EARLY EVALUATION OF DIRECT LARGE-SCALE INFINIBAND NETWORKS WITH ADAPTIVE ROUTING

March 31, 2015

Alexander Daryin
Head of Advanced Research Lab
alexander.daryin@t-platforms.ru

www.t-platforms.com
Overview

- Introduction
- Reference Configuration
- Topologies
- Suggested IB Adaptive Routing Features
- Adaptive Routing Algorithms
- Performance Evaluation

- Published paper:
Introduction

- **InfiniBand**: 45% of Nov 14 Top500 list

- Static (deterministic) routing
  - In-order packet delivery

- High-radix switches
  - Low-diameter topologies
  - High concentration
  - Need for advanced routing algorithms
Reference Configuration

- T-Platforms A-Class
  - Concentration 8 nodes/switch
  - 36 ports/switch
  - 32 switches/rack

- 32 twin racks (maximum 48):
  - 2048 switches, 16384 nodes

- Theoretical bounds:
  - Diameter: 3
  - Relative bisection:
    - upper bound $\approx 163\%$
    - Ramanujan graphs: 100%
    - practical topologies: 50%
Direct Topologies

- Flattened Butterfly
- Dragonfly
- Slim Fly
- Torus
- Hypercube
Flattened Butterfly

- Cartesian product of full graphs
- Alternative names:
  - Generalized Hypercube
  - HyperX

- Configuration:
  - Dimensions: $4 \times 8 \times 8 \times 8$
  - Link widths: $2, 1, 1, 1$
  - Diameter: 4
  - Radix: 35
  - Relative bisection: 50%
Low Diameter Topologies

- **Dragonfly** (Kim, Dally, Scott, Abts 2008)
  - Configuration:
    - 128 Groups of 16 switches
    - 8 global links/switch
    - Diameter: 3
    - Radix: 30
    - Relative bisection: 50%

- **SlimFly** (Besta, Hoefler 2014)
  - MMS graph, diameter 2
  - Configuration:

<table>
<thead>
<tr>
<th></th>
<th>Slim Fly q</th>
<th>L</th>
<th>FlatFly $K_n$</th>
<th>$L_n$</th>
<th>Degree</th>
<th>Diameter</th>
<th>Bisection</th>
</tr>
</thead>
<tbody>
<tr>
<td>SF×FF-1</td>
<td>4</td>
<td>2</td>
<td>$8 \times 8$</td>
<td>1, 1</td>
<td>34</td>
<td>4</td>
<td>50%</td>
</tr>
<tr>
<td>SF×FF-2</td>
<td>8</td>
<td>1</td>
<td>16</td>
<td>1</td>
<td>35</td>
<td>3</td>
<td>50%</td>
</tr>
</tbody>
</table>
Classical Topologies

- **Torus**: Maximum dimension: 4D (3D supported in OpenSM)

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Size</th>
<th>Link Widths</th>
<th>Degree</th>
<th>Diameter</th>
<th>Bisection</th>
</tr>
</thead>
<tbody>
<tr>
<td>3D</td>
<td>8 x 16 x 16</td>
<td>3, 5, 5</td>
<td>34</td>
<td>20</td>
<td>15,6%</td>
</tr>
<tr>
<td>4D</td>
<td>4 x 8 x 8 x 8</td>
<td>2, 4, 4, 4</td>
<td>36</td>
<td>14</td>
<td>25,0%</td>
</tr>
</tbody>
</table>

- **Hypercube**: Particular case of both Torus and Flattened Butterfly
  - Dimension (=diameter): 11D
  - Relative bisection: 25%
  - Radix: 30
Standard IB Routing

- **LID (Local ID):** 16-bit network address (48K addresses)
- **SL (Service Level):** 4-bit packet tag
- **VL (Virtual Lane):** 8 lanes per port

Routing mechanisms:
- **PathSL:** select packet SL by source and destination
- **LFT (Linear Forwarding Table):** select output port by DLID
- **SL2VL:** select output VL by SL, input and output port
- **LMC (LID Mask Control):** assign several LIDs per endpoint

Routing function:

\[
\begin{align*}
(Source, Destination) & \rightarrow (DLID, SL), \\
(Switch, DLID) & \rightarrow \text{Out Port}, \\
(Switch, SL, In Port, Out Port) & \rightarrow \text{Out VL}.
\end{align*}
\]
Adaptive Routing: Suggested Features

- **Minimal AR:**
  - Each LFT entry contains a set of ports

- **Non-minimal AR**
  - Need to track number of hops so far
    - VL is the only changing header field
  - SL2VL replaced with VL2VL
  - Multiple copies of LFT, one per input VL
  - 2 or more priority levels

- **Routing Function:**
  
  \[
  \begin{align*}
  \text{(Source, Destination)} &\rightarrow (DLID, VL), \\
  \text{(Switch, In VL, DLID)} &\rightarrow \{\text{Out Port}_1, \ldots, \text{Out Port}_k\}, \\
  \text{(Switch, In VL, In Port, Out Port)} &\rightarrow \text{Out VL}.
  \end{align*}
  \]
FlatFly Routing: Adaptive DOR

Figure 3: Routing in 2DFF: minimal DOR, adaptive DOR

• PA – port connected to adapter.
• Pn – port connected to switch along dimension n; 
• Pjn – port connected to switch j in dimension n.

Define the following Private LFTs:
0. For p2PA: select any q2Pn, where n = min{k | ik6=jk}.
1. For p2Pn, n = 1, N:
• if iin6=jn: select any q2Pjn;n;
• if iin=jn: select any q2Pm, where m = min{k | ik6=jk}.

The SL2VL mapping is
• on PA⇥Pn: VL = 1;
• otherwise VL = 0.

VLs are traversed in fixed order (010, 011, 200, 201, …, N0, N1), thus avoiding cycles.

Note the following.
1. n + 1 Private LFTs are used.
2. At most n additional hops are made, one in each dimension.
3. For C pairs of adapters connected to two arbitrary switches, the effective relative bandwidth is
   \[ \frac{L_n(S_n^1)}{C} \].
4. The number of pLFTs may be reduced if some of the dimensions are routed minimally without additional hop. However, the effective relative bandwidth then will be cut to \[ \frac{L_n}{C} \].
5.5 Twisted DOR × 1DFF
With DOR ⇥ 1DFF, if all adapters of a switch SA send packets to adapters of switch SB, these packets meet on intermediate switches after traversing each direction. The idea is to make an additional random hop in the last dimension before all other hops to provide greater distribution of paths across the fabric. The additional hop is made even if the last coordinate of source and destination is the same. This way, SN disjoint paths are provided, and if C ﬁnite SB the packets from SA to SB will only meet again on SB at the end of their paths.

Define the following Private LFTs:
0. For p2PA: select any q2Pn.
1. For p2Pn: select any q2Pn, where n = min{k | ik6=jk}.
2. For p2Pn, n = 1, N:
• if iin6=jn: select any q2Pjn;n;
• if iin=jn: select any q2Pm, where m = min{k | ik6=jk}.
FlatFly Routing: Mixed DOR

- 4 dimension orders based on destination, encoded in SL:
  - 1234, 2413, 3142, 4321
Compulsory initial extra hop in the last dimension
FlatFly Routing: Twisted Mixed DOR

- Combination of two previous routings

![Diagram of FlatFly Routing: Twisted Mixed DOR](image)
**Topology Agnostic Adaptive Routing**

- **Distance Based** $+k$
  - Set $VL = d(src, dest) + k - 1$
  - Route from `curr` do `dest` via `next`
    - with priority 1 if $d(next, dest) = d(curr, dest) - 1$
    - with priority 2 if $d(next, dest) = d(curr, dest) \leq VL$
  - Decrease $VL$ at each hop

- **Deflection** $+k$
  - Same as above, plus
  - Route from `curr` do `dest` via `next`
    - with priority 3 if $d(next, dest) = d(curr, dest) + 1 \leq VL$
Simulation Results

- Saturation levels in % of maximum bandwidth

<table>
<thead>
<tr>
<th>Routing</th>
<th>All2All</th>
<th>Bit</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Cmpl</td>
<td>Rvrs</td>
</tr>
<tr>
<td>Torus</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DOR 3D</td>
<td>26,5</td>
<td>15,6</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DOR 4D</td>
<td>48,2</td>
<td>25,0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hypercube</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DOR</td>
<td>24,2</td>
<td>24,2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mixed DOR</td>
<td>24,2</td>
<td>24,2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dragonfly</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Static</td>
<td>24,2</td>
<td>5,5</td>
</tr>
<tr>
<td>GSID</td>
<td>23,4</td>
<td>24,2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Distance</td>
<td>82</td>
<td>0,0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Distance +2</td>
<td>59,4</td>
<td>0,8</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Deflection +3</td>
<td>48,4</td>
<td>2,3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Simulation Results

<table>
<thead>
<tr>
<th>Routing</th>
<th>All2All</th>
<th>Bit</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Cmpl</td>
<td>Rvrs</td>
</tr>
<tr>
<td><strong>SlimFly × FlatFly 1</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Distance</td>
<td>80,5</td>
<td>11,7</td>
</tr>
<tr>
<td>Distance +2</td>
<td>80,5</td>
<td>19,5</td>
</tr>
<tr>
<td>Deflection +4</td>
<td>80,5</td>
<td>21,9</td>
</tr>
<tr>
<td><strong>SlimFly × FlatFly 2</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Distance</td>
<td>61,7</td>
<td>5,5</td>
</tr>
<tr>
<td>Distance +2</td>
<td>61,7</td>
<td>11,7</td>
</tr>
<tr>
<td>Deflection +4</td>
<td>74,2</td>
<td>11,7</td>
</tr>
</tbody>
</table>
## Simulation Results

<table>
<thead>
<tr>
<th>Routing</th>
<th>All2All</th>
<th>Bit</th>
<th>FlatFly</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Cmpl</td>
<td>Rvrs</td>
<td>Rotn</td>
</tr>
<tr>
<td>DOR</td>
<td>81,3</td>
<td>11,7</td>
<td>1,6</td>
</tr>
<tr>
<td>ADOR</td>
<td>52,3</td>
<td>50</td>
<td>3,1</td>
</tr>
<tr>
<td>Mixed ADOR</td>
<td>56,3</td>
<td>50,0</td>
<td>10,2</td>
</tr>
<tr>
<td>Twisted ADOR</td>
<td>24,2</td>
<td>50,0</td>
<td>21,1</td>
</tr>
<tr>
<td>Twisted Mixed</td>
<td>53,1</td>
<td>28,1</td>
<td>22,7</td>
</tr>
<tr>
<td>Distance</td>
<td>93,8</td>
<td>11,7</td>
<td>9,4</td>
</tr>
<tr>
<td>Distance +4</td>
<td>96,1</td>
<td>50,8</td>
<td>14,8</td>
</tr>
<tr>
<td>Deflection +4</td>
<td>94,5</td>
<td>50,0</td>
<td>24,2</td>
</tr>
</tbody>
</table>