Evaluating the Impact of Infiniband Routing Algorithms on Network Performance

Fabrice Mizero
Philander Smith College, SIParCS
Mentor: Dr. John Dennis

Collaborators:
Prof. Malathi Veeraraghavan, Zhengyang Liu, UVA
Dr. Robert D. Russell, Patrick MacArthur, UNH
08/01/2013
Roadmap

- Motivation
- Routing Algorithms: Emphasis on UpDn Scatter-Ports
- How routing works: Emphasis Subnet Manager
- Link Failures in Infiniband Networks
- Subnet Manager reaction to Infiniband Link failures
- Experiments
- Results and Observations
Motivation

- Application performance variability – CESM

Execution Time for ASD on Yellowstone

Execution time for non-I/O CESM day

Large Variability in execution time
CAM Scalasca Analysis

Nodes

ys5432
ys5433
ys5434
ys5436
ys5437
ys5438
ys5439
ys5440
ys5441
ys5443
ys5444
ys5448
ys5453
ys5455
ys5456
ys5458
ys5459
ys5460

Time in MPI_Waitall [s]
Possible Explanation

Execution time variability in CESM

Slow communications affecting ys5456

Routing Table Recalculations
Motivation

- A better understanding of Infiniband routing, routing algorithms, and subnet management.
- Need of low latency and high throughput for Large Scale parallel message passing applications such as Community Earth System Model (CESM)
Routing Algorithms

The Infiniband Architecture currently supports:

- UpDown
- UpDown -- Scatter Ports

Others:

- FatTree
- Minhop
- LASH
- DOR

The choice of Routing Algorithms largely depends:

- Network Topology
- Expected nature of traffic and applications demands
Routing Algorithms

Yellowstone - UpDown – Scatter Ports

How it works:

- **3 Steps:**
  - Auto-Detection of Root Nodes
  - Ranking Process
  - Minhop table setting

- **Advantages:**
  - Randomness in port selection
  - Reduces credit loops potential by reducing number of routes
  - Better adaptation to link failures. (it’s not topology-bound like Fattree)
Subnet Management in Infiniband Networks

- Subnet Manager
  - Infiniband compliant subnet manager – OpenSM
  - Tasks:
    - Initialize Infiniband Hardware
    - Local Identifiers Assignment
    - Routing Table Calculations & Distributions
    - Regularly Sweeps for changes in the Topology
    - (reassign lids) -r

Routing Recalculation is a huge task in Large Scale Networks
Link Failures in Infiniband Networks

- **Main cause:**
  - Dysfunctional cables

- **Impact on network performance:**
  - Higher latencies
  - Possible packets drop due timeouts
  - Overall poor performance of parallel message passing applications.
Topo-file Example in Use

32 nodes, 3 levels, full symmetrical FatTree
Infiniband Routing On a Healthy Subnet
Destination-Based Routing & Credit Based Flow Control

Destination LID compared to Current LID

Consult Routing Table
Find Port for Dest. LID

Request for Buffer Space Availability

Send Packet
Wait for Credits
Subnet Manager Adaptation to Link Failures

OpenSM scheduled Sweeps

Link Failure Detected

Find Directly Affected Switches

Update Routing Tables in both Switches

Subnet UP
Experiments

- Tools:
  - Infiniband Management Simulator (IBMgtSim)
  - Subnet Manager (OpenSM)

- Opensm Logs: Calculate subnet recovery times.
Observations

- Subnet Management can lower network performance.
- The more the number of dysfunctional links in the network, the more overall latency is affected by subnet management.

Future Work

- Causes of link failures?
- Evaluating ways in which OpenSM can adapt faster to changes.
Acknowledgments

- John Dennis – Mentor
- Collaborators – UVA, UNH
- Zhengyang Liu - Colleague & Friend
- Babak Behzad, Sean Fisk, Joseph Usset - For all the enriching discussions.
Thank You for Your Attention!

Q&A

Fabrice Mizero
mizero.fabrice@philander.edu
CS Junior
Philander Smith College