Clustering Oracle RAC/RDS over InfiniBand Readme

Support for IBM Power Systems with POWER6 or POWER5 processors

Readme information

Updated 11/07/2008

Overview

This Readme provides information on using Reliable Datagram Sockets (RDS) with Oracle Database 11g Release 1 (11.1.0.6 or later) for AIX.

Prerequisites

The following are the hardware and software prerequisites are requried to use RDS.

Systems

- POWER5: 9133-55A, 9117-570, 9119-590/595
- POWER6: 9117-MMA, 8203-E4A, 8204-E8A, 8234-EMA

Interconnect Hardware

- QLogic InfiniBand Switch, Models 9024, 9040, 9140, 9240
- GX Dual-port SDR Host Channel Adapter; use Feature Code 1802 for 9117-MMA and 8234-EMA, and Feature Code 5616 for 8203-E4A and 8204-E8A

Software and Firmware

- AIX 5.3 TL 8, with Service Pack 4 or later
- Please refer to Oracle MetaLink for the latest Oracle requirements for using Oracle RAC with RDS on AIX:
    - Oracle MetaLink - Certify Tab
    - Oracle MetaLink Note:282036.1- Minimum Software Versions
    - Patches Required to Support Oracle Products on IBM pSeries
- QLogic switch firmware:
    - Firmware levels can be obtained through the QLogic website. Scroll down to **IBM System p** and select **InfiniBand Switches & Management Software - Commercial / Oracle**.
    - QLogic switch firmware level:
        - 4.2.1.1.1
- POWER5 Server Firmware level:
    - SF240_358 or later
- POWER6 Server Firmware levels:
    - EL320_083 or later
    - EM320_083 or later

Back to top

Additional support

- FLRT: Fix Level Recommendation Tool

- Cluster support
- InfiniBand support
- RAC/RDS overview

Clustering Oracle RAC/RDS over InfiniBand Readme

Support for IBM Power Systems with POWER6 or POWER5 processors

Readme information

- Overview and prerequisites
- Restrictions and limitations
- RDS configuration and loading
- QLogic LMC setting
- Multipath Routing, Active DGD, and VIPA configuration
- InfiniBand port failover

- rdsctrl utility and tuning parameters
- RDS and InfiniBand reliable connection mode
- RDS keepalives
- Known problems
- Cabling and appendix
- PDFs and Readme archive

Known restrictions and limitations

The following restrictions and limitations are known for the supported hardware and software:

- Maximum of 16 nodes can be supported with RDS.
- Oracle RAC with RDS is supported on AIX partitions/systems having dedicated InfiniBand adapters.
- Oracle depends on a highly available environment. See RDS configuration and loading for details about such a configuration.
- Before bringing up the AIX partitions, the InfiniBand switch should be up and running. Failing to have the switch running might result in the InfiniBand interfaces not being setup up properly.

Additional support

- FLRT: Fix Level Recommendation Tool

- Cluster support
- InfiniBand support
- RAC/RDS overview

Clustering Oracle RAC/RDS over InfiniBand Readme

Support for IBM Power Systems with POWER6 or POWER5 processors

Readme information

- Overview and prerequisites
- Restrictions and limitations
- RDS configuration and loading
- QLogic switch LMC setting
- Multipath Routing, Active DGD, and VIPA configuration
- InfiniBand port failover

- rdsctrl utility and tuning parameters
- RDS and InfiniBand reliable connection mode
- RDS keepalives
- Known problems
- Cabling and appendix
- PDFs and Readme archive

RDS configuration and loading

If IPoIB is already configured, then skip to the step for loading RDS.

The InfiniBand interfaces and corresponding VIPA interface configuration can be performed (with root authority) from either the AIX SMIT utility or the AIX Command Line Interface (CLI).

Before you begin

Before starting the configuration steps, note the following information:

- With the current Oracle RAC (OR) VIPA implementation, two InfiniBand interfaces (ib0 and ib1) are required to be configured on separate subnets, with the corresponding VIPA interface (vi0) configured on a third subnet.
- The current *or* VIPA implementation supports a single node assigned to any IP Host Channel Adapter (HCA)
- The InfiniBand and VIPA interfaces and corresponding static routes (information on how to configure these static routes is provided in the "Multipath Routing, Active DGD and VIPA Configuration" section) are configured in the AIX ODM and are then available after this initial configuration and after subsequent reboots. Additional configuration is not required after this initial configuration.

Configuring InfiniBand and VIPA interfaces

With the above information in mind, complete the following steps to configure the InfiniBand and VIPA interfaces. This procedure needs to be done on each machine (node) in the cluster.

1. Confirm the InfiniBand HCA(s) are available to each node:
   - The current OR VIPA implementation can be configured on either 1 or ideally two InfiniBand HCAs per node. The subsequent InfiniBand and VIPA configuration steps are predicated on the availability of one or two InfiniBand HCAs per node

     To determine how many InfiniBand HCAs are currently available to the node, issue the following AIX command:
     lsdev -Cc adapter | grep iba

     Example: (**/etc/sysconfig/network/ifcfg-ib0**):
     ```
     iba0 Available InfiniBand host channel adapter
     ```

     Example output: 1 HCA
     ```
     iba0 Available InfiniBand host channel adapter
     ```

     Example output: 2 HCAs
     ```
     iba0 Available InfiniBand host channel adapter
     iba1 Available InfiniBand host channel adapter
     ```
2. Configure the InfiniBand Communication Manager (ICM):
   - One ICM is defined per node.

       Using SMIT:

       - Configuration:

smitty icm -> Add an InfiniBand Communication Manager -> Add an InfiniBand Communication Manager
or
smitty namehdr_mk_icm

- For "Name of IB Communication Manager to Add", select the default, "icm" and press Enter
- Accept all the default tunable values and press Enter
  Validation:
  smitty icm -> List All Defined IB Communication Managers
  This should show:
  icm Available InfiniBand Communication Manager
- Using AIX CLI:
  Configuration:
  mkdev -c management -s InfiniBand -t icm
  Validation:
  lsdev -CH | grep icm
  This should show:
  icm Available InfiniBand Communication Manager

3. Configure the InfiniBand interfaces:
   - Configuration:
     smitty tcpip -> Further Configuration -> Network Interfaces
     -> Network Interface Selection -> Add a Network Interface
     -> Add an IB Network Interface

     or

     smitty mkinetib

     Specify values for the following fields:
     - `INTERNET ADDRESS (dotted decimal)`
     - `Network MASK (hexadecimal or dotted decimal)`
     - `Network Interface Name (i.e ib0)`
     - `HCA Adapter`
     - `Adapter's port number`
     - `MTU and leave the default values in the remaining fields.`

     <u>Examples:</u>
     To configure the ib0 interface on the iba0 InfiniBand HCA, port 1, with IP address/netmask 192.168.1.1/255.255.255.0, specify the fields listed above with the following values:

     ```
     INTERNET ADDRESS (dotted decimal) 192.168.1.1
     Network MASK (hexadecimal or dotted decimal) 255.255.255.0
     Network Interface Name (i.e ib0) ib0
     HCA Adapter iba0
     Adapter's port number 1
     ```

```
MTU 2044
```

To configure the ib1 interface on the iba1 InfiniBand HCA, port 1, with IP address/ netmask 192.168.2.1/255.255.255.0, use these values for those same fields:

```
INTERNET ADDRESS (dotted decimal) 192.168.2.1
Network MASK (hexadecimal or dotted decimal) 255.255.255.0
Network Interface Name (i.e ib0) ib1
HCA Adapter iba1
Adapter's port number 1M
MTU 2044
```

Validation:

```
smitty tcpip -> Further Configuration -> Network Interfaces -> Network
Interface Selection
-> List All Network Interfaces
```

or

```
smitty inet -> List All Network Interfaces
```

This should show the other configured network interfaces and the InfiniBand specific (ibX) interfaces:
ib0 IP over InfiniBand Network Interface
ib1 IP over InfiniBand Network Interface

- Using AIX CLI:
  Configuration (Using the SMIT examples):

```
mkiba -a 192.168.1.1 -i ib0 -p 1 -P -1 -A iba0 -S "up" -m "255.255.255.0" -M
2044
mkiba -a 192.168.2.1 -i ib1 -p 1 -P -1 -A iba1 -S "up" -m "255.255.255.0" -M
2044
```

Validation:
lsdev -CH | grep ib | grep IP

This should show:

```
ib0 Available IP over InfiniBand Network Interface
ib1 Available IP over InfiniBand Network Interface
```

```
netstat -in
```

This should show the ibx interfaces that were created, with a non-zero datalink address. Another verification would be to ping other nodes via the InfiniBand interfaces

4. Configure the VIPA interface:
   The VIPA interface (vi0) must be configured on a separate subnet from both the ib0 and ib1 interfaces.
   - Using SMIT:
     Configuration:

smitty tcpip -> Further Configuration -> Network Interfaces -> Network Interface Selection -> Add a Network Interface -> Add a Virtual IP Address Interface

or

smitty mkinetvi

Specify values for the following fields:

- **INTERNET ADDRESS** (dotted decimal)
- **Network MASK** (hexadecimal or dotted decimal)
- **Network Interface Name** (i.e vi0)
- **ACTIVATE the Interface after Creating it?**
- **Network Interface(s) using this VIPA**

<u>Example:</u>
To configure the vi0 interface using the ib0 and ib1 interfaces, with IP address/netmask 192.168.3.1/255.255.255.0, specify the fields listed above with the following values:

```
INTERNET ADDRESS (dotted decimal) 192.168.3.1
Network MASK (hexadecimal or dotted decimal) 255.255.255.0
Network Interface vi0
ACTIVATE the Interface after Creating it? yes
Network Interface(s) using this VIPA ib0,ib1
```

Validation:

```
smitty tcpip -> Further Configuration -> Network Interfaces -> Network
Interface Selection -> List All Network Interfaces
```

or

```
smitty inet -> List All Network Interfaces
```

This should show the other configured network interfaces and the VIPA specific (viX) interface:

```
vi0 Virtual IP Address Network Interface
```

- Using AIX CLI:

  Configuration (Using the SMIT example):

  ```
  mkdev -c if -s VI -t vi -a netaddr=192.168.3.1 -a netmask='255.255.255.0' -w
  'vi0' -a state='up' -a interface_names='ib0,ib1'
  ```

  Validation:
  lsdev -CH | grep vi | grep IP

  This should show:

  ```
  vi0 Available Virtual IP Address Network Interface
  ```

5. To load RDS, complete these steps:
   - Command to load RDS is : "bypassctrl load rds"

    - This needs to be done every time after reboot. Ensure that basic IPoIB tests are successful before loading RDS.
    - Validation: "genkex | grep rds" shows RDS. Alternately run "rdsctrl stats" to check if RDS is properly loaded.
    - Currently, unloading RDS Kernel Extension is not supported.

Back to top

Additional support

- FLRT: Fix Level Recommendation Tool

- Cluster support
- InfiniBand support
- RAC/RDS overview

Clustering Oracle RAC/RDS over InfiniBand Readme

Support for IBM Power Systems with POWER6 or POWER5 processors

Readme information

QLogic Switch LMC setting

The LMC setting in QLogic switches should be set to 0. Complete the steps in this section.

Before you begin

Before starting the configuration steps, note the following information:

- With the current Oracle RAC (OR) VIPA implementation, the two InfiniBand interfaces (ib0 and ib1) are required to be configured on separate subnets, with the corresponding VIPA interface (vi0) configured on a 3rd subnet.

- The current OR VIPA implementation supports a single node assigned to any IP Host Channel Adapter (HCA)
- The InfiniBand and VIPA interfaces and corresponding static routes (information on how to configure these static routes is provided in the "Multipath Routing, Active DGD and VIPA Configuration" section) are configured in the AIX ODM and are then available after this initial configuration and after subsequent reboots. Additional configuration is not required after this initial configuration.

---

1. Login to the QLogic InfiniBand switch with userid admin and the specific password.
2. Determine current LMC setting:
   - "?" to list Main Menu
   - SubnetManagement
   - smMasterLMC
3. Set the LMC to 0 (if necessary):
   - smMasterLMC 0
   - smcontrol restart

Additional support

- FLRT: Fix Level Recommendation Tool

Clustering Oracle RAC/RDS over InfiniBand Readme

Support for IBM Power Systems with POWER6 or POWER5 processors

Readme information

Multipath Routing, Active DGD, and VIPA Configuration

Multipath routing, DGD, and VIPA are AIX features that can be used to provide interface failover, single IP address over multiple IPoIB interfaces (VIPA), and load balancing.

To configure multipath Routing, Active DGD, and VIPA, omplete the following steps:

1. Configure IPoIB interfaces, for example ib0, ib1 etc. with addresses in different subnets e.g. net-A1/net-A2.
2. Configure VIPA interface e.g. vi0 with the IP address in a different subnet, for example net-B, with the InfiniBand interfaces underneath the VIPA interfaces.
3. Define two static routes to each remote VIPA address we need to reach from this node. These

routes need to be defined to use interfaces ib0 and ib1 and have Active DGD turned on. Router to reach remote VIPA is remote nodes InfiniBand interfaces. Please see the Appendix for a sample script to create static DGD routes.

4. Since there are two routes to the remote VIPA address, we get multipath routing policy. By default, there is a weighted round robin. If an InfiniBand port is down (cable is pulled), DGD will detect it and raise cost of the route to MAX, causing the other route to be used.
5. Thus we get load balancing/failover. Note that VIPA does not need ARP support because incoming datagrams get routed to it because it is internal interface.
6. TCP caches the route to remote host. When DGD detects loss of route to remote host, it increases the cost of failed route to MAX but this does not change cached route. This issue is fixed by turning on **passive_dgd** option using **no -p -o passive_dgd=1** .
7. Other DGD parameters may need tuning but **dgd_packets_lost, dgd_ping_time and dgd_retry_time** must be > 1. The default settings are recommended for all the above tunable items. Default settings are:
    - dgd_packets_lost = 3
    - dgd_ping_time = 5
    - dgd_retry_time = 5

   For example:

   The InfiniBand interfaces use the subnet 192.168.1.255 (ib0) and 192.168.2.255 (ib1). Note that ib0 and ib1 should belong to different subnets. The VIPA subnet is 192.168.3.255.

   ```
   Interfaces on Host 1 :
   ib0 2044 192.168.1 192.168.1.1
   ib1 2044 192.168.2 192.168.2.1
   vi0 0 . 192.168.3 192.168.3.1

   Static routes on Host 1 :
   Dest Gateway Flags i/f
   192.168.3.2 192.168.1.2 UGHA ib0
   192.168.3.2 192.168.2.2 UGHA ib1

   Interfaces on Host 2 :
   ib0 2044 192.168.1 192.168.1.2
   ib1 2044 192.168.2 192.168.2.2
   vi0 0 . 192.168.3 192.168.3.2

   Static routes on Host 2 :
   Dest Gateway Flags i/f
   192.168.3.1 192.168.1.1 UGHA ib0
   192.168.3.1 192.168.2.1 UGHA ib1
   ```

8. Static routes can be created using route command using following format:
   route add -host < destination VIPA > < router address (remote ib i/f) > -if < local ib i/f > -

active_dgd

For example:
route add -host 192.168.3.1 192.168.1.1 -if ib0 -active_dgd

If route is created from TCP/IP Further Configuration, Static routes panel, ODM will be updated to make route persistent across reboot. The ODM can be updated using chdev command as well.

Please see the example script below (Appendix 1) to create these static DGD routes for each host.

A. This script requires a single command line argument of a full path filename for a file containing the hostnames of the Oracle RAC cluster hosts (1 hostname per line) that will have the DGD static routes defined.
    - This file containing the Oracle RAC cluster hostnames must be located on each of the nodes in the Oracle RAC cluster.
    - These hostnames must be defined in the /etc/hosts file on each Oracle RAC cluster node.

B. This script also requires that the ib0, ib1, and vi0 interfaces all be defined in the /etc/hosts file on each Oracle RAC cluster node.

Above network configuration provides redundant routes to (vi0 addresses) of the nodes of The cluster.

For example:
If there is connectivity loss to subnet 192.168.1.255, nodes have connectivity to vi0 addresses of the nodes via subnet 192.168.2.255.

Please note that this configuration has disjoint subnets which means Asymmetrical double faults are not covered.

For example:
If Node A loses connectivity to subnet 192.168.1.255 while node B loses connectivity to subnet 192.168.2.255, we lose connectivity between node A and B.

Additional support

- FLRT: Fix Level Recommendation Tool

- Cluster support
- InfiniBand support
- RAC/RDS overview

Clustering Oracle RAC/RDS over InfiniBand Readme

Support for IBM Power Systems with POWER6 or POWER5 processors

Readme information

InfiniBand port failover

The keepalives help in knowing if the other end of the connection went down. If so, RDS can then flush the stale connection. There is another use for keepalives. It helps identify InfiniBand port failure. In this case, RDS can switch to an alternative port, if available.

Each InfiniBand port has a corresponding IPoIB interface defined. To enable Port Failover, a VIPA interface needs to be configured for the IP-o-IB interfaces. Refer to Section 2 above or the online AIX documentation for details on configuring VIPA.

The VIPA address should be used in the sendmsg calls. In Oracle RAC configuration use the VIPA address for the private interconnect (cluster_interconnect).

The RDS keepalive mechanism detects a port failure the same way it detects node failure. If an alternative InfiniBand port (and a corresponding IPoIB interface) is active, RDS transitions the existing InfiniBand connection to use the other port.

Additional support

- FLRT: Fix Level Recommendation Tool

- Cluster support
- InfiniBand support
- RAC/RDS overview

Clustering Oracle RAC/RDS over InfiniBand Readme

Support for IBM Power Systems with POWER6 or POWER5 processors

Readme information

rdsctrl utility and tuning parameters

For RDS statistics, for modifying the tunable items and for diagnostics, the rdsctrl utility is provided (/usr/sbin/rdsctrl). The utility can be used after RDS is loaded (bypassctrl load rds). Running the command "rdsctrl" with no arguments provides the usage.

Statistics

```
# rdsctrl stats
```

displays various RDS statistics.

The statistics can be reset using: # rdsctrl stats reset

Tuning Parameters

The following RDS parameters can be tuned after RDS is loaded, but before any RDS application is run.

- rds_sendspace: This refers to the high-water mark of the per-flow sendbuffer. There may be multiple flows per socket. The default value is 524288 bytes (512KB). The value is set using the command:
  # rdsctrl set rds_sendspace= < value in bytes >
- rds_recvspace: This refers to the per-flow high-water mark of the per-socket receive-buffer. For every additional flow to this socket, the receive high-water mark is bumped up by this value. The default value is 524288 bytes (512 KB). The value is set using the command:
  # rdsctrl set rds_recvspace= < value in bytes >
- rds_mclustsize: This refers to the size of the individual memory cluster, which is also the message fragment size. The default size is 16384 bytes (16 KB). The value, always a multiple of 4096, is set using the command:
  # rdsctrl set rds_mclustsize= < multiple of 4096, in bytes >

Warning: The rds_mclustsize value must be the same on all machines (nodes) in the cluster. Changing this value also has performance implications.

The current values for the parameters above can be retrieved using the following command:
# rdsctrl get < parameter >

To view the list of tunables, run the following command:
# rdsctrl get

The default values for these tunable items are recommended for use with Oracle RAC.

Data-structure dumps

Various RDS structures can be dumped for troubleshooting purposes. The command to use is:
# rdsctrl dump < structure >

The structures include:

- ibc (the details of the InfiniBand Reliable Connection)
- sendcb (the flow details)
- pcb (the RDS socket PCB details)

Note: RDS packets do not show up under commands like "netstat -i" and "netstat -s". These commands only show IP traffic.

Back to top

Additional support

- FLRT: Fix Level Recommendation Tool

- Cluster support
- InfiniBand support
- RAC/RDS overview

Clustering Oracle RAC/RDS over InfiniBand Readme

Support for IBM Power Systems with POWER6 or POWER5 processors

Readme information

- Overview and prerequisites
- Restrictions and limitations
- RDS configuration and loading
- QLogic switch LMC setting
- Multipath Routing, Active DGD, and VIPA configuration
- InfiniBand port failover

- rdsctrl utility and tuning parameters
- RDS and InfiniBand reliable connection mode
- RDS keepalives
- Known problems
- Cabling and appendix
- PDFs and Readme archive

RDS and InfiniBand RC mode

RDS uses the Reliable Connection (RC) mode of InfiniBand at the Network Layer for data transfer. For every remote node, one RC is created. A "remote node" is defined by a unique external IP address. If a particular remote node has 2 ports both active and configured with an IP-over-IB interface, then two RCs are created.

Per-RC Memory:

For every InfiniBand RC set up, the following private memory pools are created:

- InfiniBand Transmit pool: 1024 clusters of 16K each.
- InfiniBand Receive pool: 1024 clusters of 16K each.
- RDS Ack pool: 1024 clusters of 512B each.
- RDS Soname pool: 1024 clusters of 512B each.

Total per-RC memory: 35 MB

The details of the pools can be viewed using the following command:
# netstat -M

No RC is created for a loopback IP, i.e., for interfaces on the same machine. Loopback data transfer bypasses InfiniBand.

Additional support

- FLRT: Fix Level Recommendation Tool

Clustering Oracle RAC/RDS over InfiniBand Readme

Support for IBM Power Systems with POWER6 or POWER5 processors

Readme information

RDS keepalives

Once an InfiniBand reliable connection is established between two nodes (this happens on the first sendmsg to that node), a keepalive probe is sent every 5 seconds on the active side (the side that sends the first message) or 10 seconds on the passive side of inactivity (no current message traffic).

If there is no activity for 60 seconds on the active side or 30 seconds on the passive side, RDS infers that the other side went down, rebooted or crashed. So it destroys the InfiniBand connection since it is now obsolete. A new connection will be established on the next send or on receiving a request from the remote node. Currently, a keepalive is a message from and to port 33000.

Currently, the keepalive parameters can not be tuned.

Additional support

- FLRT: Fix Level Recommendation Tool

Clustering Oracle RAC/RDS over InfiniBand Readme

Support for IBM Power Systems with POWER6 or POWER5 processors

Readme information

Known problems

1. **Abstract:** Exec format error

   Symptom: Exec format error on "bypassctrl load rds"

   Workaround:
   InfiniBand has not been configured. Perform steps 1, 2, and 3 in RDS configuration and loading.

2. **Abstract:** socket: Addr family not supported by protocol

   Symptom: socket: Addr family not supported by protocol

Workaround:
RDS has not been loaded. Perform steps 5 in RDS configuration and loading.

3. **Abstract:** No route to host.

   Symptom: No route to host.

   Workaround:
   A non-existent remote IP was used in a sendmsg. Or the host is not reachable via the InfiniBand interface.

4. **Abstract:** sendmsg: No buffer space available.

   Symptom: Temporarily there are no buffers available.

   Workaround:
   This is a non-fatal error. RDS expects that the application retry after sleeping for approximately 5 seconds.

5. **Abstract:** ibX interface is down

   Symptom: If an ibX interface is down for more than 30 seconds using the "ifconfig ibX down" command, this ibx interface can not be restored to network connectivity using the "ifconfig ibX up" command.

   Workaround:
   The current workaround to restore this ibX interface to network connectivity is to perform the following steps:
   1. Remove the ibX interface using the rmdev command.
   2. Recreate the ibX interface using the mkiba command and the same configuration parameters it was previously configured with.
   3. Add the ibX interface to the vi0 iflist using the chdev command and the interface_names option.
   4. Recreate the ibX interface ADGD static routes as described in Multipath Routing, Active DGD and VIPA Configuration.
6. **Abstract:** IP address may result in RDS connection failures.

   Symptom: In some customer clusters, chosen IP address may result in RDS connections failures. Following discussion can be ignored, if you are not experiencing connection issues.

   Workaround:
   RDS uses values of least significant byte (LSBs) of node IP addresses to avoid simultaneous connection attempts by prioritizing nodes that have higher value compared with the destinations. When node with lower value tries to connect, the connection requests (up to

rds_conn_block_limit) are dropped. rds_conn_block_limit can be tuned using rdsctrl set option. Default value is 8000. To disable this feature, set rds_conn_block_limit to 0.

7. **Abstract:** Oracle RAC will evicts node after using **ifconfig ibX down** command.

   Symptom: If both the ib0 and ib1 interfaces are down for more than 10 seconds using the **ifconfig ibX down** command for a node, Oracle RAC will evict this node.

   Workaround:
   The current workaround is to ifconfig down only 1 ibX interface at a time (and for no more than 30 seconds). An alternate workaround is to remove only 1 interface at a time (using "rmdev -dl ibX") if the interface needs to be down longer than 30 seconds, then when done with the related maintenance for this ibX interface, recreate the interface with the following steps:
   1. Recreate the ibX interface using the mkiba command and the same configuration parameters it was previously configured with
   2. Add the ibX interface to the vi0 iflist using the chdev command and the interface_names option.
   3. Recreate the ibX interface ADGD static routes as described in Multipath Routing, Active DGD, and VIPA Configuration.

- **Abstract:** Oracle RAC will evicts node after using **rmdev -dl ibX** command.

Symptom: If both the ib0 and ib1 interfaces are removed at the same time using the **rmdev -dl ibX** interface command, Oracle RAC will evict this node.

Workaround:
The current workaround is to remove only 1 ibX interface at a time (using "rmdev -dl ibX"), then when done with the related maintenance for this ibX interface, recreate the interface with the following steps:

1. Recreate the ibX interface using the mkiba command and the same configuration parameters it was previously configured with
2. Add the ibX interface to the vi0 iflist using the chdev command and the interface_names option
3. Recreate the ibX interface ADGD static routes as described in Multipath Routing, Active DGD, and VIPA Configuration.

Back to top

Additional support

- FLRT: Fix Level Recommendation Tool

- Cluster support
- InfiniBand support
- RAC/RDS overview

Clustering Oracle RAC/RDS over InfiniBand Readme

Support for IBM Power Systems with POWER6 or POWER5 processors

Readme information

- Overview and prerequisites
- Restrictions and limitations
- RDS configuration and loading
- QLogic switch LMC setting
- Multipath Routing, Active DGD, and VIPA configuration
- InfiniBand port failover

- rdsctrl utility and tuning parameters
- RDS and InfiniBand reliable connection mode
- RDS keepalives
- Known problems
- Cabling and appendix
- PDFs and Readme archive

Cabling recommendations

In addition to consulting the "IBM System p HPC Clusters Fabric Guide using InfiniBand Hardware" for basic InfiniBand cabling information, please note the following recommendations for plugging and pulling InfiniBand cables.

1. Confirm InfiniBand HCA(s) properly installed:
   Prior to plugging or pulling an InfiniBand cable, confirm that the associated InfiniBand HCA (adapter) is firmly seated in the Managed Server's InfiniBand HCA slot. If the InfiniBand HCA appears loose or otherwise not completely and firmly seated, do not remove the associated cable (s) unless the Managed Server has been powered off and the affected InfiniBand HCA can be re-seated by the appropriate service personnel (most likely during a maintenance window).

2. Pulling cable(s):
   When pulling an InfiniBand cable from an InfiniBand HCA, pull the cable straight out as opposed to an up/down direction for vertically installed InfiniBand HCAs (for example, the 9117-570 Managed Server) or as opposed to a left/right direction for horizontally installed InfiniBand HCAs (for example, the 9119-590/595 Managed Servers).

   To confirm the InfiniBand HCA does not move (even slightly during the cable pull), hold the InfiniBand HCA at the base of the adapter.
3. Plugging cable(s)
   When plugging an InfiniBand cable from an InfiniBand HCA, plug the cable straight in and confirm that it is firmly attached.

   To confirm the InfiniBand HCA does not move (even slightly during the cable plug), hold the InfiniBand HCA at the base of the adapter.
4. Unplugging all cables from an InfiniBand switch. When unplugging all InfiniBand cables from an InfiniBand switch (for example, during InfiniBand switch hardware maintenance), the recommendation is to first power off the InfiniBand switch before performing the cable pulls.
5. Replacing an existing InfiniBand switch or combining a second cable connection onto a single InfiniBand HCA with an existing cable connection For an InfiniBand switch replacement and/or for any node that will have a cable change that combines multiple ibX interfaces onto the same InfiniBand HCA, please perform the following steps.
   A. Shutdown all applications. For Oracle RAC, stop the workloads and shutdown the DB instance(s).
   B. Remove all InfiniBand (ibX) and VIPA (vi0) interfaces for the cables to be changed These commands are to be issued on each node with an InfiniBand switch replacement and/or with a cable change performed.
      Example:
      ```
      rmdev -dl vi0
      rmdev -dl ib0
      rmdev -dl ib1
      ```
   C. Power off the target node(s) with the cable change(s) to be performed Perform this step from the CSM Management Server or HMC.
      Example:
      From the Management Server issue
      ```
      rpower -n < target node 1 > off
      rpower -n < target node 2 > off
      ```
   D. Perform the actual InfiniBand cable changes.
      1. Relocate the existing cable from the source InfiniBand adapter to the target InfiniBand adapter on the same node with an existing ibX/cable connection, and/or
      2. Relocate the InfiniBand cable from the old switch to the replacement InfiniBand switch.
      3. If performing an InfiniBand switch replacement, power on the replacement

InfiniBand switch and wait at least 5 minutes for it to complete its power on sequence.

6. Power on the node(s) that had the cable change(s) performed. Perform this step from the CSM Management Server or HMC.

   Example:
   From the Management Server issue the following:
   ```
   rpower -n < target node 1 > on
   rpower -n < target node 2 > on
   ```

7. Recreate the InfiniBand (ibX) and VIPA (vi0) interfaces for the node(s) with cable changes. These commands are to be issued on each node with an InfiniBand switch replacement and/or with a cable change performed.
   A. For nodes with an InfiniBand switch replacement only, the **mkiba** command should be run with the same ibX configuration parameters as were used for this InfiniBand interface's previous configuration.
   B. For nodes with a cable change from a second InfiniBand adapter to a single InfiniBand adapter with an existing InfiniBand interface, the mkiba command must be updated with the new ibaX adapter and port location information.

      Example:
      If the ib1 interface was previously configured for the iba1 adapter and port 1, and the ib1 cable has been moved to the iba0 adapter and port 2, then the mkiba command must specify the iba0 adapter and port 2 for the new ib1 interface configuration.

Appendix

```
# Purpose: Using a command line argument of a full path file name
containing a list of cluster
#
# hostnames, this script creates static DGD routes across the ib0 and
ib1 interfaces
#
# through the vi0 VIPA interface.
#
# #
# Example: make.static.dgd.routes /etc/hosts.oracle.rac.cluster
#
# #
##########################################

function create_dgd_route
{ unset mask
```

```
unset interface
mask=${5#-m}
interface=${6#-i}
if [ $1 = "host" -a -n "$mask" ] ; then
echo "netmask not allowed when adding a route with TYPE host"
exit 1
fi
if [ $mask ] ; then
if [ "`echo $mask | cut -c 1,2`" != "0x" ] ; then
unset i
unset j
i=5
if [ "`echo $mask | cut -f $i -d .`" != "" ] ; then
echo "Invalid netmask"
exit 1
else
i=`expr $i - 1`
while [ $i -gt 0 ] ; do
j=`echo $mask | cut -f $i -d .`
if [ "$j" = "" ] ; then
echo "Invalid netmask"
exit 1
fi
if [ $j -lt 0 -o $j -gt 255 ] ; then
echo "Warning : The netmask is not valid and may result to ambiguity"
fi
i=`expr $i - 1`
done
fi
fi
fi
if test -z 'lsdev -C -S1 -F name -l inet0'
then
mkdev -t inet
fi
unset arg2
if [ $mask ] ; then
arg2=-netmask,$mask
else
arg2=
fi
unset arg
if [ $8 = "yes" ] ; then
```

```
arg=-interface
else
arg=-hopcount,$4
fi
unset arg3
if [[ $interface != '' && $interface != 'any Use any available
interface' ]] ; then
arg3=`echo $interface | awk '{ print $1 }'`
arg3=-if,$arg3
else
arg3=
fi
unset arg4
if [ $7 = "yes" ] ; then
arg4=-active_dgd
else
arg4=
fi
unset arg5
if [ $9 = "0" ] ; then
arg5=
else
arg5=-policy,$9
fi
unset arg6
if [ ${10} = "1" ] ; then
arg6=
else

arg6=-weight,${10}
fi
unset arg7
if [ ${11} = "no" ] ; then
arg7=
else
arg7=-P
fi
chdev -l inet0 $arg7 -a route=$1,$arg,$arg2,$arg3,$arg4,$arg5,$arg6,
$2,$3
unset arg
unset arg2
unset arg3
unset arg4
unset arg5
unset arg6
unset arg7
unset mask
unset interface
}

##########################################

hosts_dgd_routes=$1

host_name=`hostname -s`
let i=0
while read c1 c2 c3
do

if [ $c2 != $host_name ]
then
let i=$i+1

echo
echo " $i. Creating static DGD routes for the ib0 and ib1 interfaces
for node $c2 ... 'date' "

ib0_interface=`grep $c2 /etc/hosts | grep ib0 | awk '{print $1}'`
ib1_interface=`grep $c2 /etc/hosts | grep ib1 | awk '{print $1}'`
vi0_interface=`grep $c2 /etc/hosts | grep vi0 | awk '{print $1}'`
create_dgd_route 'host' $vi0_interface $ib0_interface '0' -m'' -i'ib0
IP over InfiniBand Network Interface' 'yes' 'no' '0'
'1' 'no'
create_dgd_route 'host' $vi0_interface $ib1_interface '0' -m'' -i'ib1
IP over InfiniBand Network Interface' 'yes' 'no' '0'
'1' 'no'

fi

done < $hosts_dgd_routes
```

Back to top

Additional support

- FLRT: Fix Level Recommendation Tool

- Cluster support
- InfiniBand support
- RAC/RDS overview

Clustering Oracle RAC/RDS over InfiniBand Readme

Support for IBM Power Systems with POWER6 or POWER5 processors

Readme information

- Overview and prerequisites
- Restrictions and limitations
- RDS configuration and loading
- QLogic switch LMC setting
- Multipath Routing, Active DGD, and VIPA configuration
- InfiniBand port failover

- rdsctrl utility and tuning parameters
- RDS and InfiniBand reliable connection mode
- RDS keepalives
- Known problems
- Cabling and appendix
- PDFs and Readme archive

PDFs and Readme archives

Clustering Oracle RAC/RDS over InfiniBand Readme - 11/06/08 **[placeholder]**

Additional support

- FLRT: Fix Level Recommendation Tool