Network Contention and Congestion Control: Lustre Fine-Grained Routing

Matt Ezell
HPC Systems Administrator

International Workshop on The Lustre Ecosystem: Challenges and Opportunities
Annapolis, Maryland  March 3-4, 2015
Congestion  Performance
Fat Tree Network (Folded Clos Network)

Image from http://clusterdesign.org/fat-trees/60_nodes_labeled/
Fat Tree Network (Folded Clos Network)

• Network can be
  – Non-blocking (1:1)
  – Blocking (x:1)

• Congestion and hot spots can (significantly) reduce performance
  – Even in an expensive full-bisection-bandwidth setup
  – Worse in under-provisioned or irregular networks
InfiniBand Routing

• InfiniBand uses a Linear Forwarding Table
  – A given destination always uses the same output port

<table>
<thead>
<tr>
<th>Destination LID</th>
<th>Output Port</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>59</td>
<td>3</td>
</tr>
<tr>
<td>66</td>
<td>3</td>
</tr>
<tr>
<td>101</td>
<td>5</td>
</tr>
<tr>
<td>121</td>
<td>11</td>
</tr>
<tr>
<td>200</td>
<td>11</td>
</tr>
<tr>
<td>232</td>
<td>20</td>
</tr>
<tr>
<td>299</td>
<td>27</td>
</tr>
<tr>
<td>341</td>
<td>32</td>
</tr>
<tr>
<td>351</td>
<td>35</td>
</tr>
</tbody>
</table>
What happens when
node 1 => node 3
node 2 => node 5
Gemini 3D Torus

Image from http://clusterdesign.org/torus/
Gemini Dimension Ordered Routing

X → Y → Z
How fast are Gemini Links?

• It depends!
• Link speed depends on link type
• Protocol overhead is around 35% for large messages

<table>
<thead>
<tr>
<th>Link Type</th>
<th>Link Data Rate</th>
<th>Number Links</th>
<th>Raw Bitrate</th>
<th>Data Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y-Mezzanine</td>
<td>6.25 gbps</td>
<td>12</td>
<td>9.375 GB/s</td>
<td>~ 6 GB/s</td>
</tr>
<tr>
<td>Z-Backplane</td>
<td>5.0 gbps</td>
<td>24</td>
<td>15 GB/s</td>
<td>~ 9.75 GB/s</td>
</tr>
<tr>
<td>X, Z Cable</td>
<td>3.125 gbps</td>
<td>24</td>
<td>9.375 GB/s</td>
<td>~ 6 GB/s</td>
</tr>
<tr>
<td>Y Cable</td>
<td>3.125 gbps</td>
<td>12</td>
<td>4.6875 GB/s</td>
<td>~ 3 GB/s</td>
</tr>
</tbody>
</table>
Geometric Bandwidth Reductions

Titan is a 25 x 16 x 24 torus
Dragonfly Network

Group (384 Nodes)

Group (384 Nodes)

Group (384 Nodes)

Group (384 Nodes)

Group (384 Nodes)

Group (384 Nodes)

Group (384 Nodes)

Group (384 Nodes)
Dragonfly Network

• In an Aries network, global bandwidth can be increased by purchasing more links

• Adaptive routing can ease congestion
  – But it makes it difficult to predict the path that a packet will take
The Million Dollar Question
LNET Routing
LNET Routing

identifier@network
10.10.10.101@o2ib0
10.10.10.101@o2ib201
9409@gni0
9409@gni101
LNET Routing

options Inet routes="remoteNet hops id@localNet"

lctl --net remoteNet add_route id@localNet hops

<table>
<thead>
<tr>
<th>net</th>
<th>o2ib225</th>
<th>hops 10</th>
<th>gw</th>
<th>18329@gni102 up</th>
</tr>
</thead>
<tbody>
<tr>
<td>net</td>
<td>o2ib225</td>
<td>hops 1</td>
<td>gw</td>
<td>10871@gni102 up</td>
</tr>
<tr>
<td>net</td>
<td>o2ib234</td>
<td>hops 10</td>
<td>gw</td>
<td>5991@gni102 up</td>
</tr>
<tr>
<td>net</td>
<td>o2ib234</td>
<td>hops 10</td>
<td>gw</td>
<td>18247@gni102 up</td>
</tr>
<tr>
<td>net</td>
<td>o2ib234</td>
<td>hops 1</td>
<td>gw</td>
<td>10921@gni102 up</td>
</tr>
<tr>
<td>net</td>
<td>o2ib208</td>
<td>hops 10</td>
<td>gw</td>
<td>15218@gni102 up</td>
</tr>
<tr>
<td>net</td>
<td>o2ib208</td>
<td>hops 10</td>
<td>gw</td>
<td>2946@gni102 up</td>
</tr>
<tr>
<td>net</td>
<td>o2ib208</td>
<td>hops 1</td>
<td>gw</td>
<td>7788@gni102 up</td>
</tr>
<tr>
<td>net</td>
<td>o2ib217</td>
<td>hops 10</td>
<td>gw</td>
<td>15212@gni102 up</td>
</tr>
<tr>
<td>net</td>
<td>o2ib217</td>
<td>hops 10</td>
<td>gw</td>
<td>2908@gni102 up</td>
</tr>
</tbody>
</table>
LNET Routing

Start

remote net == local net?

Yes → Deliver Directly

No → Find Best Router

Deliver to Router

End
Why Route?

- Control *bandwidth* by varying the number of routers
- Control *I/O paths* by selecting which routers to use
- Control *route computation* by partitioning fabrics

... and if you are using multiple network types, you have to route traffic
Initial Implementation for Jaguar

- Focused on relieving contention within the IB network only
Atlas Layout
Atlas Layout

InfiniBand Switch A
InfiniBand Switch B
InfiniBand Switch C
InfiniBand Switch D
InfiniBand Switch E
InfiniBand Switch F
InfiniBand Switch G
InfiniBand Switch H
InfiniBand Switch I

OSSs
OSS Connections

- DDN 1i1
- DDN 1i2
- InfiniBand Switch I

- OSS1i1
- OSS1i2
- OSS1i3
- OSS1i4
- OSS1i5
- OSS1i6
- OSS1i7
- OSS1i8

- DDN 1a1
- DDN 1a2
- InfiniBand Switch A

- OSS1a1
- OSS1a2
- OSS1a3
- OSS1a4
- OSS1a5
- OSS1a6
- OSS1a7
- OSS1a8

- DDN 1b1
- DDN 1b2
- InfiniBand Switch B

- OSS1b1
- OSS1b2
- OSS1b3
- OSS1b4
- OSS1b5
- OSS1b6
- OSS1b7
- OSS1b8

- InfiniBand Switch C
How many routers?

Performance Goal: > 1TB/s

÷

Router Performance: 2.6GB/s

= 

Required Router Count: > 385

Deployed Router Count: > 432
FGR Group

InfiniBand Switch B

8:12 Ratio
Router Module

Node 0  Gemini 0
Node 1  Gemini 1
Node 2
Node 3

Atlas1
InfiniBand Switch 1A
InfiniBand Switch 2A
InfiniBand Switch 3A
InfiniBand Switch 4A

Atlas2
Router Layout
Router Placement
Fine-Grained Routing Configuration

InfiniBand

36

Cray

12
Routing Group
FGR On an XC30

- Our XC30 is only two groups
  - We have not measured significant contention in the Aries network
- 9 routers with 4 interfaces each, to provide connectivity to all 36 TOR switches
- Routers provide higher percentage of peak bandwidth, compared to our XK7
FGR on an IB Cluster

• Compute nodes and routers connect to a 648-port core switch
• 36 routers connect directly to the 36 TOR switches
FGR Best Practices

• Understand your topology
• Understand your application I/O workload
• Create large FGR groups
• Create FGR groups in both directions
• Verify all settings and physical connections
Questions?

ezellma@ornl.gov