linux Kernel 2.6.9-34.ELsmp
RDS:Found 1 HCAs
RDS: ops_proto_register success!!
ip_tables: (C) 2000-2002 Netfilter core team
IB Port State Change: Hca 1 Port 1 New State: Active PhysState: LinkUp
IPOIB: ib1: Using Primary path
```

- Verify that RAC is using desired IPC protocols from **alert<sid>.log** file
  Check for the string "**cluster interconnect IPC version:Oracle RDS/IP (generic)"** in the alert<sid>.log file.

  Example:

```
.............
.............
Starting ORACLE instance (normal)
LICENSE_MAX_SESSION = 0
LICENSE_SESSIONS_WARNING = 0
Interface type 1 ib1 192.128.96.0 configured from OCR for use as a cluster interconnect
Interface type 1 eth0 10.35.58.0 configured from OCR for use as  a public interface
.................................

SYS auditing is disabled
ksdpec: called for event 13740 prior to event group initialization
Starting up ORACLE RDBMS Version: 10.2.0.3.0.
.........................
.........................
  pga_aggregate_target     = 387973120
Cluster communication is configured to use the following interface(s) for this instance
  192.128.96.46
Mon Oct 23 11:31:07 2006
cluster interconnect IPC version:Oracle RDS/IP (generic)
IPC Vendor 1 proto 3
  Version 1.0
PMON started with pid=2, OS id=18765
DIAG started with pid=3, OS id=18796
.........................
```

- Verify that the correct network is used for database cluster interconnect

```
SQL> SELECT INST_ID, NAME, IP_ADDRESS, IS_PUBLIC, SOURCE
       FROM GV$CLUSTER_INTERCONNECTS
       ORDER BY INST_ID;
```

Example:

INST_ID	NAME	IP_ADDRESS	IS_PUBLIC	SOURCE
2	ib1	192.128.96.45	NO	Oracle Cluster Repository
4	ib1	192.128.96.46	NO	Oracle Cluster Repository
5	ib1	192.128.96.47	NO	Oracle Cluster Repository

…….
…….

or use oradebug like below

SQL> oradebug setmypid
SQL> oradebug ipc
SQL> oradebug tracefile_name

Example:

 # cat /home/ractest/oracle/OraHome/admin/RAC/udump/rac7_ora_30741.trc

……..
wait delta 333 sec (333729 msec) ctx ts 0x85807 last ts 0x34194
user cpu time since last wait -171798692 sec 42949672 ticks
system cpu time since last wait -1 sec 0 ticks
locked 1
blocked 0
timed wait receives 0
admno 0x3c7ecb0f admport:
SSKGXPT 0xcd8c160 flags socket no 7 IP 192.128.96.46 RDS 63065
context timestamp 0x85807
no ports
sconno accono ertt state seq# sent async sync rtrans acks
0x043afd78 0x15cfefe8 64 3 32764 1 1 0 0 0
0x043afd79 0x4a030532 64 3 32771 8 8 0 0 0
…….

SQL> SELECT INST_ID, PUB_KSXPIA, PICKED_KSXPIA, NAME_KSXPIA, IP_KSXPIA FROM X$KSXPIA ORDER BY INST_ID;

Example:

INST_ID	PUB_KSXPIA	PICKED_KSXPIA	NAME_KSXPIA	IP_KSXPIA
2	Y	OCR	eth0	10.35.58.74
2	N	OCR	ib1	192.128.96.46

RDS PERFORMANCE MONITORING

Monitor RDS statistics using the following OS commands:

- When RDS is running, an associated directory structure will be present in the /proc file system **(/proc/driver/rds**)

- Start **'/sbin/iba_mon'** as a daemon on all the nodes with values set in the "**iba_mon.conf**. " configuration file.

 # cat /etc/sysconfig/iba/iba_mon.conf
 # /sbin/iba_mon –d

Example:

```
# cat /etc/sysconfig/iba/iba_mon.conf
Interval                        10      # monitoring interval in seconds
SyslogFacility                  local6  # syslog facility code, or disable
PortXmitData                    0       # as MB/second
PortRcvData                     0       # as MB/second
PortXmitPkts                    0       # as packets/second
PortRcvPkts                     0       # as packets/second
SymbolErrorCounter              100
LinkErrorRecoveryCounter        3
LinkDownedCounter               3
PortRcvErrors                   100
PortRcvRemotePhysicalErrors     100
PortRcvSwitchRelayErrors        100
PortXmitDiscards                100
PortXmitConstraintErrors        10
PortRcvConstraintErrors         10
LocalLinkIntegrityErrors        3
ExcessiveBufferOverrunErrors    3
VL15Dropped                     100

# /sbin/iba_mon -d

iba_mon: Starting
iba_mon: Settings:
iba_mon:   Interval                     10
iba_mon:   SyslogFacility               local6
iba_mon:   SymbolErrorCounter           100
iba_mon:   LinkErrorRecoveryCounter     3
iba_mon:   LinkDownedCounter            3
iba_mon:   PortRcvErrors                100
iba_mon:   PortRcvRemotePhysicalErrors  100
iba_mon:   PortRcvSwitchRelayErrors     100
iba_mon:   PortXmitDiscards             100
iba_mon:   PortXmitConstraintErrors     10
iba_mon:   PortRcvConstraintErrors      10
iba_mon:   LocalLinkIntegrityErrors     3
iba_mon:   ExcessiveBufferOverrunErrors 3
iba_mon:   VL15Dropped                  100
[root@node1]# iba_mon: Port 0x0002c9020021ccb9 Active
iba_mon: Port 0x0002c9020021ccba Down
```

- To Monitor RDS usage statistics

 # cat /proc/driver/rds/stats

Example:

```
# cat /proc/driver/rds/stats
Rds Statistics:
   Sockets open:                96
   End Nodes connected:         14

   Performance Counters: ON
   Transmit:
```

```
  Xmit bytes              1353187764
  Xmit packets               3417817
  Xmit errors                      3
  Loopback packets dropped         0

Receive:
  Recv bytes              1265875795
  Recv packets               3401421
  Recv packets pending             0
  Recv packets dropped         27864
  Recv errors                   5814

Stalled Ports:                     0
  Stalls   Sent                    0
  Unstalls Sent                    0
  Stalls   Recvd                   0
  Unstalls Recvd                   0

Debug Stats:
  ENOBUFs (105) returned           0
  EWOULDBLOCKs(11) returned       71
  Rx pkts pending HWM           5607
  Rx pkts coalescing              50
  Rx buf alloc failed              0
  Rx post_thread_wakeups           0
  Stall events ignored             0
  Session failovers                3
```

- To view RDS connections in use

  # cat /proc/driver/rds/info

  Example:

```
# cat /proc/driver/rds/info
Session Info:
IP              State        Rx bufs   Rx Cache
192.128.96.45   ACTIVE         499       0
192.128.96.46   ACTIVE         468       0
……….
……….
192.168.100.51   ACTIVE        489       0
192.168.100.16   IDLE            0       0
----------------------
Socket Info:
Port  Rx pending    State
```

- View IB network statistics and monitor cluster interconnect for any collisions, errors and lost packets on the ib Interface

  # /bin/netstat –i

  Example:

```
# netstat -i
Kernel Interface table
Iface      MTU Met    RX-OK RX-ERR RX-DRP RX-OVR    TX-OK TX-ERR TX-DRP TX-OVR Flg
eth0      1500      0  1127673      0      0      0   711366      0      0      0 BMRU
```

```
eth0:1    1500    0     - no statistics available -                    BMRU
eth1      1500    0  829395     2    0    0   331282    0    0    0 BMRU
ib1       2044    0 10020829    0    0    0  9898208    0    2    0 BMRU
lo        16436   0  3991905    0    0    0  3991905    0    0    0 LRU
```

- Measure Interconnect Traffic

  # /usr/bin/sar –n DEV

  Example:

```
# /usr/bin/sar -n DEV|more
Linux 2.6.9-34.ELsmp (node1)  11/30/2006

10:10:01 PM   IFACE   rxpck/s   txpck/s   rxbyt/s   txbyt/s   rxcmp/s   txcmp/s  rxmcst/s
10:10:01 AM     ib1     29.53     28.86   5092.14   3393.59      0.00      0.00      0.07
10:20:01 AM     ib1     29.49     28.46   5101.32   3373.69      0.00      0.00      0.06
10:30:01 AM     ib1     29.45     28.35   5086.49   3364.08      0.00      0.00      0.06
10:40:01 AM     ib1     29.49     28.18   5100.64   3357.71      0.00      0.00      0.07
10:50:01 AM     ib1     29.45     28.35   5075.19   3360.91      0.00      0.00      0.06
11:00:01 AM     ib1     29.42     28.22   5077.42   3355.07      0.00      0.00      0.06
11:10:01 AM     ib1     29.46     28.23   5099.52   3360.73      0.00      0.00      0.07
11:20:01 AM     ib1     29.48     28.24   5100.54   3361.19      0.00      0.00      0.06
11:30:01 AM     ib1     29.43     28.18   5085.15   3354.46      0.00      0.00      0.06
11:40:01 AM     ib1     29.43     28.27   5084.64   3358.86      0.00      0.00      0.07
Average:        ib1     30.32     29.70   7218.50   4601.99      0.00      0.00      0.06
```

- Get the IB port1 and port2 statistics

  # /sbin/p1stats
  # /sbin/p2stats
  # /sbin/showallports –h '<host>'

  Example:

```
Port 1 Counters
  Performance: Transmit
    Xmit Data                3039 MB (796858998 Quads)
    Xmit Pkts                17023482
  Performance: Receive
    Rcv Data                 3627 MB (950912764 Quads)
    Rcv Pkts                 17092369
  Errors:                      Async Events:
    Symbol Errors            0     State Change            0
    Link Error Recovery      0     Traps:
    Link Downed              0     Link Integrity          0
    Port Rcv Errors          0     Exc. Buffer Overrun     0
    Port Rcv Rmt Phys Err    0     Flow Control Watchdog   0
    Port Rcv Sw Relay Err    0     Capability Mask Chg     0
    Port Xmit Discards       0     Platform Guid Chg       0
    Port Xmit Constraint     0     Bad M-Key               0
    Port Rcv Constraint      0     Bad P-Key               0
    Local Link Integrity     0     Bad Q-Key               0
    Exc. Buffer Overrun      0     Other                   0
    VL15 Dropped             0
```

- Clear the IB Port Statistic history

  # /sbin/clear_p1stats
  # /sbin/clear_p2stats

- View HCA and Port level statistics from "/proc/iba/mt23108/" location

- View memory usage by RDS

  # cat /proc/slabinfo |grep RDS; cat /proc/slabinfo|grep IPOIB

  Example:

```
RDS:control      2200   2205   256   15   1 : tunables  120   60    8 : slabdata    147    147      0
RDS:data         8105   8105  4352    1   2 : tunables    8    4    0 : slabdata   8105   8105      0
IPOIB:ports        14    135    28  135   1 : tunables  120   60    8 : slabdata      1      1      0
IPOIB:mac          14     65    60   65   1 : tunables  120   60    8 : slabdata      1      1      0
IPOIB:av           19    214    36  107   1 : tunables  120   60    8 : slabdata      2      2      0
```

- View IB Card or Port failover log

  # cat /proc/iba/log

  Example:

```
.....
.....
0004294697: IPOIB: Initializing SilverStorm Technologies Inc. IP over IB Driver, version
3.3.0.10.1
0004294697: IPOIB: Built for Linux Kernel 2.6.9-34.ELsmp
0004294697: Initializing rds version 3.3.0.10.1
for SilverStorm Technologies Inc. Infiniband(tm) Rds , version 3.3.0.10.1
0004294697: Built for Linux Kernel 2.6.9-34.ELsmp
0004294697: RDS:Found 1 HCAs
0004294697: IB Port State Change: Hca 1 Port 1 New State: Active PhysState: LinkUp
0004294697: IPOIB: ib1: Using Primary path
.....
.....
```

- Measure the RDS traffic throughput finding the packet counts in the **"/proc/driver/rds/stats"** file

- From AWR / statspack report, monitor the "**global_cache_statistics**" wait events

  Example:

```
INST_ID NAME                                             VALUE
------- -------------------------------------------------- ----------
2       gc cr blocks served                              25434
2       gc cr block build time                              36
2       gc cr block flush time                            1676
2       gc cr block send time                               63
2       gc current blocks served                        110207
2       gc current block pin time                           58
2       gc current block flush time                        104
```

2	gc current block send time	346
2	gc cr blocks received	11082
2	gc cr block receive time	703
2	gc current blocks received	30782
2	gc current block receive time	1384
2	gc local grants	13624
2	gc remote grants	99393
2	gc blocks lost	0
2	gc claim blocks lost	0
2	gc blocks corrupt	0
2	gc CPU used by this session	5493

- Verify that instances are running:

  - **SQL> SELECT * FROM V$ACTIVE_INSTANCES;**

Example:

INST_NUMBER	INST_NAME
2	node1: rac1
4	node2: rac2
5	node3: rac3
…	……..
…	……

  - **SQL> SELECT INSTANCE_NAME, HOST_NAME, VERSION, STARTUP_TIME, STATUS, LOGINS, DATABASE_STATUS, BLOCKED FROM GV$INSTANCE;**

Example:

INSTANCE_NAME	HOST_NAME	VERSION	STARTUP_T	STATUS	LOGINS	DATABASE_STATUS	BLO
rac1	node1	10.2.0.3.0	28-NOV-06	OPEN	ALLOWED	ACTIVE	NO
rac2	node2	10.2.0.3.0	28-NOV-06	OPEN	ALLOWED	ACTIVE	NO
rac3	node3	10.2.0.3.0	28-NOV-06	OPEN	ALLOWED	ACTIVE	NO
……..	……..	…….	…….	……..…….	…….	….	
……..	……..	…….	…….	……..…….	…….	….	

DIAGNOSIS

In case of node hung or crash, for debugging purpose,

- Enable the magic keys at OS level:

  # sysctl –w kernel.sysrq=1

- Enable and start the NMI watchdog timer on nodes to capture the stack by setting the parameter "**nmi_watchdog=1**" **in /boot/grub/grub.conf** file

- Disable SM sweep and Discovery Scan Time at switch level to avail CPU for InfiniBand host management processes

- Collect the HCA and IB ports statistics dump for diagnosis purpose

  # iba_capture <filename>.tgz

REFERENCES

- Oracle 10g RAC and SilverStorm's RDS Installation Guide
- Oracle 10g Performance and Deployment Guide
- QuickSilver Fast Fabric User's Guide

Oracle 10gR2 RAC on Linux Cluster using RDS over IB Interconnect Installation and Configuration
December 2006

Author: Badrinath Tripathy
Contributing Author: Amit Das, Richard Frank, Roland Knapp, Daniel Dibbets

Version 1.9

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
oracle.com
Copyright © 2006, Oracle. All rights reserved.
This document is provided for information purposes only and the contents hereof are subject to change without notice.
This document is not warranted to be error-free, nor subject to any other warranties or conditions, whether expressed orally or implied in law, including implied warranties and conditions of merchantability or fitness for a particular purpose. We specifically disclaim any liability with respect to this document and no contractual obligations are formed either directly or indirectly by this document. This document may not be reproduced or transmitted in any form or by any means, electronic or mechanical, for any purpose, without our prior written permission.
Oracle, JD Edwards, PeopleSoft, and Siebel are registered trademarks of Oracle Corporation and/or its affiliates. Other names may be trademarks of their respective owners.