IP over IB Protocol

The following sections appear in this chapter:

- Introduction, page 3-1
- Manually Configure IPoIB for Default IB Partition, page 3-2
- Subinterfaces, page 3-2
- Verify IPoIB Functionality, page 3-5
- IPoIB Performance, page 3-6
- Sample Startup Configuration File, page 3-7

Introduction

Configuring IPoIB requires that you follow similar steps to those for configuring IP on an Ethernet network. When you configure IPoIB, you assign an IP address and a subnet mask to each HCA port. The first HCA port on the first HCA in the host is the ib0 interface, the second port is ib1, and so on.

Note

To enable these IPoIB settings across reboots, you must explicitly add these settings to the networking interface startup configuration file. For a sample configuration file, see the “Sample Startup Configuration File” section on page 3-7.

Refer to your Linux distribution documentation for additional information about configuring IP addresses.

Note

See the “Root and Non-root Conventions in Examples” section on page 1-4 for details about the significance of prompts used in the examples in this chapter.
Manually Configure IPoIB for Default IB Partition

To configure IPoIB on your Linux host, perform the following steps:

**Step 1** Log in to your Linux host.

**Step 2** To configure the interface, enter the `ifconfig` command with the following:

- The appropriate IB interface (`ib0` or `ib1` on a host with one HCA)
- The IP address that you want to assign to the interface
- The `netmask` keyword
- The subnet mask that you want to assign to the interface

This example shows how to configure an IB interface:

```
host1# ifconfig ib0 192.168.0.1 netmask 255.255.255.0
```

**Step 3** (Optional) Enter the `ifconfig` command with the appropriate port identifier `ib#` argument to verify the configuration.

This example shows how to verify the configuration:

```
host1# ifconfig ib0
ib0        Link encap:UNSPEC  HWaddr 00-00-00-00-00-00-00-00-00-00-00-00-00-00-00-00
inet addr:192.168.0.1  Bcast:192.168.0.255  Mask:255.255.255.0
inet6 addr: fe80::205:ad00:20:849/64 Scope:Link
      UP BROADCAST RUNNING MULTICAST  MTU:2044  Metric:1
      RX packets:46 errors:0 dropped:0 overruns:0 frame:0
      TX packets:47 errors:0 dropped:0 overruns:0 carrier:0
      collisions:0 txqueuelen:128
      RX bytes:45056 (44.0 KiB) TX bytes:3011 (2.9 KiB)
```

**Step 4** Repeat step 2 and step 3 on the remaining interface(s).

Subinterfaces

Subinterfaces divide primary (parent) interfaces to provide traffic isolation. Partition assignments distinguish subinterfaces from parent interfaces. The default Partition Key (`p_key`), ff:ff, applies to the primary (parent) interface.

This section includes the following topics:

- Create a Subinterface Associated with a Specific IB Partition, page 3-3
- Remove a Subinterface Associated with a Specific IB Partition, page 3-4
Create a Subinterface Associated with a Specific IB Partition

This section describes how to create a subinterface associated with a specific IB partition.

To create a subinterface, perform the following steps:

---

**Step 1**
Create a partition on an IB server switch. Alternatively, you can choose to create the partition of the InfiniBand interface first, and then create the partition for the ports on the IB server switch. Refer to the Cisco SFS 7000 Series Product Family Element Manager User Guide for information regarding partitions on the IB server switch.

**Step 2**
Log in to your host.

**Step 3**
Add the value of the partition key to the file as root user.

This example shows how to add partition 80:02 to the primary interface ib0:

```
host1# echo 0x8002 >> /sys/class/net/ib0/create_child
```

**Step 4**
Verify that the interface is set up by ensuring that ib0.8002 is displayed.

This example shows how to set up the interface:

```
host1# ls /sys/class/net
eth0  ib0  ib0.8002  ib1  lo  sit0
```

**Step 5**
Enter the `ifconfig -a` command to verify that the interface was created, as shown in this example:

```
host1# ifconfig -a
eth0  Link encap:Ethernet  HWaddr 00:30:48:20:D5:D1
      inet addr:172.29.237.206  Bcast:172.29.239.255  Mask:255.255.252.0
      inet6 addr: fe80::230:48ff:fe20:d5d1/64 Scope:Link
      UP BROADCAST RUNNING MULTICAST  MTU:1500  Metric:1
      RX packets:17591 errors:0 dropped:0 overruns:0 frame:0
      TX packets:4831 errors:0 dropped:0 overruns:0 carrier:0
      collisions:0 txqueuelen:1000
      RX bytes:8704196 (8.3 MiB)  TX bytes:446771 (436.2 KiB)
      Base address:0x3040 Memory:dd420000-dd440000

ib0  Link encap:UNSPEC  HWaddr 00-00-00-00-00-00-00-00-00-00-00-00-00-00-00-00
     BROADCAST MULTICAST  MTU:2044  Metric:1
     RX packets:17591 errors:0 dropped:0 overruns:0 frame:0
     TX packets:4831 errors:0 dropped:0 overruns:0 carrier:0
     collisions:0 txqueuelen:128
     RX bytes:8704196 (8.3 MiB)  TX bytes:446771 (436.2 KiB)
     Base address:0x3040 Memory:dd420000-dd440000

ib0.8002 Link encap:UNSPEC  HWaddr 00-00-00-00-00-00-00-00-00-00-00-00-00-00-00-00
          BROADCAST MULTICAST  MTU:2044  Metric:1
          RX packets:0 errors:0 dropped:0 overruns:0 frame:0
          TX packets:0 errors:0 dropped:0 overruns:0 carrier:0
          collisions:0 txqueuelen:128
          RX bytes:0 (0.0 b)  TX bytes:0 (0.0 b)

ib1  Link encap:UNSPEC  HWaddr 00-00-00-00-00-00-00-00-00-00-00-00-00-00-00-00
     BROADCAST MULTICAST  MTU:2044  Metric:1
     RX packets:0 errors:0 dropped:0 overruns:0 frame:0
     TX packets:0 errors:0 dropped:0 overruns:0 carrier:0
     collisions:0 txqueuelen:128
     RX bytes:0 (0.0 b)  TX bytes:0 (0.0 b)

lo   Link encap:Local Loopback
     inet addr:127.0.0.1  Mask:255.0.0.0
     inet6 addr: ::1/128 Scope:Host
     UP LOOPBACK RUNNING  MTU:16436  Metric:1
     RX packets:136 errors:0 dropped:0 overruns:0 frame:0
```
Verify that you see the ib0.8002 output.

**Step 6** Configure the new interface just as you would the parent interface. (See the “Manually Configure IPoIB for Default IB Partition” section on page 3-2.)

This example shows how to configure the new interface:

```
host1# ifconfig ib0.8002 192.168.12.1 netmask 255.255.255.0
```

### Remove a Subinterface Associated with a Specific IB Partition

This section describes how to remove a subinterface.

To remove a subinterface, perform the following steps:

**Step 1** Take the subinterface offline. You cannot remove a subinterface until you bring it down.

This example shows how to take the subinterface offline:

```
host1# ifconfig ib0.8002 down
```

**Step 2** Remove the value of the partition key to the file as root user.

This example shows how to remove the partition 80:02 from the primary interface ib0:

```
host1# echo 0x8002 >> /sys/class/net/ib0/delete_child
```

**Step 3** (Optional) Enter the `ifconfig -a` command to verify that the subinterface no longer appears in the interface list.

This example shows how to verify that the subinterface no longer appears in the interface list:

```
host1# ifconfig -a
```

```
eth0  Link encap:Ethernet  HWaddr 00:30:48:20:D5:D1
  inet addr:172.29.237.206  Bcast:172.29.239.255  Mask:255.255.252.0
  inet6 addr: fe80::230:48ff:fe20:d5d1/64 Scope:Link
  UP BROADCAST RUNNING MULTICAST  MTU:1500  Metric:1
  RX packets:21431 errors:0 dropped:0 overruns:0 frame:0
  TX packets:5474 errors:0 dropped:0 overruns:0 carrier:0
  collisions:0 txqueuelen:1000
  RX bytes:9542238 (9.1 MiB)  TX bytes:562793 (549.6 KiB)
  Base address:0x3040 Memory:dd420000-dd440000

ib0  Link encap:UNSPEC  HWaddr 00:00:00:00:00:00:00:00:00:00:00:00:00:00:00
  inet addr:192.168.0.1  Bcast:192.168.0.255  Mask:255.255.255.0
  inet6 addr: fe80::205:ad00:20:849/64 Scope:Link
  UP BROADCAST RUNNING MULTICAST  MTU:2044  Metric:1
  RX packets:46 errors:0 dropped:0 overruns:0 frame:0
  TX packets:47 errors:0 dropped:0 overruns:0 carrier:0
  collisions:0 txqueuelen:128
  RX bytes:952238 (9.1 MiB) TX bytes:562793 (549.6 KiB)
  Base address:0x30340 Memory:dd420000-dd440000

ib1  Link encap:UNSPEC  HWaddr 00:00:00:00:00:00:00:00:00:00:00:00:00:00:00
  inet addr:192.168.0.1  Bcast:192.168.0.255  Mask:255.255.255.0
  inet6 addr: fe80::205:ad00:20:849/64 Scope:Link
  UP BROADCAST RUNNING MULTICAST  MTU:2044  Metric:1
  RX packets:46 errors:0 dropped:0 overruns:0 frame:0
  TX packets:47 errors:0 dropped:0 overruns:0 carrier:0
  collisions:0 txqueuelen:128
  RX bytes:952238 (9.1 MiB) TX bytes:562793 (549.6 KiB)
  Base address:0x30340 Memory:dd420000-dd440000
```
Verify IPoIB Functionality

This section displays how to verify your configuration and your IPoIB functionality.

**Step 1**
Log in to your hosts.

**Step 2**
Verify the IPoIB functionality by using the `ifconfig` command.

This example shows how two IB nodes are used to verify IPoIB functionality. In this example, IB node 1 is at 192.168.0.1, and IB node 2 is at 192.168.0.2.

```
host1# ifconfig ib0 192.168.0.1 netmask 255.255.252.0
host2# ifconfig ib0 192.168.0.2 netmask 255.255.252.0
```

**Step 3**
Enter the `ping` command from 192.168.0.1 to 192.168.0.2.

```
host1# ping -c 5 192.168.0.2
PING 192.168.0.2 (192.168.0.2) 56(84) bytes of data.
64 bytes from 192.168.0.2: icmp_seq=0 ttl=64 time=0.079 ms
64 bytes from 192.168.0.2: icmp_seq=1 ttl=64 time=0.044 ms
64 bytes from 192.168.0.2: icmp_seq=2 ttl=64 time=0.055 ms
64 bytes from 192.168.0.2: icmp_seq=3 ttl=64 time=0.049 ms
64 bytes from 192.168.0.2: icmp_seq=4 ttl=64 time=0.065 ms

--- 192.168.0.2 ping statistics ---
5 packets transmitted, 5 received, 0% packet loss, time 3999ms rtt min/avg/max/mdev = 0.044/0.058/0.079/0.014 ms, pipe 2
```
IPoIB Performance

This section describes how to verify IPoIB performance by running the Bandwidth test and the Latency test. These tests are described in detail at this URL:

http://www.netperf.org/netperf/training/Netperf.html

**Step 1** Download Netperf from this URL:


**Step 2** Follow the instructions at http://www.netperf.org/netperf/NetperfPage.html to compile Netperf.

**Step 3** Start the Netperf server.

This example shows how to start the Netperf server:

```
host1% netserver
Starting netserver at port 12865
Starting netserver at hostname 0.0.0.0 port 12865 and family AF_UNSPEC
host1%
```

**Step 4** Run the Netperf client. The default test is the Bandwidth test.

This example shows how to run the Netperf client, which starts the Bandwidth test by default:

```
host2% netperf -H 192.168.0.1 -c -C -- -m 65536
TCP STREAM TEST from 0.0.0.0 (0.0.0.0) port 0 AF_INET to 192.168.0.1 (192.168.0.1) port 0 AF_INET
Recv  Send  Send                          Utilization       Service Demand
      Send  Socket Socket  Message  Elapsed  Send     Recv     Send    Recv
      Size  Size   bytes  bytes  secs.  10^6bits/s  % S      % S      us/KB   us/KB
87380  16384  65536    10.00      2903.14   25.29    25.64    2.855   2.894
```

**Note** You must specify the IPoIB IP address when running the Netperf client.

The following list describes parameters for the **netperf** command:

- **-H** Where to find the server
- **192.168.0.1** IPoIB IP address
- **-c** Client CPU utilization
- **-C** Server CPU utilization
- **--** Separates the global and test-specific parameters
- **-m** Message size, which is 65536 in the example above

The notable performance values in the example above are as follows:

Throughput is 2.90 gigabits per second.
Client CPU utilization is 25.29 percent of client CPU.
Server CPU utilization is 25.64 percent of server CPU.

**Step 5** Run the Netperf Latency test.

Run the test once, and stop the server so that it does not repeat the test.

This example shows how to run the Latency test and then stop the Netperf server:

```
host2% netperf -H 192.168.0.1 -c -C -t TCP_RR -- -r 1,1
```
TCP REQUEST/RESPONSE TEST from 0.0.0.0 (0.0.0.0) port 0 AF_INET to 192.168.0.1
(192.168.0.1) port 0 AF_INET
Local /Remote

<table>
<thead>
<tr>
<th>Socket Size</th>
<th>Request Size</th>
<th>Resp. Size</th>
<th>Elapsed Time</th>
<th>Rate</th>
<th>local CPU</th>
<th>remote CPU</th>
<th>local S.dem</th>
<th>remote S.dem</th>
</tr>
</thead>
<tbody>
<tr>
<td>Send bytes</td>
<td>bytes</td>
<td>bytes</td>
<td>secs.</td>
<td>per sec</td>
<td>% S</td>
<td>% S</td>
<td>us/Tr</td>
<td>us/Tr</td>
</tr>
<tr>
<td>16384</td>
<td>87380</td>
<td>1</td>
<td>10.00</td>
<td>30927.36</td>
<td>13.06</td>
<td>13.82</td>
<td>16.896</td>
<td>17.878</td>
</tr>
</tbody>
</table>

Stop netperf server.

```
host1% pkill netserver
```

The following list describes parameters for the `netperf` command:

- **-H** Where to find the server
- **192.168.0.1** IPoIB IP address
- **-c** Client CPU utilization
- **-C** Server CPU utilization
- **-t** Test type
- **TCP_RR** TCP required response test
- **--** Separates the global and test-specific parameters
- **-r 1,1** The request size sent and how many bytes requested back

The notable performance values in the example above are as follows:

- Client CPU utilization is 13.06 percent of client CPU.
- Server CPU utilization is 13.82 percent of server CPU.
- Latency is 16.896 microseconds and 17.878 microseconds.

**Sample Startup Configuration File**

IP addresses that are configured manually are not persistent across reboots. You must use a configuration file to configure IPoIB when the host boots. The sample configuration below shows an example file named `ifcfg-ib0` that resides on a Linux host in `/etc/sysconfig/network-scripts/` on RHEL4 and in `/etc/syconfig/network/` on SLES10. The configuration file configures an IP address at boot time.

```
host1# cat > /etc/sysconfig/network-scripts/ifcfg-ib0 << EOF
> DEVICE=ib0
> BOOTPROTO=static
> IPADDR=192.168.0.1
> NETMASK=255.255.255.0
> ONBOOT=yes
> EOF
```