Network topologies for large-scale compute centers: It's the diameter, stupid!

Torsten Hoefler

with support of Maciej Besta @ SPCL
presented at Hot Interconnects 2016, San Jose, CA, USA
A BRIEF HISTORY OF NETWORK TOPOLOGIES

copper cables, small radix switches

Mesh

Torus

Butterfly

Clos/Benes

~2005

Kautz

Hypercube

Fat Trees

1980’s

2000’s

2007

2008

2008

2014

2007

Flat Fly

Random

Dragonfly

Slim Fly

fiber, high-radix switches

Fat Trees

Trees

Tobacco

????
A BRIEF HISTORY OF NETWORK TOPOLOGIES

copper cables, small radix switches

fiber, high-radix switches

Bandwidth \( = 2^{\frac{d}{2} \sqrt{N^{d-1}}} \)
Latency \( = \frac{d}{2} \sqrt{N} \)
Radix \( = 2d \)
A BRIEF HISTORY OF NETWORK TOPOLOGIES

- **Bandwidth**: $\frac{N}{2}$
- **Latency**: $\log_2 N$
- **Radix**: $\log_2 N$

- **Mesh**
- **Torus**
- **Butterfly**
- **Clos/Benes**
- **Kautz**
- **Dragonfly**
- **Slim Fly**
- **Hypercube**
- **Trees**
- **Fat Trees**
- **Flat Fly**
- **Random**

1980’s

- copper cables, small radix switches

~2005

- fiber, high-radix switches

2007

- 2008

2014

- 2008
A BRIEF HISTORY OF NETWORK TOPOLOGIES

copper cables, small radix switches
fiber, high-radix switches

Mesh

1980’s

Butterfly

Clos/Benes

2000’s

Kautz

~2005

Hypercube

Torus

Bandwidth \(= 1\)
Latency \(= 2 \log_2 N\)
Radix \(= 2\)

2007

Flat Fly

2008

Fat Trees

2014

Random

???
A BRIEF HISTORY OF NETWORK TOPOLOGIES

Bandwidth \(= \frac{N}{2}\)
Latency \(= 2 \log_2 N\)
Radix \(= 4\)
A BRIEF HISTORY OF NETWORK TOPOLOGIES

- Mesh
- Torus
- Butterfly
- Clos/Benes
- Kautz
- Hypercube
- Trees
- Fat Trees
- Dragonfly
- Slim Fly
- 1980's
- 2000's
- ~2005
- Bandwidth: \( \frac{N}{4} \)
- Latency: \( \log_k N \)
- Radix: \( k \)

- Copper cables, small radix switches
- Fiber, high-radix switches
- Copper cables, small radix switches
A BRIEF HISTORY OF NETWORK TOPOLOGIES

**Bandwidth** \[ \approx \frac{N}{4} \]

**Latency** \[ = 3 - 5 \]

**Radix** \[ = 48 - 64 \]

2010 18th IEEE Symposium on High Performance Interconnects

**The PERCS High-Performance Interconnect**

Baba Aminilli *, Ravi Aminilli *, Vicente Chung *, Scott Clark *, Wolfgang Denzel 1, Ben Dreup *, Torsten Hoeffer 1, Jody Joyner *, Jerry Lewis *, Jian Li 1, Nan Ni 2 and Ram Rajamony 1

* IBM Systems and Technology Group, 11700 Burnet Road, Austin, TX 78758
1 IBM Research (Austin, Zurich), 11700 Burnet Road, Austin, TX 78758
2 Blue Waters Directorate, NCSA, University of Illinois at Urbana-Champaign, Urbana, IL 61801

E-mail: aminilli@us.ibm.com, rajamony@us.ibm.com, hber@illinois.edu

Abstract—The PERCS system was designed by IBM in response to a DARPA challenge that called for a high-productivity, high-performance computing system. A major innovation in the PERCS design is the network that is built using Hub chips that are integrated into the compute nodes. Each Hub chip is about 580 mm² in size, has over 3700 signal I/Os, and is packaged in a module that also contains I/O-attached optical electronic devices.

The Hub module implements five types of high-bandwidth interconnects with multiple links that are fully connected with a high-performance internal crossbar switch. These links provide over 9 Tbit/second of raw bandwidth and are used to construct a two-level direct-connect topology, amounting to a total of 300 bandwidths do not scale accordingly. For instance, while High Performance Linpack performance [5, 10] shows a steady improvement over time, interconnect-intensive metrics such as G-RandomAccess and G-FFTE [5] show very little improvement.

The challenge of building a high-performance, highly productive, multi-Petaflop system forced us to recognize early on that the entire infrastructure had to scale along with the microprocessor’s capabilities. A significant component of our scaling solution is a new switchless interconnect with very high port count integrated into a two-level direct connect...
A BRIEF HISTORY OF NETWORK TOPOLOGIES

Key ideas:

“It’s the diameter, stupid”

Lower diameter:
→ Less cables traversed
→ Less cables needed
→ Less routers needed

Cost and energy savings:
→ Up to 50% over Fat Tree
→ Up to 33% over Dragonfly

Bandwidth \( \approx \frac{N}{4} \)
Latency \( = 2 - 4 \)
Radix \( = k \)
DESIGNING AN EFFICIENT NETWORK TOPOLOGY

EXAMPLE: FULL-BANDWIDTH FAT TREE VS HOFFMAN-SINGLETON GRAPH

3-level fat tree:

Slim Fly based on the Hoffman-Singleton Graph [1]:

diameter = 2
> ~50% fewer routers
> ~30% fewer cables

diameter = 4

DESIGNING AN EFFICIENT NETWORK TOPOLOGY

Key method

Optimize towards the Moore Bound [1]: the upper bound on the number of vertices in a graph with given diameter $D$ and radix $k$.

$$MB(D, k) = 1 + k + k(k - 1) + k(k - 1)^2 + \cdots$$

$$MB(D, k) = 1 + k \sum_{i=0}^{D-1} (k - 1)^i$$

DESIGNING AN EFFICIENT NETWORK TOPOLOGY

CONNECTING ROUTERS: DIAMETER 2

Example Slim Fly design for diameter = 2: MMS graphs [1]

DESIGNING AN EFFICIENT NETWORK TOPOLOGY

CONNECTING ROUTERS: DIAMETER 2

Groups form a fully-connected bipartite graph
Designing an Efficient Network Topology

Connecting Routers: Diameter 2

1. Select a prime power $q$
   
   $q = 4w + \delta$;
   
   $w \in \mathbb{N}$  $\delta \in \{-1, 0, 1\}$,

   A Slim Fly based on $q$:

   Number of routers: $2q^2$

   Network radix: $(3q - \delta)/2$

2. Construct a finite field $\mathcal{F}_q$.

   Assuming $q$ is prime:

   $\mathcal{F}_q = \mathbb{Z}/q\mathbb{Z} = \{0, 1, \ldots, q - 1\}$

   with modular arithmetic.

3. Example: $q = 5$

   50 routers

   network radix: 7

   $\mathcal{F}_5 = \{0, 1, 2, 3, 4\}$
DESIGNING AN EFFICIENT NETWORK TOPOLOGY

CONNECTING ROUTERS: DIAMETER 2

3 Label the routers

Set of routers:
\( \{0,1\} \times \mathcal{F}_q \times \mathcal{F}_q \)

Example: \( q = 5 \)

- (0,0,0)
- (0,0,1)
- (0,0,2)
- (0,0,3)
- (0,0,4)

- (1,0,0)
- (1,0,1)
- (1,0,2)
- (1,0,3)
- (1,0,4)

- (0,1,0)
- (0,1,1)
- (0,1,2)
- (0,1,3)
- (0,1,4)

- (0,2,0)
- (0,2,1)
- (0,2,2)
- (0,2,3)
- (0,2,4)

- (0,3,0)
- (0,3,1)
- (0,3,2)
- (0,3,3)
- (0,3,4)

- (0,4,0)
- (0,4,1)
- (0,4,2)
- (0,4,3)
- (0,4,4)

- (1,1,0)
- (1,1,1)
- (1,1,2)
- (1,1,3)
- (1,1,4)

- (1,2,0)
- (1,2,1)
- (1,2,2)
- (1,2,3)
- (1,2,4)

- (1,3,0)
- (1,3,1)
- (1,3,2)
- (1,3,3)
- (1,3,4)

- (1,4,0)
- (1,4,1)
- (1,4,2)
- (1,4,3)
- (1,4,4)
**DESIGNING AN EFFICIENT NETWORK TOPOLOGY**

**CONNECTING ROUTERS: DIAMETER 2**

1. **Find primitive element** $\xi$

   $\xi \in \mathbb{F}_q$ generates $\mathbb{F}_q$:
   
   All non-zero elements of $\mathbb{F}_q$ can be written as $\xi^i; \ i \in \mathbb{N}$

2. **Build Generator Sets**

   $X = \{1, \xi^2, \ldots, \xi^{q-3}\}$
   
   $X' = \{\xi, \xi^3, \ldots, \xi^{q-2}\}$

**Example:** $q = 5$

- $\mathbb{F}_5 = \{0,1,2,3,4\}$
- $\xi = 2$
- $1 = \xi^4 \mod 5 = 16 \mod 5$
- $X = \{1, 4\}$
- $X' = \{2, 3\}$
**DESIGNING AN EFFICIENT NETWORK TOPOLOGY**

**CONNECTING ROUTERS: DIAMETER 2**

6 *Intra-group connections*

Two routers in one group are connected iff their “vertical Manhattan distance” is an element from:

\[ X = \{1, \xi^2, \ldots, \xi^{q-3}\} \quad \text{(for subgraph 0)} \]

\[ X' = \{\xi, \xi^3, \ldots, \xi^{q-2}\} \quad \text{(for subgraph 1)} \]

**Example: \(q = 5\)**

Take Routers \((0,0,.)\)

\[ X = \{1, 4\} \]
DESIGNING AN EFFICIENT NETWORK TOPOLOGY

CONNECTING ROUTERS: DIAMETER 2

**6 Intra-group connections**

Two routers in one group are connected iff their “vertical Manhattan distance” is an element from:

\[ X = \{1, \xi^2, ..., \xi^{q-3}\} \] (for subgraph 0)

\[ X' = \{\xi, \xi^3, ..., \xi^{q-2}\} \] (for subgraph 1)

**Example: q = 5**

Take Routers (1, 4, .)

\[ X' = \{2, 3\} \]
DESIGNING AN EFFICIENT NETWORK TOPOLOGY

CONNECTING Routers: DIAMETER 2

7 Inter-group connections

Router \((0, x, y) \leftrightarrow (1, m, c)\)

iff \(y = mx + c\)

Example: \(q = 5\)

- Take Router \((1,0,0)\)
  - \((1,0,0) \leftrightarrow (0, x, 0)\)

- Take Router \((1,1,0)\)
  - \((1,0,0) \leftrightarrow (0, x, x)\)

\(m = 0, c = 0\)

\(m = 1, c = 0\)
DESIGNING AN EFFICIENT NETWORK TOPOLOGY

ATTACHING ENDPOINTS: DIAMETER 2

- How many endpoints do we attach to each router?
- As many to ensure full global bandwidth:
  - Global bandwidth: the theoretical cumulative throughput in all-to-all in a steady state
DESIGNING AN EFFICIENT NETWORK TOPOLOGY

ATTACHING ENDPOINTS: DIAMETER 2

1. Get load $l$ per router-router channel (average number of routes per channel)

$$l = \frac{\text{total number of routes}}{\text{total number of channels}}$$

2. Make the network balanced, i.e.,:
   each endpoint can inject at full capacity

local uplink load = number of endpoints = $l$

- 67% of ports lead to other routers
- 33% of ports lead to endpoints
Cost Comparison

Cost Models: Variants

**Variant 1:** Routers and servers together

**Variant 2:** Routers and servers separately
**COST COMPARISON**

**CABLE COST MODEL**

- Cable cost as a function of distance
  - The functions obtained using linear regression*
  - Optical transceivers considered
  - Cables used:
    - Mellanox IB FDR10 40Gb/s QSFP

- Other 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:

- Mellanox IB QDR 56Gb/s QSFP
- Mellanox Ethernet 40Gb/s QSFP
- Mellanox Ethernet 10Gb/s SFP+
- Elpeus Ethernet 10Gb/s SFP+
COST COMPARISON

ROUTER COST MODEL

- Router cost as a function of radix
  - The function obtained using linear regression*
  - Routers used:
    - Mellanox IB FDR10
    - Mellanox Ethernet 10/40 Gb

*Prices based on:
COST COMPARISON

RESULTS

Variant 1:
SF less expensive than DF by ~13% (Mellanox IB routers) up to ~39% (Mellanox Ethernet routers)

Variant 2:
Cost & Power Comparison

Detailed Case-Study

- A Slim Fly with:
  - \( N = 10,830 \)
  - \( k = 43 \)
  - \( N_r = 722 \)
## Cost & Power Comparison
### Detailed Case-Study: High-Radix Topologies

<table>
<thead>
<tr>
<th>Topology</th>
<th>Fat tree</th>
<th>Random</th>
<th>Flat. Butterfly</th>
<th>Dragonfly</th>
<th>Slim Fly</th>
</tr>
</thead>
<tbody>
<tr>
<td>Endpoints ($N$)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19,876</td>
<td>40,200</td>
<td>20,736</td>
<td>58,806</td>
<td></td>
<td>10,830</td>
</tr>
<tr>
<td>Routers ($N_r$)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2,311</td>
<td>4,020</td>
<td>1,728</td>
<td>5,346</td>
<td></td>
<td>722</td>
</tr>
<tr>
<td>Radix ($k$)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>43</td>
<td>43</td>
<td>43</td>
<td>43</td>
<td>43</td>
<td>43</td>
</tr>
<tr>
<td>Electric cables</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19,414</td>
<td>32,488</td>
<td>9,504</td>
<td>56,133</td>
<td></td>
<td>6,669</td>
</tr>
<tr>
<td>Fiber cables</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>40,215</td>
<td>33,842</td>
<td>20,736</td>
<td>29,524</td>
<td></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>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>8.02</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Topology</th>
<th>Fat tree</th>
<th>Random</th>
<th>Flat. Butterfly</th>
<th>Dragonfly</th>
<th>Slim Fly</th>
</tr>
</thead>
<tbody>
<tr>
<td>Endpoints ($N$)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10,718</td>
<td>9,702</td>
<td>10,000</td>
<td>9,702</td>
<td></td>
<td>10,830</td>
</tr>
<tr>
<td>Routers ($N_r$)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1,531</td>
<td>1,386</td>
<td>1,000</td>
<td>1,386</td>
<td></td>
<td>722</td>
</tr>
<tr>
<td>Radix ($k$)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>35</td>
<td>28</td>
<td>33</td>
<td>27</td>
<td>43</td>
<td></td>
</tr>
<tr>
<td>Electric cables</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7,350</td>
<td>6,837</td>
<td>4,500</td>
<td>9,009</td>
<td></td>
<td>6,669</td>
</tr>
<tr>
<td>Fiber cables</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24,806</td>
<td>7,716</td>
<td>10,000</td>
<td>4,900</td>
<td></td>
<td>6,869</td>
</tr>
<tr>
<td>Cost per node [$]</td>
<td>2,315</td>
<td>1,566</td>
<td>1,535</td>
<td>1,342</td>
<td>1,033</td>
</tr>
<tr>
<td>Power per node [W]</td>
<td>14.0</td>
<td>11.2</td>
<td>10.8</td>
<td>10.8</td>
<td>8.02</td>
</tr>
</tbody>
</table>
## STRUCTURE ANALYSIS

### RESILIENCY

- Disconnection metrics
- Other studied metrics:
  - Average path length (increase by 2);
  - SF is 10% more resilient than DF

<table>
<thead>
<tr>
<th>Number of endpoints</th>
<th>Torus3D</th>
<th>Torus5D</th>
<th>Hypercube</th>
<th>Long Hop</th>
<th>Fat tree</th>
<th>Dragonfly</th>
<th>Flat. Butterfly</th>
<th>Random</th>
<th>Slim Fly</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>65%</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>

"-" means that a given topology does not have a variant of a given size
**PERFORMANCE & ROUTING**

**MINIMUM ROUTING**

1. *Intra-group connections*
   - ∃ Path of length 1 or 2 between two routers

2. *Inter-group connections (different types of groups)*
   - ∃ Path of length 1 or 2 between two routers

3. *Inter-group connections (identical types of groups)*
   - ∃ Path of length 2 between two routers
**Performance & Routing**

**Random Uniform Traffic**

![Graph 500 Network Diagram](image)

**Graph 500**

**Routing Protocol**

- Slim Fly (Valiant)
- Slim Fly (Minimum)
- Slim Fly (UGAL–L)
- Slim Fly (UGAL–G)
- Dragonfly (UGAL–L)
- Fat Tree (ANCA)

**Latency vs. Offered Load**

- latency [cycles]
- Offered load
OTHER RESULTS

Table 5: (§ 5.3)

<table>
<thead>
<tr>
<th>Strategy</th>
<th>N ≈ 200</th>
<th>N ≈ 1,296</th>
</tr>
</thead>
<tbody>
<tr>
<td>CM</td>
<td>FBF</td>
<td>128</td>
</tr>
<tr>
<td>SF-VAL</td>
<td>128</td>
<td>128</td>
</tr>
<tr>
<td>SF-INL</td>
<td>128</td>
<td>128</td>
</tr>
<tr>
<td>SF-N</td>
<td>128</td>
<td>128</td>
</tr>
<tr>
<td>SF-S</td>
<td>128</td>
<td>128</td>
</tr>
<tr>
<td>SF-B</td>
<td>128</td>
<td>128</td>
</tr>
<tr>
<td>SF-L</td>
<td>128</td>
<td>128</td>
</tr>
<tr>
<td>Only-EB</td>
<td>128</td>
<td>128</td>
</tr>
</tbody>
</table>

Area due to:
- i-routers
- a-routers
- RRg-wires
- RRa-wires

Power due to:
- routers
- RR-wires

Buffer size (flits):
- 512
- 256
- 128

Dynamic power/node [W]
- wires
- crossbars
- buffers

Latency [cycles]
- Load due to rstart
- Throughput [flits/J]
- Power consumption [W]

Network size [endpoints]

Topology:
- Dragonfly
- Slim Fly
A LOWEST-DIAMETER TOPOLOGY
- Approaching the Moore Bound
- Resilient

A COST & POWER EFFECTIVE TOPOLOGY
- 25% less expensive than Dragonfly,
- 26% less power-hungry than Dragonfly

A HIGH-PERFORMANCE TOPOLOGY
- Lowest latency
- Full global bandwidth

Thank you for your attention
http://spcl.inf.ethz.ch/SlimFly