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

TORSTEN HOEFLER

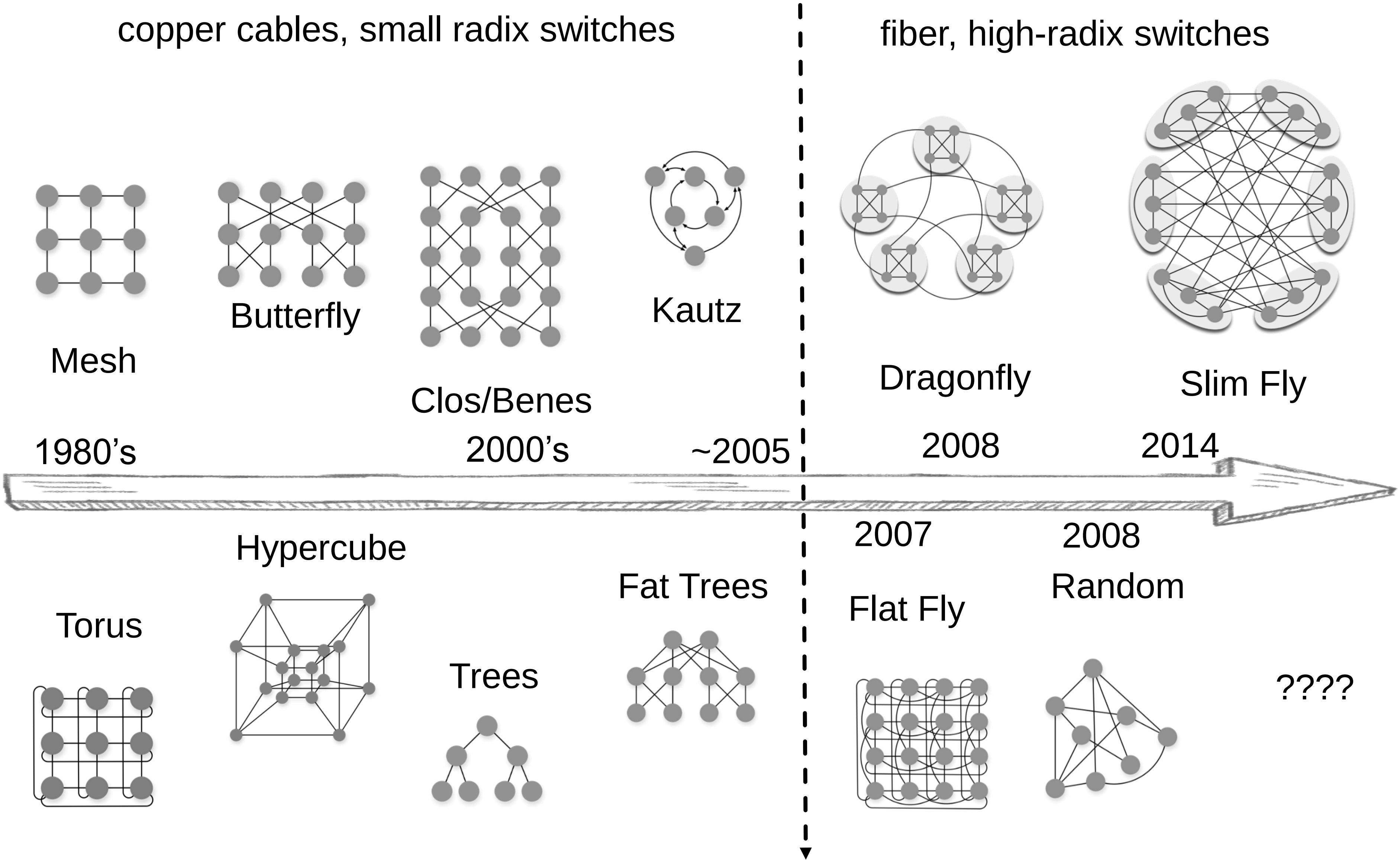
with support of Maciej Besta @ SPCL
presented at Hot Interconnects 2016, San Jose, CA, USA

50% [1]

33% [2]

[1] D. Abts et al. (2010), *Energy Proportional Datacenter Networks*, ISCA'10
[2] J. Kim et al. (2007), *Flattened Butterfly: A Cost-Efficient Topology for High-Radix Networks*, ISCA'07

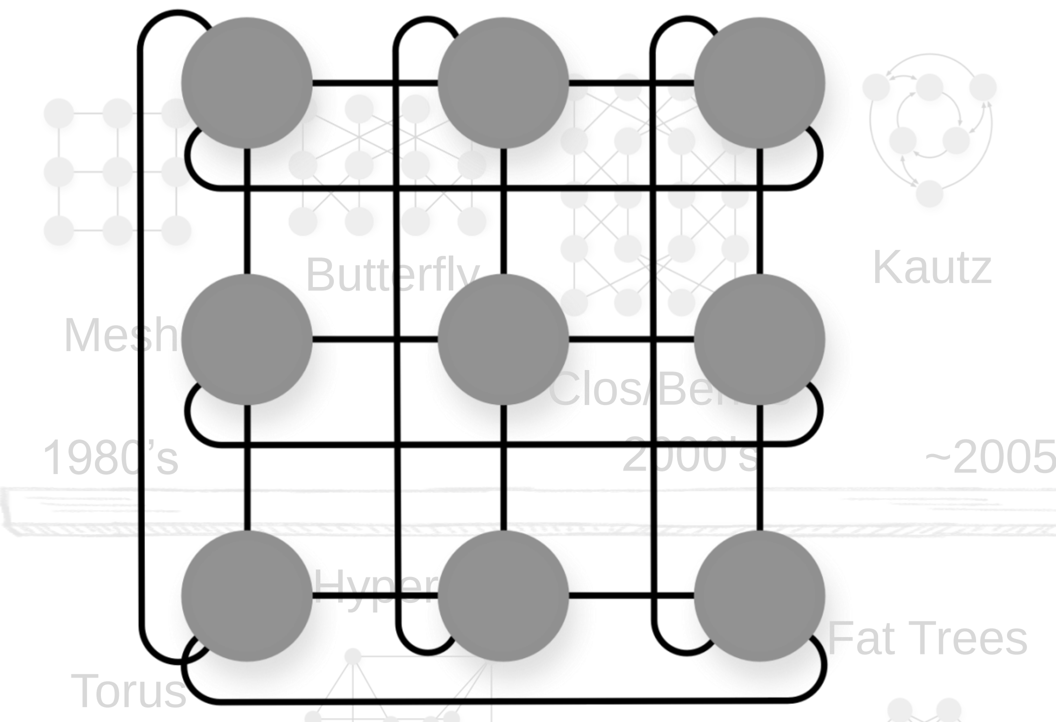
A BRIEF HISTORY OF NETWORK TOPOLOGIES



A BRIEF HISTORY OF NETWORK TOPOLOGIES

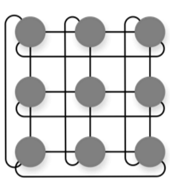
copper cables, small radix switches

fiber, high-radix switches



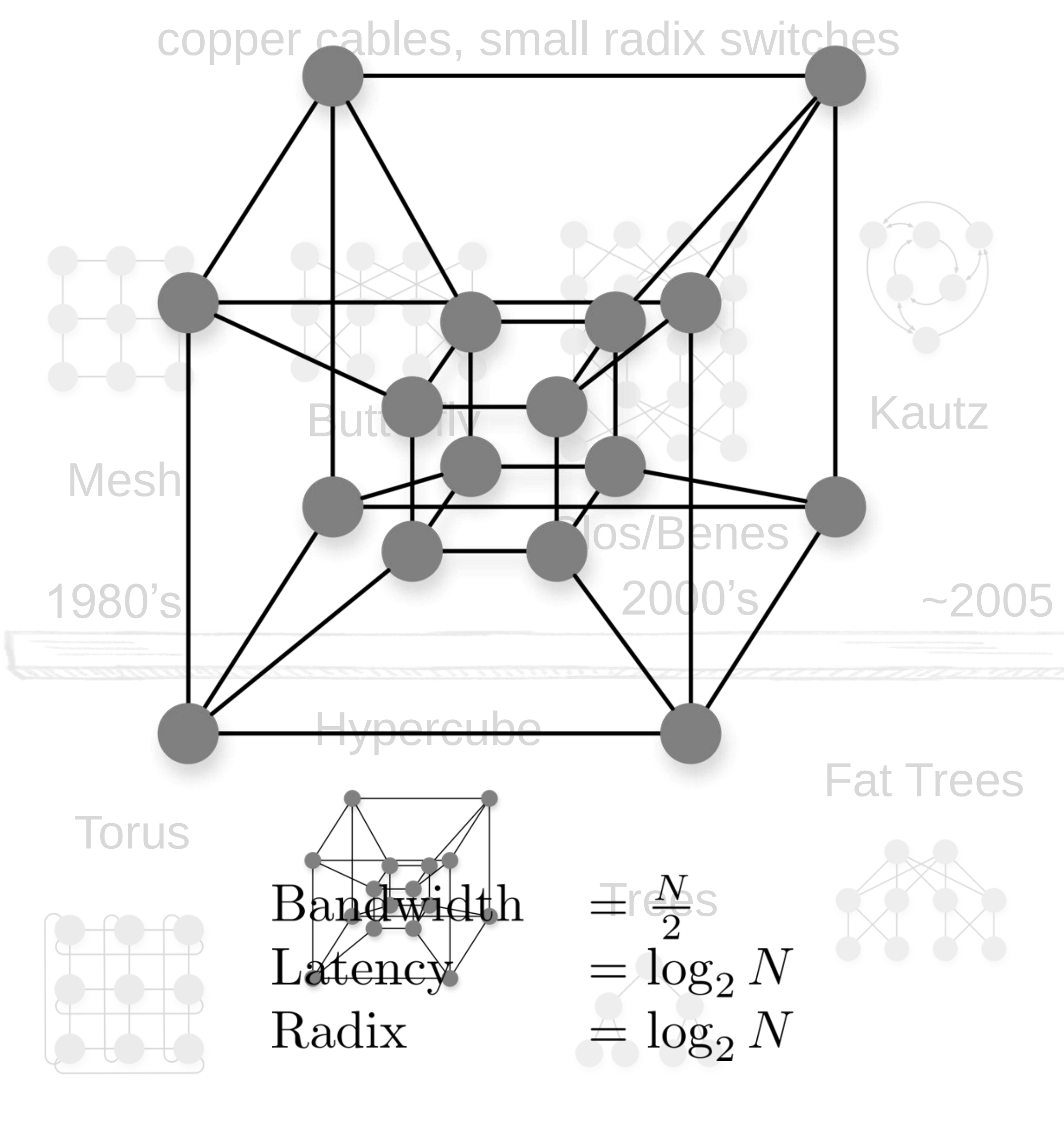
2008

2014



$$\begin{aligned} \text{Bandwidth} &= 2\sqrt[d]{N^{d-1}} \\ \text{Latency} &= \tfrac{d}{2}\sqrt[d]{N} \\ \text{Radix} &= 2d \end{aligned}$$

A BRIEF HISTORY OF NETWORK TOPOLOGIES

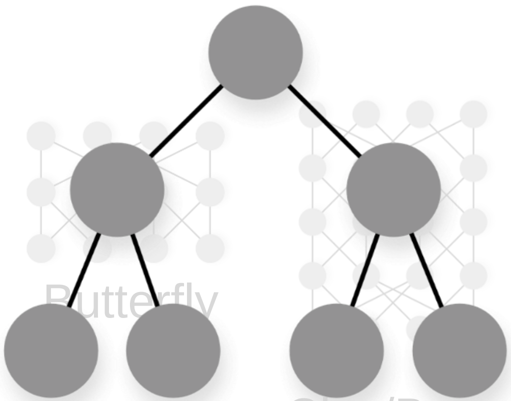


$$\text{Bandwidth} = \frac{N}{2}$$
$$\text{Latency} = \log_2 N$$
$$\text{Radix} = \log_2 N$$

A BRIEF HISTORY OF NETWORK TOPOLOGIES

copper cables, small radix switches

fiber, high-radix switches



Mesh

Butterfly

Clos/Benes

Kautz

1980's

2000's

~2005

2008

2014

2007

2008

Hypercube

Fat Trees

Random

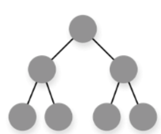
Flat Fly

Torus

Trees

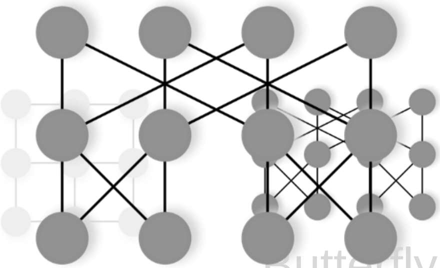
????

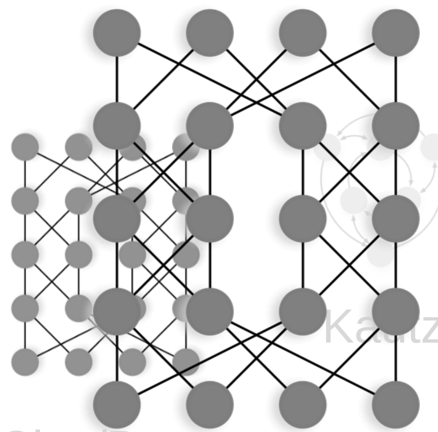
$$\begin{aligned} \text{Bandwidth} &= 1 \\ \text{Latency} &= 2\log_2 N \\ \text{Radix} &= 2 \end{aligned}$$



A BRIEF HISTORY OF NETWORK TOPOLOGIES

copper cables, small radix switches





Butterfly

Kautz

Mesh

Clos/Benes

$$\begin{aligned} \text{Bandwidth} &= \frac{N}{2} \\ \text{Latency} &= 2\log_2 N \\ \text{Radix} &= 4 \end{aligned}$$

1980's

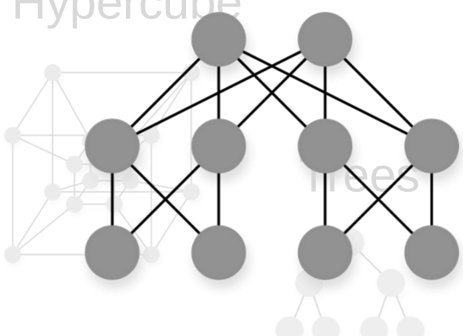
2000's

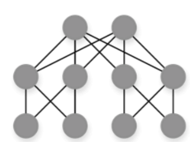
~2005

Hypercube

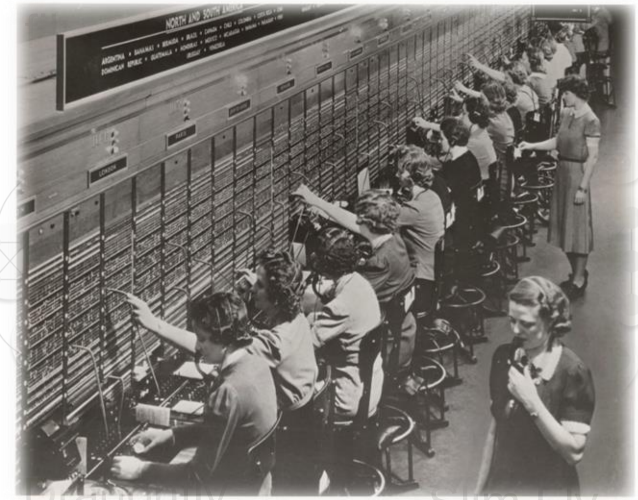
Fat Trees

Torus





Trees



Dragonfly

Slim Fly

2008

2014

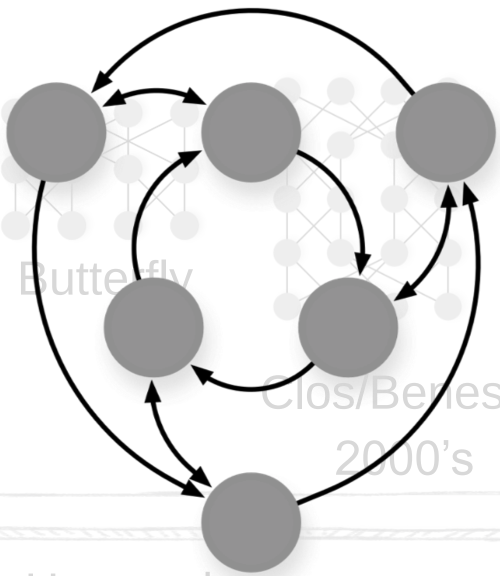
A BRIEF HISTORY OF NETWORK TOPOLOGIES

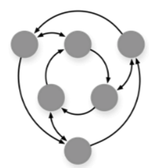
copper cables, small radix switches

fiber, high-radix switches

Mesh

Butterfly





Kautz

Clos/Benes

Dragonfly

Slim Fly

1980's

2000's

~2005

20

Hypercube

Fat Trees

Fla

Torus

Trees

????

$$
\begin{aligned}
\text{Bandwidth} &= \rightarrow \frac{N}{4} \\
\text{Latency} &= \log_k N \\
\text{Radix} &= k
\end{aligned}
$$

A BRIEF HISTORY OF NETWORK TOPOLOGIES

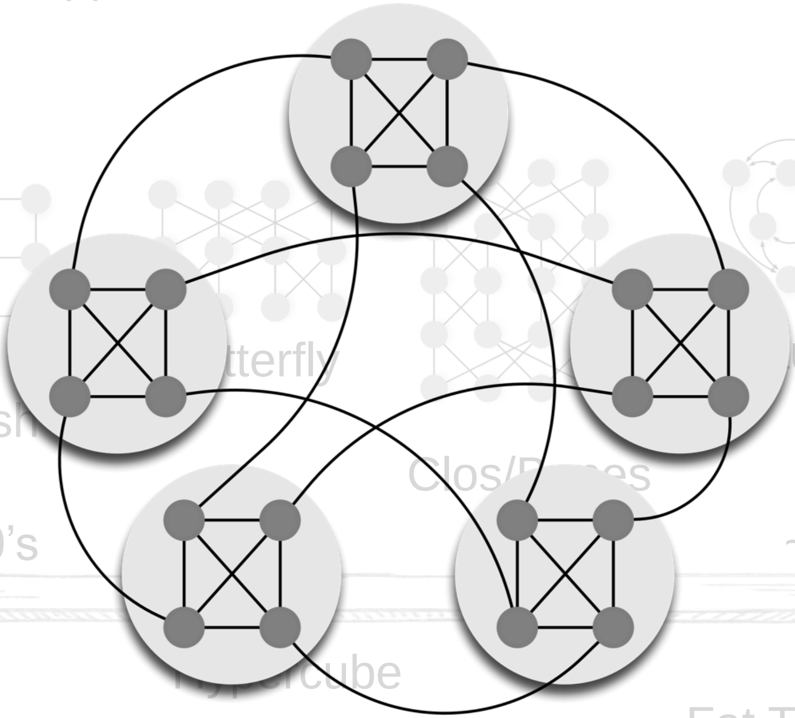
copper cables, small radix switches

fiber, high-radix switches

Mesh

1980's

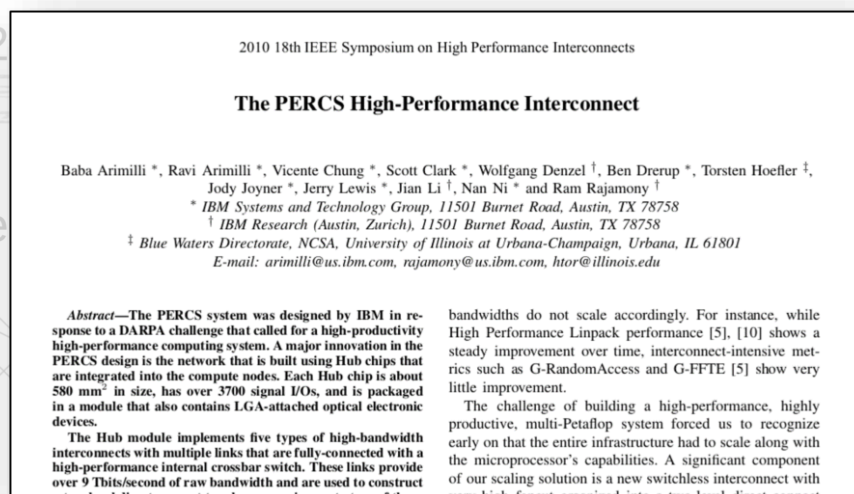
Torus



Dragonfly

Slim Fly

$$
\begin{aligned}
\text{Bandwidth} &\approx \frac{N}{4} \\
\text{Latency} &= 3-5 \\
\text{Radix} &= 48-64
\end{aligned}
$$



2010 18th IEEE Symposium on High Performance Interconnects

The PERCS High-Performance Interconnect

Baba Arimilli *, Ravi Arimilli *, Vicente Chung *, Scott Clark *, Wolfgang Denzel †, Ben Drerup *, Torsten Hoefler ‡, Jody Joyner *, Jerry Lewis *, Jian Li †, Nan Ni * and Ram Rajamony †
* IBM Systems and Technology Group, 11501 Burnet Road, Austin, TX 78758
† IBM Research (Austin, Zurich), 11501 Burnet Road, Austin, TX 78758
‡ Blue Waters Directorate, NCSA, University of Illinois at Urbana-Champaign, Urbana, IL 61801
E-mail: arimilli@us.ibm.com, rajamony@us.ibm.com, htor@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 mm^2 in size, has over 3700 signal I/Os, and is packaged in a module that also contains LGA-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 Tbits/second of raw bandwidth and are used to construct a two-level direct-connect topology spanning up to tens of thou-

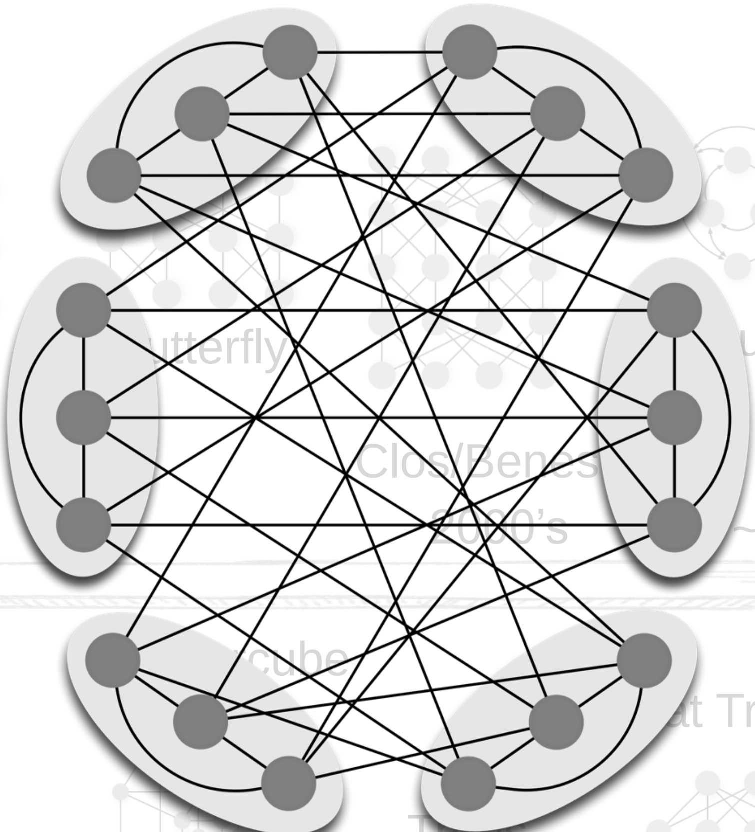
bandwidths do not scale accordingly. For instance, while High Performance Linpack performance [5], [10] shows a steady improvement over time, interconnect-intensive metrics such as G-RandomAccess and G-FFTE [5] show very little improvement.

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 fanout organized into a two-level direct connect

A Brief History of Network Topologies

copper cables, small radix switches

fiber, high-radix switches



$$
\begin{aligned}
\text{Bandwidth} &\approx \frac{N}{4} \\
\text{Latency} &= 2-4 \\
\text{Radix} &= k
\end{aligned}
$$

Key ideas:

“It’s the diameter, stupid”

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

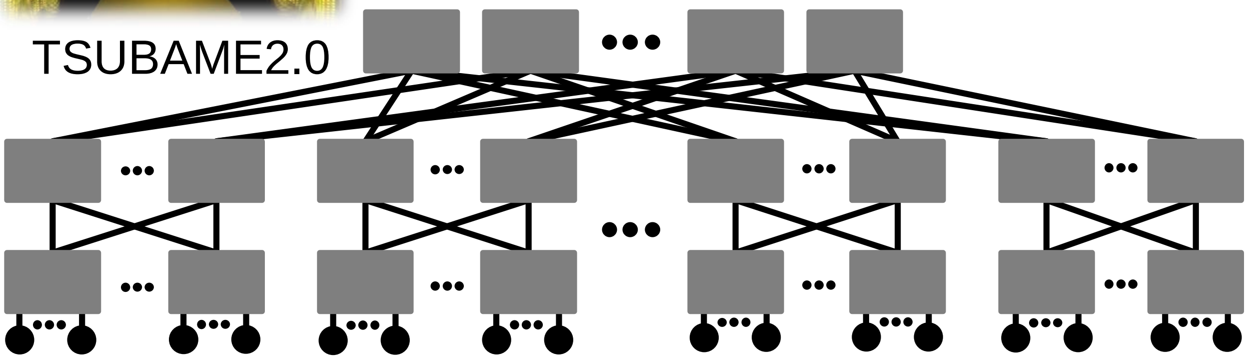
Cost and energy savings:
→ Up to 50% over Fat Tree
→ Up to 33% over Dragonfly

Designing an Efficient Network Topology

Example: full-bandwidth fat tree vs Hoffman-Singleton graph

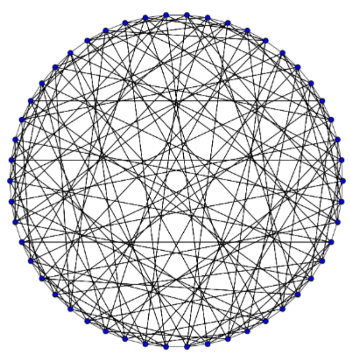
TSUBAME2.0

3-level fat tree:



diameter = 4

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



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

[1] Hoffman, Alan J.; Singleton, Robert R. (1960), *Moore graphs with diameter 2 and 3*, IBM Journal of Research and Development

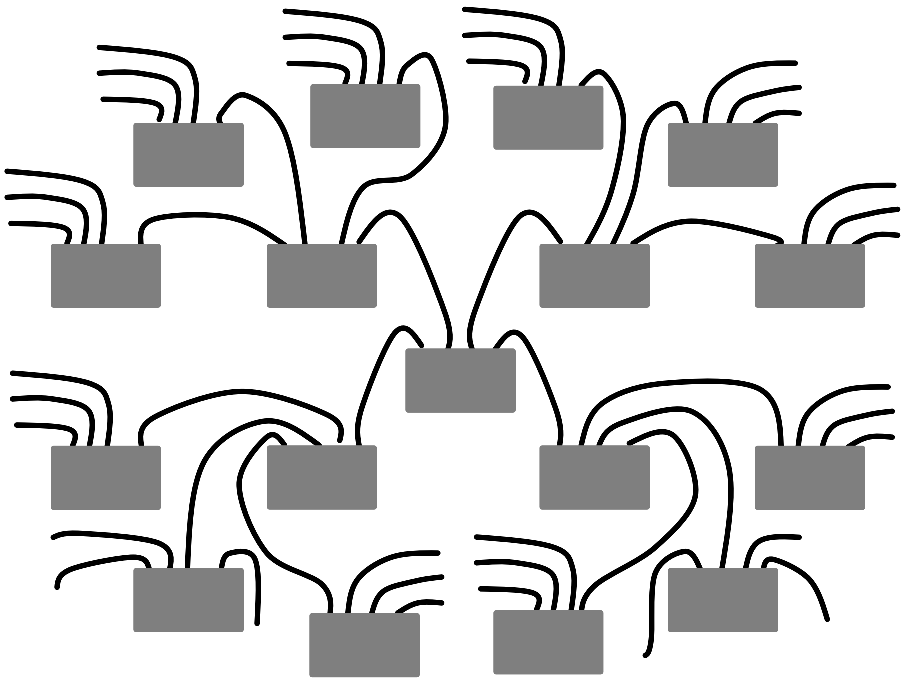
Designing an Efficient Network Topology

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.*

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

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



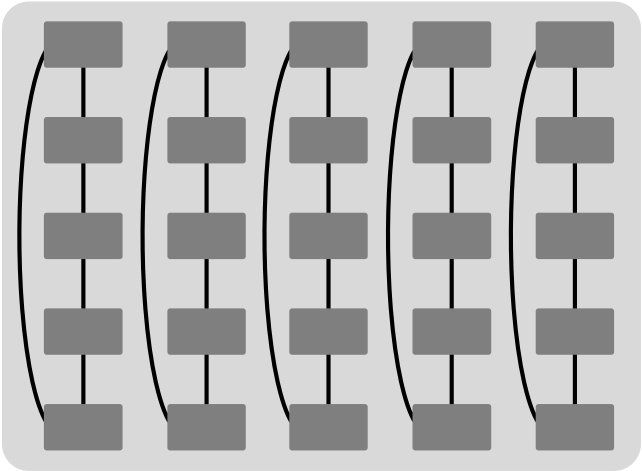
[1] M. Miller, J. Siráň. Moore graphs and beyond: A survey of the degree/diameter problem, Electronic Journal of Combinatorics, 2005.

Designing An Efficient Network Topology

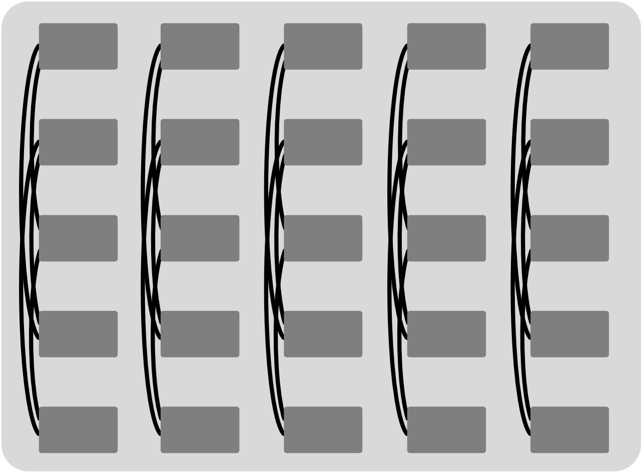
Connecting routers: Diameter 2

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

A subgraph with
identical groups of routers



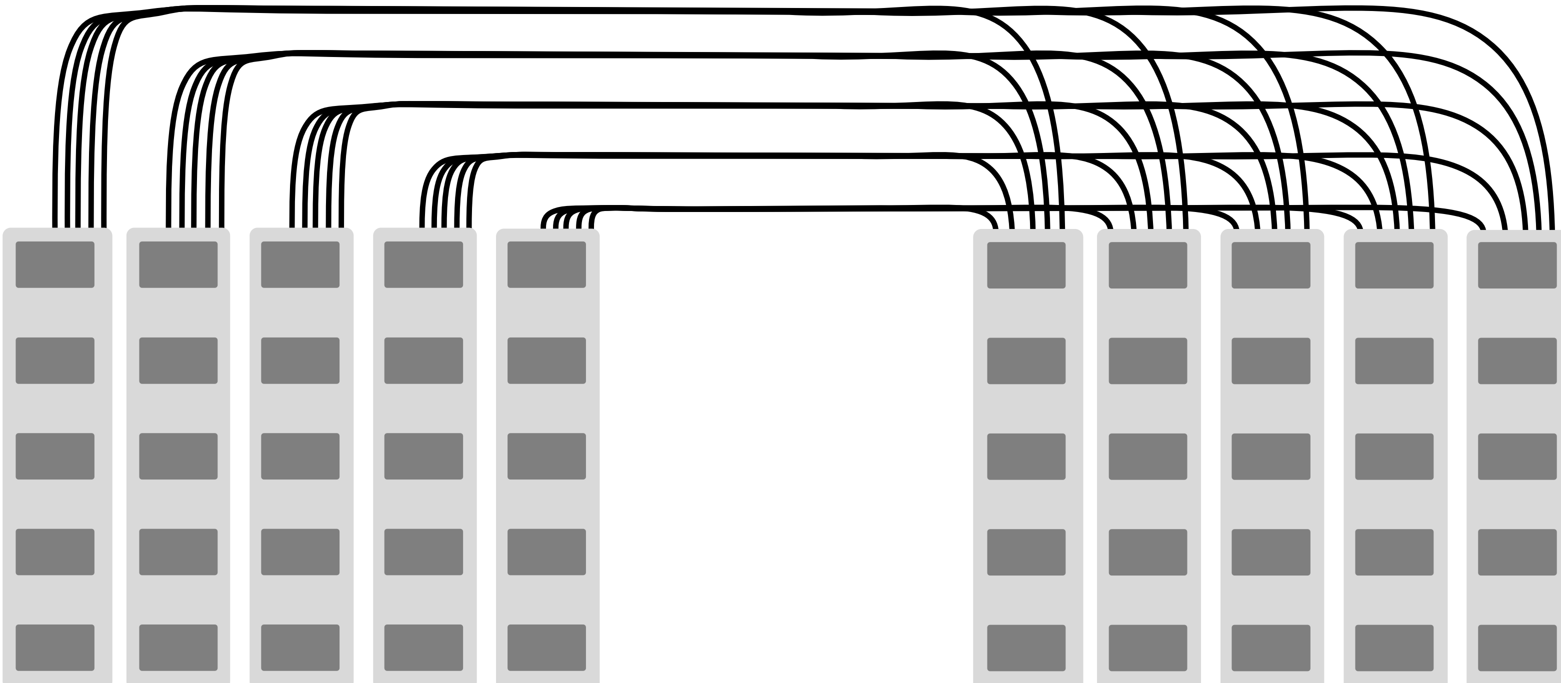
A subgraph with
identical groups of routers



[1] B. D. McKay, M. Miller, and J. Siráň. A note on large graphs of diameter two and given maximum degree. Journal of Combinatorial Theory, Series B, 74(1):110 – 118, 1998

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*

$$q = 4w + \delta;$$

$w \in \mathbb{N} \quad \delta \in \{-1,0,1\},$

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:

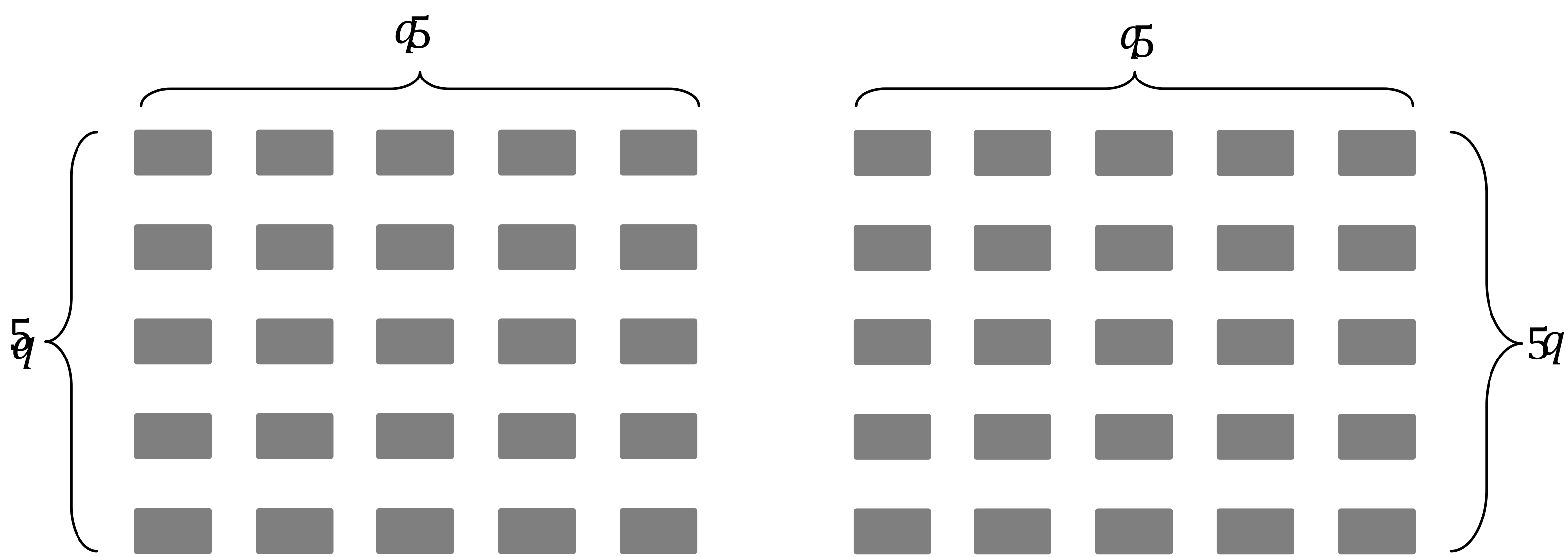
$$\mathcal{F}_q = \mathbb{Z}/q\mathbb{Z} = \{0,1,\ldots,q-1\}$$

with modular arithmetic.

E *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

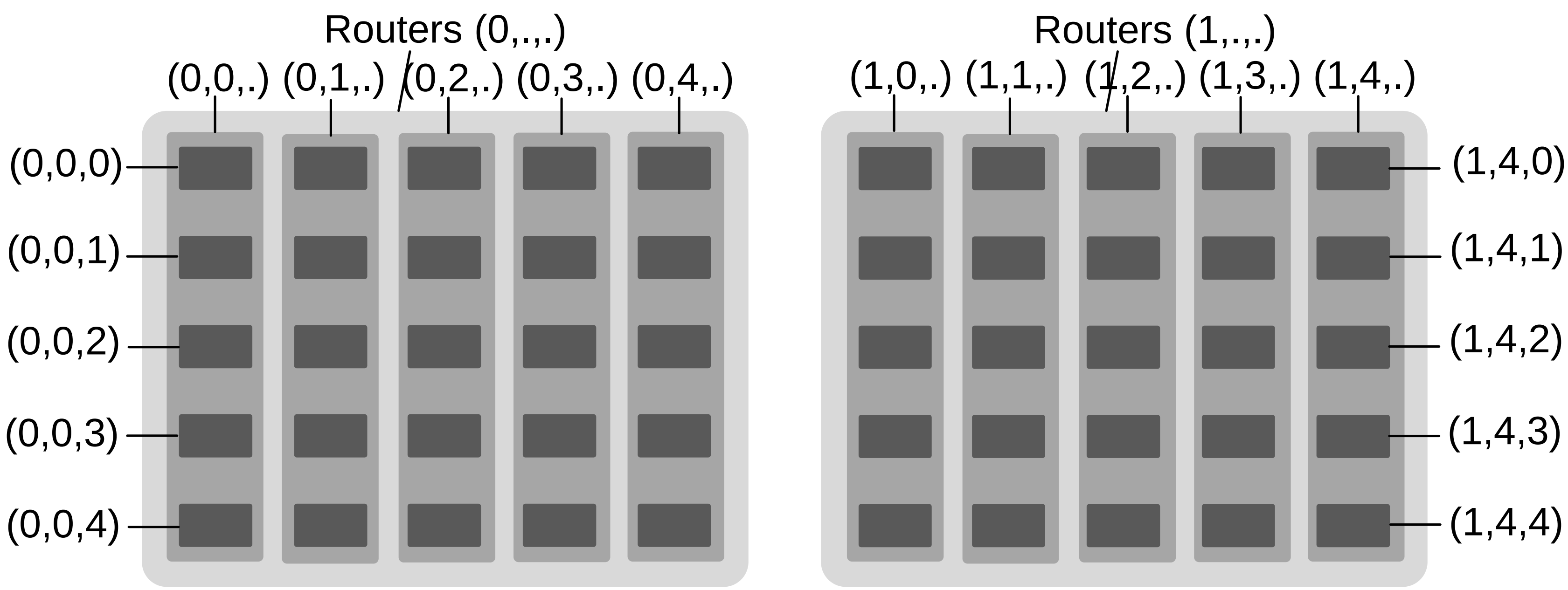
3 *Label the routers*

Set of routers:

$$\boxed{\{0,1\}} \times \boxed{\mathcal{F}_q \times \mathcal{F}_q}$$

E *Example:* $q = 5$

…



Designing an Efficient Network Topology

Connecting routers: Diameter 2

4 *Find primitive element ξ*

$\xi \in \mathcal{F}_q$ generates $\mathcal{F}_q$:

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

5 *Build Generator Sets*

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

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

E *Example:* $q = 5$

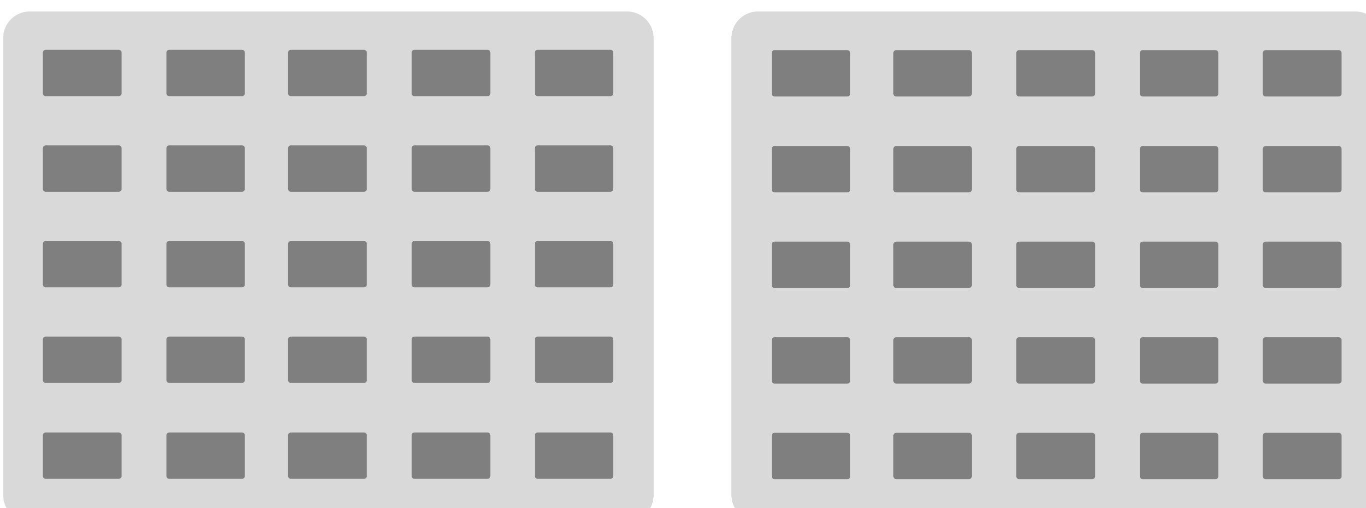
$$\mathcal{F}_5 = \{0,1,2,3,4\}$$

$$\xi = 2$$

$$1 = \xi^4 \, mod \, 5 = 2^4 \, mod \, 5 = 16 \, mod \, 5$$

$$X = \{1,4\}$$

$$X' = \{2,3\}$$



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:

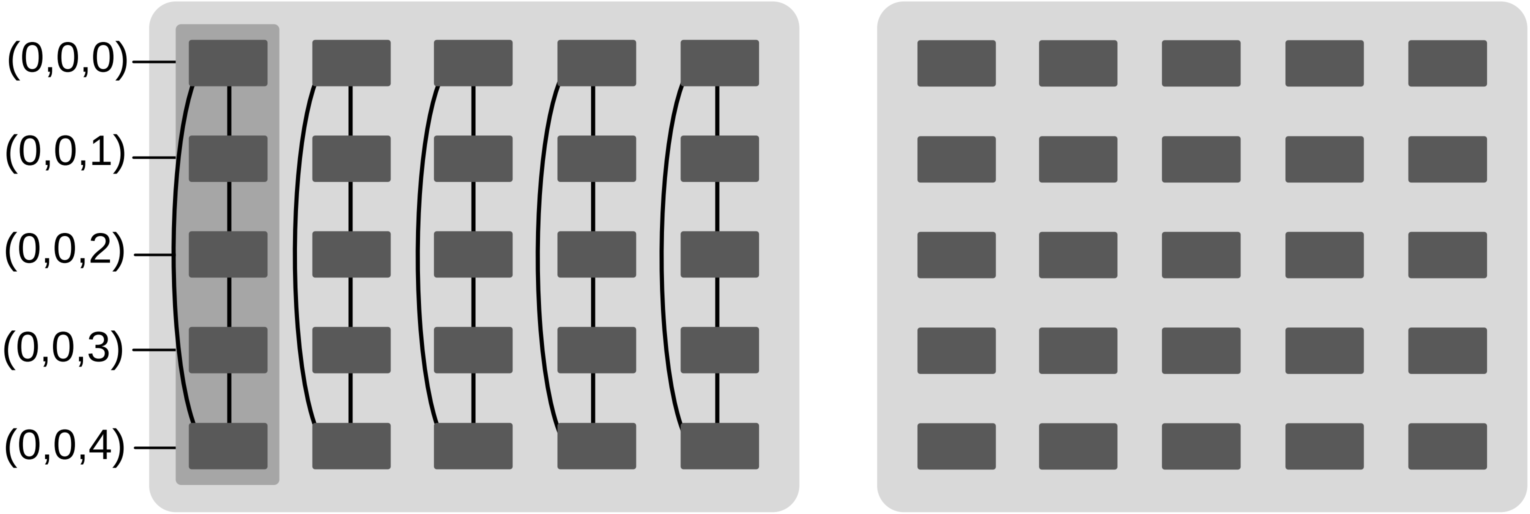
$X = \{1, \xi^2, \dots, \xi^{q-3}\}$ (for subgraph 0)

$X' = \{\xi, \xi^3, \dots, \xi^{q-2}\}$ (for subgraph 1)

E *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:

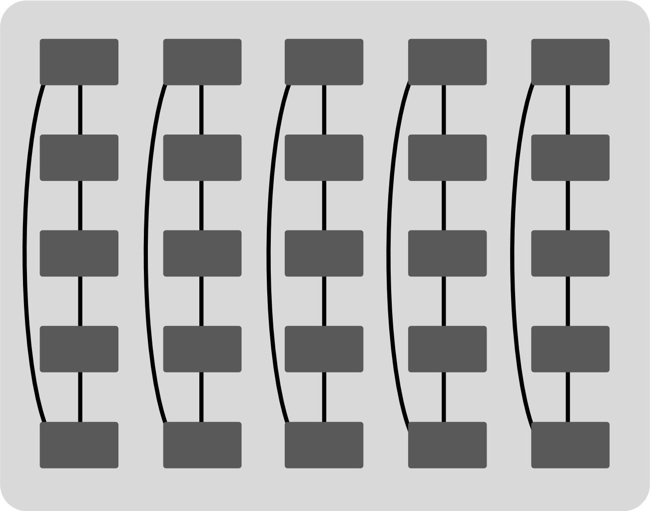
$X = \{1, \xi^2, \ldots, \xi^{q-3}\}$ (for subgraph 0)

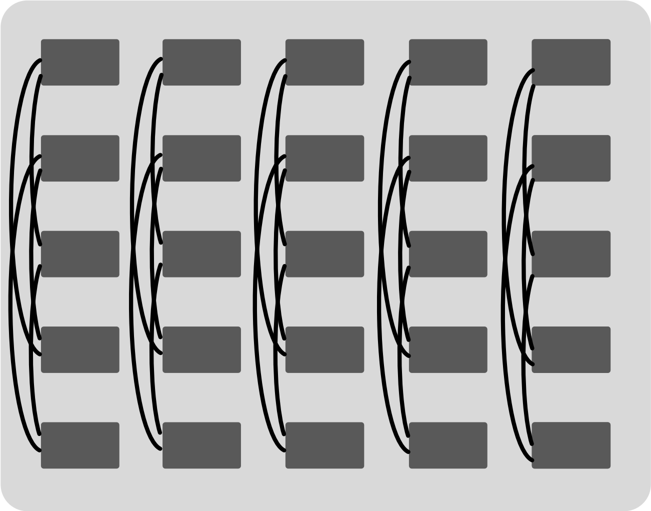
$X' = \{\xi, \xi^3, \ldots, \xi^{q-2}\}$ (for subgraph 1)

E *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 $y = mx + c$

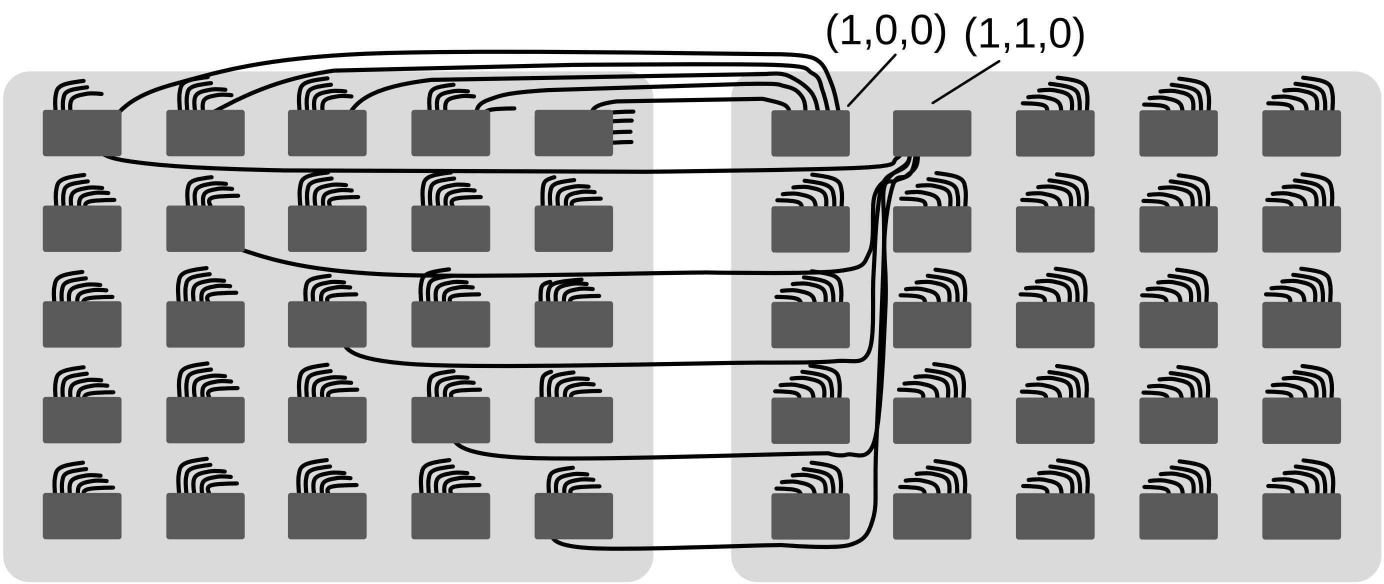
E *Example:* $q = 5$

Take Router $(1,0,0)$ $m = 0, c = 0$

$(1,0,0) \leftrightarrow (0, x, 0)$

Take Router $(1,1,0)$ $m = 1, c = 0$

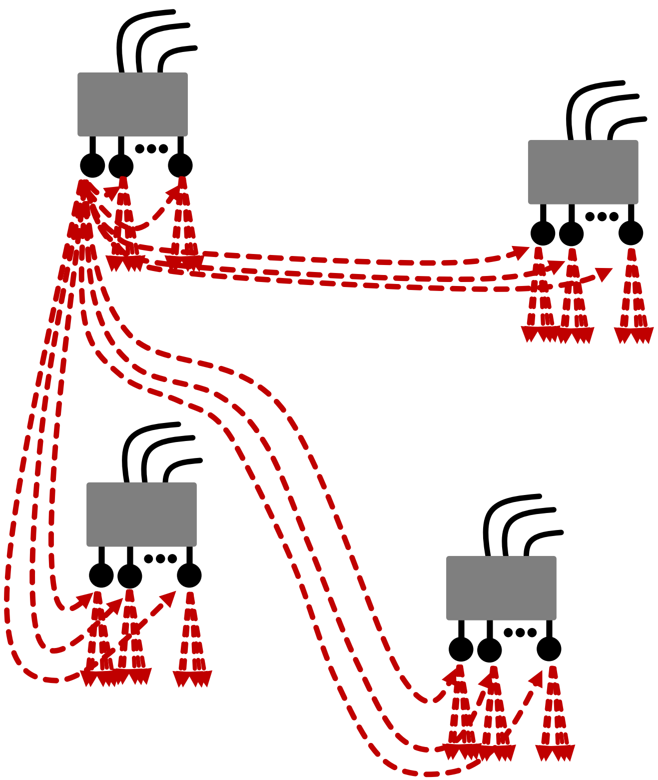
$(1,0,0) \leftrightarrow (0, x, x)$

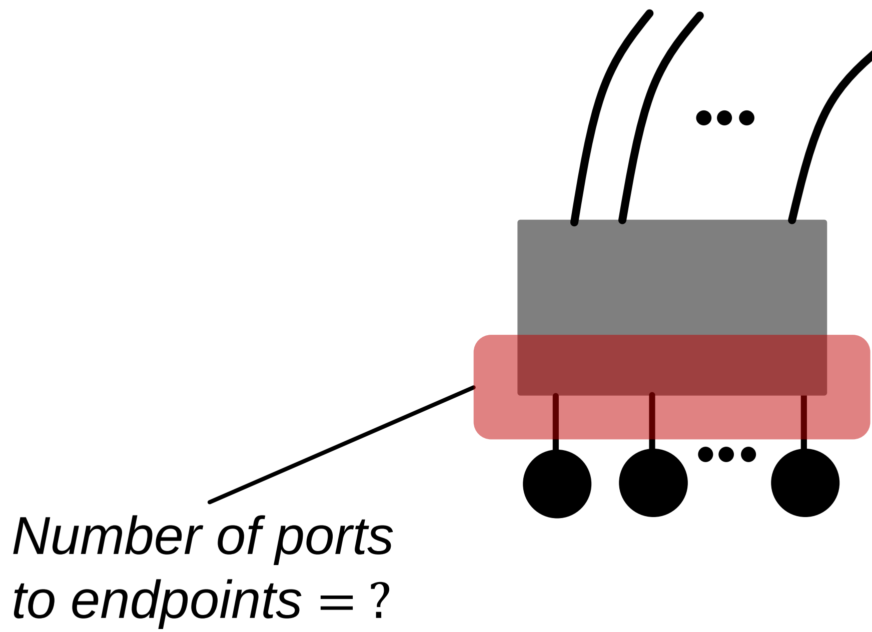


Designing an Efficient Network Topology

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

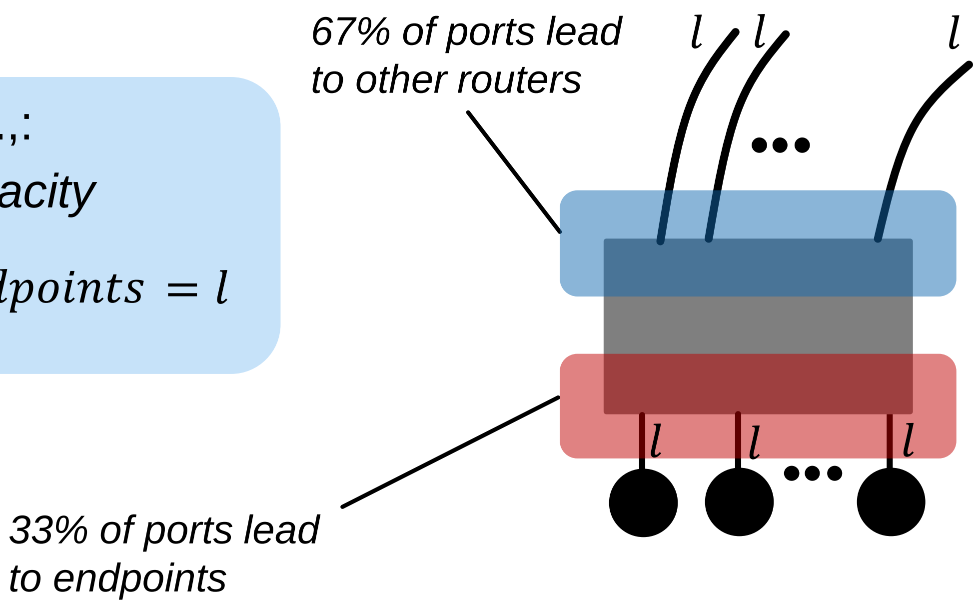
1 Get load *l* per router-router channel (average number of routes per channel)

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

2 Make the network balanced, i.e.,:

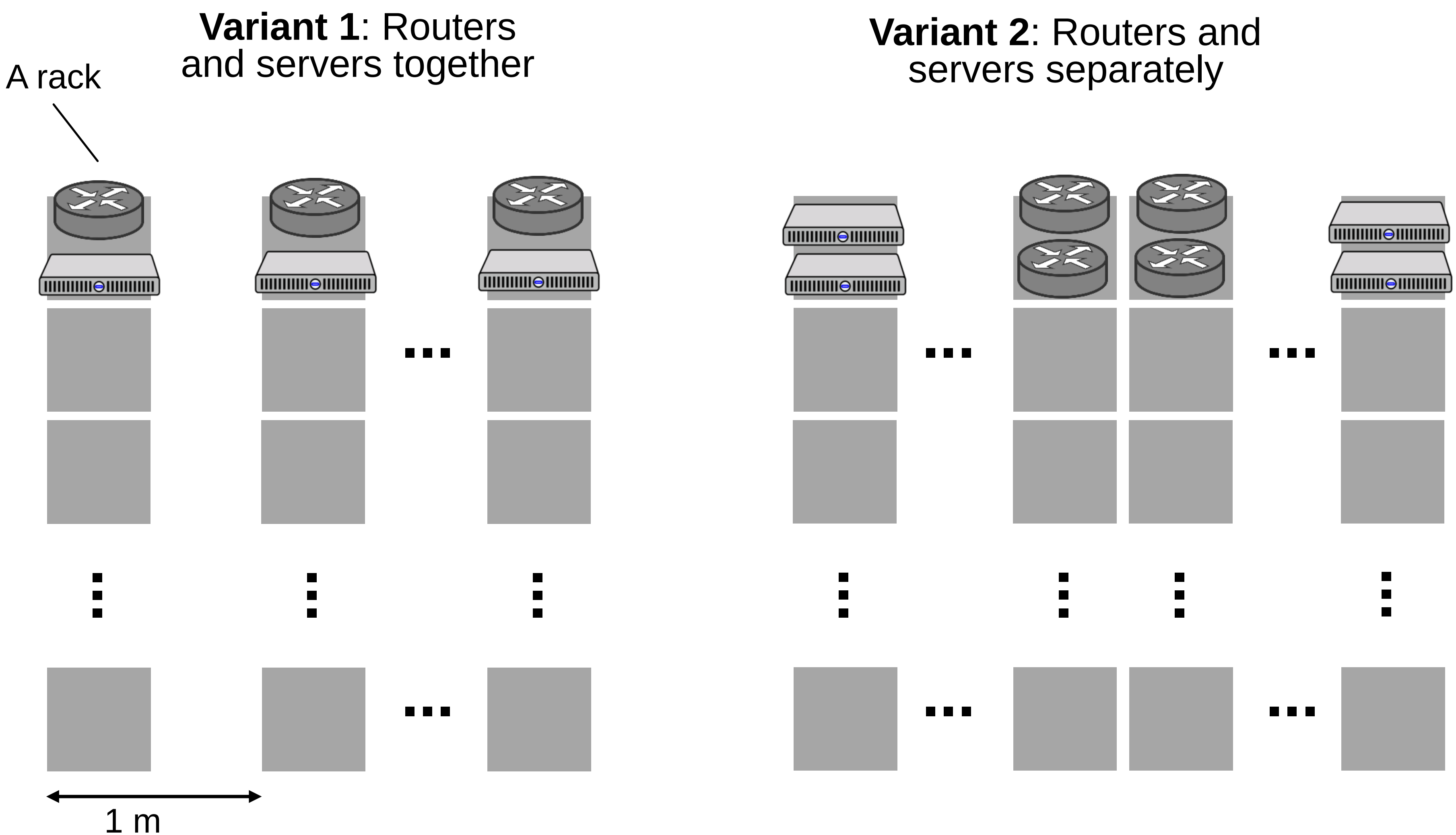
each endpoint can inject at full capacity

$$local\ uplink\ load = number\ of\ endpoints\ = l$$



Cost Comparison

Cost Models: variants



Cost Comparison

Cable Cost Model

*Prices based on: COLFAX DIRECT HPC and Data Center Gear

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

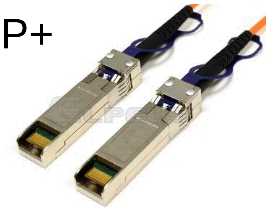
- Other used cables:

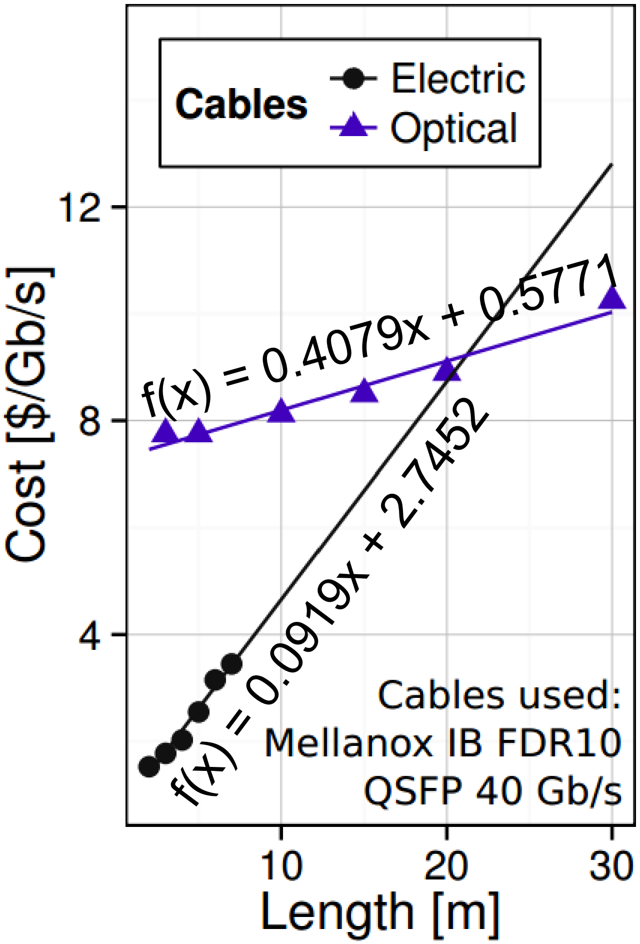
Mellanox IB QDR 56Gb/s QSFP

Mellanox Ethernet 40Gb/s QSFP

Mellanox Ethernet 10Gb/s SFP+

Elpeus Ethernet 10Gb/s SFP+





Cost Comparison

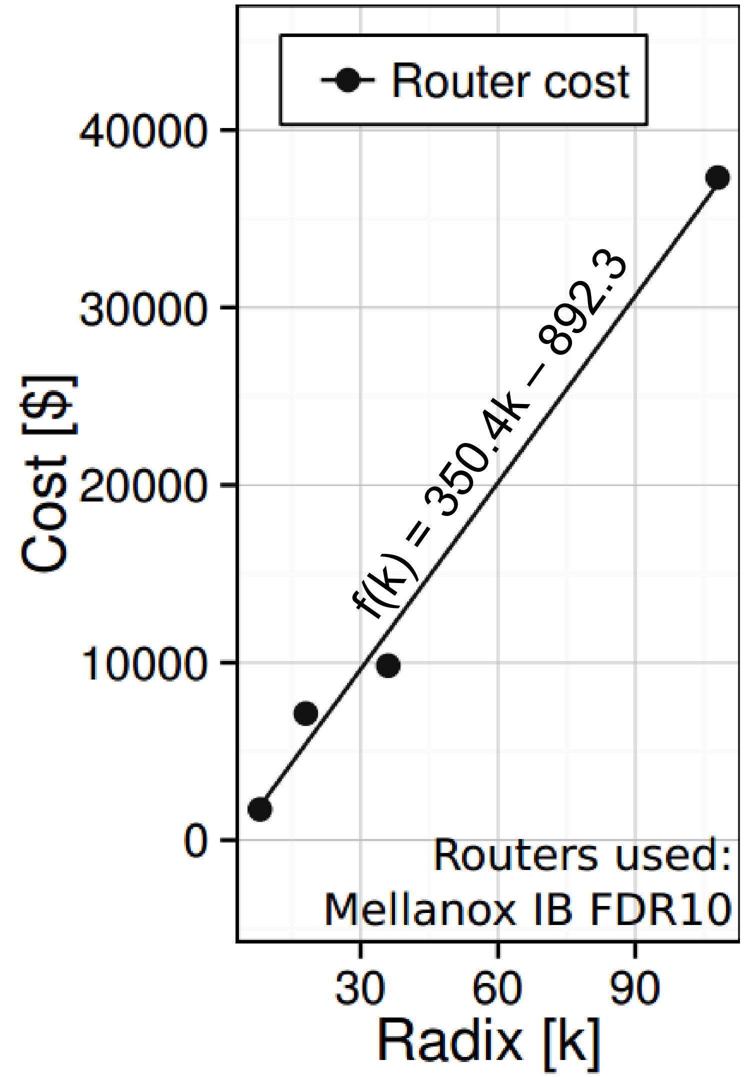
Router Cost Model

*Prices based on: COLFAX DIRECT HPC and Data Center Gear

- Router cost as a function of radix
  - The function obtained using linear regression*
  - Routers used:

Mellanox IB FDR10

Mellanox Ethernet 10/40 Gb



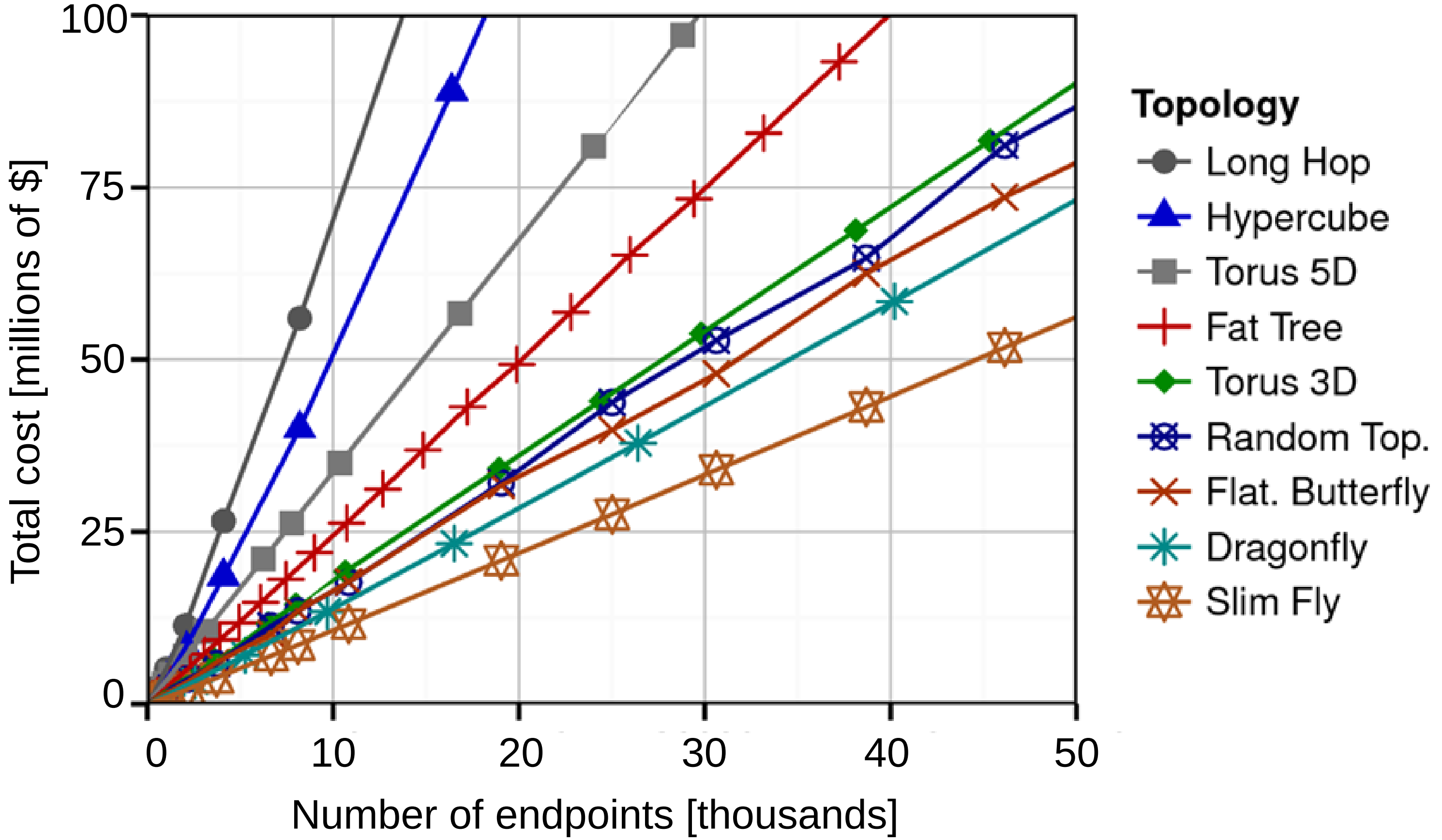
Cost Comparison

Results

Variant 2:

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

Variant 1:



Cost & Power Comparison

Detailed Case-Study

- A Slim Fly with;
  - N = 10,830
  - k = 43
  - N_r = 722

Cost & Power Comparison

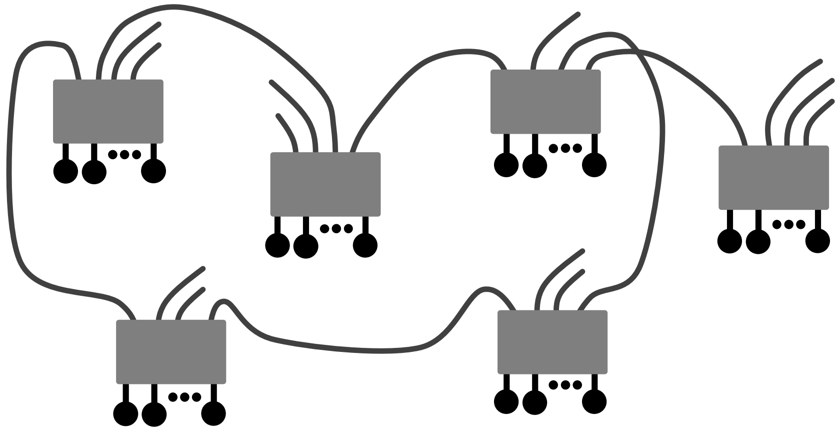
Detailed Case-Study: High-Radix Topologies

Topology	Fat tree	Random	Flat. Butterfly	Dragonfly	**Slim Fly**
Endpoints (N)	19,876	40,200	20,736	58,806	**10,830**
Routers (N_r)	2,311	4,020	1,728	5,346	**722**
Radix (k)	**43**	**43**	**43**	**43**	**43**
Electric cables	19,414	32,488	9,504	56,133	**6,669**
Fiber cables	40,215	33,842	20,736	29,524	**6,869**
Cost per node [$]	2,346	1,743	1,570	1,438	**1,033**
Power per node [W]	14.0	12.04	10.8	10.9	**8.02**

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

STRUCTURE ANALYSIS

RESILIENCY



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

Number of endpoints	Torus3D	Torus5D	Hypercube	Long Hop	Fat tree	Dragonfly	Flat. Butterfly	Random	**Slim Fly**
512	30%	-	40%	55%	35%	-	55%	60%	**60%**
1024	25%	40%	40%	55%	40%	50%	60%	-	**-**
2048	20%	-	40%	55%	40%	55%	65%	65%	**65%**
4096	15%	-	45%	55%	55%	60%	70%	70%	**70%**
8192	10%	35%	45%	55%	60%	65%	-	75%	**75%**

"-" means that a given topology does not have a variant of a given size

Performance & routing

Minimum routing

1 *Intra-group connections*

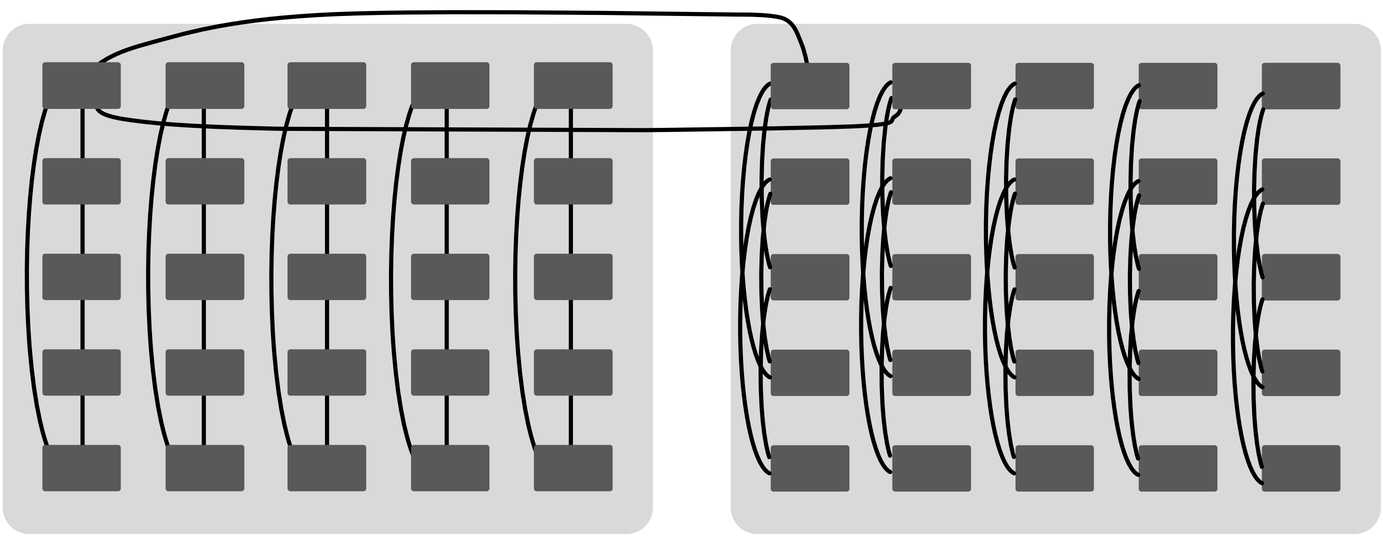
∃ Path of length 1 or 2 between two routers

2 *Inter-group connections (different types of groups)*

∃ Path of length 1 or 2 between two routers

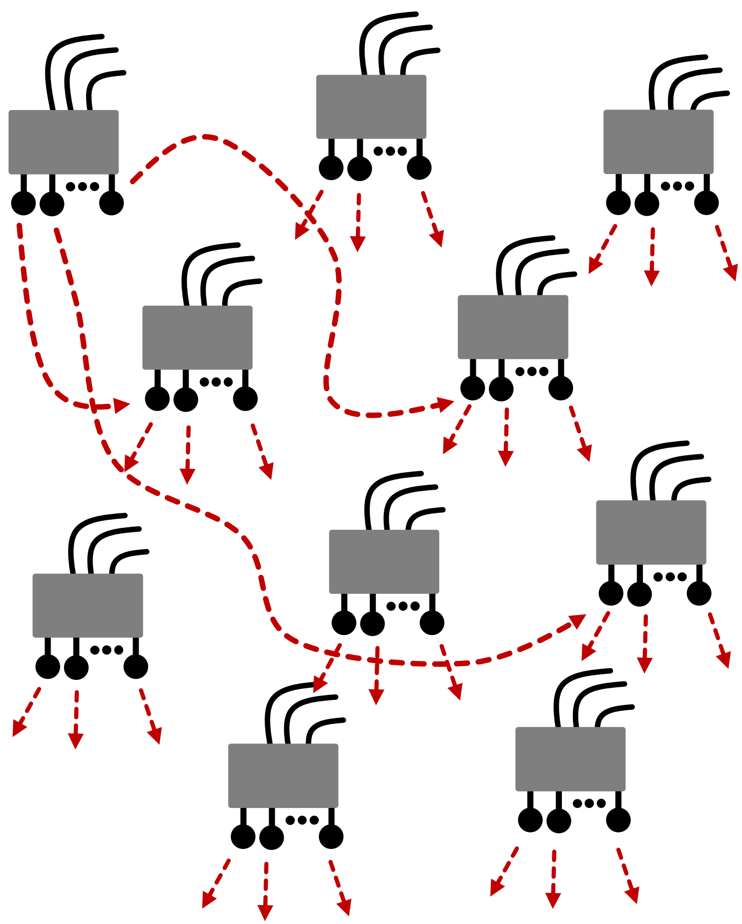
3 *Inter-group connections (identical types of groups)*

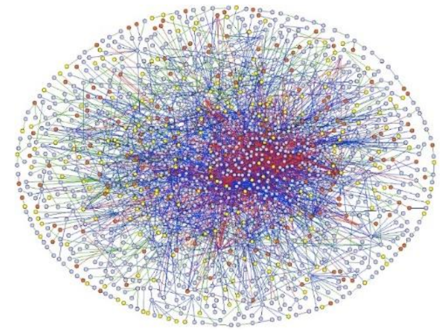
∃ Path of length 2 between two routers

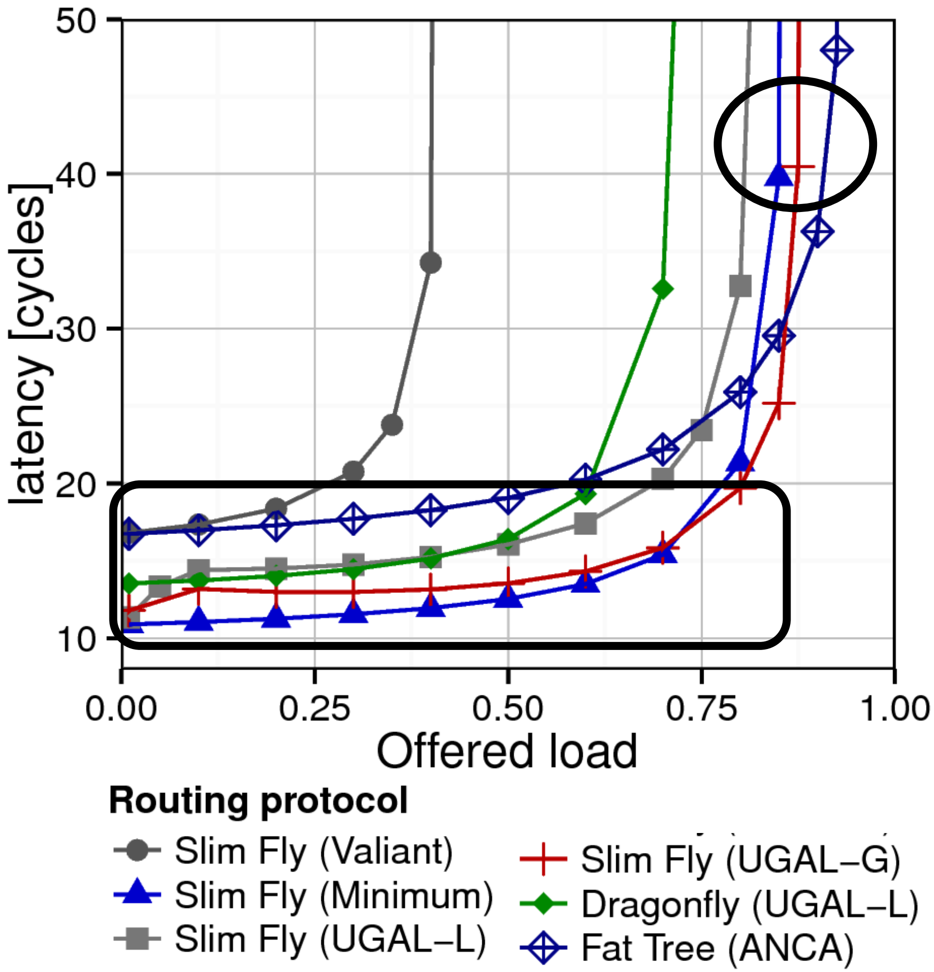


Performance & routing

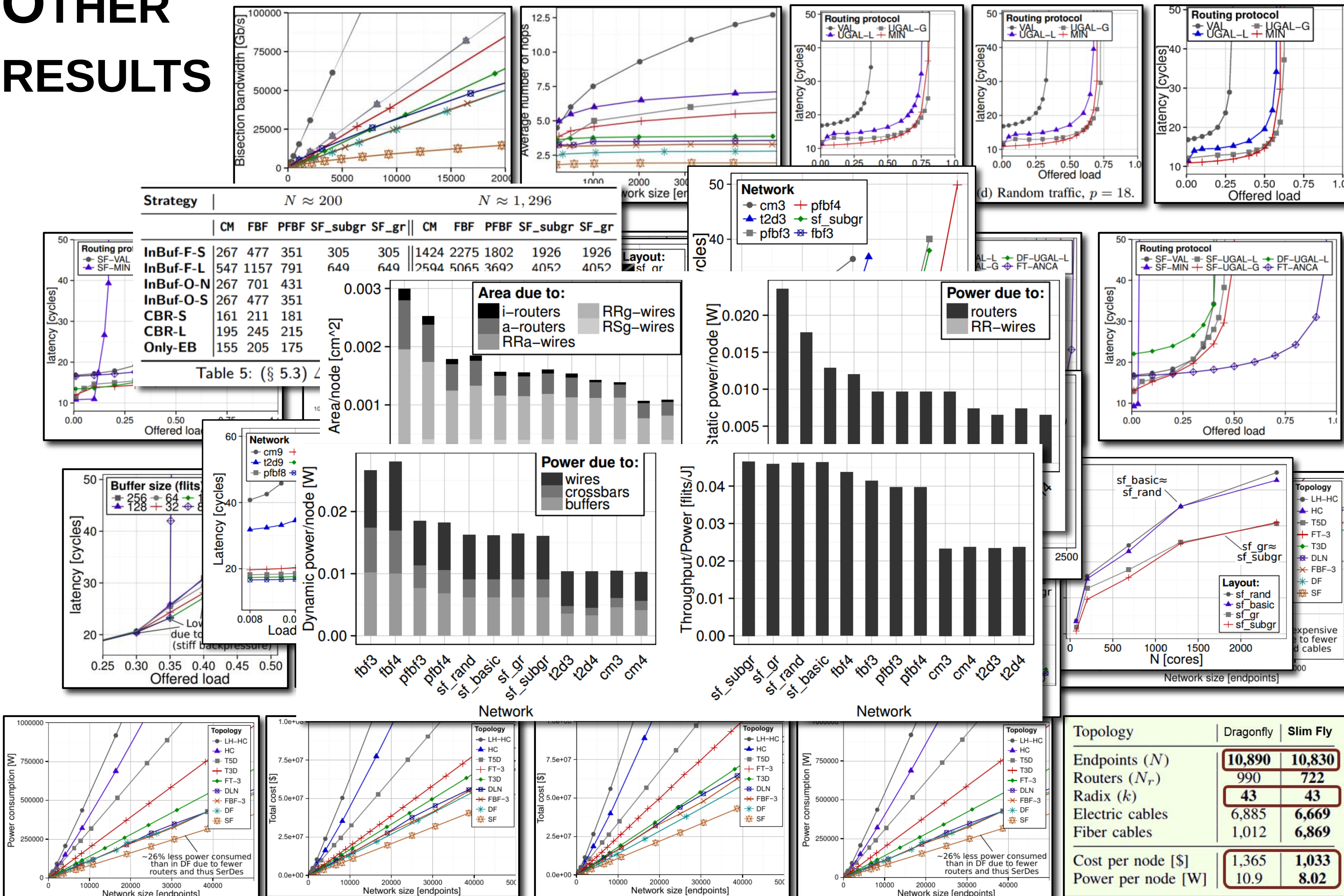
Random uniform traffic







OTHER RESULTS



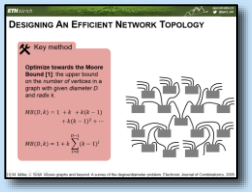
Strategy	N ≈ 200 CM	FBF	PFBF	SF_subgr	SF_gr	N ≈ 1,296 CM	FBF	PFBF	SF_subgr	SF_gr
InBuf-F-S	267	477	351	305	305	1424	2275	1802	1926	1926
InBuf-F-L	547	1157	791	649	649	2594	5065	3692	4052	4052
InBuf-O-N	267	701	431							
InBuf-O-S	267	477	351							
CBR-S	161	211	181							
CBR-L	195	245	215							
Only-EB	155	205	175							

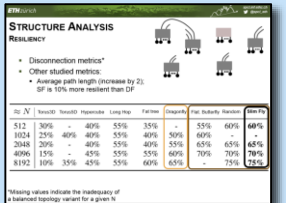
Table 5: (§ 5.3)

Topology	Dragonfly	Slim Fly
Endpoints (N)	10,890	10,830
Routers (N_r)	990	722
Radix (k)	43	43
Electric cables	6,885	6,669
Fiber cables	1,012	6,869
Cost per node [$]	1,365	1,033
Power per node [W]	10.9	8.02

A LOWEST-DIAMETER TOPOLOGY

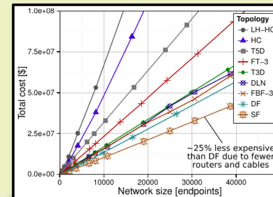
- → Approaching the Moore Bound
- → Resilient

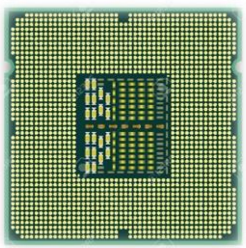




A COST & POWER EFFECTIVE TOPOLOGY

- → 25% less expensive than Dragonfly,
- → 26% less power-hungry than Dragonfly





Thank you for your attention

http://spcl.inf.ethz.ch/SlimFly

A HIGH-PERFORMANCE TOPOLOGY

- → Lowest latency
- → Full global bandwidth