ETH zürich

spcl.inf.ethz.ch
@spcl_eth

TORSTEN HOEFLER, MACIEJ BESTA

Slim Fly: A Cost Effective Low-Diameter Network Topology

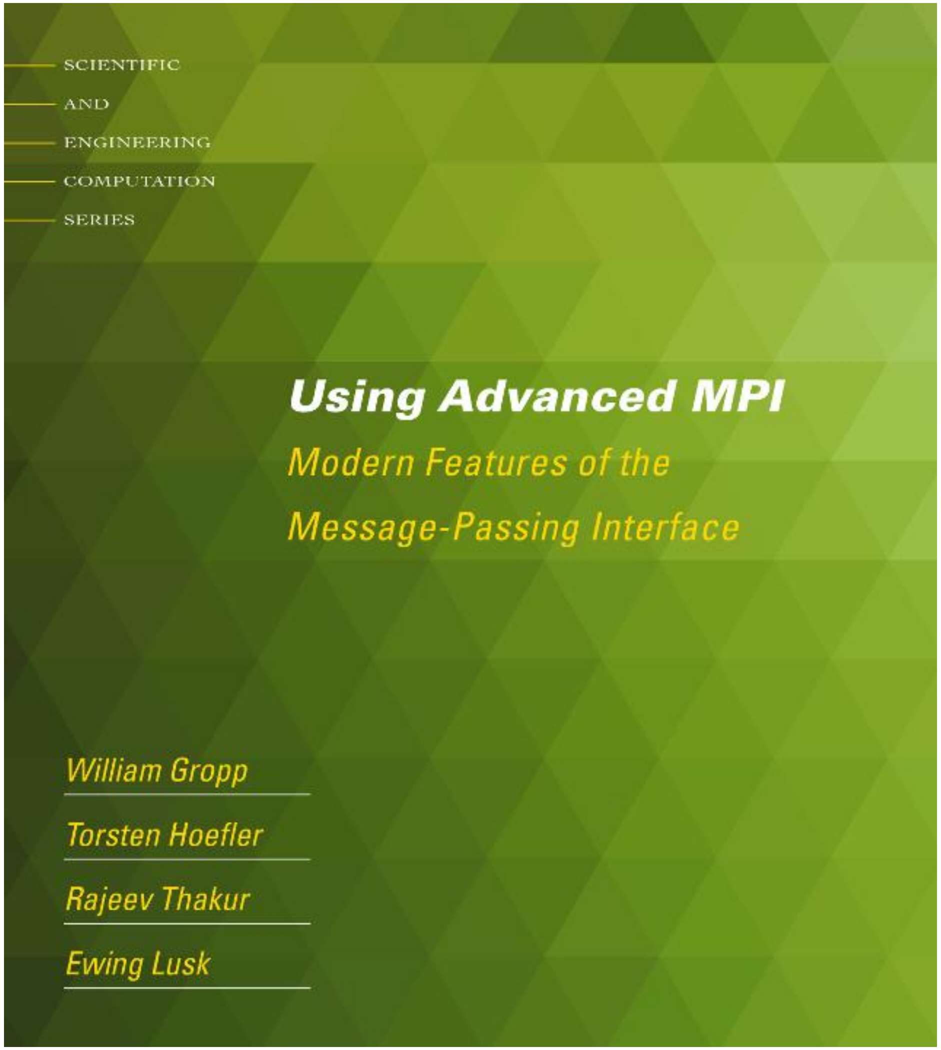
Images belong to their creator!

Background

- **I'm an HPC (systems) guy**

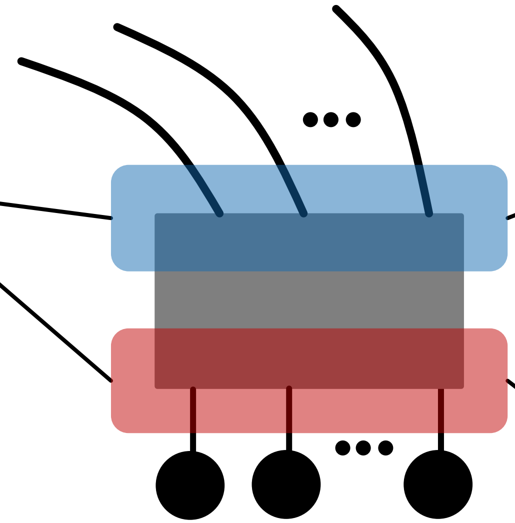
INFINIBAND℠
TRADE ASSOCIATION MEMBER

- **New to the DC area but very interested and motivated!**
  - Several projects (see last slide)

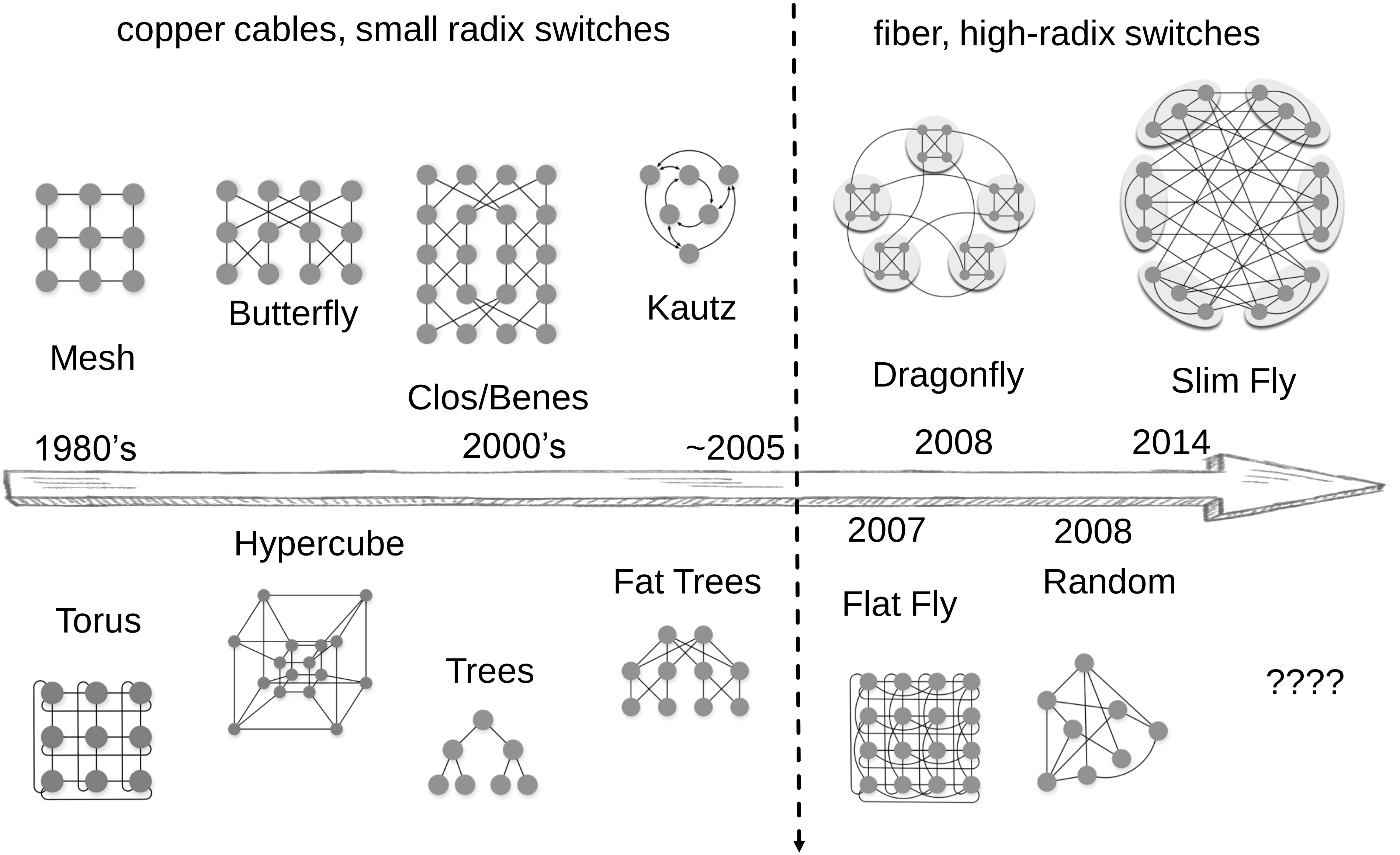


Networks, Limits, and Design Space

- **Networks cost 25-30% of a large compute cluster**
  - How much at rack-scale?
- **Hard limits:**
  - Router radix
  - Cable length
- **Soft limits:**
  - Cost
  - Performance



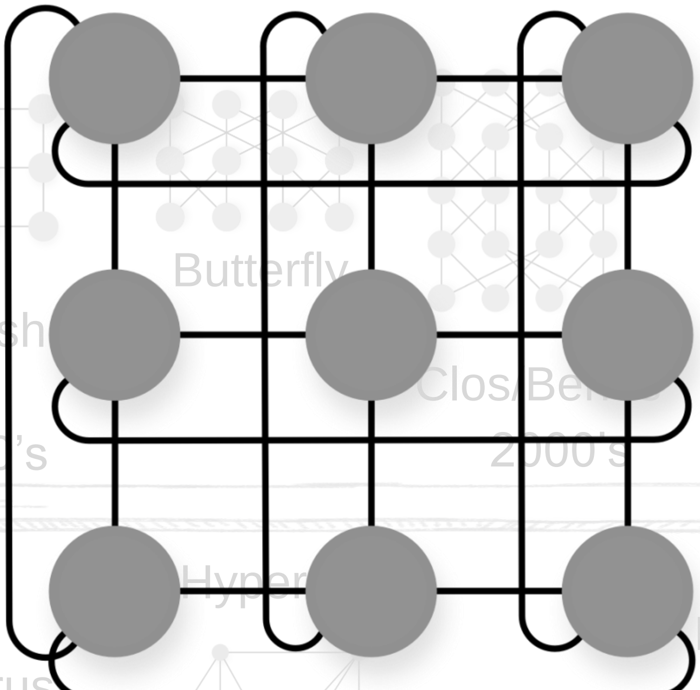
A BRIEF HISTORY OF NETWORK TOPOLOGIES

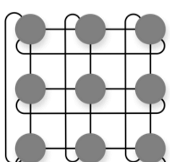


A BRIEF HISTORY OF NETWORK TOPOLOGIES

copper cables, small radix switches

fiber, high-radix switches



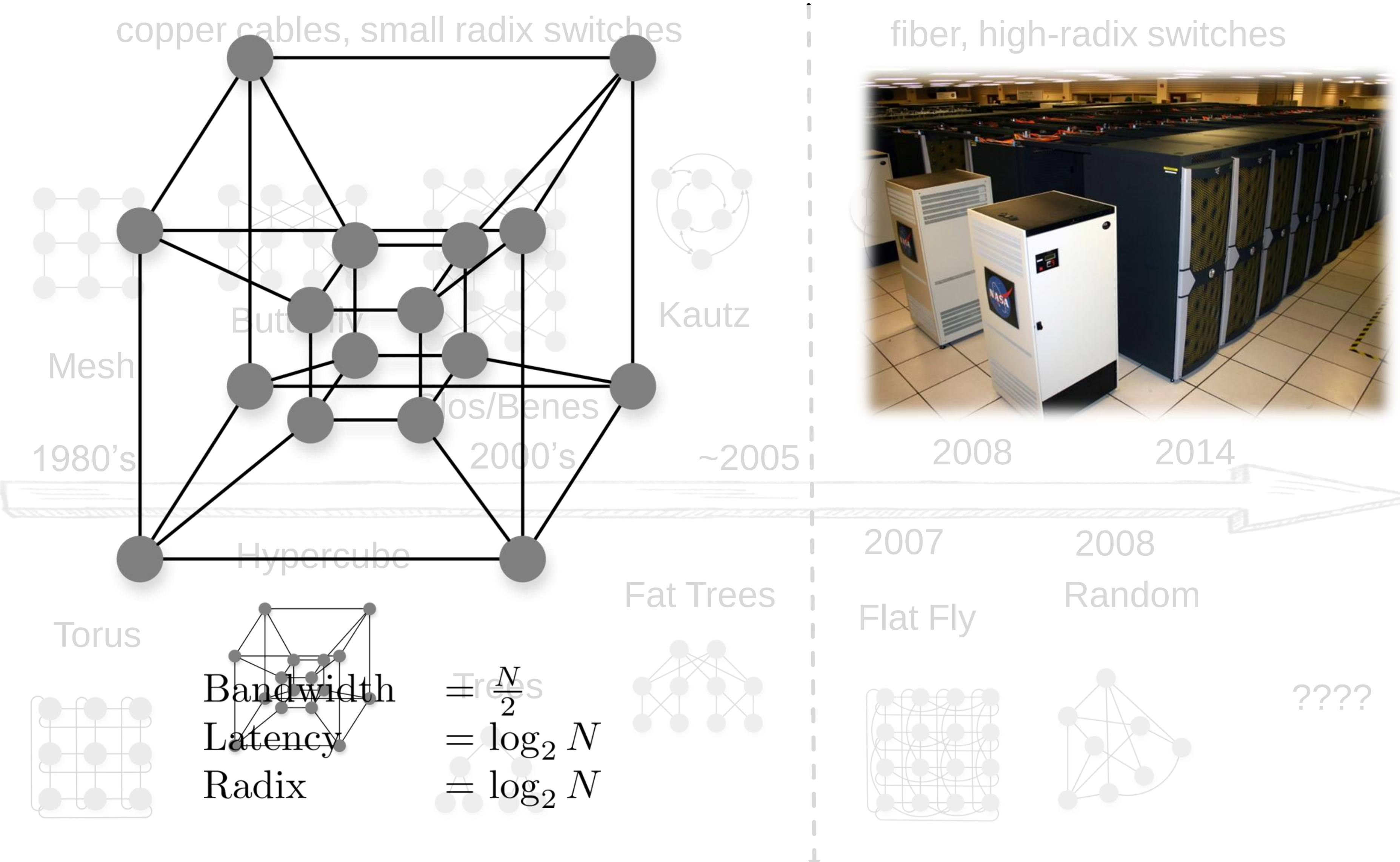


$$\text{Bandwidth} = 2\sqrt[d]{N^{d-1}}$$
$$\text{Latency} = \frac{d}{2}\sqrt[d]{N}$$
$$\text{Radix} = 2d$$

Mesh · Butterfly · Kautz · Torus · Fat Trees · Trees

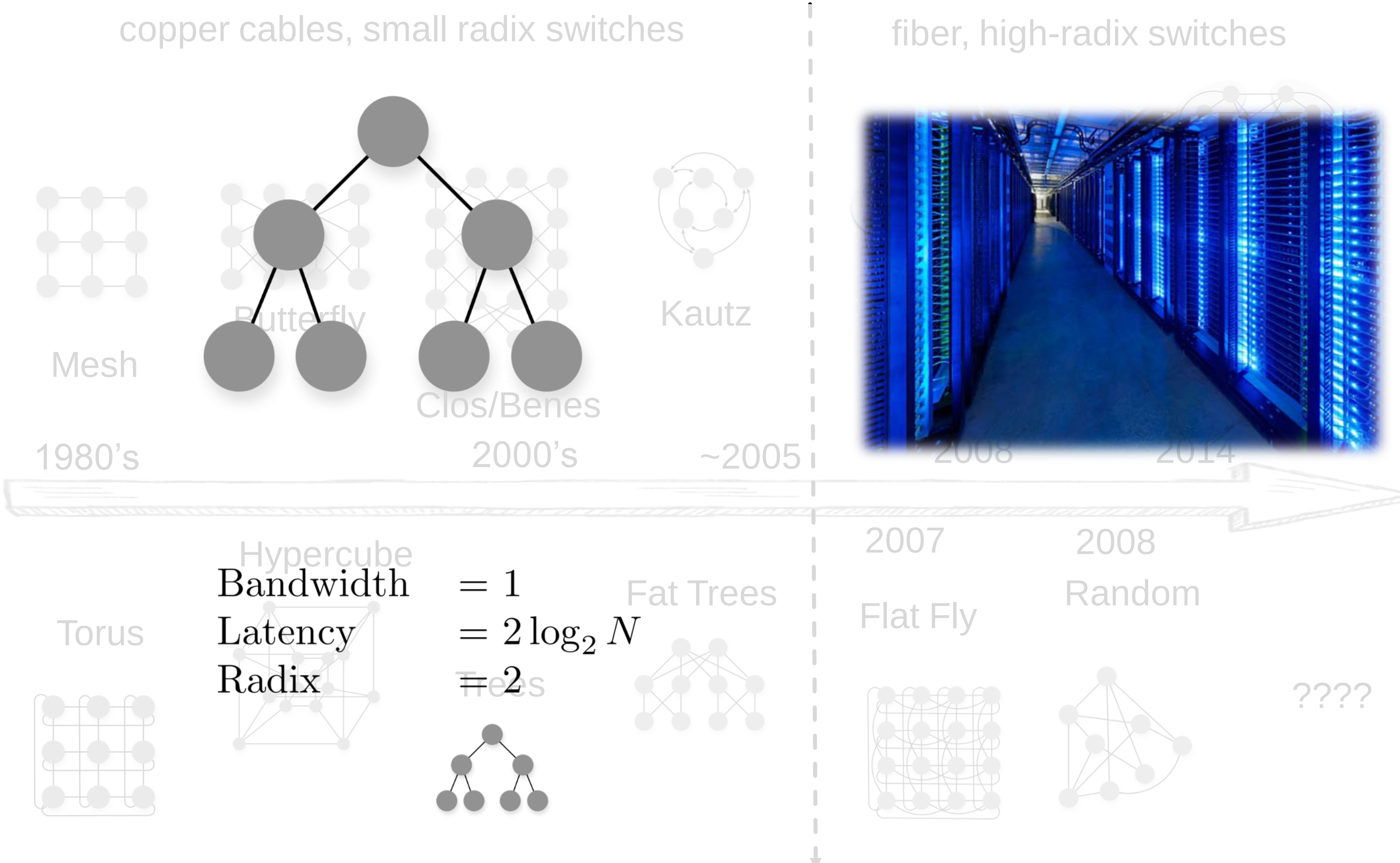
1980's · ~2005 · 2008 · 2014

A BRIEF HISTORY OF NETWORK TOPOLOGIES



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

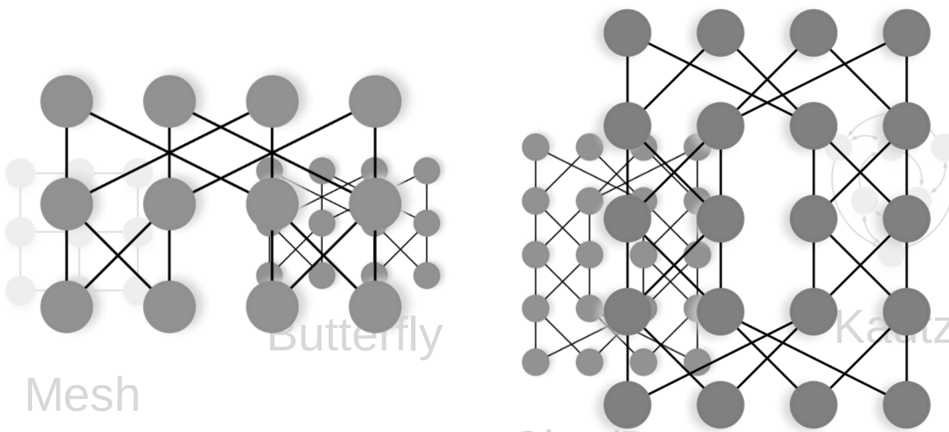
A BRIEF HISTORY OF NETWORK TOPOLOGIES



$$\text{Bandwidth} = 1$$
$$\text{Latency} = 2\log_2 N$$
$$\text{Radix} = 2$$

A BRIEF HISTORY OF NETWORK TOPOLOGIES

copper cables, small radix switches

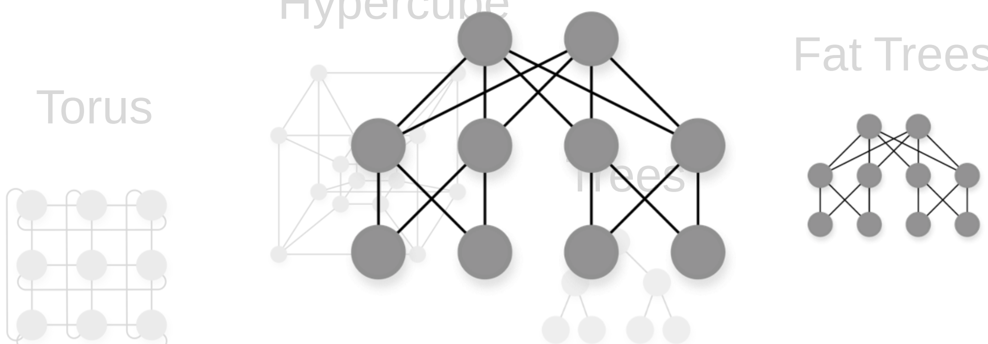


$$\text{Bandwidth} = \frac{N}{2}$$

$$\text{Latency} = 2\log_2 N$$

$$\text{Radix} = 4$$

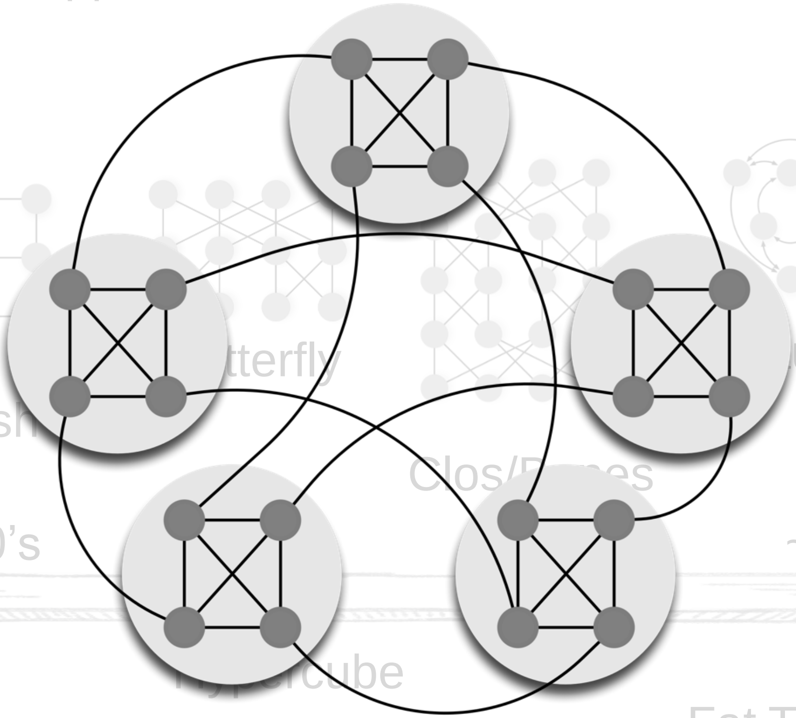
1980's 2000's ~2005 2008 2014



A BRIEF HISTORY OF NETWORK TOPOLOGIES

copper cables, small radix switches

fiber, high-radix switches



Mesh

Clos/Benes

1980's

~2005

2008

2014

Dragonfly

Slim Fly

Hypercube

Torus

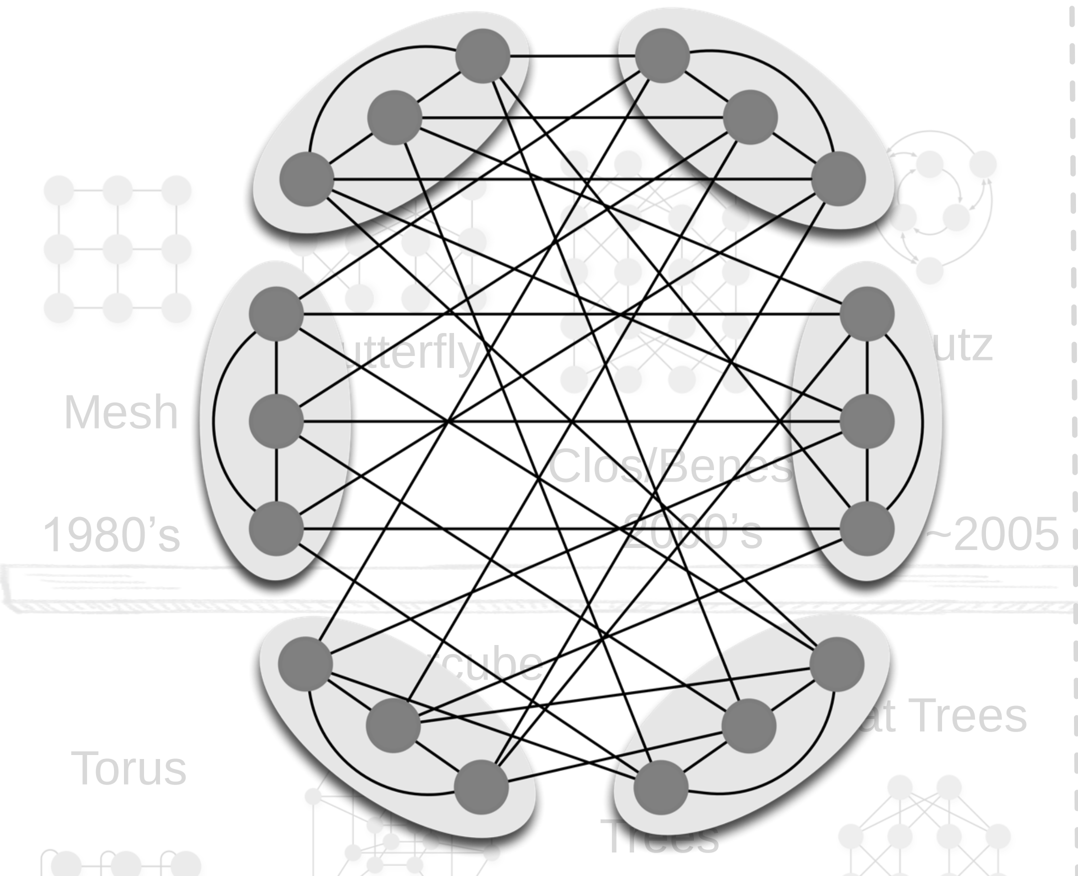
Trees

Fat Tree

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

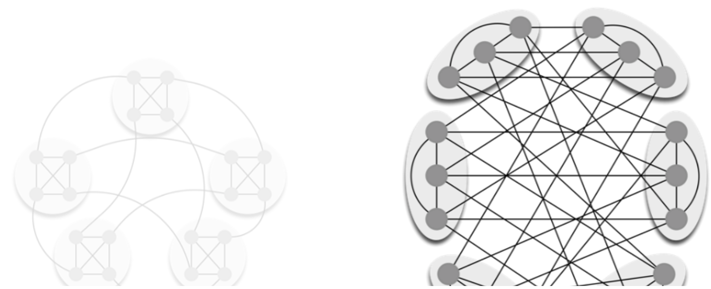
A BRIEF HISTORY OF NETWORK TOPOLOGIES

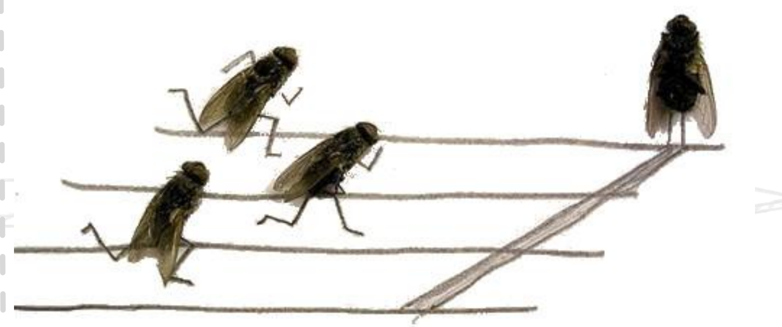
copper cables, small radix switches



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

fiber, high-radix switches





????

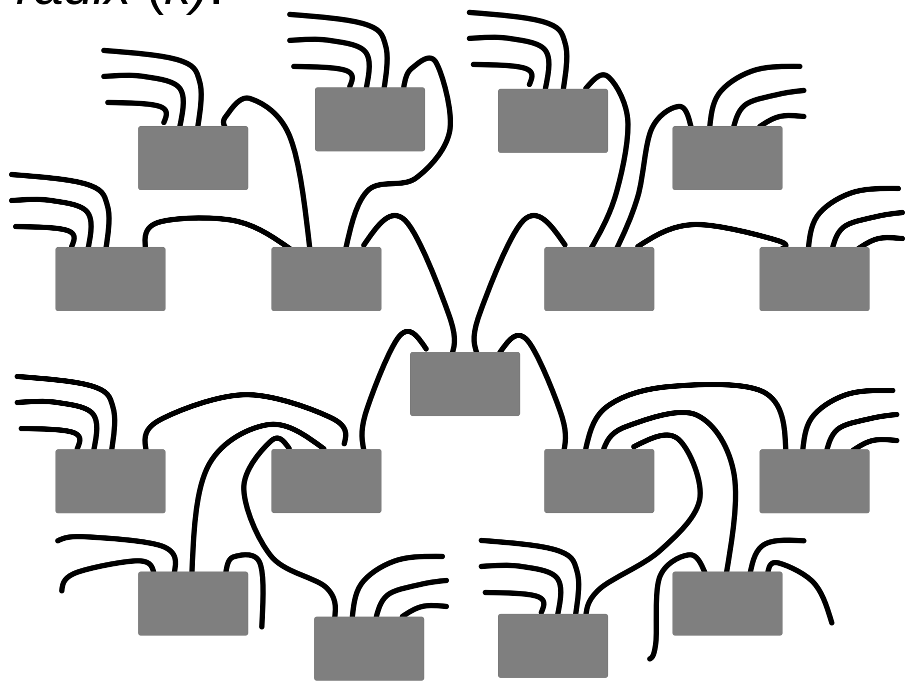
DESIGNING AN EFFICIENT NETWORK TOPOLOGY

CONNECTING ROUTERS

- **Intuition: lower average distance → lower resource needs**
  - A new view as primary optimization target!
- Moore Bound [1]: upper bound on the *number of routers* in a graph with given *diameter* (*D*) and *network radix* (*k*).

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

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

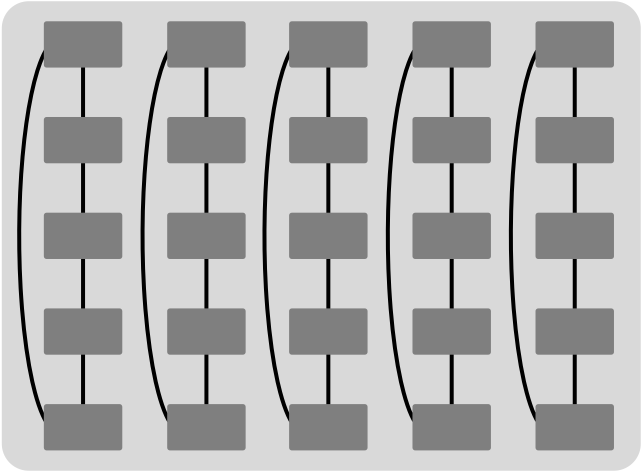


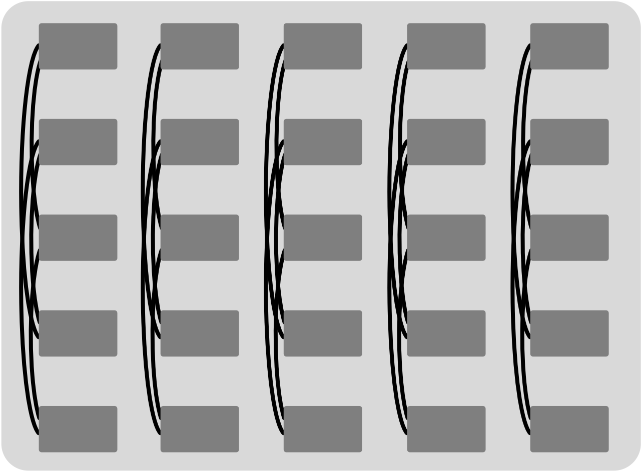
[1] M. Miller, J. Siráň. Moore graphs and beyond: A survey of the degree/diameter problem, Electronic Journal of Combinatorics, 2005.

Designing an Efficient Network Topology

Connecting routers: Diameter 2

- Example Slim Fly design for *diameter* = 2: *MMS graphs* [1] (utilizing graph covering)

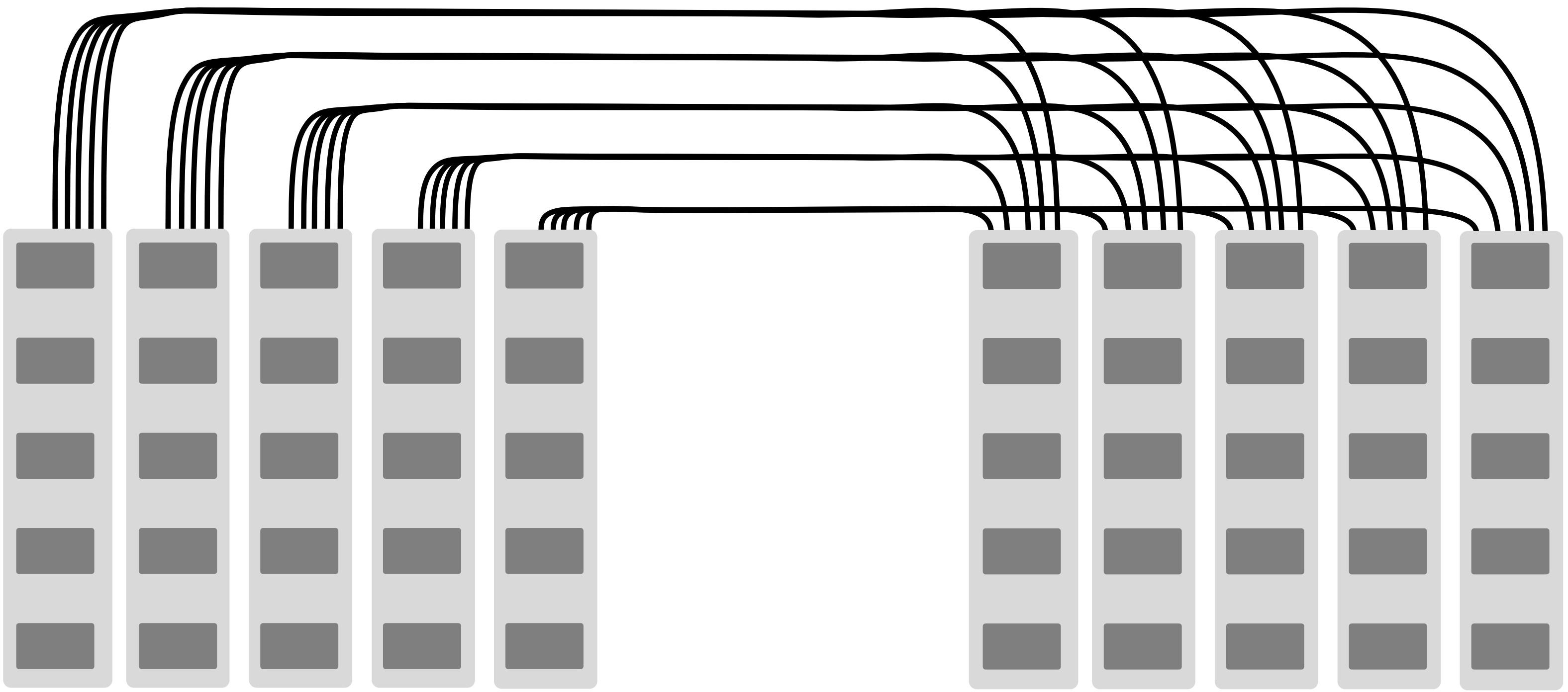




[1] B. D. McKay, M. Miller, and J. Siráň. A note on large graphs of diameter two and given maximum degree. Journal of Combinatorial Theory, Series B, 74(1):110 – 118, 1998

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:

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

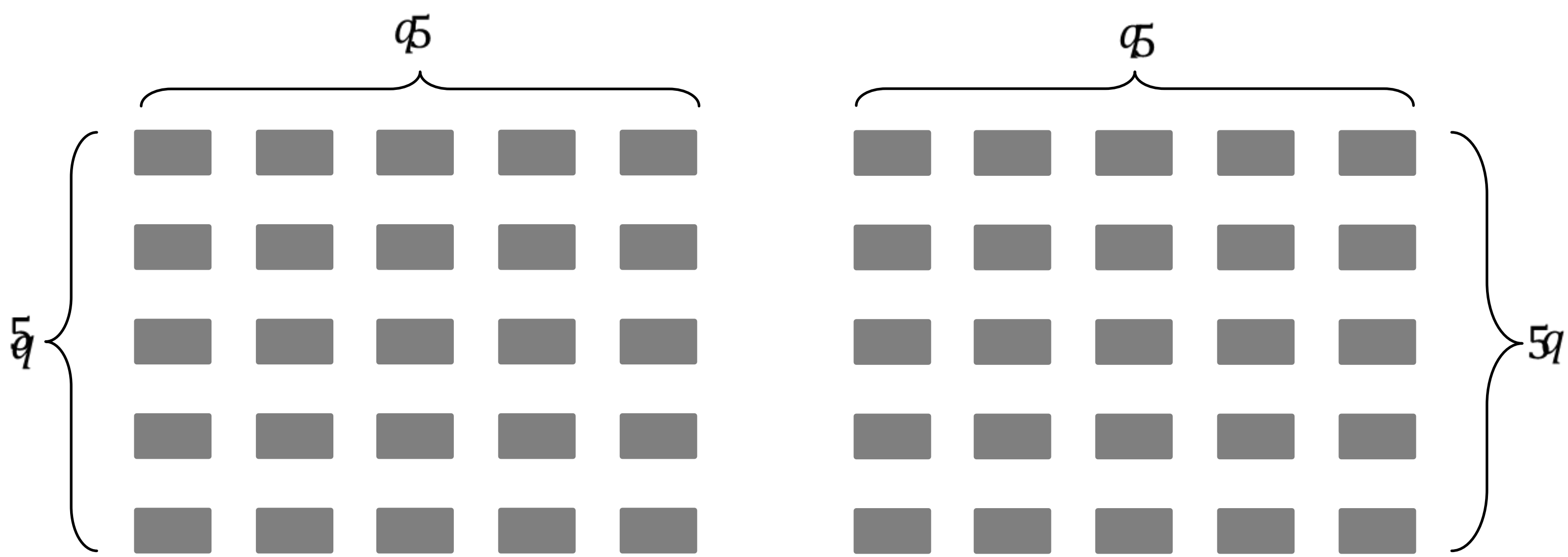
with modular arithmetic.

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

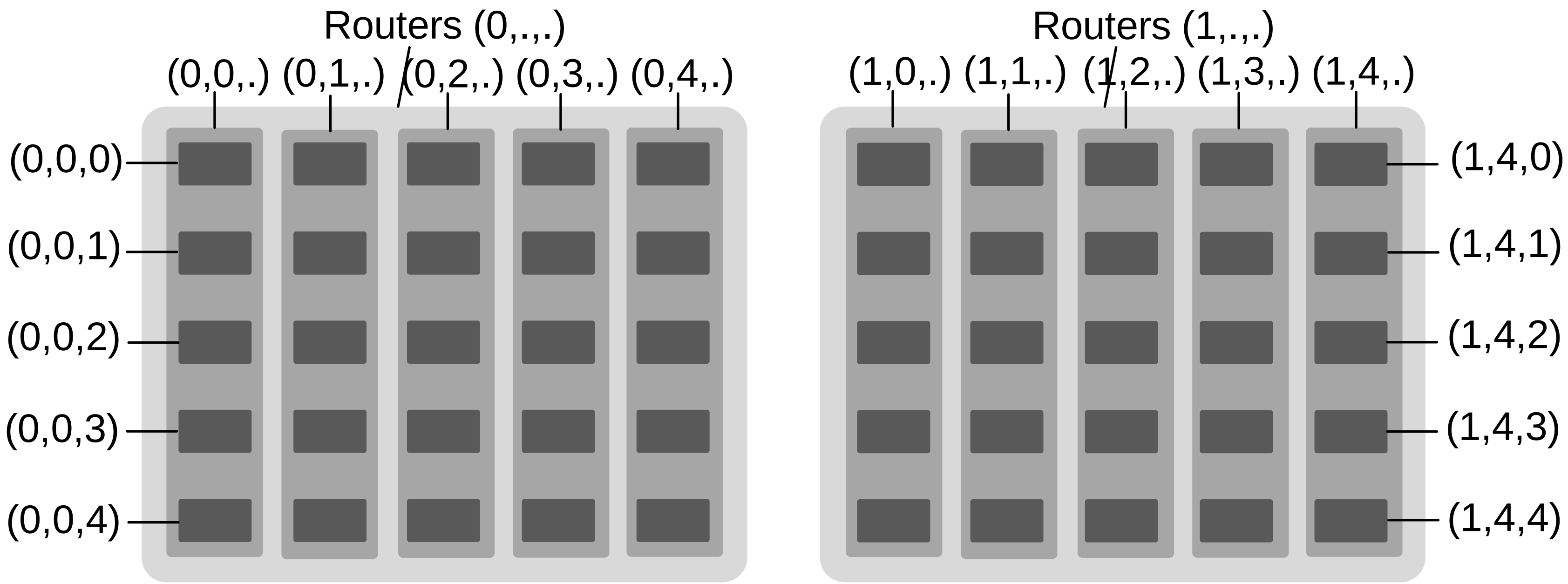
3 *Label the routers*

Set of routers:

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

E *Example:* $q = 5$

…



Designing An Efficient Network Topology

Connecting routers: Diameter 2

E *Example:* $q = 5$

$$\mathcal{F}_5 = \{0,1,2,3,4\}$$
$$\xi = 2$$
$$1 = \xi^4 \, mod \, 5 = 2^4 \, mod \, 5 = 16 \, mod \, 5$$

$$X = \{1,4\}$$
$$X' = \{2,3\}$$

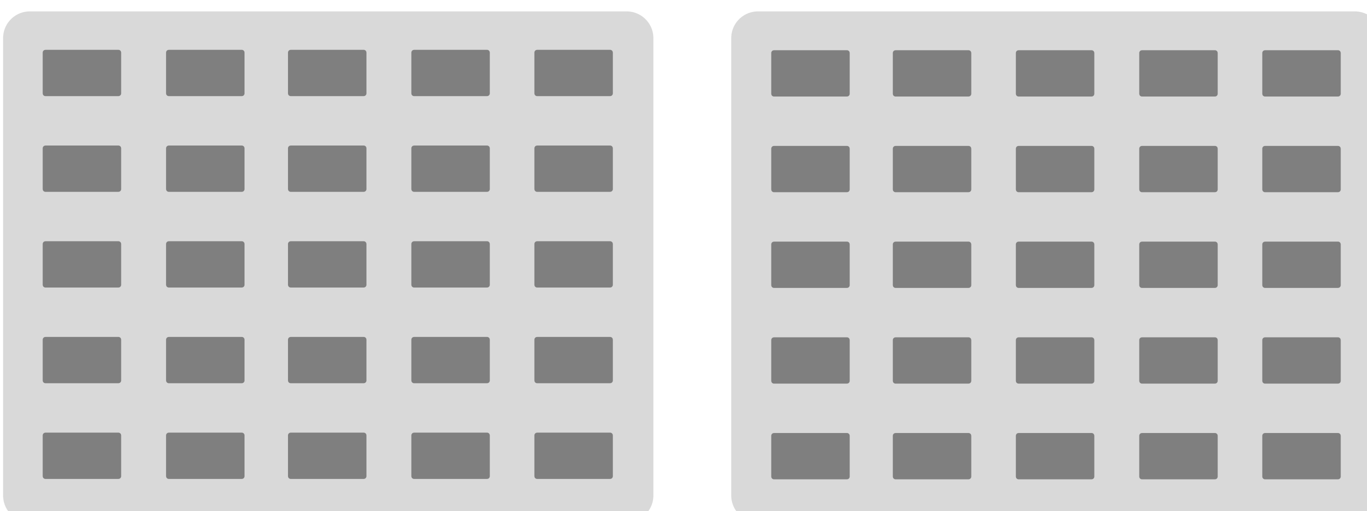
4 *Find primitive element* ξ

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

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

5 *Build Generator Sets*

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



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:

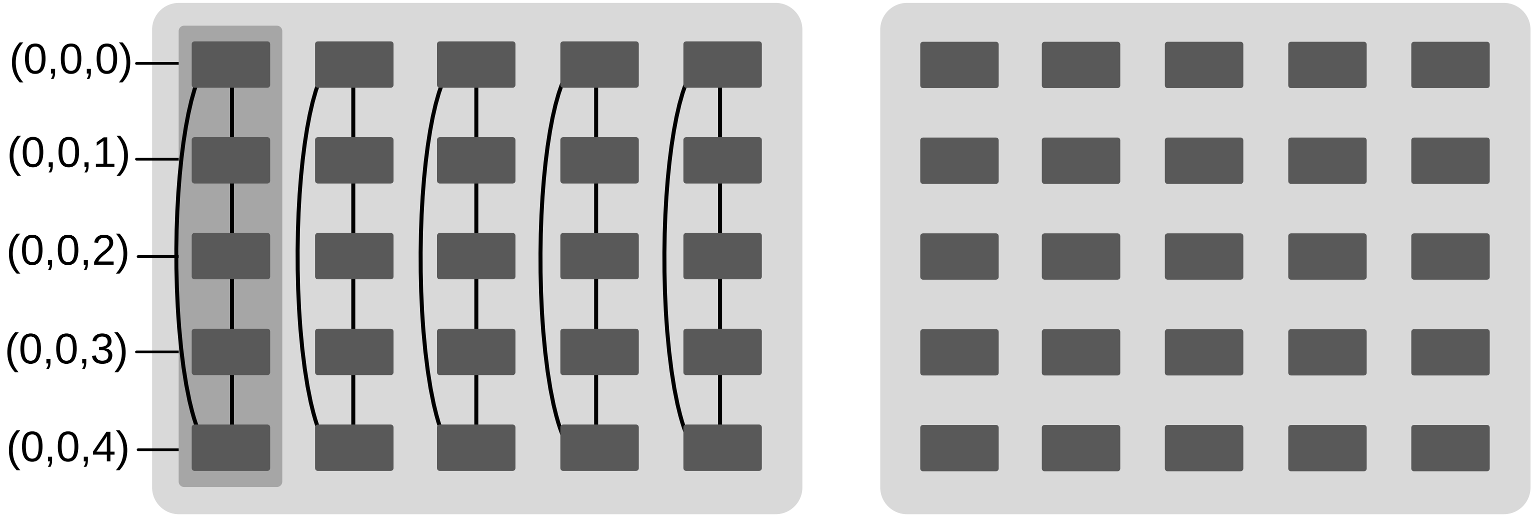
$X = \{1, \xi^2, \dots, \xi^{q-3}\}$ (for subgraph 0)

$X' = \{\xi, \xi^3, \dots, \xi^{q-2}\}$ (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:

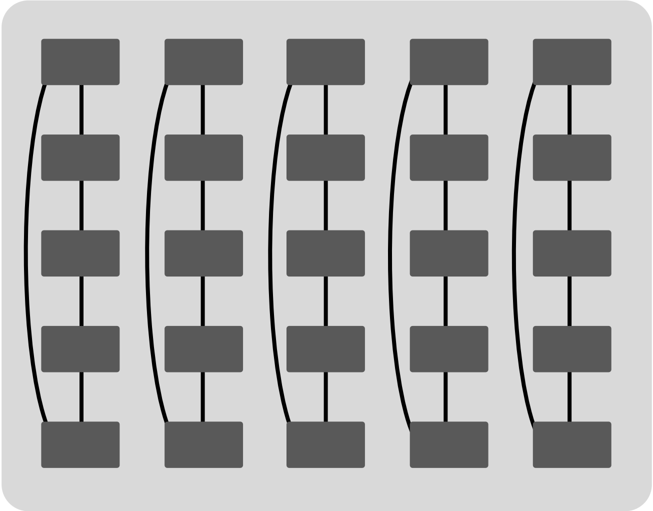
$X = \{1, \xi^2, \dots, \xi^{q-3}\}$ (for subgraph 0)

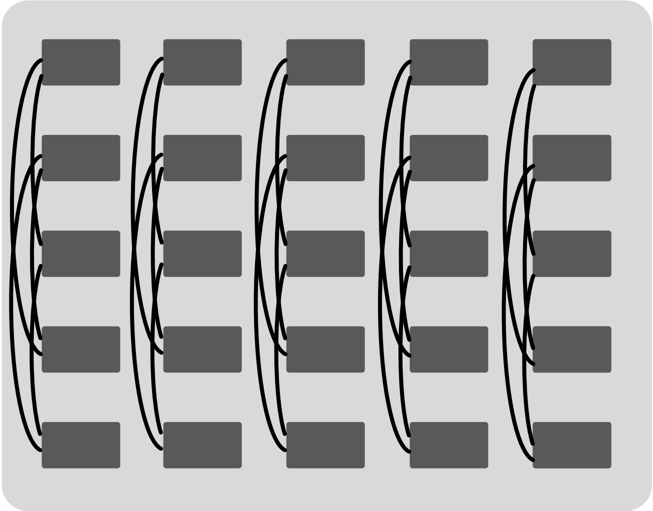
$X' = \{\xi, \xi^3, \dots, \xi^{q-2}\}$ (for subgraph 1)

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 $y = mx + c$

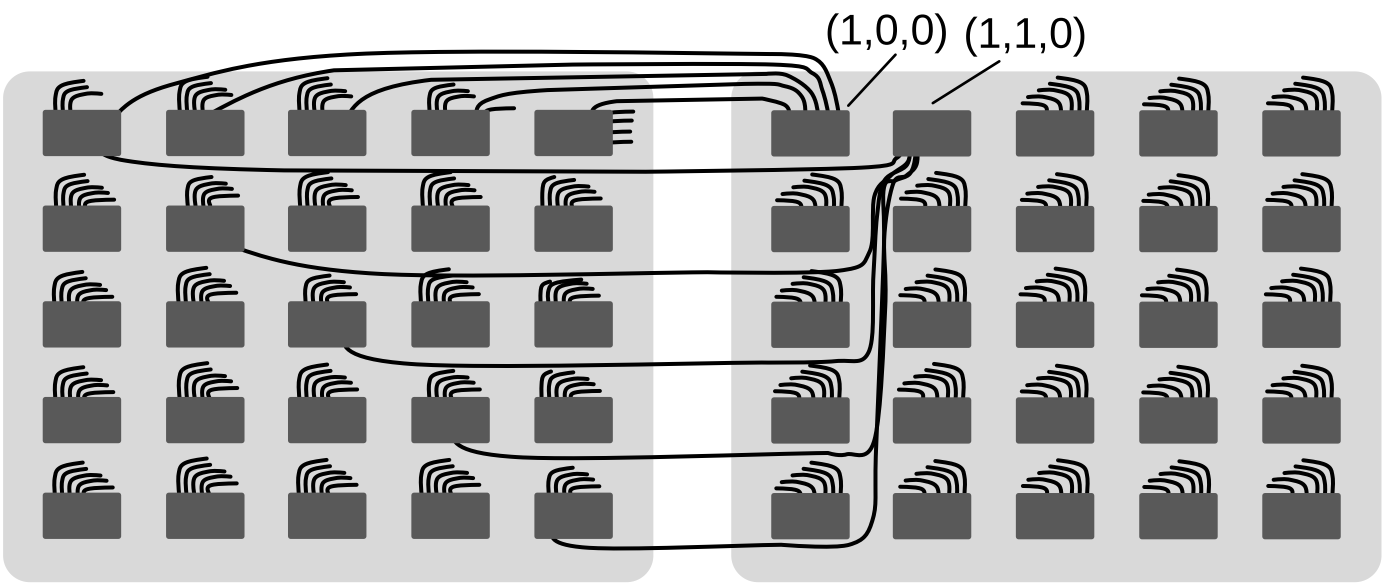
E *Example:* $q = 5$

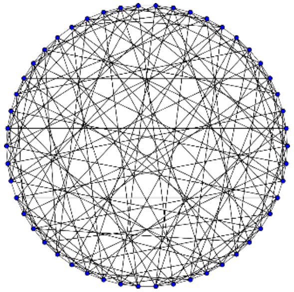
Take Router $(1,0,0)$ $m = 0, c = 0$

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

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

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

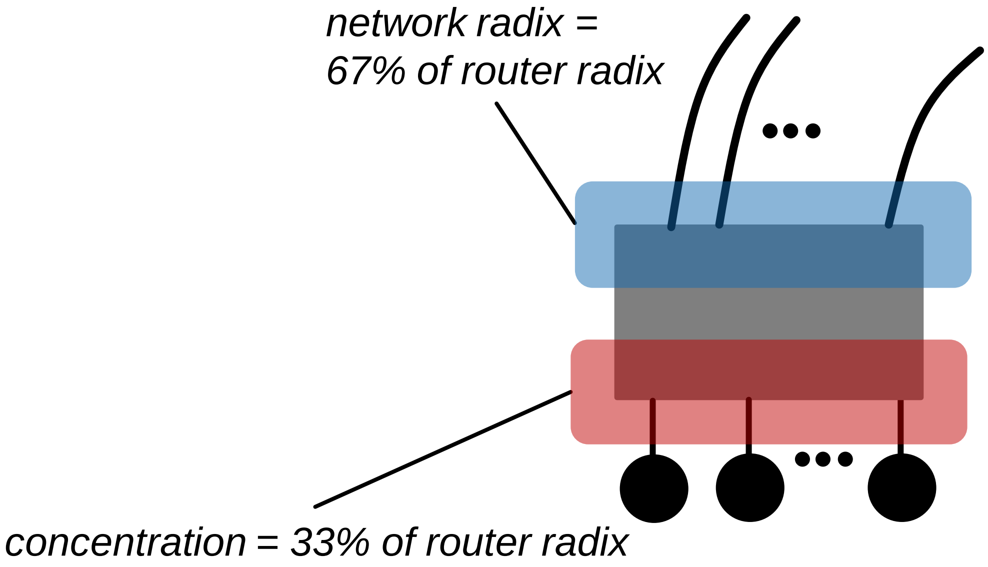




Designing An Efficient Network Topology

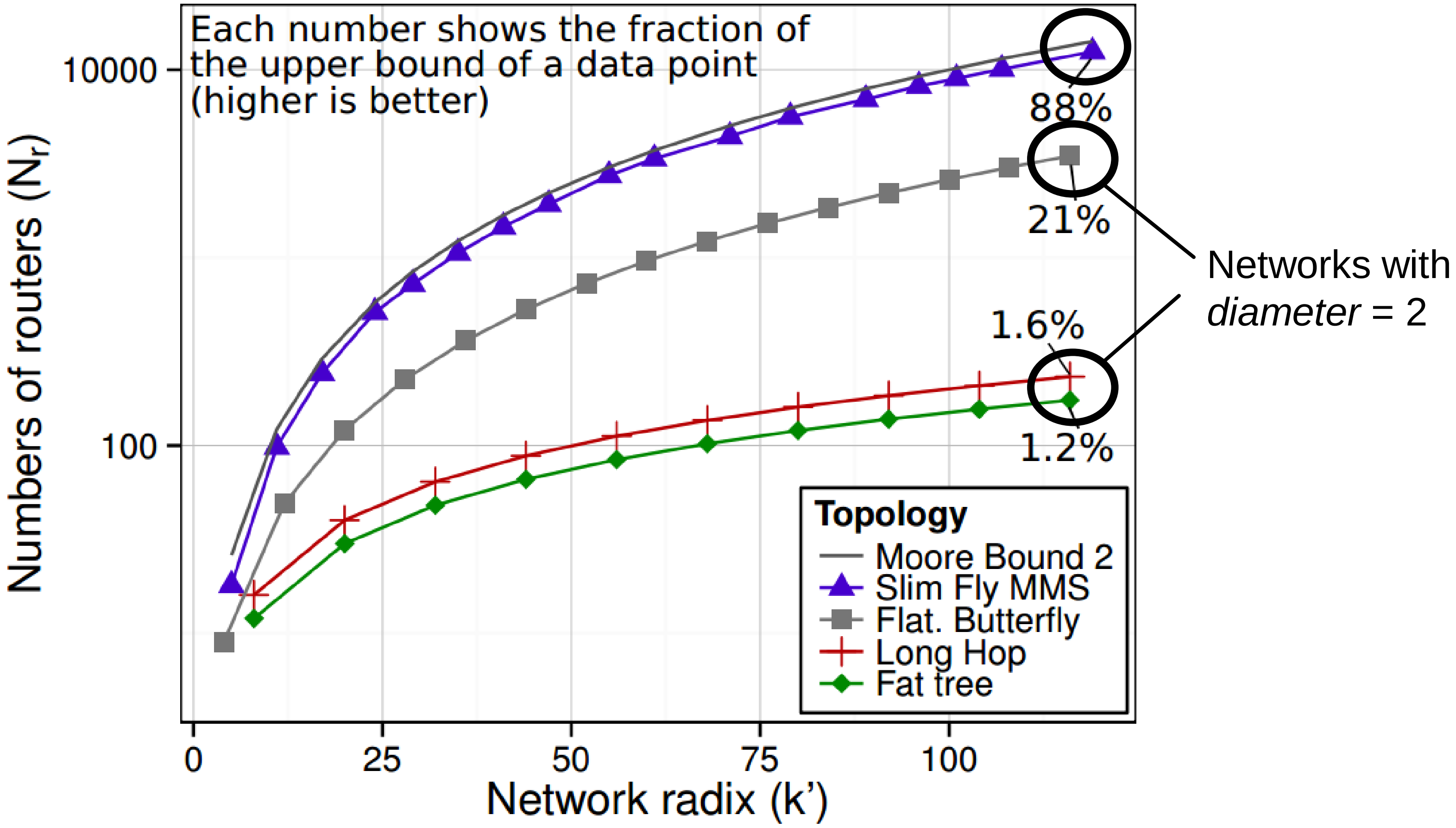
Attaching endpoints: Diameter 2

- How many endpoints do we attach to each router?
- As many to ensure *full global bandwidth:*
  - Global bandwidth: the theoretical cumulative throughput if all endpoints simultaneously communicate with all other endpoints in a steady state



COMPARISON TO OPTIMALITY

- How close is the presented Slim Fly network to the Moore Bound?



Overview of our Research

Topology design

Optimizing towards Moore Bound

Attaching endpoints

Comparison of optimality

Cost, power, resilience analysis

Physical layout

Cost model

Comparison targets

Cost & power results

Detailed case-study

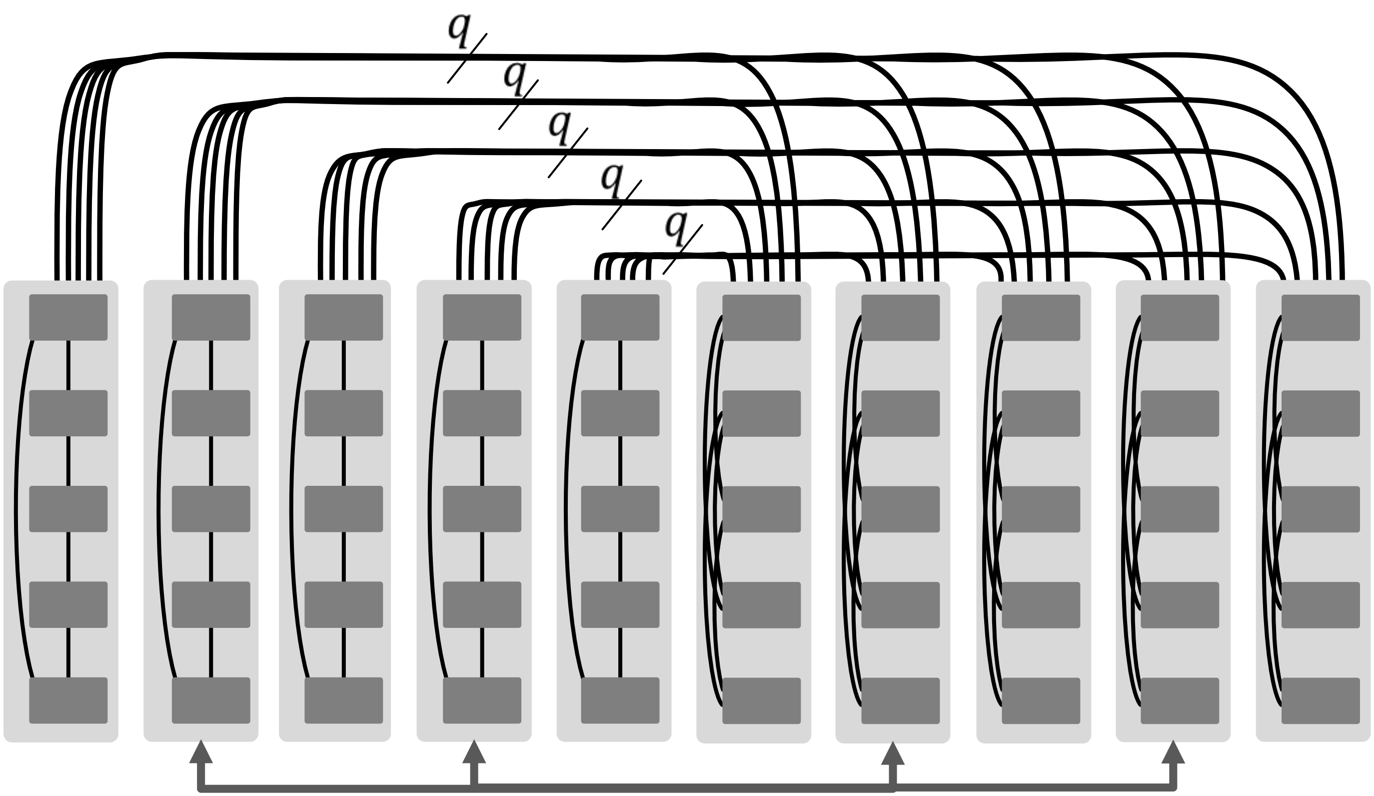
Resilience

Routing and performance

Routing

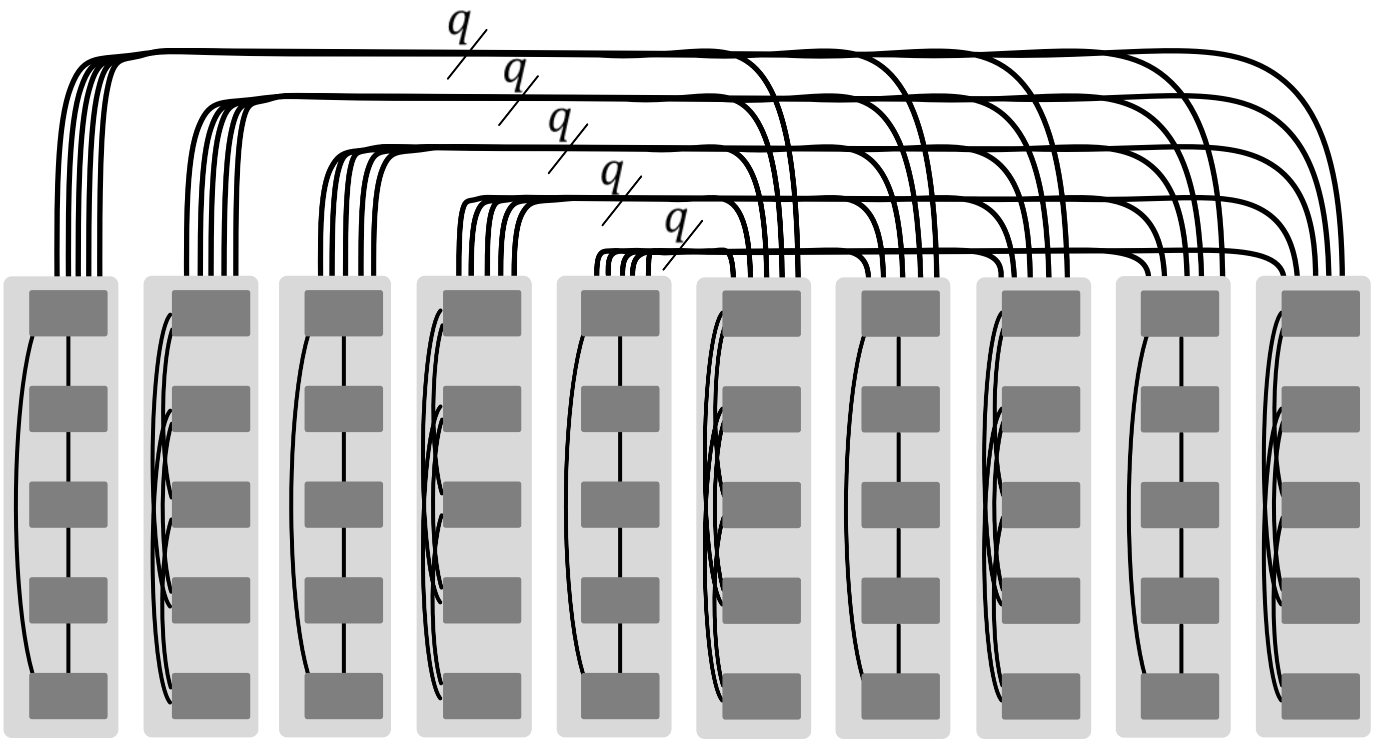
Performance, latency, bandwidth

PHYSICAL LAYOUT

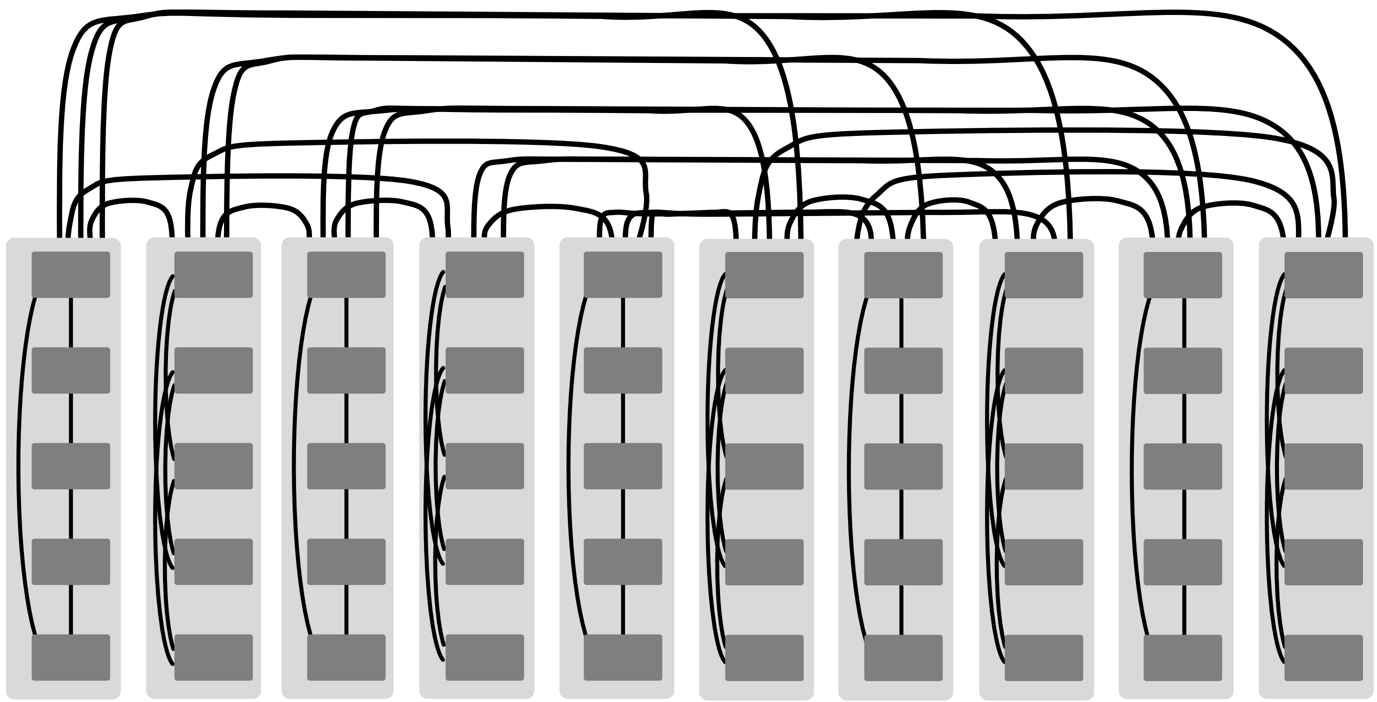


Mix (pairwise) groups
with different cabling patterns
to shorten inter-group cables

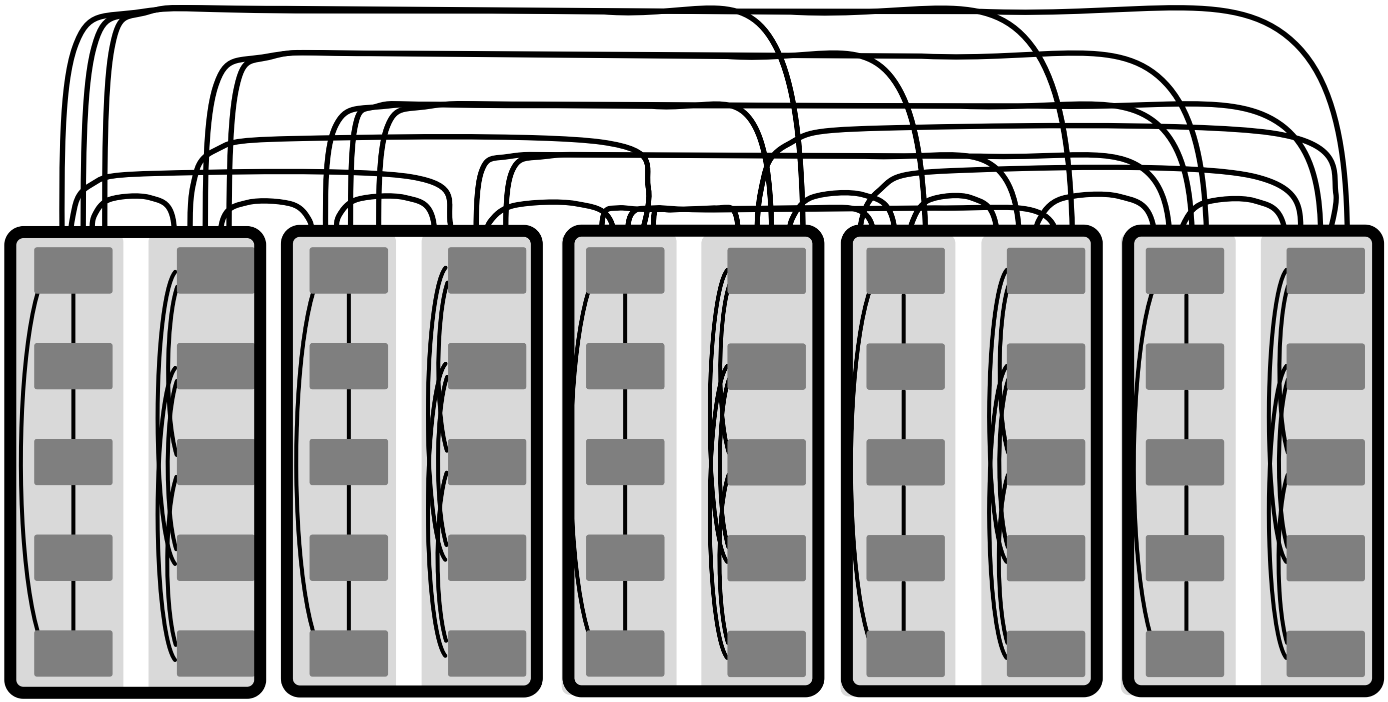
Physical Layout



PHYSICAL LAYOUT

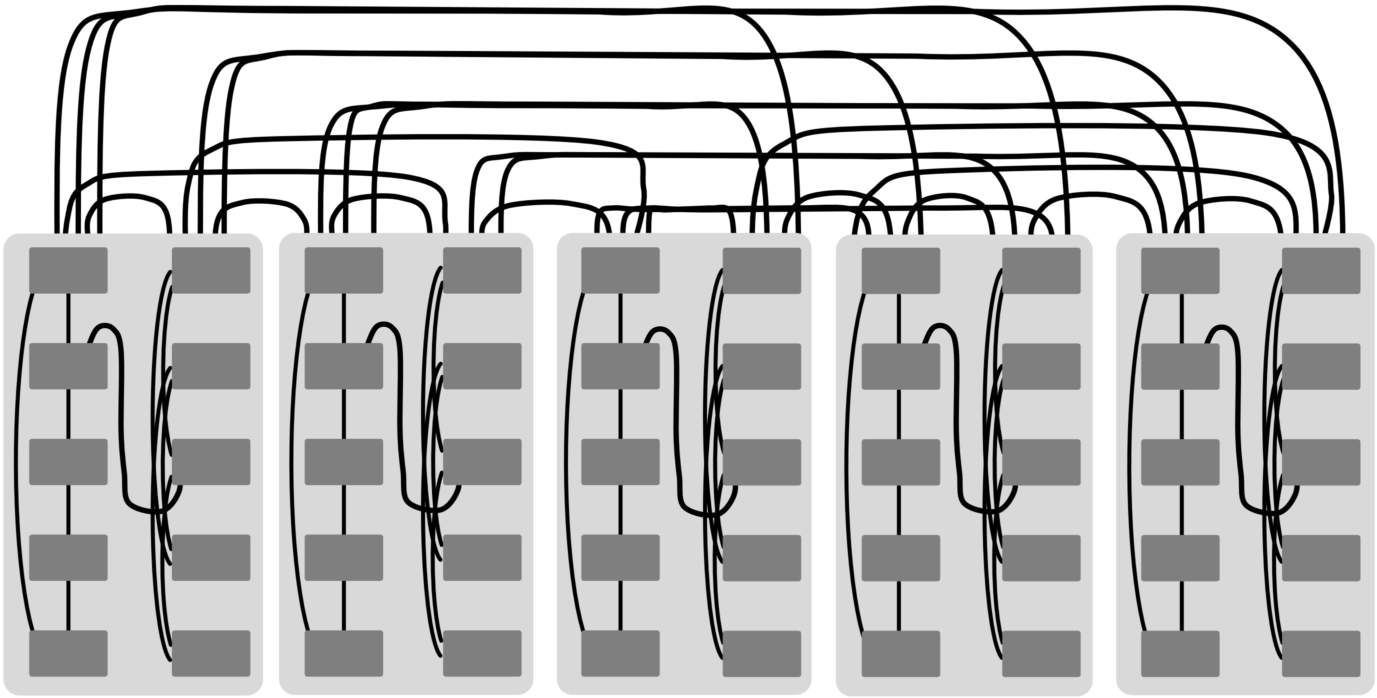


PHYSICAL LAYOUT

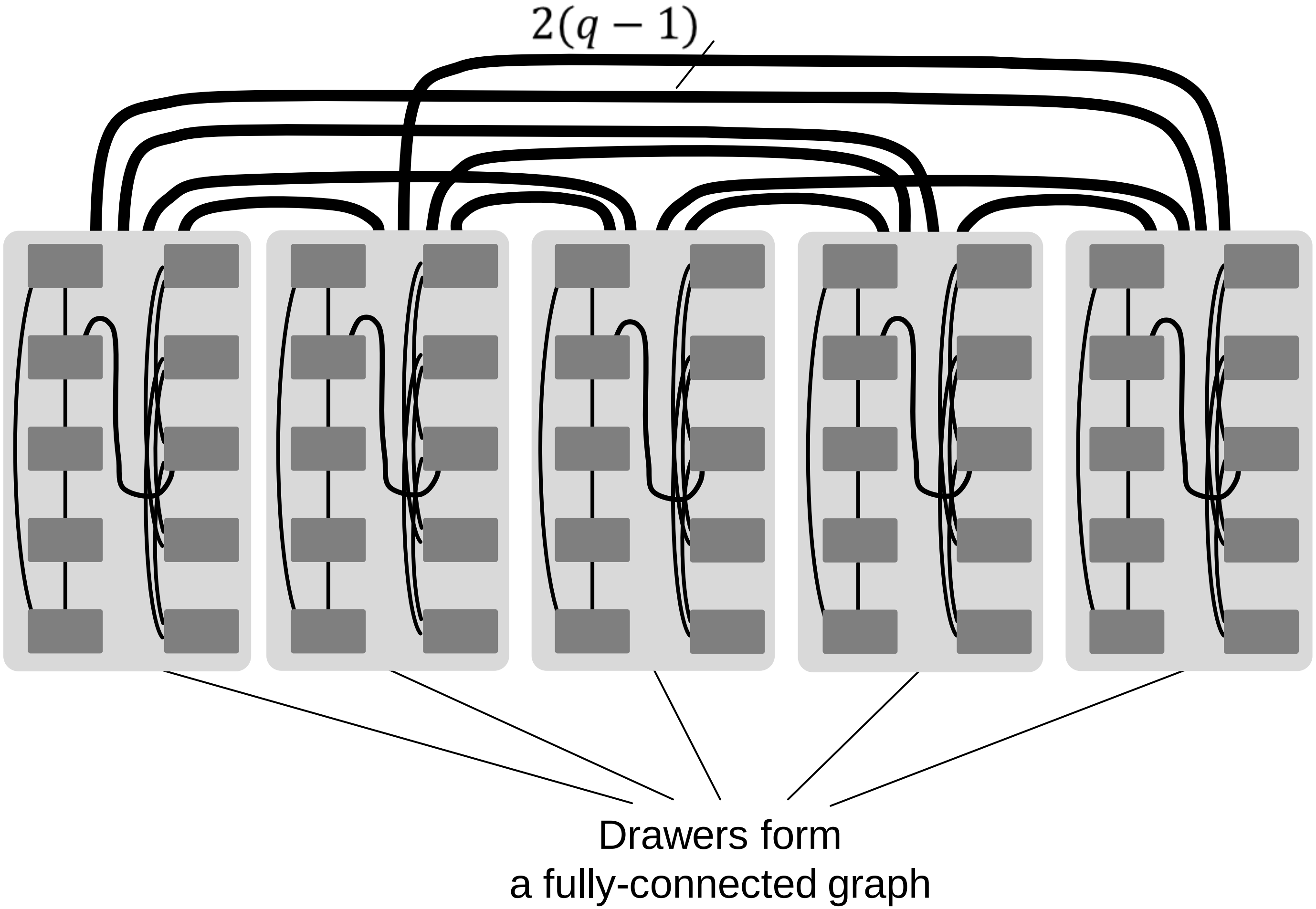


Merge groups pairwise
to create drawers

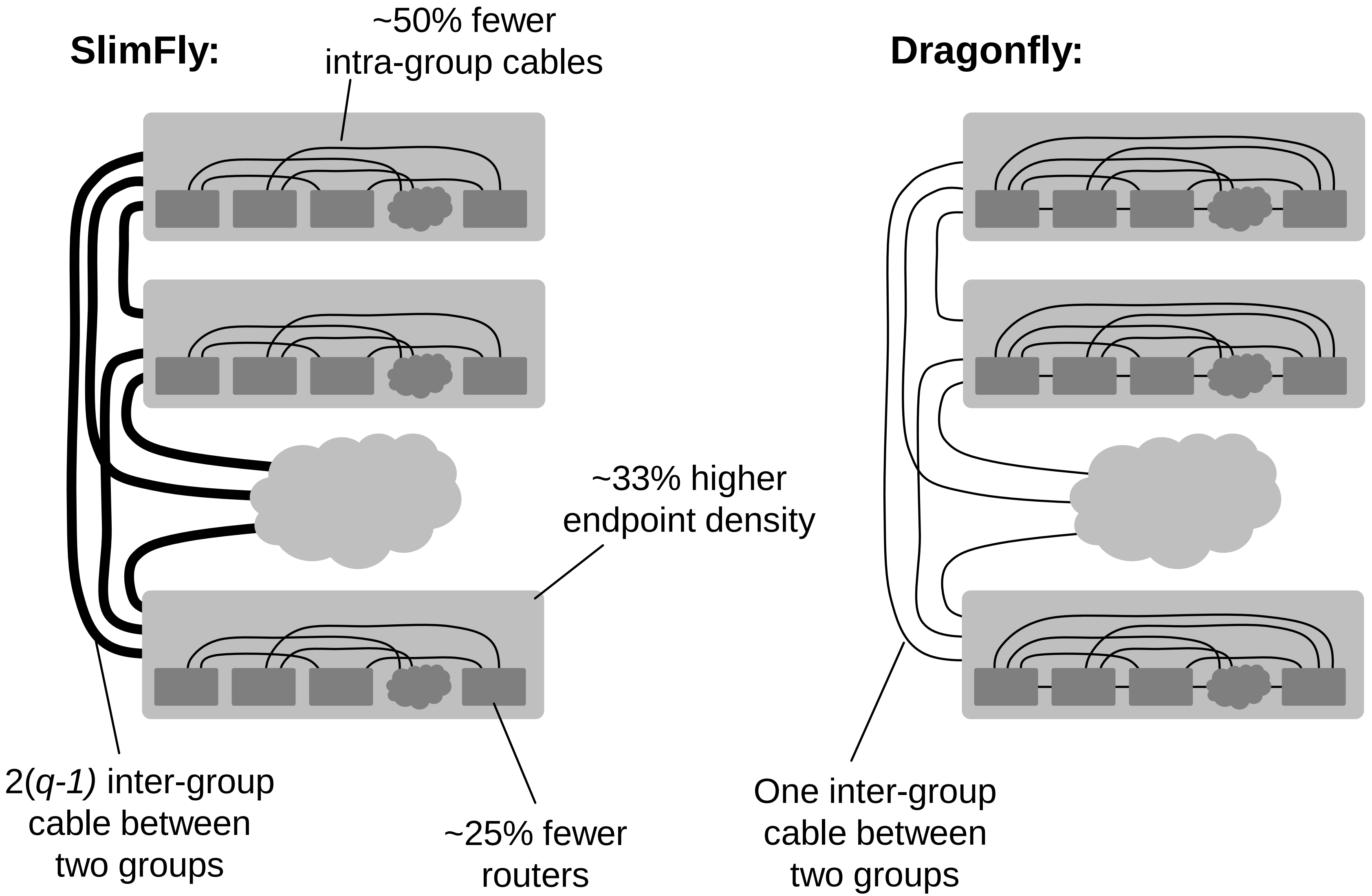
PHYSICAL LAYOUT



PHYSICAL LAYOUT



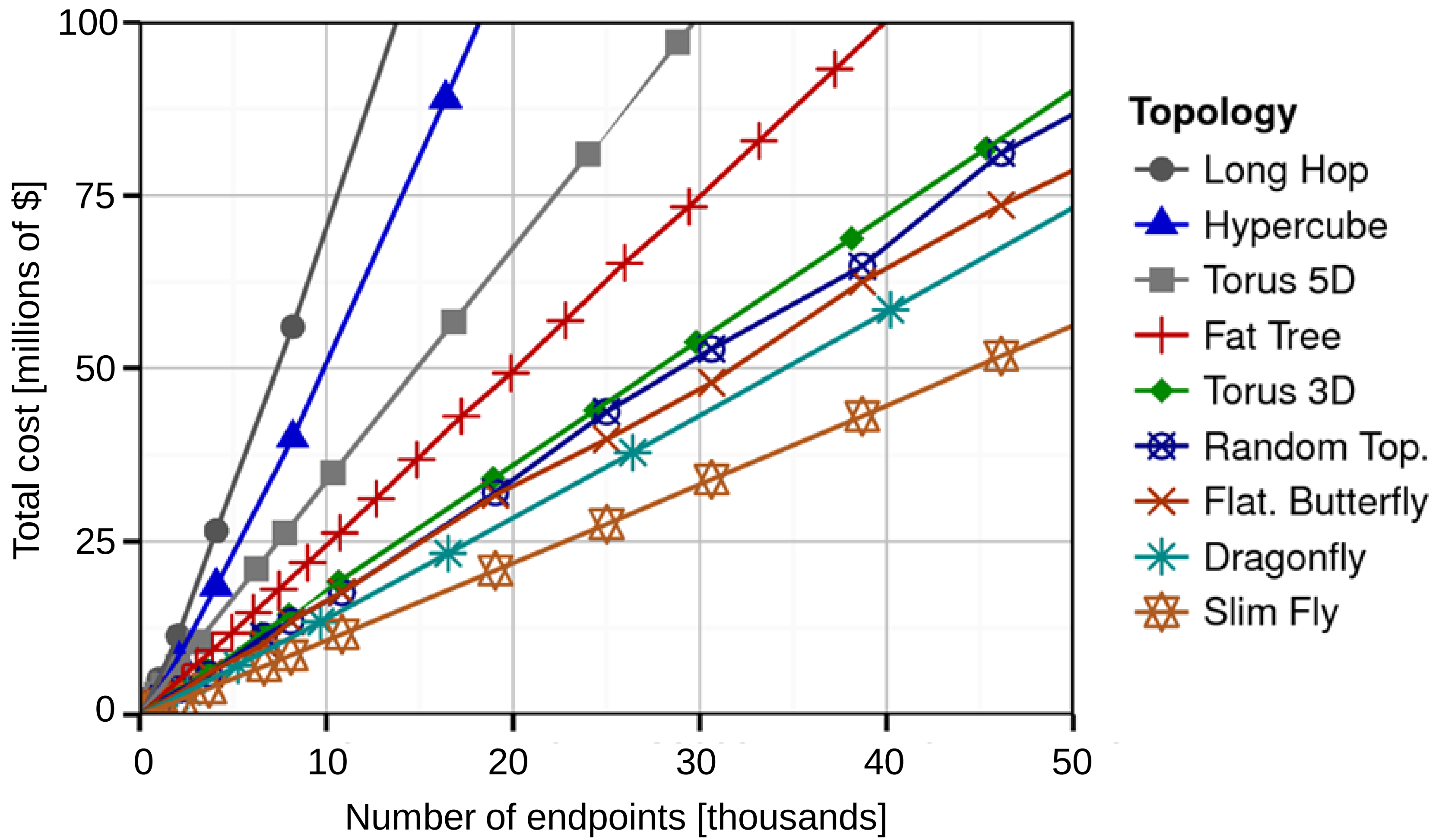
Physical Layout



Cost Comparison

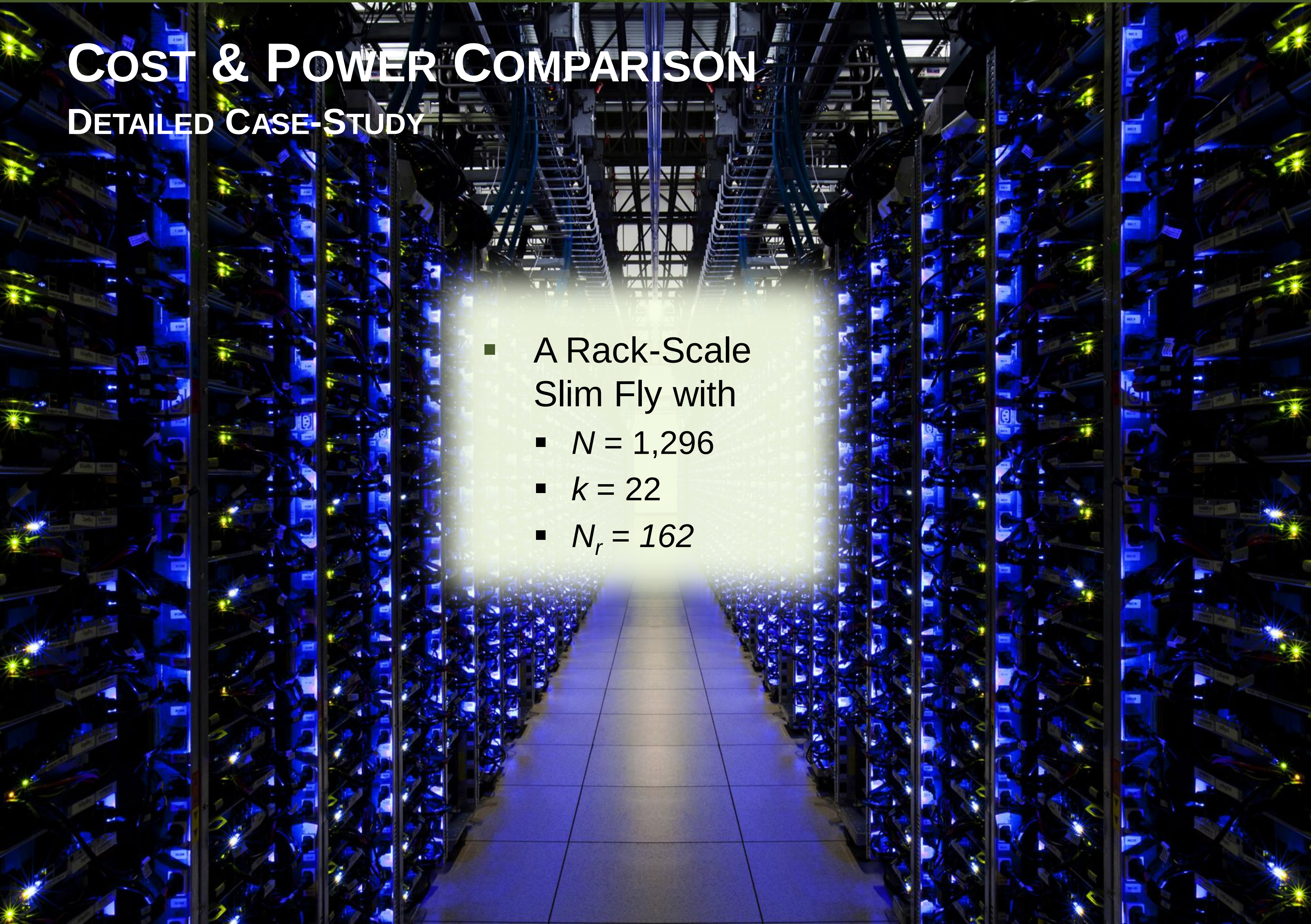
Results

Assuming COTS material costs and best known layout for each topology!



Cost & Power Comparison

Detailed Case-Study



- A Rack-Scale Slim Fly with
  - N = 1,296
  - k = 22
  - N_r = 162

Cost & Power Comparison

Detailed Case-Study: High-Radix Topologies

	Low-radix		High-radix				
Topology	3D Torus	5D Torus	Fat tree	Random	Dragfly	Dragfly	SF
Endpoints (N)	**1,200**	**1,280**	**1,024**	**1,296**	**1,056**	**1,200**	**1,296**
Routers (N_r)	1,200	1,280	320	260	264	240	**162**
Radix (k)	7	11	16	22	15	**20**	**22**
Electric cables	3,600	6,400	2,048	2,210	1,452	1800	**1134**
Cost per node [$]	1,802	3,364	1,634	1,504	1,201	1,343	**922**
Power per node [W]	19.6	30.8	14.0	12.35	10.50	11.20	**7.70**

	Low-radix		High-radix				
Topology	3D Torus	5D Torus	Fat tree	Random	Dragfly	Dfly	SF
Endpoints (N)	**216**	**243**	**250**	**250**	**342**	**270**	**250**
Routers (N_r)	216	243	125	84	114	90	**50**
Radix (k)	7	11	10	13	11	**12**	**13**
Electric cables	648	1,215	500	419	456	405	**200**
Cost per node [$]	1,802	3,364	1,466	1,366	1,094	1,224	**797**
Power per node [W]	19.6	30.8	14.0	12.23	10.26	11.20	**7.28**

Overview of our Research

Topology design

Optimizing towards Moore Bound

Attaching endpoints

Comparison of optimality

Cost, power, resilience analysis

Physical layout

Cost model

Comparison targets

Cost & power results

Detailed case-study

Resilience

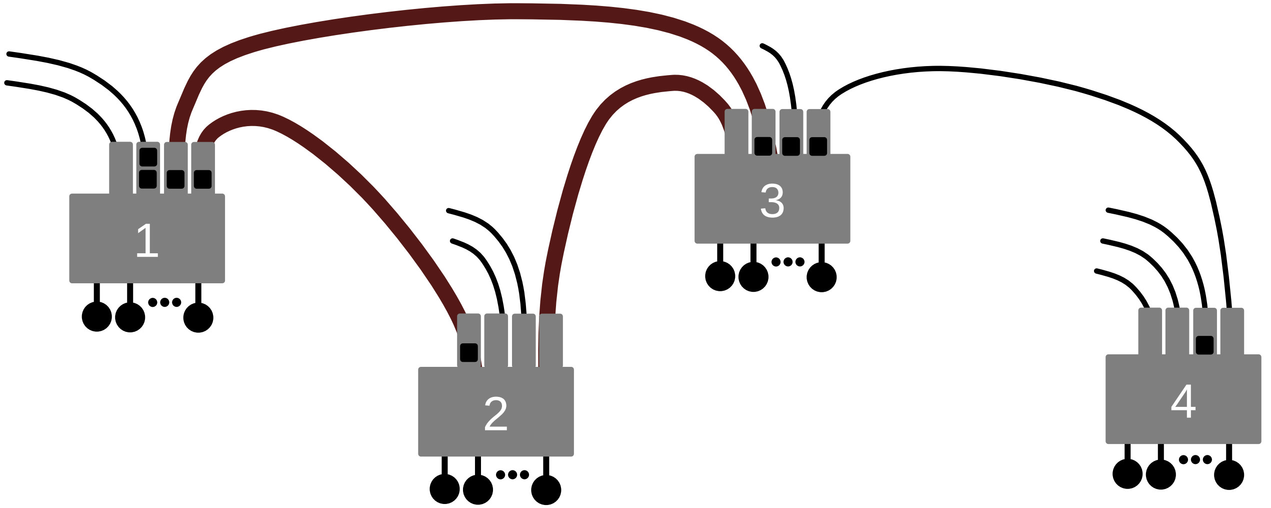
Routing and performance

Routing

Performance, latency, bandwidth

PERFORMANCE & ROUTING

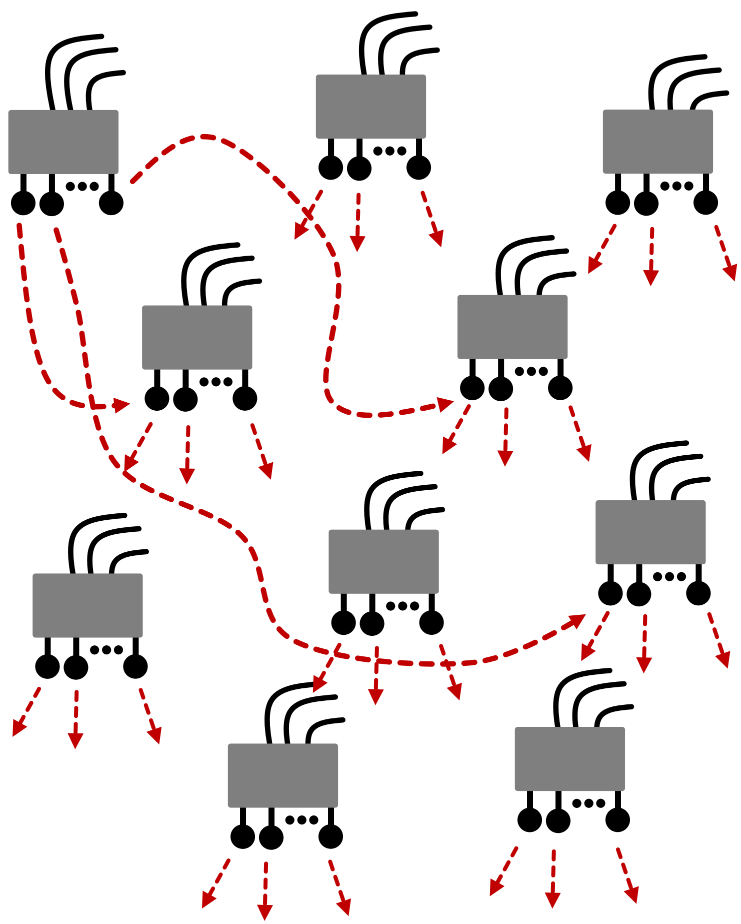
- Cycle-accurate simulations [1]
- Routing protocols:
  - Minimum static routing
  - Valiant routing [2]
  - Universal Globally-Adaptive Load-Balancing routing [3]
    UGAL-L: each router has access to its local output queues
    UGAL-G: each router has access to the sizes of all router queues in the network

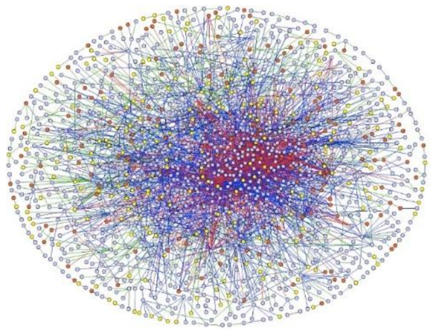


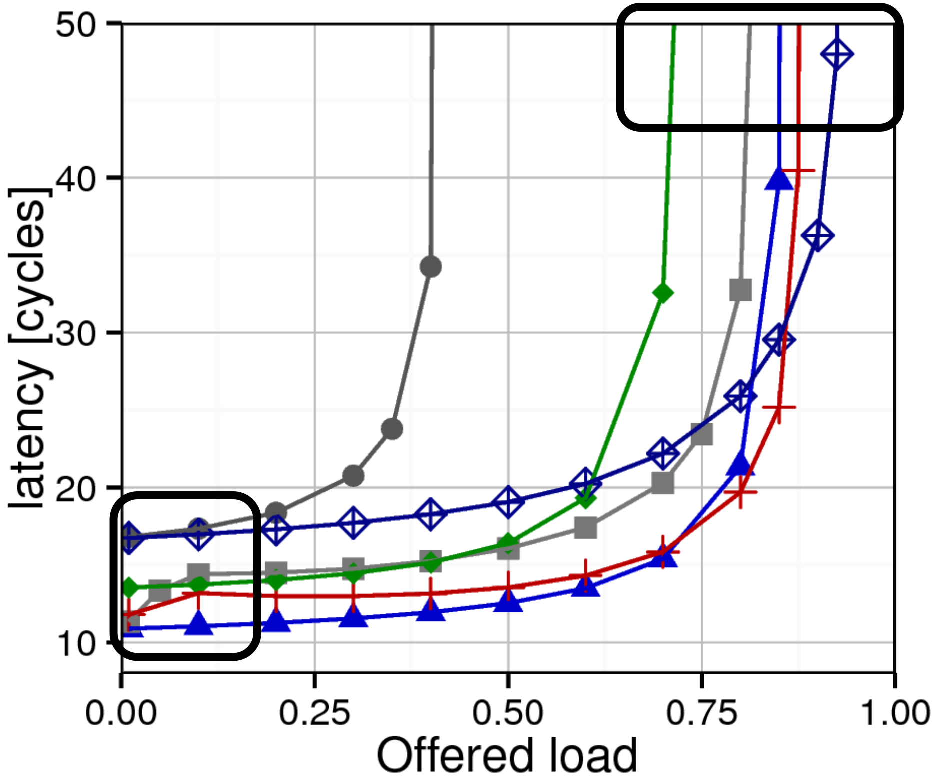
[1] N. Jiang et al. A detailed and flexible cycle-accurate Network-on-Chip simulator. ISPASS'13
[2] L. Valiant. A scheme for fast parallel communication. SIAM journal on computing, 1982
[3] A. Singh. Load-Balanced Routing in Interconnection Networks. PhD thesis, Stanford University, 2005

PERFORMANCE & ROUTING

RANDOM UNIFORM TRAFFIC



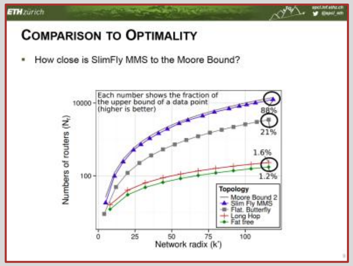




SUMMARY

Topology design

Optimizing towards the Moore Bound reduces expensive network resources

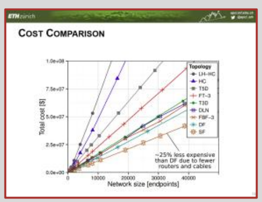


Credits

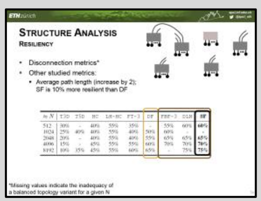
Maciej Besta
(PhD Student @SPCL)

Advantages of SlimFly

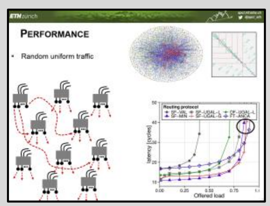
Cost & power



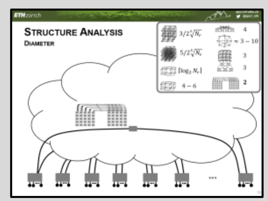
Resilience



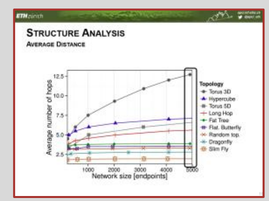
Performance



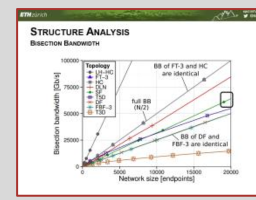
Diameter



Avg. distance

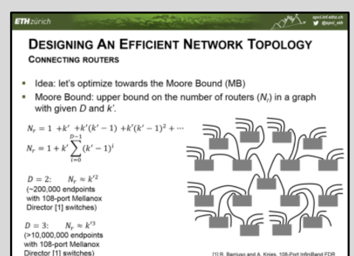


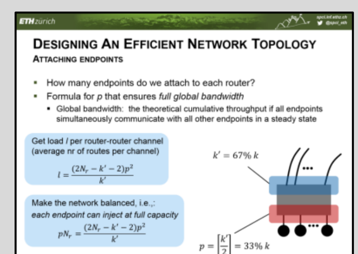
Bandwidth



Optimization approach

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

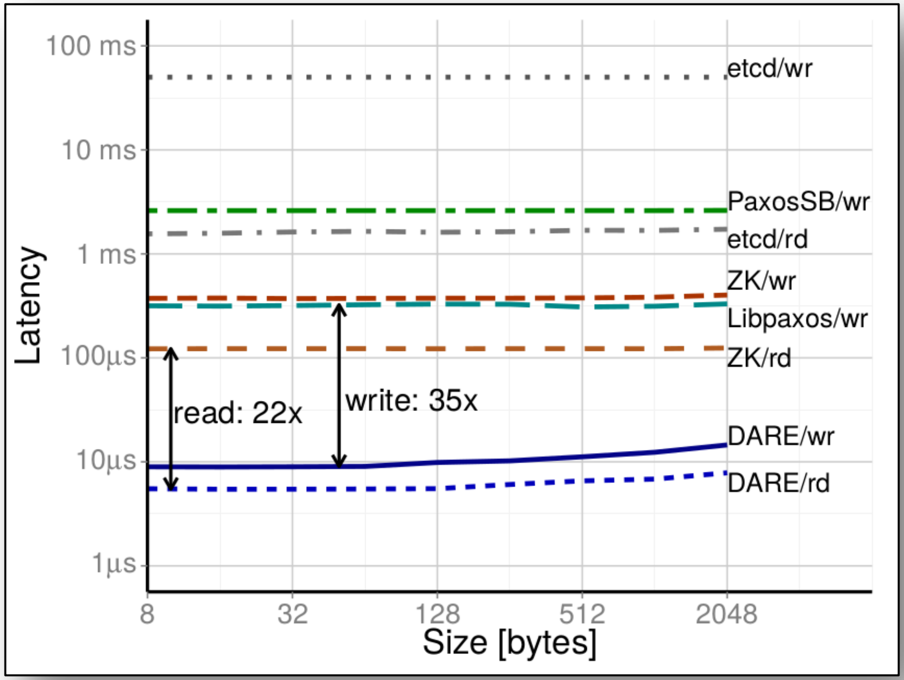




M. Besta, TH: "Slim Fly: A Cost Effective Low-Diameter Network Topology", SC15

Related projects at SPCL@ETH

- **DARE - Fast RDMA replicated state machines [1]**
  - Access latency: 6/9 us
    (22-35x faster than Zookeeper)
  - Request throughput : 720/460kreq/s
    (1.7x faster than Zookeeper)
  - Available within 30ms of leader crash
    no interruption for server failure
  - All strongly consistent (linearizable)



- **HTM for distributed memory graph analytics [2]**
  - Accelerates Graph500 & Galois by 10-50%, beats Hama by 100-1000x

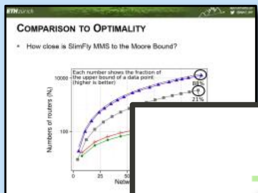
- **Ethernet routing for low-diameter topologies [in progress]**
  - Make Slim Fly practical in Ethernet settings

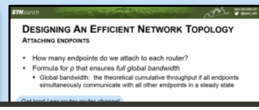
[1]: M. Poke, TH: "DARE: High-Performance State Machine Replication on RDMA Networks", HPDC'15
[2]: M. Besta, TH: "Accelerating Irregular Computations with Hardware Transactional Memory and Active Messages", HPDC'15

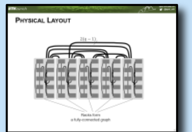
A LOWEST-DIAMETER TOPOLOGY

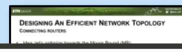
→ Viable set of configurations

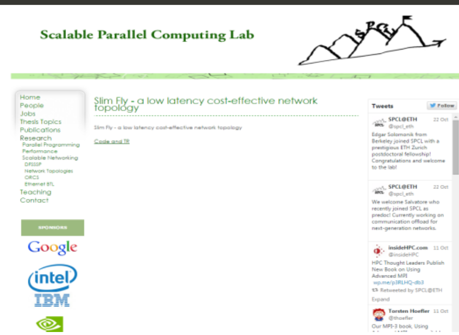
→ Resilient









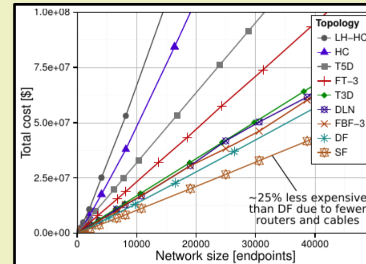


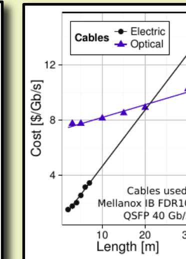
http://spcl.inf.ethz.ch/Research/Scalable_Networking/SlimFly

A COST & POWER EFFECTIVE TOPOLOGY

→ 25% less expensive than Dragonfly,

→ 26% less power-hungry than Dragonfly





AWAY M

Thank you for your attention

A HIGH-PERFORMANCE TOPOLOGY

→ Lowest latency

→ Full global bandwidth