Abstract

As Infiniband has evolved and bandwidth available to Infiniband-attached HPC clusters has climbed, many HPC systems find themselves with compute nodes attached to both an ethernet infrastructure (for management and network storage access) and an Infiniband infrastructure (for node IPC traffic). The most current infiniband specification makes 32 Gb/s of bandwidth available to each compute node, yet many HPC cluster nodes are still using 1 Gb/s Ethernet to attach to network storage resources. In this paper, we will examine the viability of using whitebox Infiniband-to-ethernet "gateway" systems to allow HPC clusters to more efficiently use the bandwidth that Infiniband provides, especially as it relates to storage traffic. We will explore the configuration, performance characteristics, and architectural consequences of using IB-to-ethernet gateways.

Topology Overview

(diagram needs lots of cleanup)
Host Configuration

- Install OFED
- Install 10GbE NIC drivers
- disable OFED driver load for 10GbE NIC, if present
- configure ECMP for default gateways

Gateway Configuration

Kernel IP Routing

Turn on IPv4 Kernel Routing

1. Turn on forwarding

```
shell# echo 1 > /proc/sys/net/ipv4/ip_forward
```

2. Make the setting persistent

Edit /etc/sysctl.conf and add the following line:

```
net.ipv4.ip_forward = 1
```
Add routes to the kernel routing table

To add static routes, use the `route` command. For a minimal configuration, you would only need 1 route for each external network you want to reach. You’ll also need a route on the external network pointing to the IPoIB network behind the gateway.

```
shell# route add -net <destination> netmask <destination mask, dotted decimal format> gw <next-hop>
```

High Availability

- Two approaches: (1) failure detection in the gateways (2) failure detection on individual hosts

Approach 1: Failure detection in the gateways

- RHEL Cluster Suite Configuration Steps here

Approach 2: Failure detection on individual hosts

- Panasas check_routers script example here

Performance Characteristics

- Maximum throughput per Gateway (with RRD graphs)
- Throughput in the face of gateway failure

Infiniband Subnet Manager Placement

Infiniband Subnet Manager Configuration Notes

Hardware Specs and Notes

Conclusions

Testing Procedures

iperf parameters

iozone parameters

OFED tools syntax

To find a device lid:

```
# ibndetdiscover | grep <hostname>
```

```
[root@up156 ~]# ibnetdiscover | grep gateway2
[8] "H-0002c9020028bf54"[1](2c9020028bf55)          # "gateway2 HCA-1" lid 295 4xDDR
Ca      2 "H-0002c9020028bf54"          # "gateway2 HCA-1"
```

To view counters on an IB port:

```
# perfquery <lid> <port>
```
To zero counters on an IB port:

```
# perfquery -R <lid> <port>
```

Batch script for automated iperf tests

Sources and Acknowledgements