SlimFly: A Cost Effective Low-Diameter Network Topology

Torsten Hoefler
With Maciej Besta
Edison’s vs. Pasteur’s quadrant

<table>
<thead>
<tr>
<th>Quest for fundamental understanding?</th>
<th>High</th>
<th>Low</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>BOHR QUADRANT</td>
<td>$100M FOX QUADRANT?</td>
</tr>
<tr>
<td>High</td>
<td>PASTEUR QUADRANT</td>
<td>EDISON QUADRANT</td>
</tr>
</tbody>
</table>

Use-inspired basic research

Pure basic research

Consideration of use?
DESIGNING AN EFFICIENT NETWORK TOPOLOGY

- Main intuition/idea: decrease network diameter
  - lower latency
  - smaller cost (fewer routers and cables for same bandwidth)
  - lower power consumption (packets traverse fewer SerDes)

Fat tree: Diameter == 4

Dragonfly: Diameter == 3

...still high 😞

...can we do better?
DESIGNING AN EFFICIENT NETWORK TOPOLOGY

- Goal: find a close-to-optimal topology that maximizes the number of endpoints ($N$) for a given diameter ($D$) and degree ($k$)

- Moore Bound: upper bound on the number of routers ($N_r$) in a graph with given $D$ and $k'$.

\[
N_r = 1 + k' \sum_{i=0}^{D-1} (k' - 1)^i
\]

$D = 2$: \hspace{1cm} $N_r \approx k'^2$

(~200,000 endpoints with 108-port switches)

$D = 3$: \hspace{1cm} $N_r \approx k'^3$

(>10,000,000 endpoints with 108-port switches)
DESIGNING AN EFFICIENT NETWORK TOPOLOGY

- Degree-Diameter problem

<table>
<thead>
<tr>
<th>Degree (k)</th>
<th>Diameter (D)</th>
<th>Graph with the maximum $N_r$ for a given $D$ and $k$</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>10 $100%$</td>
<td>20 $100%$</td>
</tr>
<tr>
<td>4</td>
<td>15 $100%$</td>
<td>52 $78.84%$</td>
</tr>
<tr>
<td>5</td>
<td>24 $100%$</td>
<td>104 $69.23%$</td>
</tr>
<tr>
<td>6</td>
<td>32 $100%$</td>
<td>186 $59.67%$</td>
</tr>
<tr>
<td>7</td>
<td>50 $100%$</td>
<td>301 $55.81%$</td>
</tr>
<tr>
<td>8</td>
<td>63 $90.47%$</td>
<td>456 $55.48%$</td>
</tr>
<tr>
<td>9</td>
<td>80 $92.50%$</td>
<td>657 $89.04%$</td>
</tr>
<tr>
<td>10</td>
<td>99 $91.91%$</td>
<td>910 $71.42%$</td>
</tr>
</tbody>
</table>
DESIGNING AN EFFICIENT NETWORK TOPOLOGY

- Degree-Diameter problem
  - Construct a graph with the maximum $N_r$ for a given $D$ and $k'$
- We use a result from McKay, Miller, Siran (MMS graphs) [1]; $D = 2$

A subgraph with identical groups of routers

Connections between subgraphs
(details skipped for clarity)

**ATTACHING ENDPOINTS**

- How many endpoints do we attach to each router?
- Maximize for \( p \) while maintaining full global bandwidth
  - Global bandwidth: the theoretical cumulative throughput if all processes simultaneously communicate with all other processes in a steady state
  - Result:
    \[
    p = \left\lfloor \frac{k'}{2} \right\rfloor
    \]
COMPARISON TO OPTIMALITY

- How close is SlimFly MMS to the Moore Bound (D=2)?
STRUCTURE ANALYSIS

COMPARISON TARGETS

Torus 3D

Torus 5D

Hypercube

Long Hop [1]

Fat tree

Dragonfly

Flattened Butterfly

Random networks

## Structure Analysis

**Diameter**

<table>
<thead>
<tr>
<th>Topology</th>
<th>Symbol</th>
<th>Example System</th>
<th>Diameter</th>
</tr>
</thead>
<tbody>
<tr>
<td>3-dimensional torus</td>
<td>T3D</td>
<td>Cray Gemini</td>
<td>$\left\lceil \frac{3}{2} \sqrt[3]{N_r} \right\rceil$</td>
</tr>
<tr>
<td>5-dimensional torus</td>
<td>T5D</td>
<td>IBM BlueGene/Q</td>
<td>$\left\lceil \frac{5}{2} \sqrt[5]{N_r} \right\rceil$</td>
</tr>
<tr>
<td>Hypercube</td>
<td>HC</td>
<td>NASA Pleiades</td>
<td>$\log_2 N_r$</td>
</tr>
<tr>
<td>3-level fat tree</td>
<td>FT-3</td>
<td>Tianhe-2</td>
<td>4</td>
</tr>
<tr>
<td>3-level Flat. Butterfly</td>
<td>FBF-3</td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>Dragonfly topologies</td>
<td>DF</td>
<td>Cray Cascade</td>
<td>3</td>
</tr>
<tr>
<td>Random topologies</td>
<td>DLN</td>
<td></td>
<td>3–10</td>
</tr>
<tr>
<td>Long Hop topologies</td>
<td>LH–HC</td>
<td>Infinetics Systems</td>
<td>4–6</td>
</tr>
<tr>
<td><strong>Slim Fly MMS</strong></td>
<td>SF</td>
<td></td>
<td>2</td>
</tr>
</tbody>
</table>
**Structure Analysis**

**Average Distance**

Random uniform traffic using minimum path routing

![Graph showing average number of hops vs. network size](image)

**Topology**
- Torus 3D
- Hypercube
- Torus 5D
- Long Hop
- Fat Tree
- Flat. Butterfly
- Random top.
- Dragonfly
- Slim Fly
STRUCTURE ANALYSIS

RESILIENCY

- Disconnection metrics*
- Other studied metrics ($N \approx 8192$):
  - Diameter (increase by 2) [1]; SF: 40%, DF: 25%, DLN: 60%
  - Average path length (increase by 2); SF: 55%, DF: 45%, DLN: 60%

*Missing values indicate the inadequacy of a balanced topology variant for a given $N$

---

<table>
<thead>
<tr>
<th>$N$</th>
<th>T3D</th>
<th>T5D</th>
<th>HC</th>
<th>LH–HC</th>
<th>FT–3</th>
<th>DF</th>
<th>FBF–3</th>
<th>DLN</th>
<th>SF</th>
</tr>
</thead>
<tbody>
<tr>
<td>512</td>
<td>30%</td>
<td>-</td>
<td>40%</td>
<td>55%</td>
<td>35%</td>
<td>-</td>
<td>55%</td>
<td>60%</td>
<td>60%</td>
</tr>
<tr>
<td>1024</td>
<td>25%</td>
<td>40%</td>
<td>40%</td>
<td>55%</td>
<td>40%</td>
<td>50%</td>
<td>60%</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>2048</td>
<td>20%</td>
<td>-</td>
<td>40%</td>
<td>55%</td>
<td>40%</td>
<td>55%</td>
<td>65%</td>
<td>65%</td>
<td>65%</td>
</tr>
<tr>
<td>4096</td>
<td>15%</td>
<td>-</td>
<td>45%</td>
<td>55%</td>
<td>55%</td>
<td>60%</td>
<td>70%</td>
<td>70%</td>
<td>70%</td>
</tr>
<tr>
<td>8192</td>
<td>10%</td>
<td>35%</td>
<td>45%</td>
<td>55%</td>
<td>60%</td>
<td>65%</td>
<td>-</td>
<td>75%</td>
<td>75%</td>
</tr>
</tbody>
</table>

PHYSICAL LAYOUT

Step 1: generate an MMS graph

Edges defined by Eq. (3)

Groups with edges defined by Eq. (2)

Pairs of subgroups of vertices become racks
**Comparison to Dragonfly**

**SlimFly:**
- Groups not necessarily fully connected
- ~50% fewer intra-group cables
- 2$q$ inter-group cable between two groups (see paper for details)
- ~25% fewer routers
- ~33% higher endpoint density

**Dragonfly:**
- Groups fully connected
- One inter-group cable between two groups
**Cost Comparison**

**Cost Model**

- Electric cables, avg length: 1m
- Top-of-rack routers
- 1 meter rack
- 1 meter rack
- 2m of overhead for each global link
- Optic cables, length: Manhattan distance
- Racks arranged as close to a square as possible

*Most cables skipped for clarity*

---

J. Kim, W. J. Dally, and D. Abts. Flattened Butterfly: A Cost-efficient Topology for High-radix Networks. ISCA '07
COST COMPARISON
CABLE COST MODEL

- Bandwidth cost as a function of distance
  - The functions obtained using linear regression*
- Used cables:
  - Mellanox IB QDR 56Gb/s QSFP
  - Mellanox Ethernet 40Gb/s QSFP
  - Mellanox Ethernet 10Gb/s SFP+
  - Elpeus Ethernet 10Gb/s SFP+

*Prices based on ColfaxDirect, June 2014
COST COMPARISON

ROUTER COST MODEL

- Router cost as a function of radix
  - The function obtained using linear regression*
- Used routers:
  - Mellanox Ethernet 10/40Gb

*Prices based on ColfaxDirect, June 2014
COST COMPARISON

RESULTS

~25% less expensive than DF due to fewer routers and cables
POWER COMPARISON

POWER MODEL

- Model similar to [1],
  - Each router port has four lanes,
  - Each lane has one SerDes,
  - Each SerDes consumes 0.7 W
  - Other parameters as in the cost model

# Cost & Power Comparison

## Detailed Case-Study

<table>
<thead>
<tr>
<th></th>
<th>Low-radix topologies</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Topology</td>
<td>T3D</td>
<td>T5D</td>
<td>HC</td>
<td>LH-HC</td>
<td>SF</td>
</tr>
<tr>
<td>Endpoints (N)</td>
<td>10,648</td>
<td>10,368</td>
<td>8,192</td>
<td>8,192</td>
<td>10,830</td>
</tr>
<tr>
<td>Routers (N_r)</td>
<td>10,648</td>
<td>10,368</td>
<td>8,192</td>
<td>8,192</td>
<td>722</td>
</tr>
<tr>
<td>Radix (k)</td>
<td>7</td>
<td>11</td>
<td>14</td>
<td>19</td>
<td>43</td>
</tr>
<tr>
<td>Electric cables</td>
<td>31,900</td>
<td>50,688</td>
<td>32,768</td>
<td>53,248</td>
<td>6,669</td>
</tr>
<tr>
<td>Fiber cables</td>
<td>0</td>
<td>0</td>
<td>12,288</td>
<td>12,288</td>
<td>6,869</td>
</tr>
<tr>
<td>Cost per node [$]</td>
<td>1,682</td>
<td>3,176</td>
<td>4,631</td>
<td>6,481</td>
<td>1,033</td>
</tr>
<tr>
<td>Power per node [W]</td>
<td>19.6</td>
<td>30.8</td>
<td>39.2</td>
<td>53.2</td>
<td>8.02</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>High-radix topologies</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Topology</td>
<td>FT-3</td>
<td>DLN</td>
<td>FBF-3</td>
<td>DF</td>
<td>FT-3</td>
<td>DLN</td>
<td>FBF-3</td>
<td>DF</td>
<td>DF</td>
<td>SF</td>
</tr>
<tr>
<td>Endpoints (N)</td>
<td>19,876</td>
<td>40,200</td>
<td>20,736</td>
<td>58,806</td>
<td>10,718</td>
<td>9,702</td>
<td>10,000</td>
<td>9,702</td>
<td>10,890</td>
<td>10,830</td>
</tr>
<tr>
<td>Routers (N_r)</td>
<td>2,311</td>
<td>4,020</td>
<td>1,728</td>
<td>5,346</td>
<td>1,531</td>
<td>1,386</td>
<td>1,000</td>
<td>1,386</td>
<td>990</td>
<td>722</td>
</tr>
<tr>
<td>Radix (k)</td>
<td>43</td>
<td>43</td>
<td>43</td>
<td>43</td>
<td>35</td>
<td>28</td>
<td>33</td>
<td>27</td>
<td>43</td>
<td>43</td>
</tr>
<tr>
<td>Electric cables</td>
<td>19,414</td>
<td>32,488</td>
<td>9,504</td>
<td>56,133</td>
<td>7,350</td>
<td>6,837</td>
<td>4,500</td>
<td>9,009</td>
<td>6,885</td>
<td>6,669</td>
</tr>
<tr>
<td>Fiber cables</td>
<td>40,215</td>
<td>33,842</td>
<td>20,736</td>
<td>29,524</td>
<td>24,806</td>
<td>7,716</td>
<td>10,000</td>
<td>4,900</td>
<td>1,012</td>
<td>6,869</td>
</tr>
<tr>
<td>Cost per node [$]</td>
<td>2,346</td>
<td>1,743</td>
<td>1,570</td>
<td>1,438</td>
<td>2,315</td>
<td>1,566</td>
<td>1,535</td>
<td>1,342</td>
<td>1,365</td>
<td>1,033</td>
</tr>
<tr>
<td>Power per node [W]</td>
<td>14.0</td>
<td>12.04</td>
<td>10.8</td>
<td>10.9</td>
<td>14.0</td>
<td>11.2</td>
<td>10.8</td>
<td>10.8</td>
<td>10.9</td>
<td>8.02</td>
</tr>
</tbody>
</table>
**DEADLOCK FREEDOM**

**MINIMUM STATIC ROUTING**

- Assign two virtual channels (VC0 and VC1) to each link
- For a 1-hop path use VC0
- For a 2-hop path use VC0 (hop 1) and VC1 (hop 2)
- One can also use the DFSSSP scheme [1]

**PERFORMANCE**

- Cycle-based flit-level simulations (Booksim)
- Routing protocols:
  - Minimum static routing
  - Valiant’s random routing
  - Universal Globally-Adaptive Load-Balancing routing
    - **UGAL-L**: each router has access to its local output queues
    - **UGAL-G**: each router has access to the sizes of all router queues in the network
PERFORMANCE

- Random uniform traffic
- Bit permutation (reverse) traffic
CONCLUSIONS

Topology design
Optimizing towards the Moore Bound reduces expensive network resources

Advantages of SlimFly
- Diameter
- Avg. distance
- Bisection bandwidth
- Resilience
- Performance
- Cost & power

Optimization approach
Combining mathematical optimization and current technology trends effectively tackles challenges in networking

PhD fellowship for parallel computing