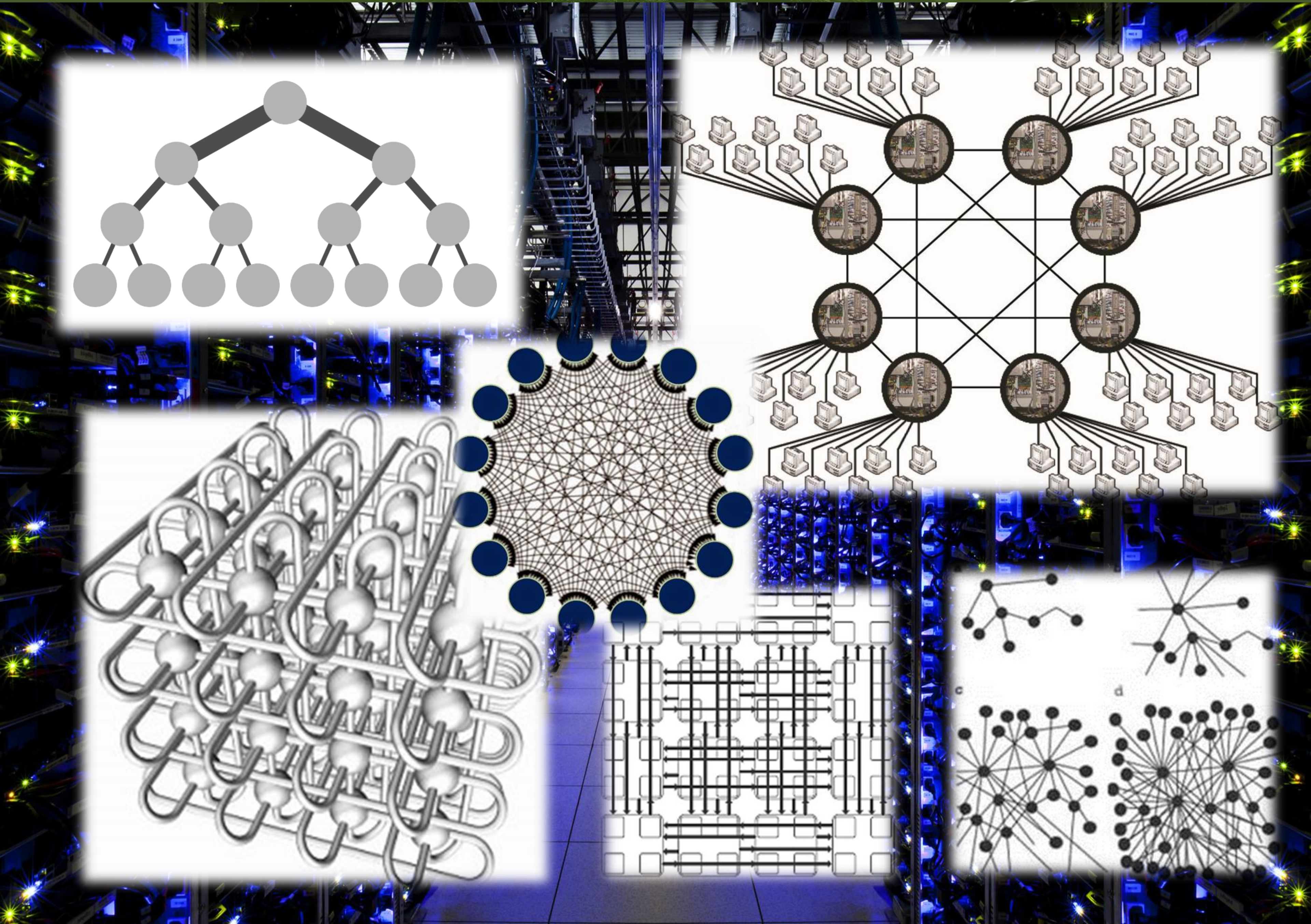
SlimFly: A Cost Effective Low-Diameter Network Topology

MACIEJ BESTA, TORSTEN HOEFLER


c
d

Designing An Efficient Network Topology

- Goals:
  - Decrease network cost & power consumption
  - Preserve high bandwidth

- How can the cost/power consumption be reduced?

  By lowering diameter!

- Intuition: lower diameter means:

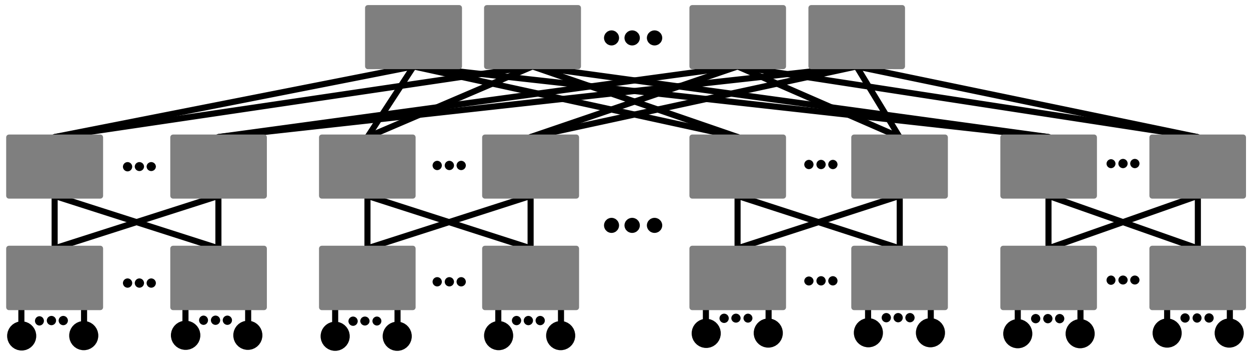
Fewer router buffers and thus SerDes (Serializers/Deserializers) traversed
➔ reduces power consumption

Lower average path length
➔ reduces the number of necessary cables and routers

Designing An Efficient Network Topology

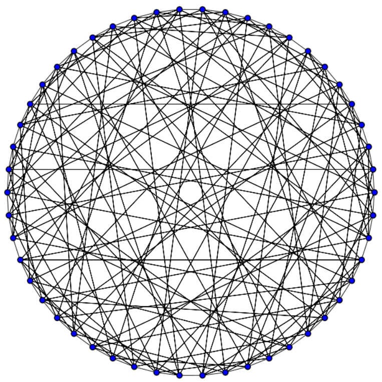
Example: full-bandwidth fat tree vs Hoffman-Singleton graph

3-level fat tree:



diameter = 4

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

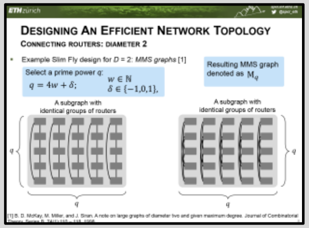


diameter = 2
> ~50% fewer routers
> ~30% fewer cables

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

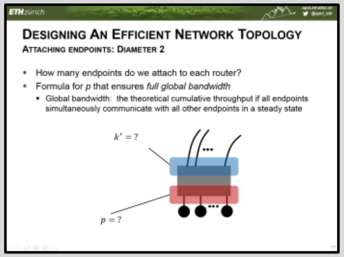
Overview of our Research

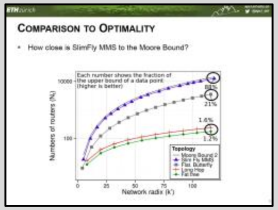
Topology design



Optimizing towards Moore Bound

Attaching endpoints





Comparison of optimality

Cost, power, resilience analysis

Physical layout

Cost model

Cost & power results

Detailed case-study

Resilience

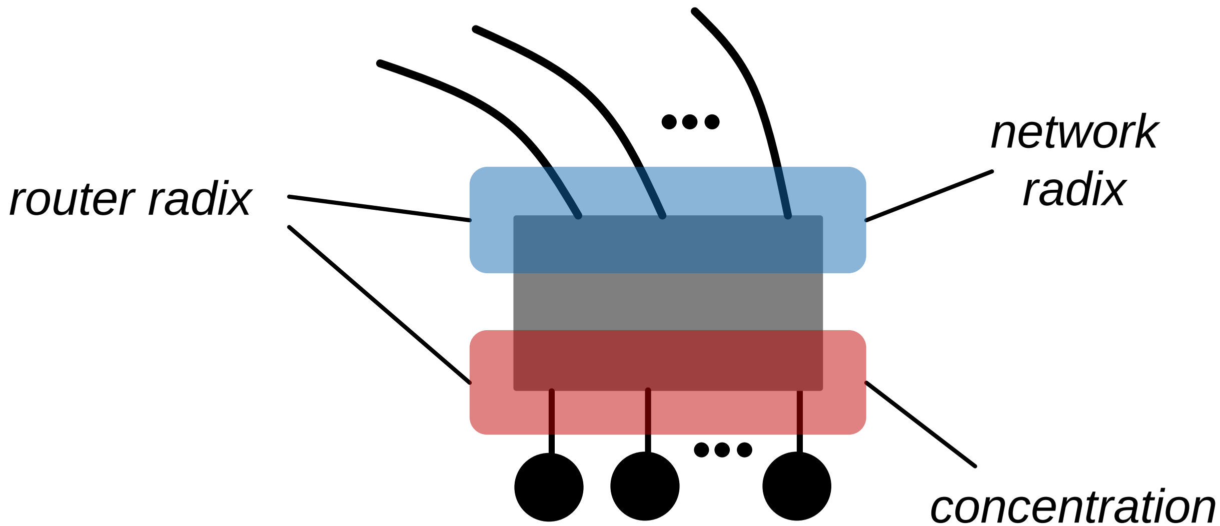
Routing and performance

Routing

Performance, latency, bandwidth

Designing An Efficient Network Topology

TERMINOLOGY



Designing An Efficient Network Topology

General construction scheme

- We establish a general construction approach with two phases:

Connect routers:

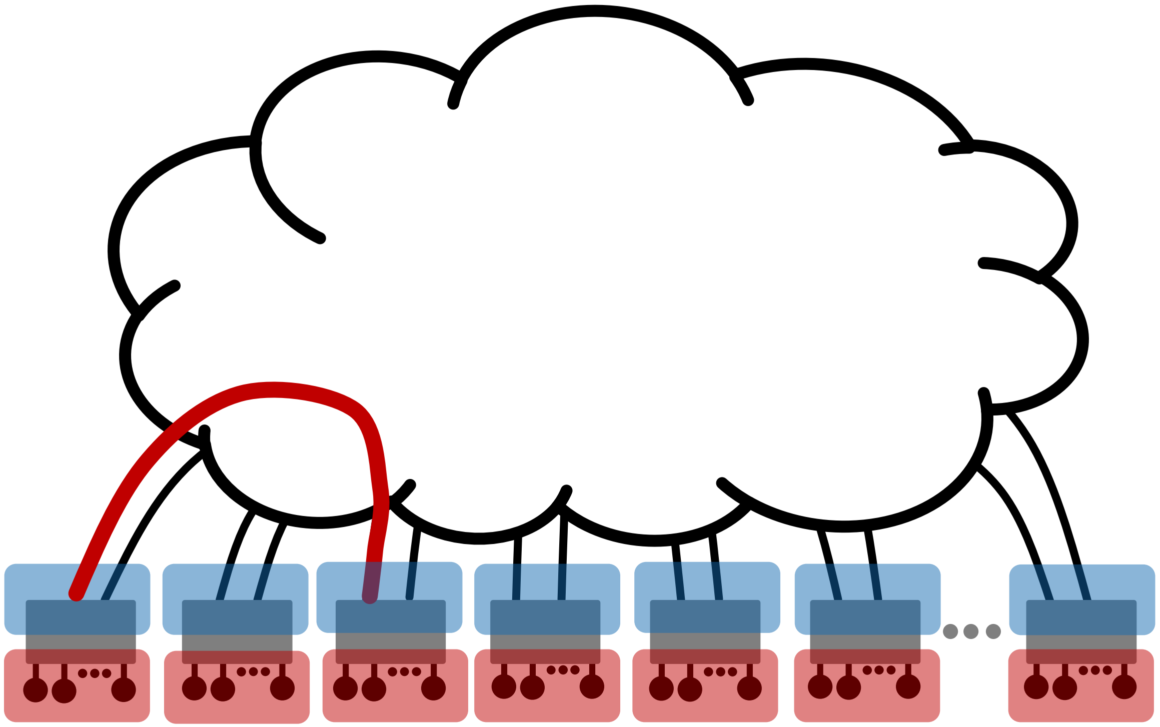
select *diameter*

select *network radix*

maximize *number of routers*

Attach endpoints

Derive *concentration* that provides full global bandwidth



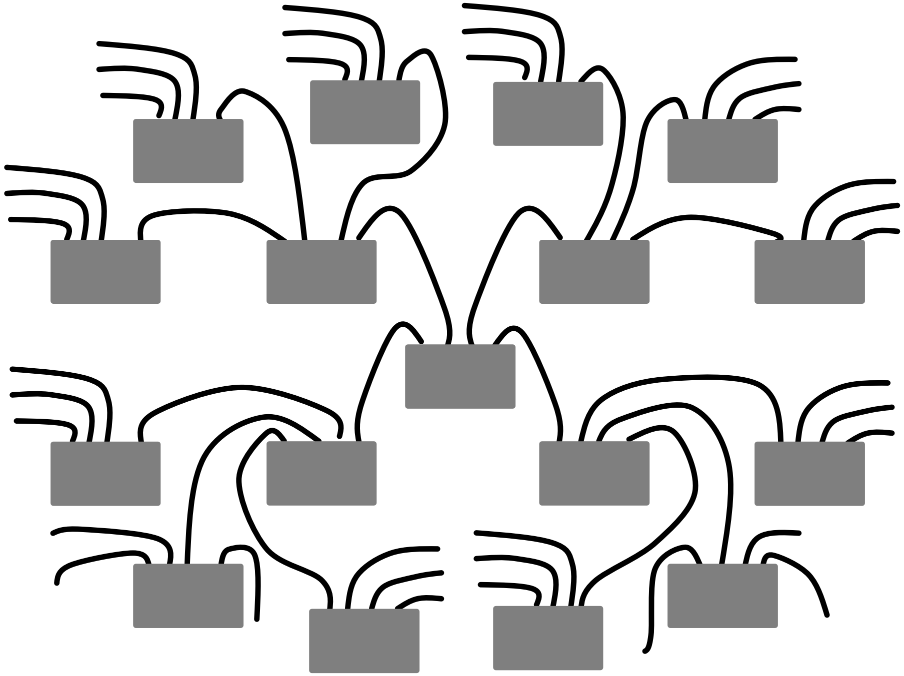
Designing An Efficient Network Topology

Connecting routers

- Idea: optimize towards the Moore Bound (MB)
- Moore Bound [1]: upper bound on the *number of routers* in a graph with given *diameter* (*D*) and *network radix* (*k*).

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

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



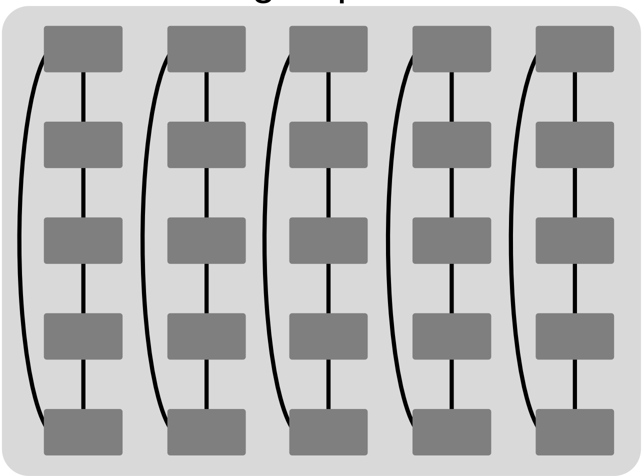
[1] M. Miller, J. Siráň. Moore graphs and beyond: A survey of the degree/diameter problem, Electronic Journal of Combinatorics, 2005.

Designing An Efficient Network Topology

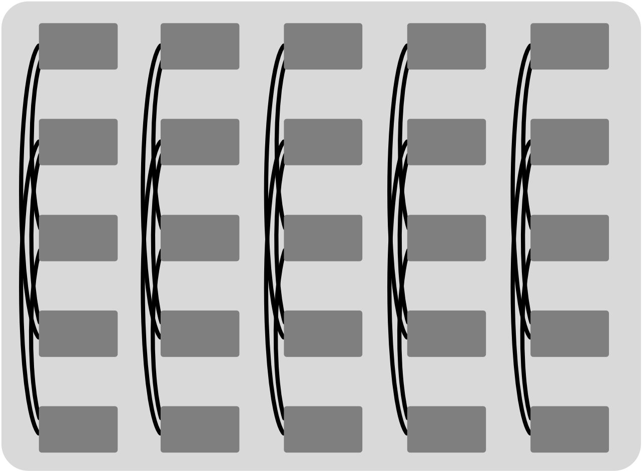
Connecting routers: Diameter 2

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

A subgraph with
identical groups of routers



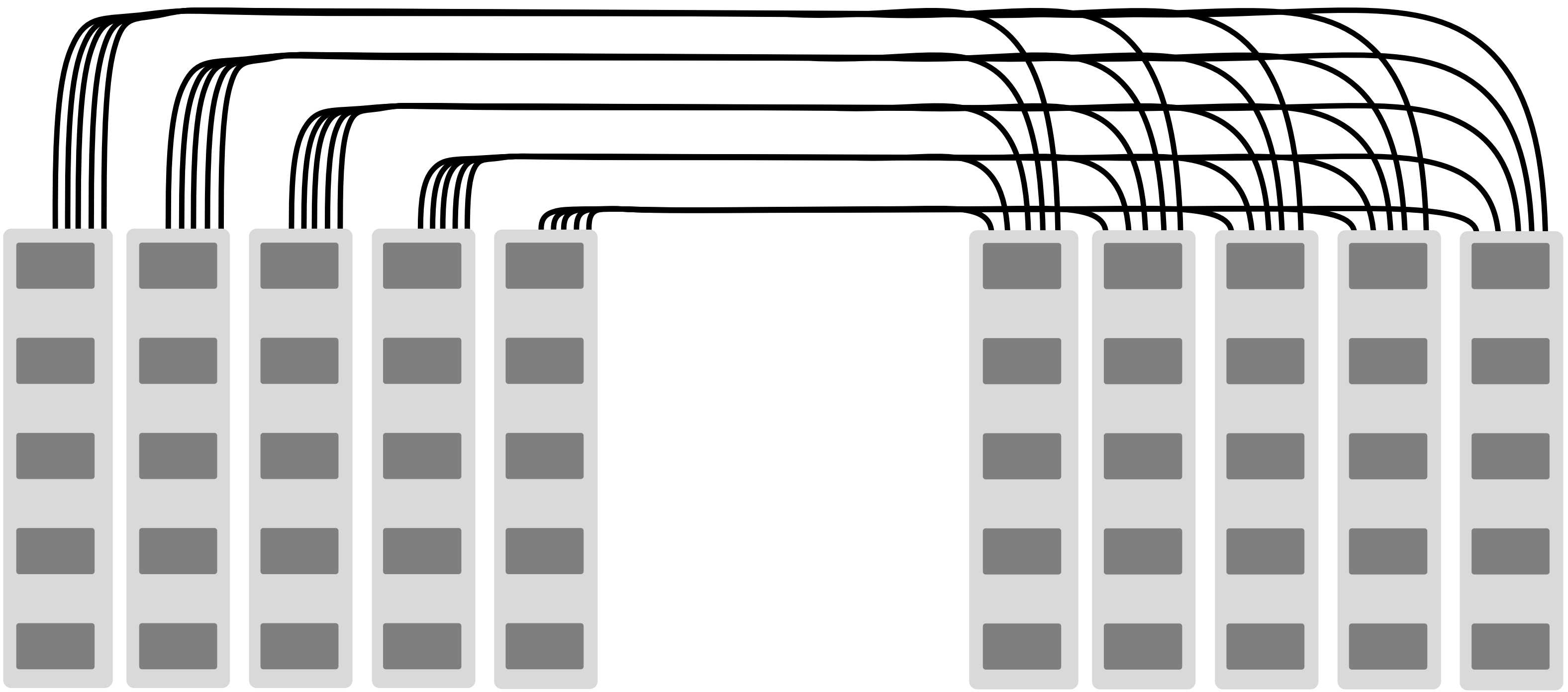
A subgraph with
identical groups of routers



[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:

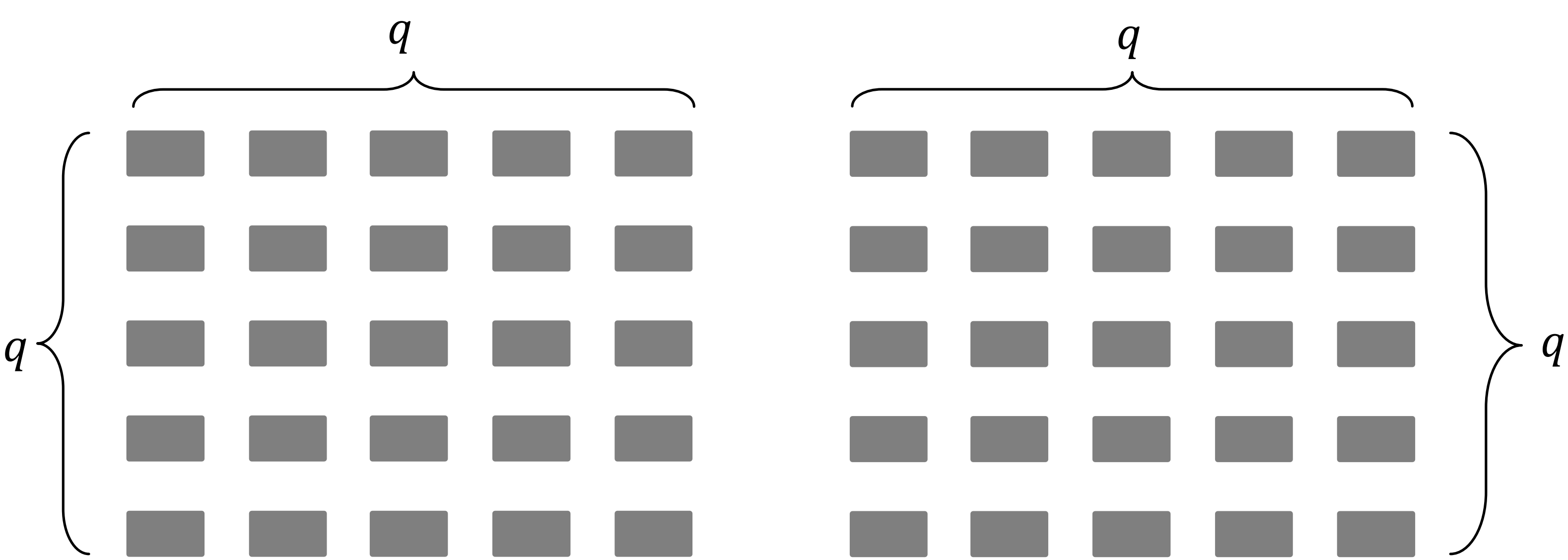
$$\mathcal{F}_q = \mathbb{Z}/q\mathbb{Z} = \{0,1,\dots,q-1\}$$

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

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:

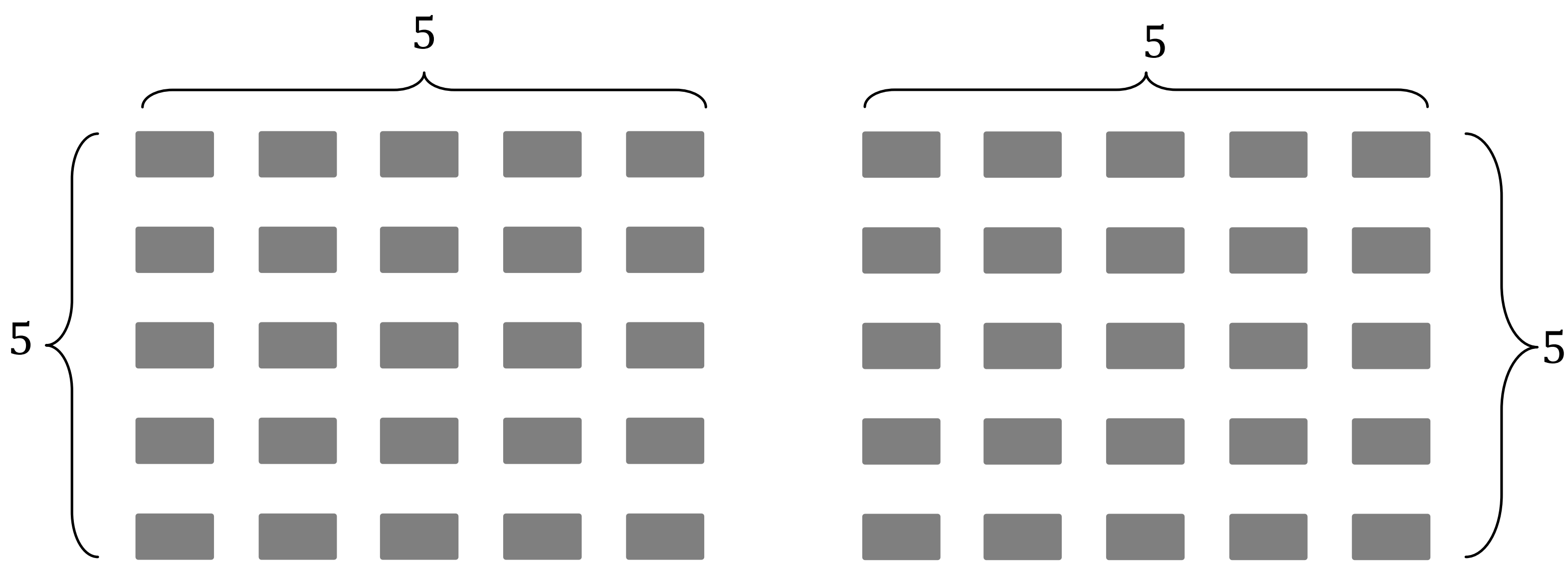
$$\mathcal{F}_q = \mathbb{Z}/q\mathbb{Z} = \{0,1,\dots,q-1\}$$

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

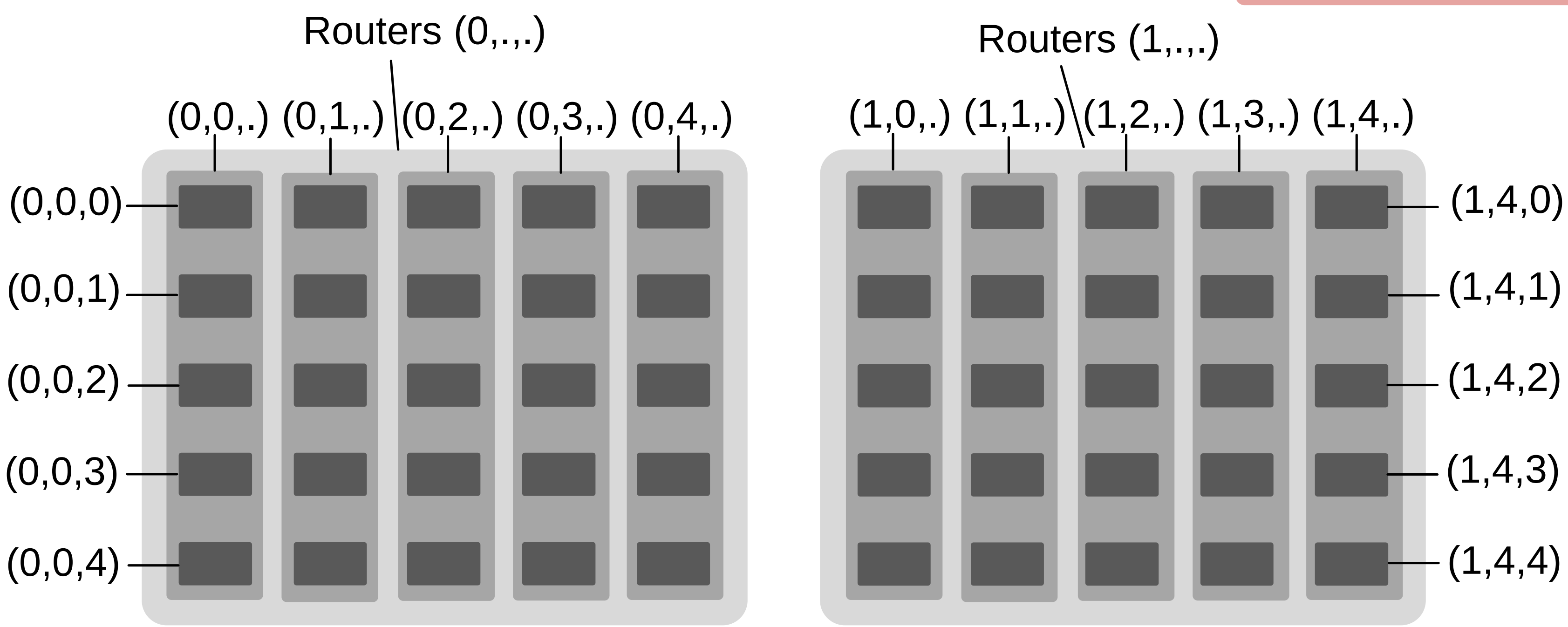
3 *Label the routers*

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

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: $q = 5$*

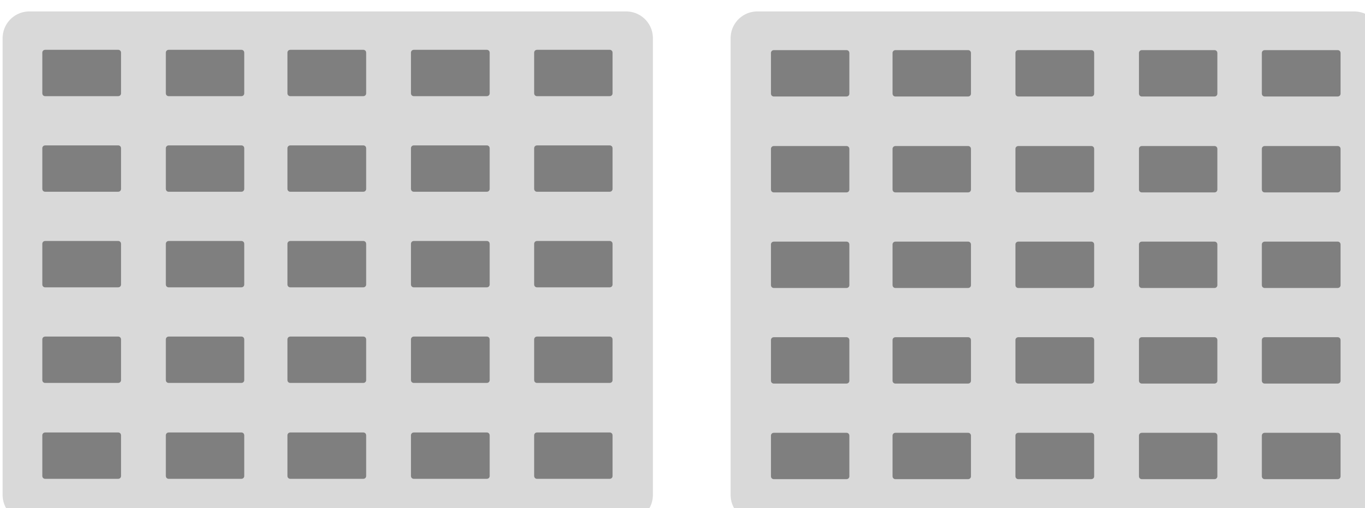
$$\mathcal{F}_5 = \{0,1,2,3,4\}$$

$$\xi = 2$$

$$1 = \xi^4 \mathit{mod}\ 5 = 2^4 \mathit{mod}\ 5 = 16\ \mathit{mod}\ 5$$

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

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



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:

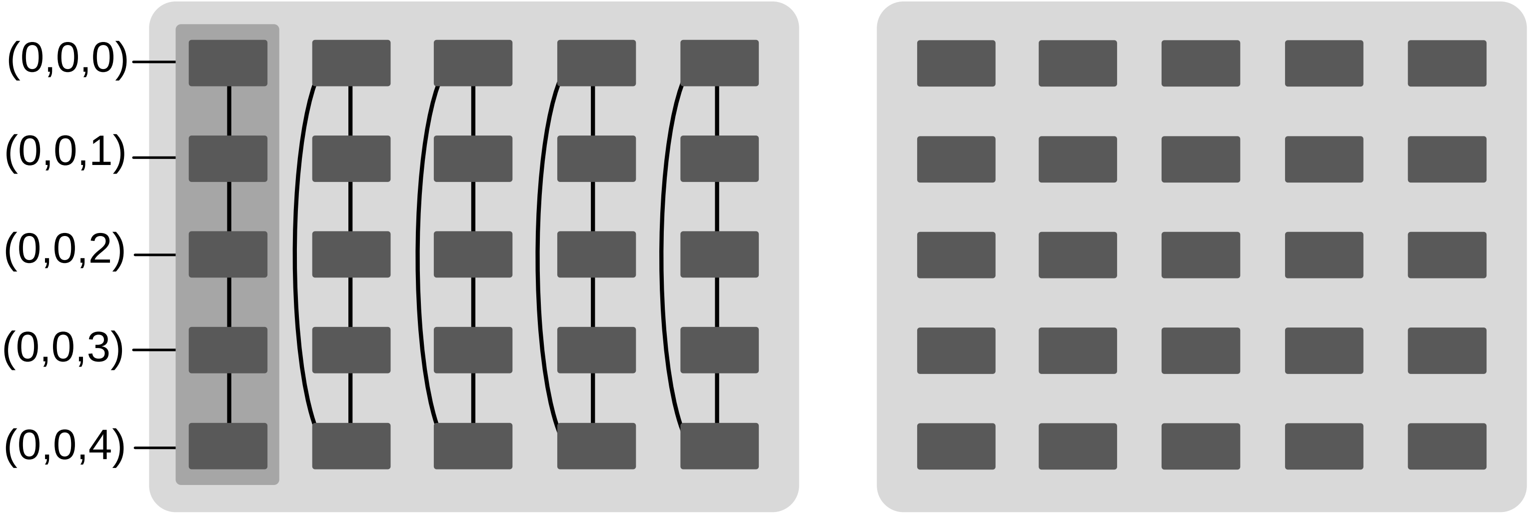
$X = \{1, \xi^2, \ldots, \xi^{q-3}\}$ (for subgraph 0)

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

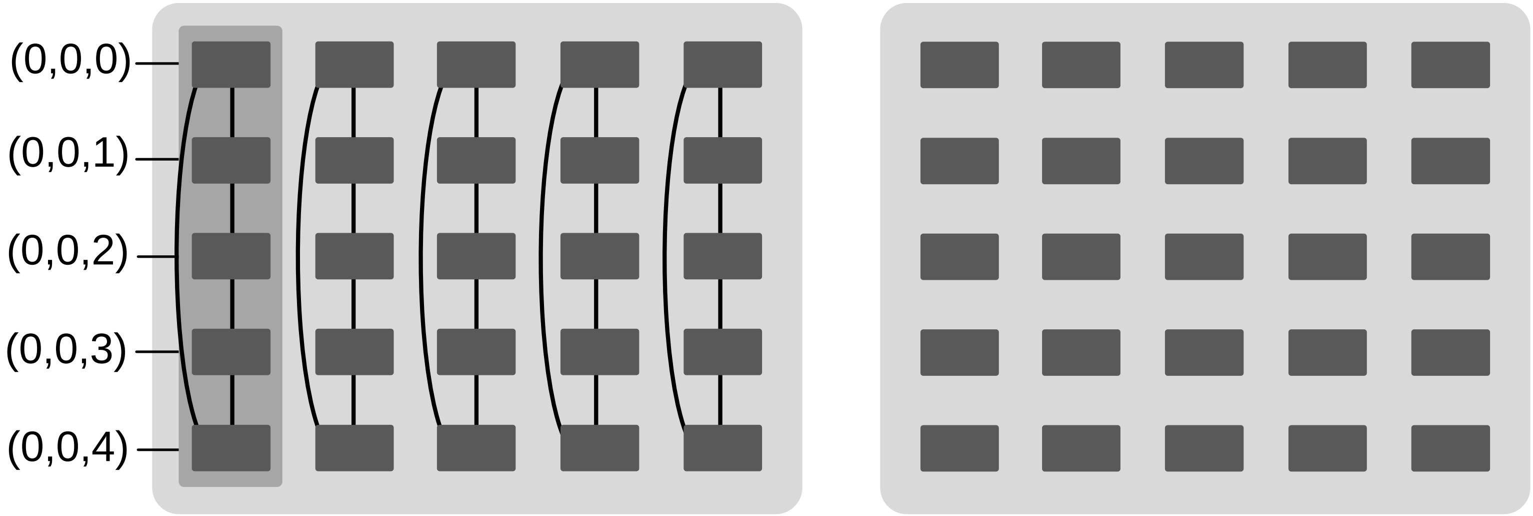
$X = \{1, \xi^2, \dots, \xi^{q-3}\}$ (for subgraph 0)

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

E *Example:* $q = 5$

Take Routers $(0,0,.)$

$X = \{1, 4\}$



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:

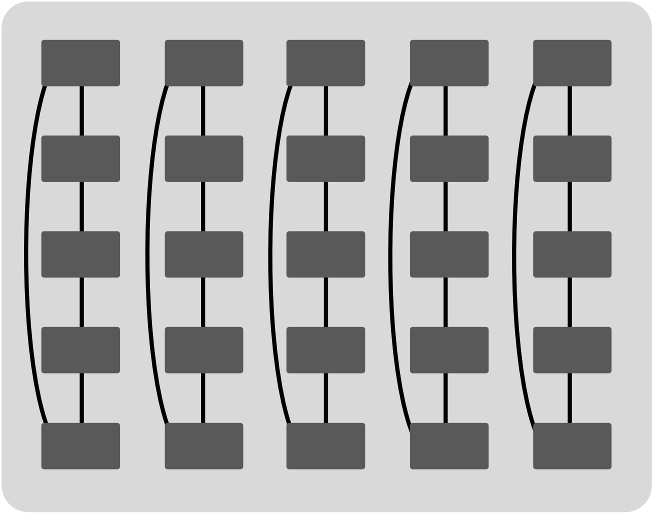
$X = \{1, \xi^2, \dots, \xi^{q-3}\}$ (for subgraph 0)

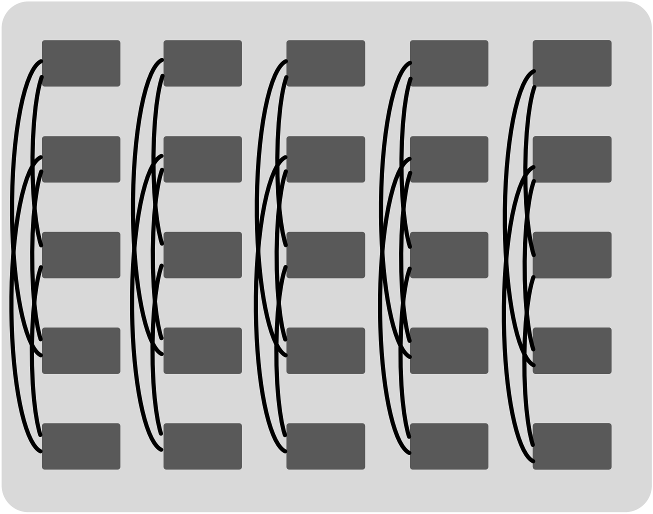
$X' = \{\xi, \xi^3, \dots, \xi^{q-2}\}$ (for subgraph 1)

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 $y = mx + c$

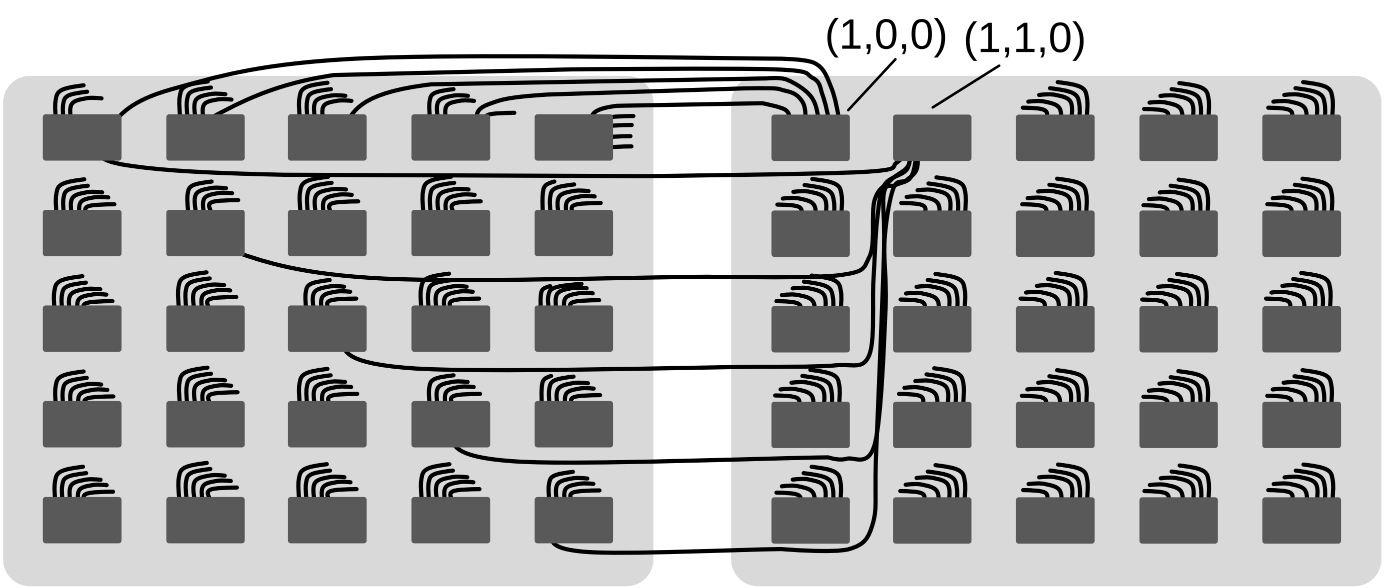
E *Example:* $q = 5$

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

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

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

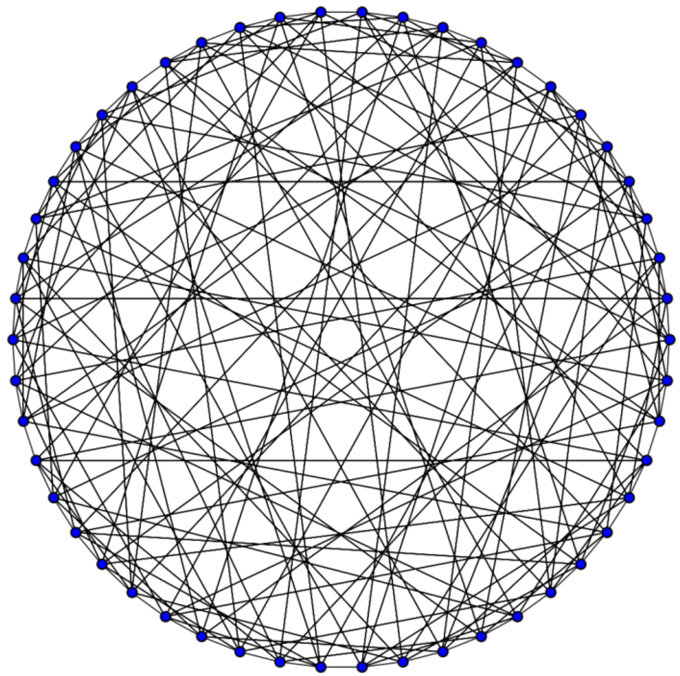
$(1,0,0) \leftrightarrow (0, x, x)$



Designing An Efficient Network Topology

Connecting routers: Diameter 2

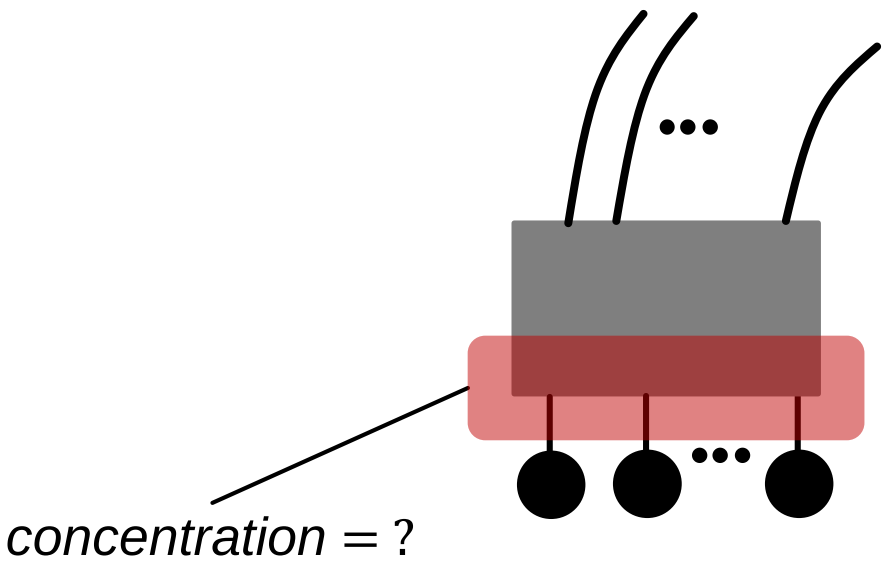
- Viable set of configurations
  - 10 SF networks with *the number of endpoints* < 11,000 (compared to 6 balanced Dragonflies [1])
- Let's pick *network radix* = 7...
  - ... We get the Hoffman-Singleton graph (attains the Moore Bound)



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



Designing an Efficient Network Topology

Attaching endpoints: Diameter 2

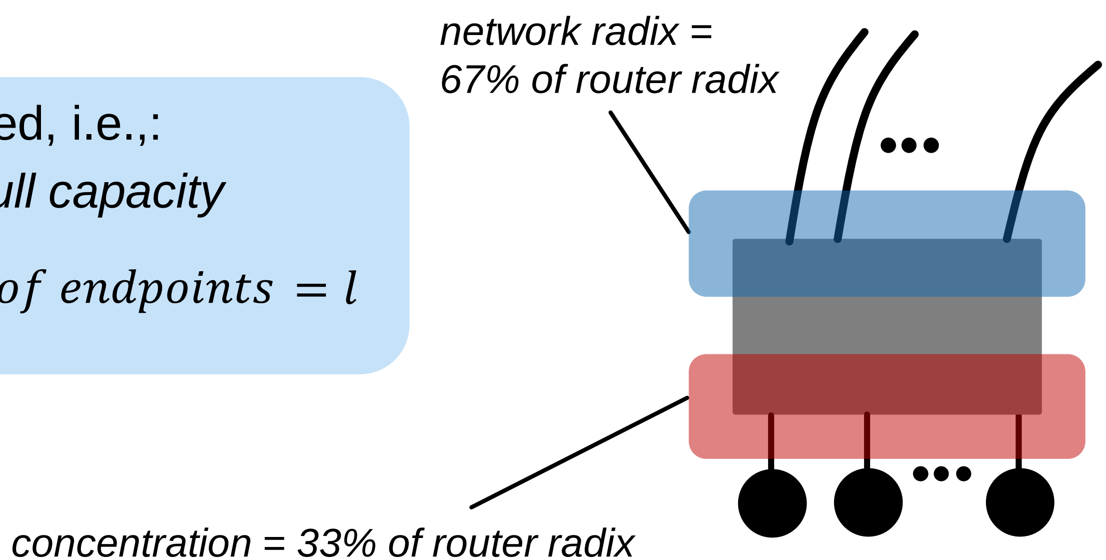
1 Get load *l* per router-router channel (average number of routes per channel)

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

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

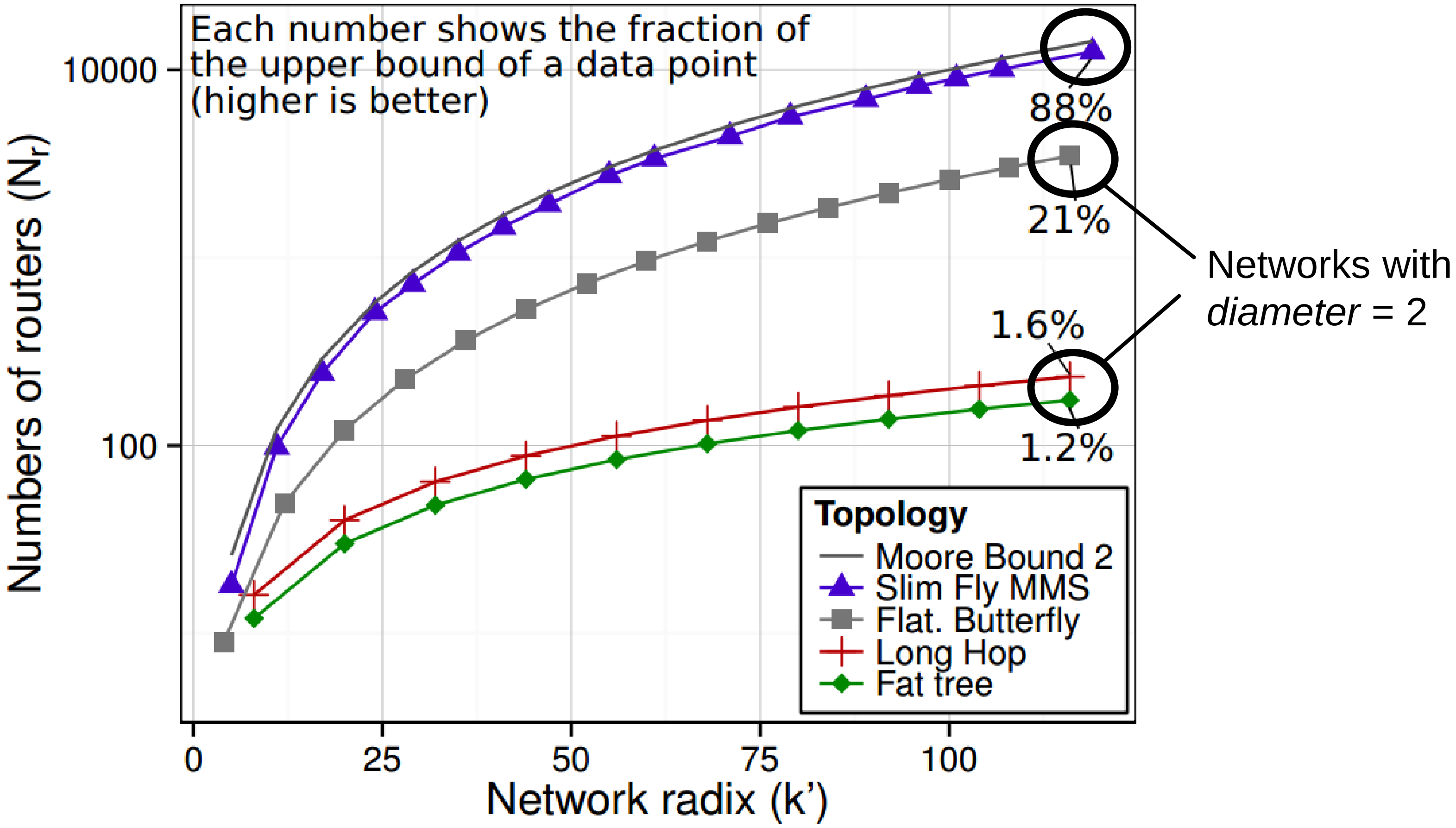
each endpoint can inject at full capacity

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



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

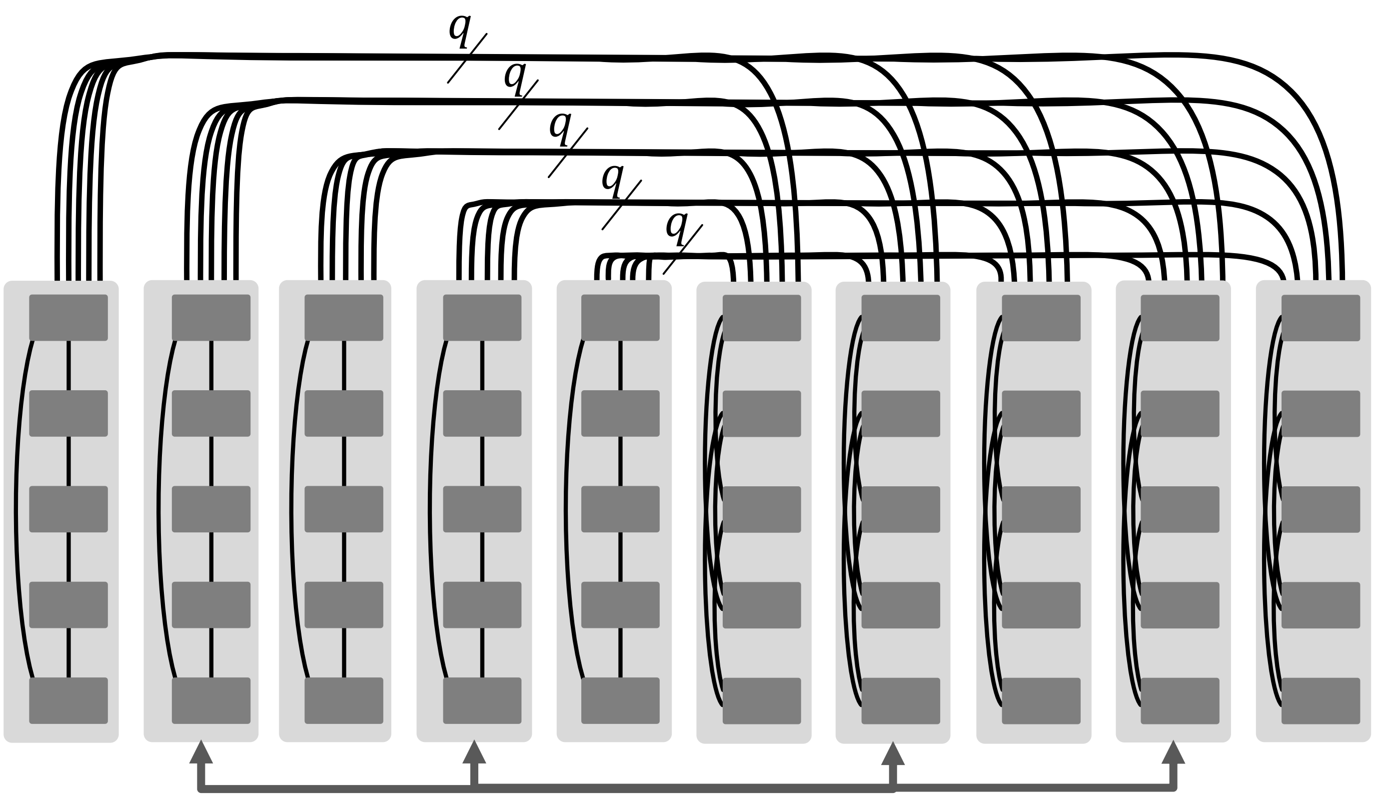
Resilience

Routing and performance

Routing

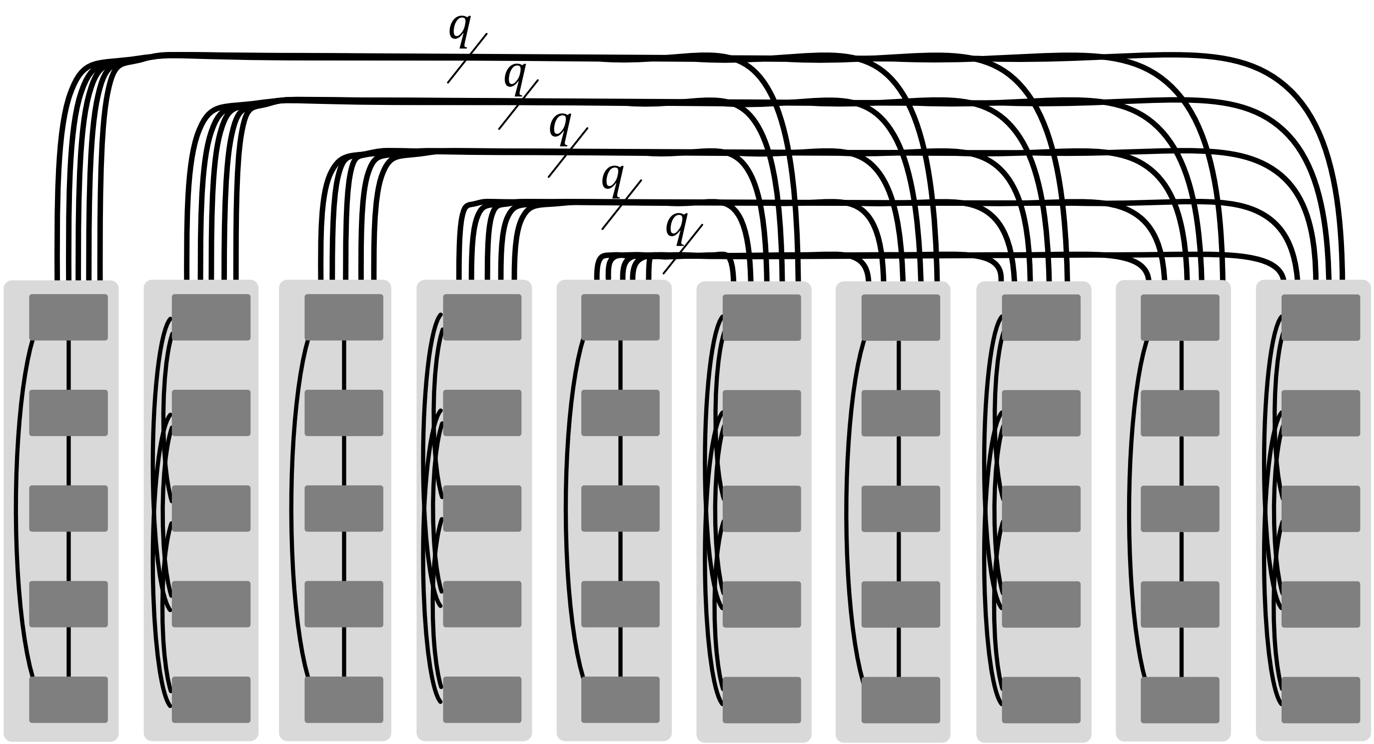
Performance, latency, bandwidth

PHYSICAL LAYOUT

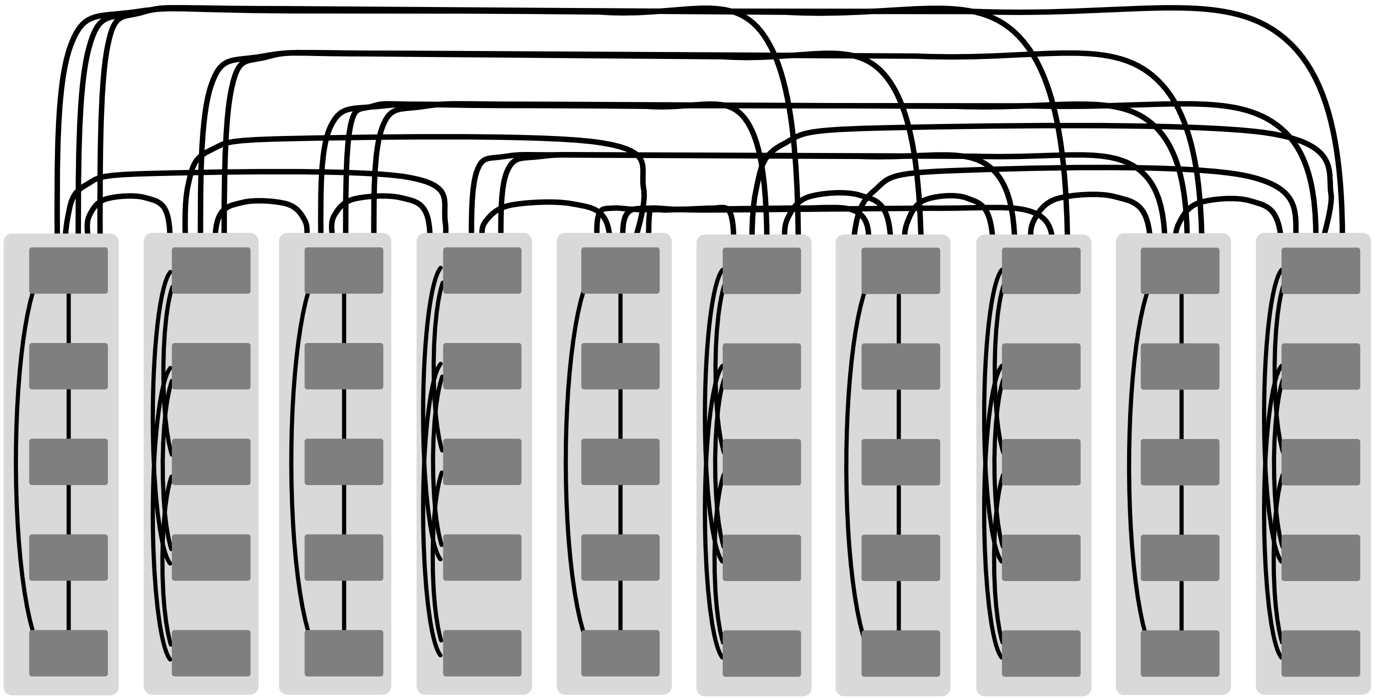


Mix (pairwise) groups
with different cabling patterns
to shorten inter-group cables

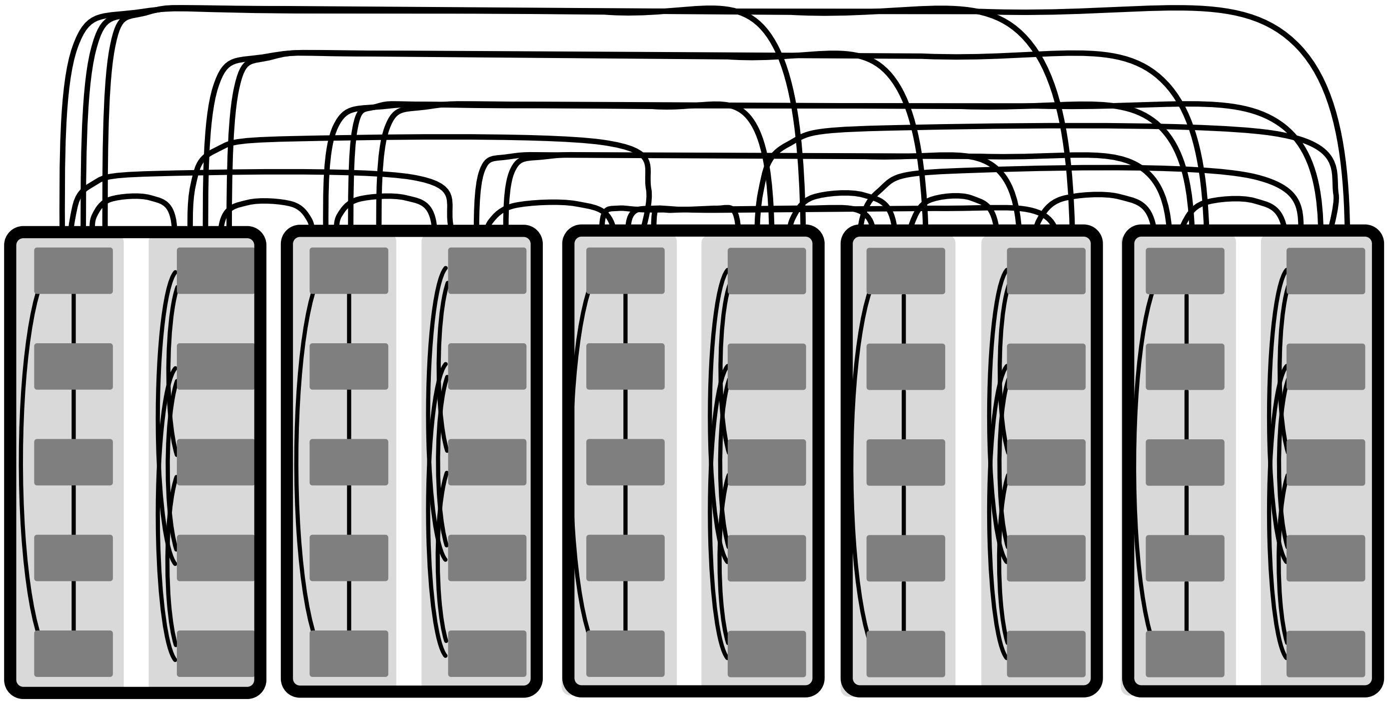
PHYSICAL LAYOUT



Physical Layout

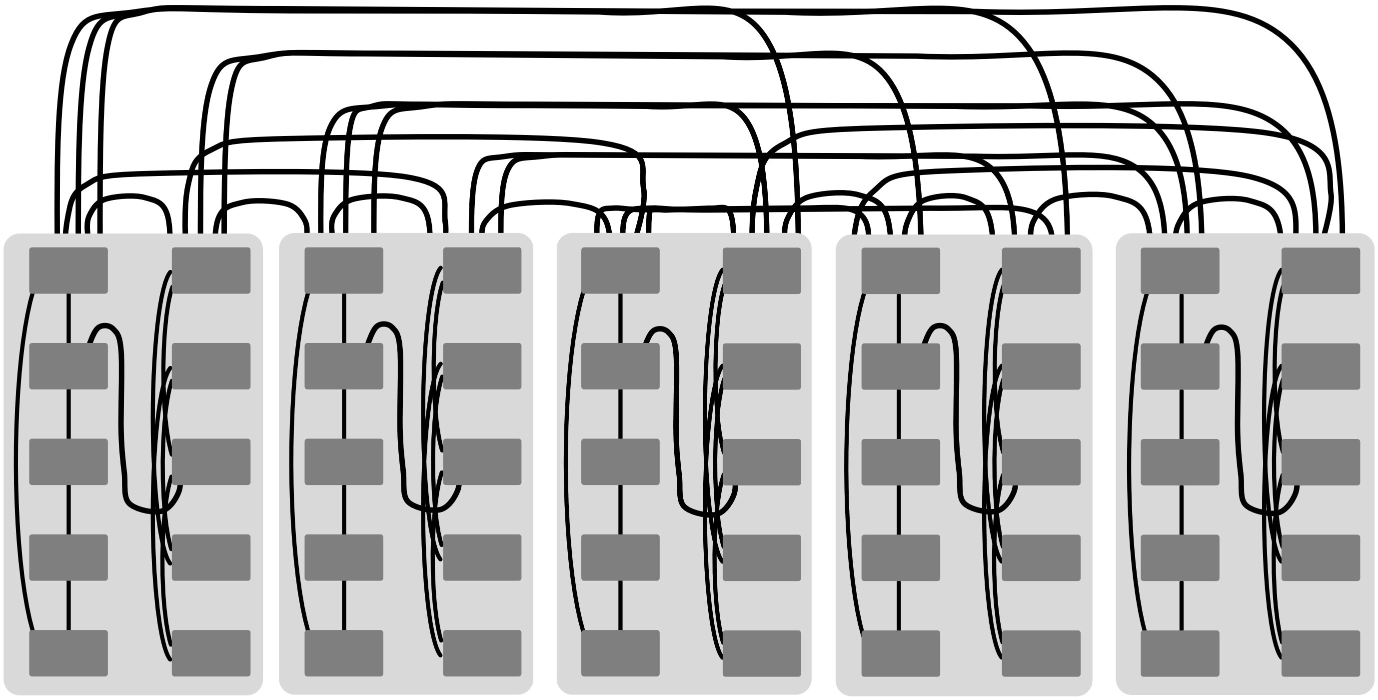


Physical Layout

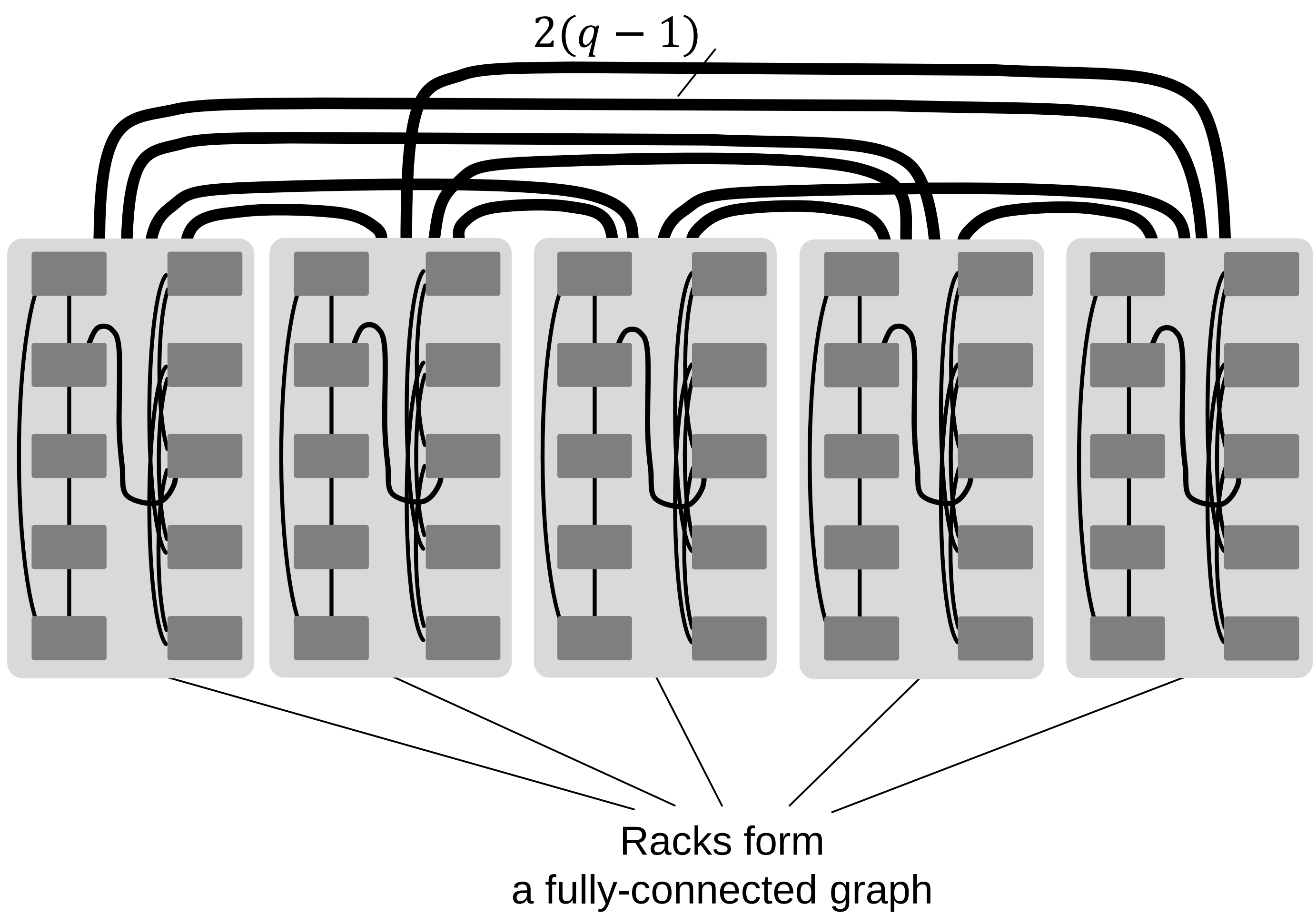


Merge groups pairwise
to create racks

Physical Layout

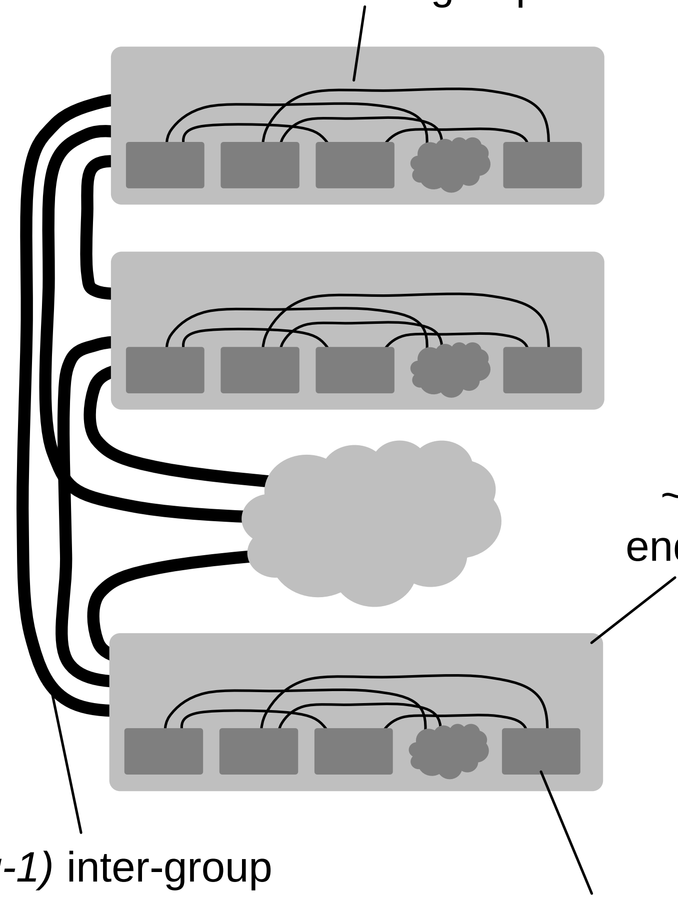


Physical Layout

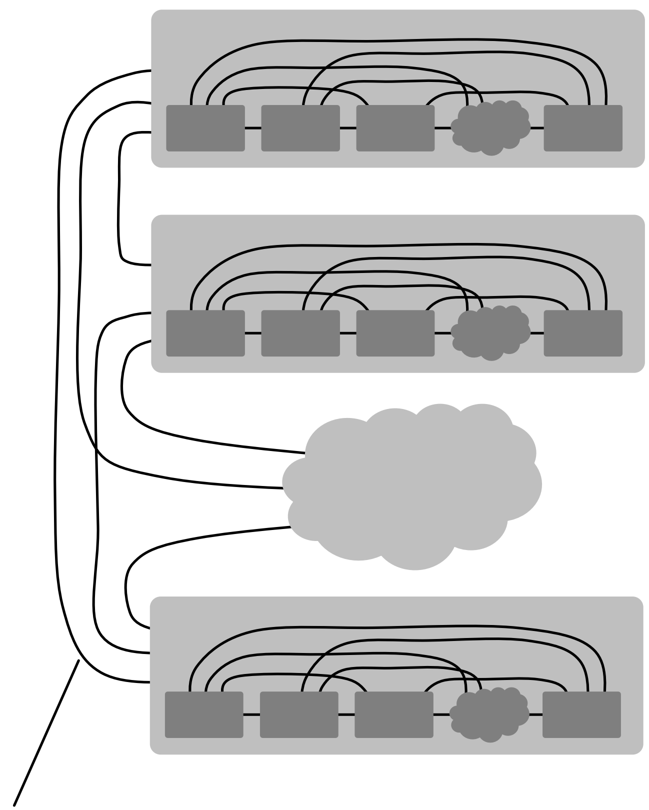


Physical Layout

SlimFly:



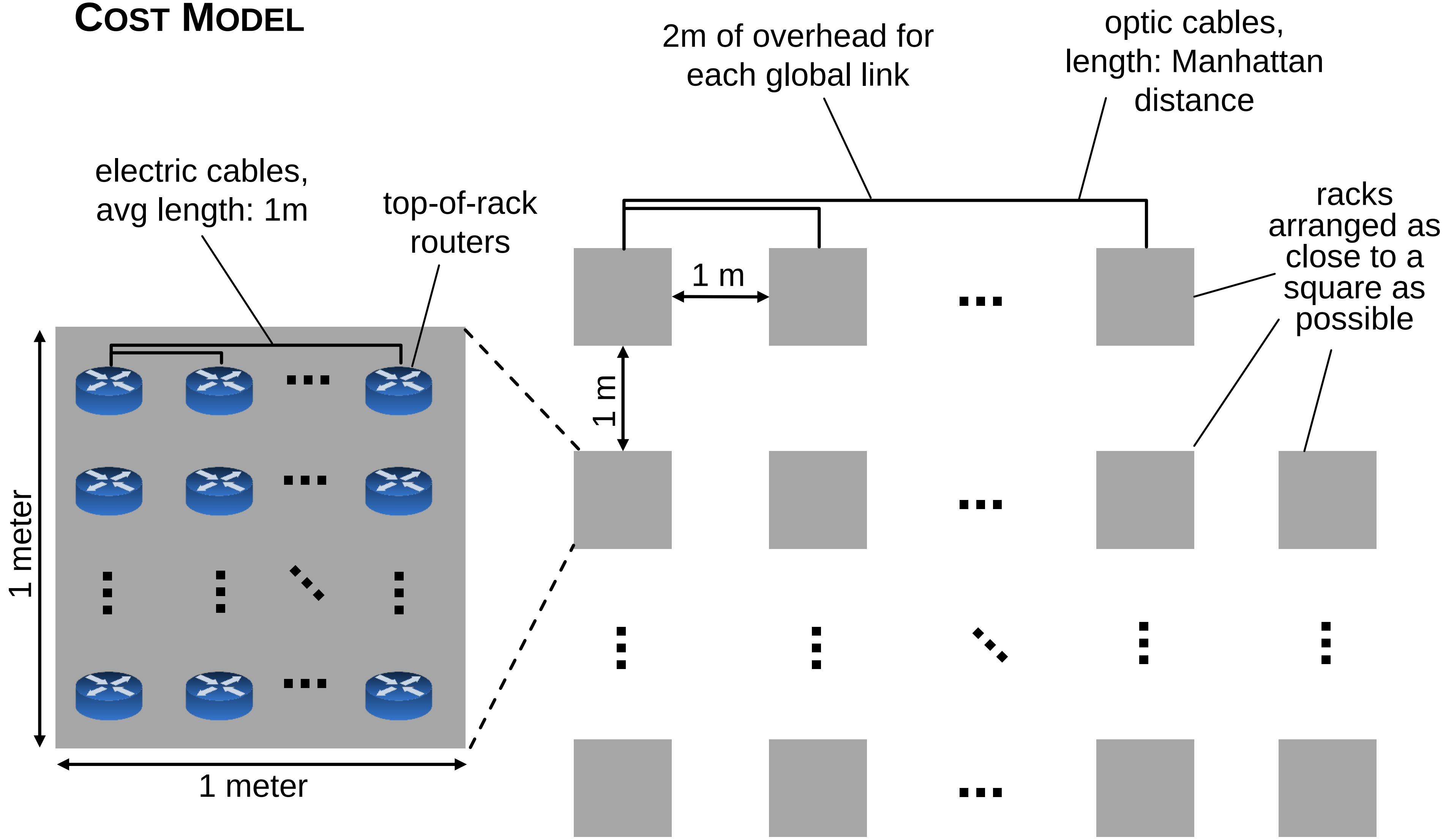
Dragonfly:



Cost Comparison

Cost Model

*Most cables skipped for clarity

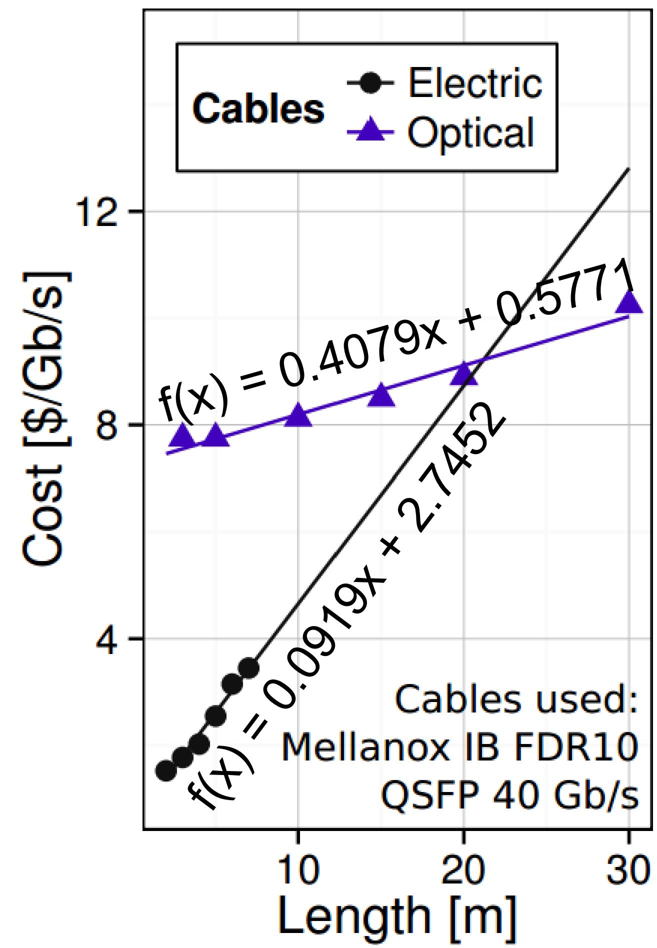


Cost Comparison

Cable Cost Model

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

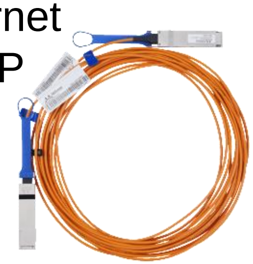
- Cable cost as a function of distance
  - The functions obtained using linear regression*
  - Cables used:
    Mellanox IB FDR10 40Gb/s QSFP



- 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

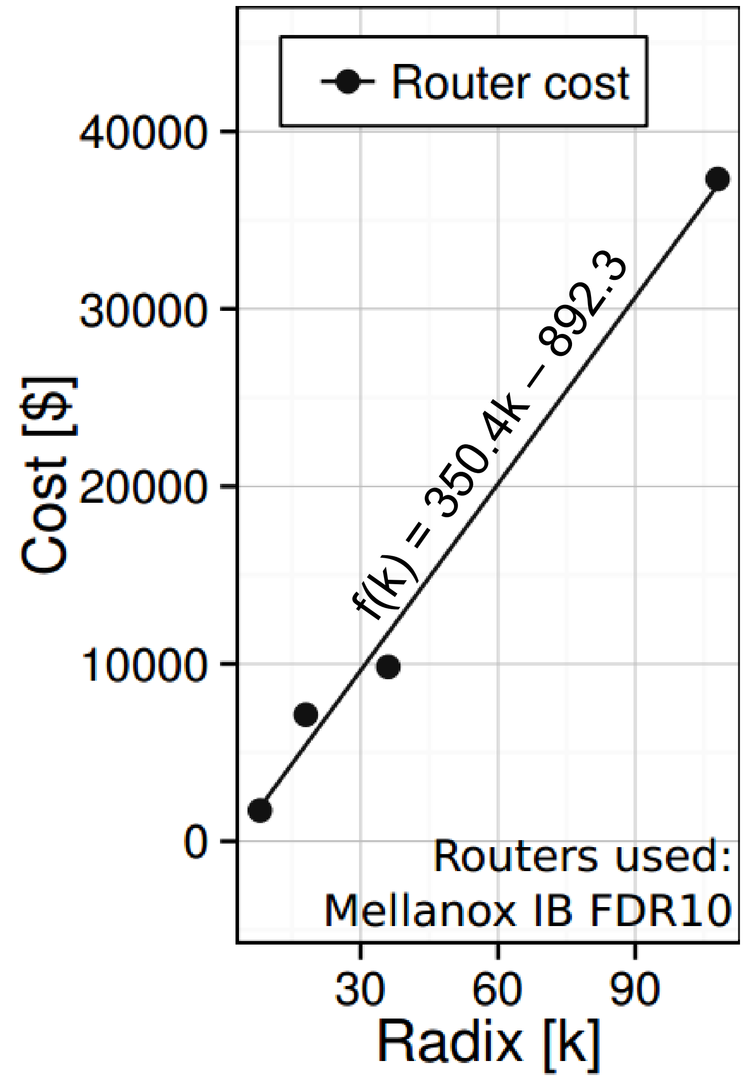
Router Cost Model

*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

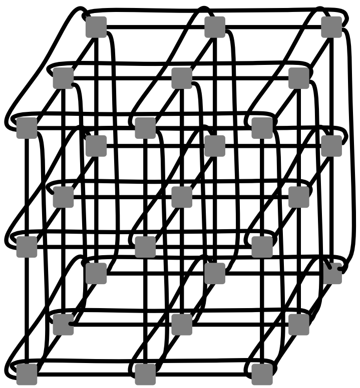
Mellanox Ethernet 10/40 Gb



Comparison Targets

Low-radix topologies

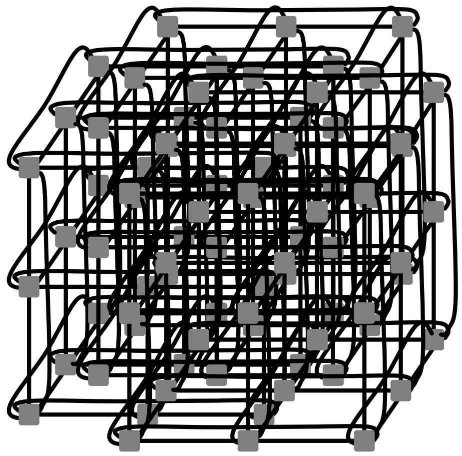
Torus 3D



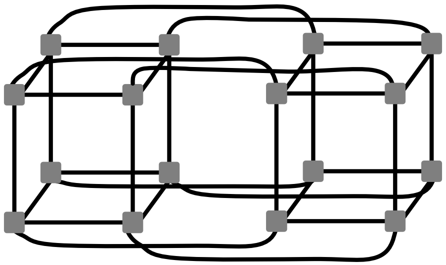
Cray XE6

IBM BG/Q

Torus 5D



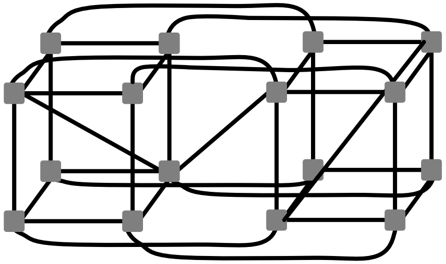
Hypercube



NASA Pleiades

Infinetics

Long Hop [1]

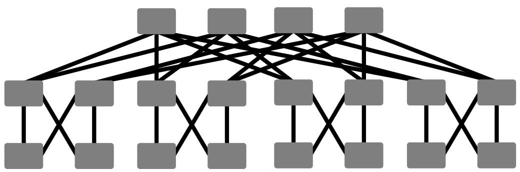


[1] Tomic, Ratko V. Optimal networks from error correcting codes. 2013 ACM/IEEE Symposium on Architectures for Networking and Communications Systems (ANCS)

Comparison Targets

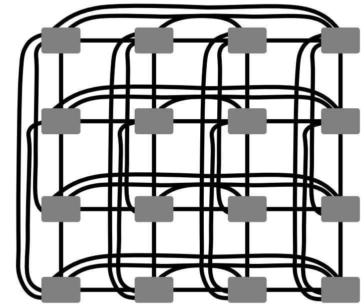
High-radix topologies

Fat tree [1]

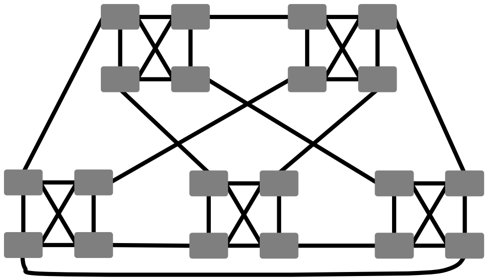


TSUBAME2.0

Flattened Butterfly [2]

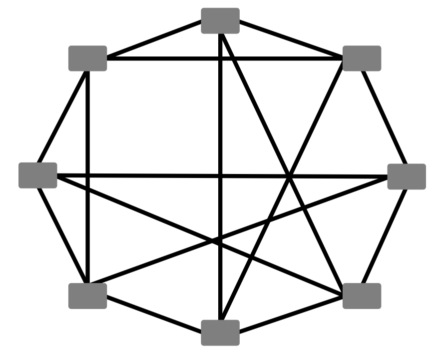


Dragonfly [3]



Cray Cascade

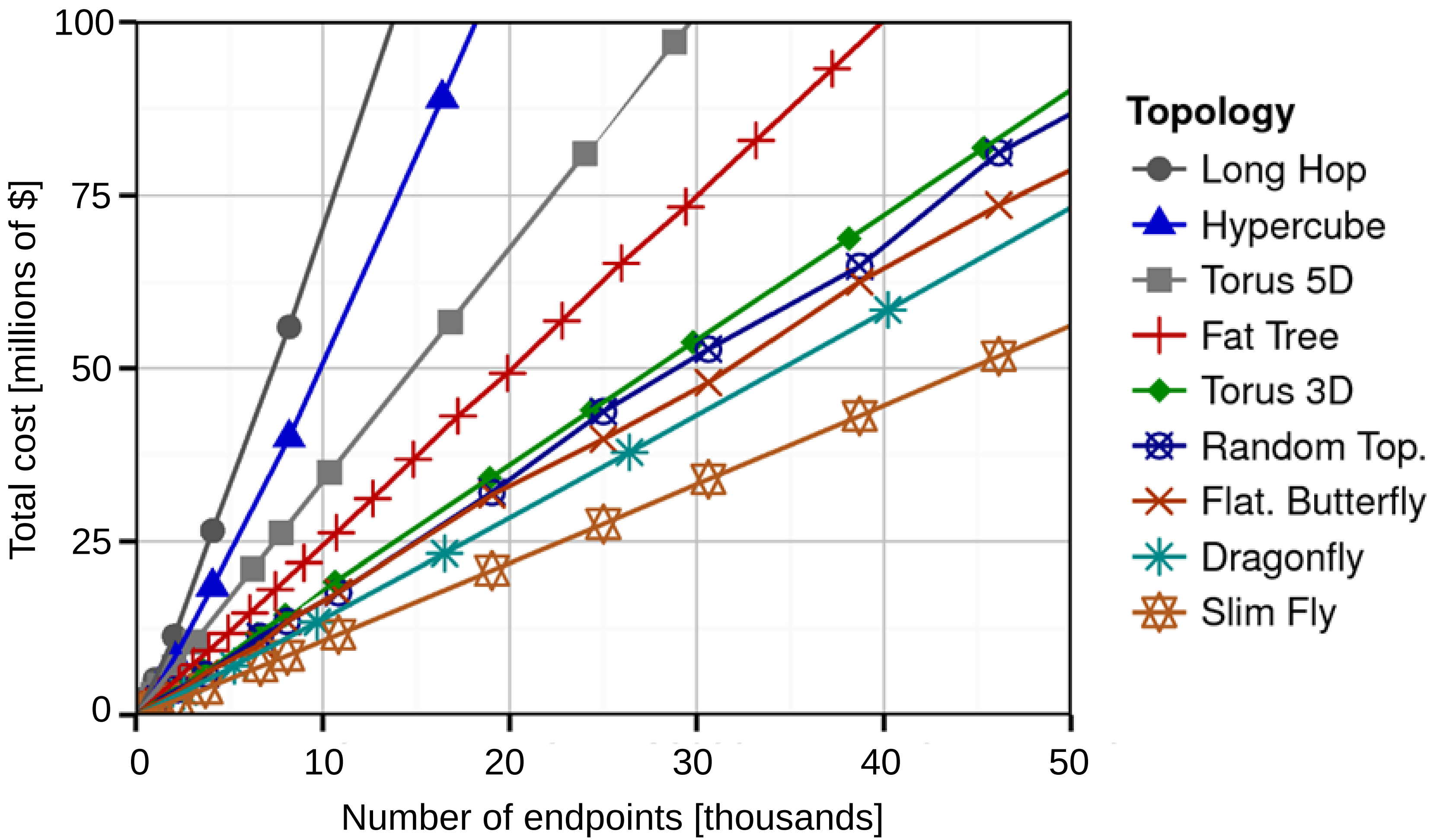
Random Topologies [4,5]



[1] C. E. Leiserson. Fat-trees: universal networks for hardware-efficient supercomputing. IEEE Transactions on Computers. 1985
[2] J. Kim, W. J. Dally, D. Abts. Flattened butterfly: a cost-efficient topology for high-radix networks. ISCA'07
[3] J. Kim, W. J. Dally, S. Scott, D. Abts. Technology-Driven, Highly-Scalable Dragonfly Topology. ISCA'08
[4] A. Singla, C. Hong, L. Popa, P. B. Godfrey. Jellyfish: Networking Data Centers Randomly. NSDI'12
[5] M. Koibuchi, H. Matsutani, H. Amano, D. F. Hsu, H. Casanova. A case for random shortcut topologies for HPC interconnects. ISCA'12

Cost Comparison

Results



Cost & Power Comparison

Detailed Case-Study

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

Cost & Power Comparison

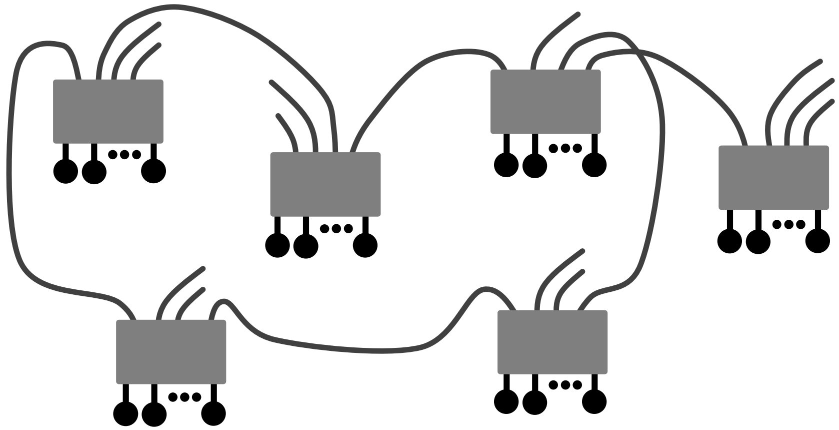
Detailed Case-Study: High-Radix Topologies

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

$\approx N$	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%**

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

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

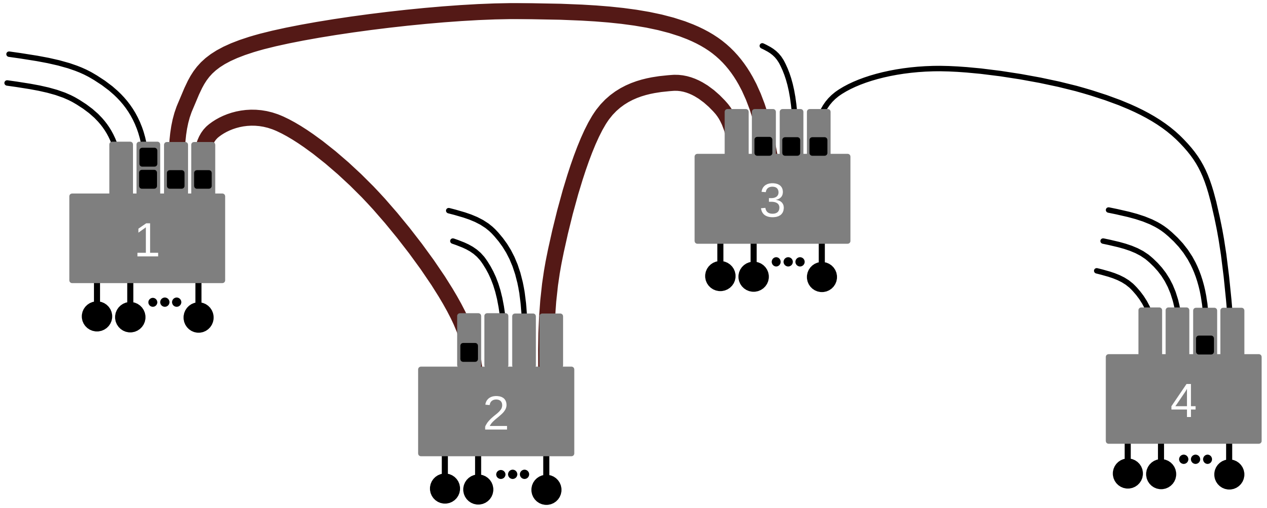
Routing and performance

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

MINIMUM ROUTING

1 *Intra-group connections*

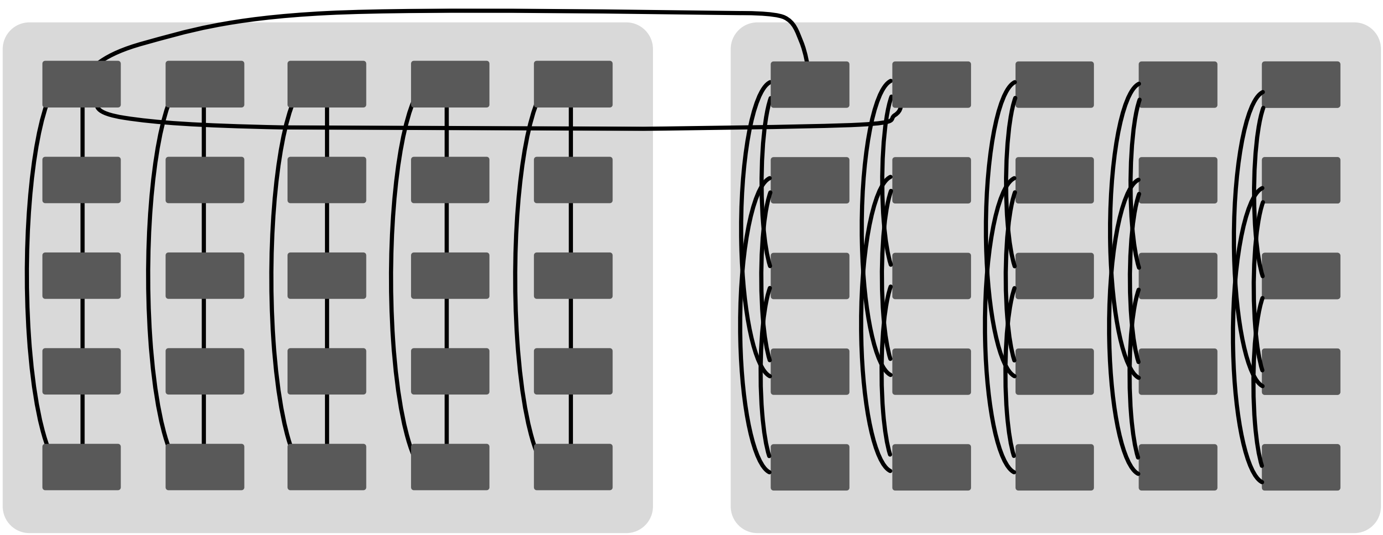
∃ Path of length 1 or 2 between two routers

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

∃ Path of length 1 or 2 between two routers

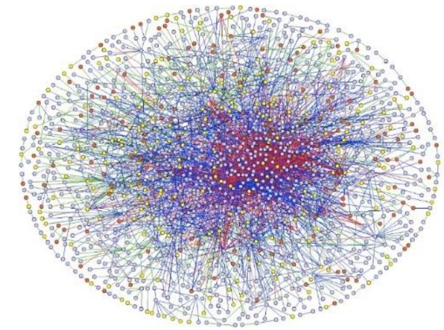
3 *Inter-group connections (identical types of groups)*

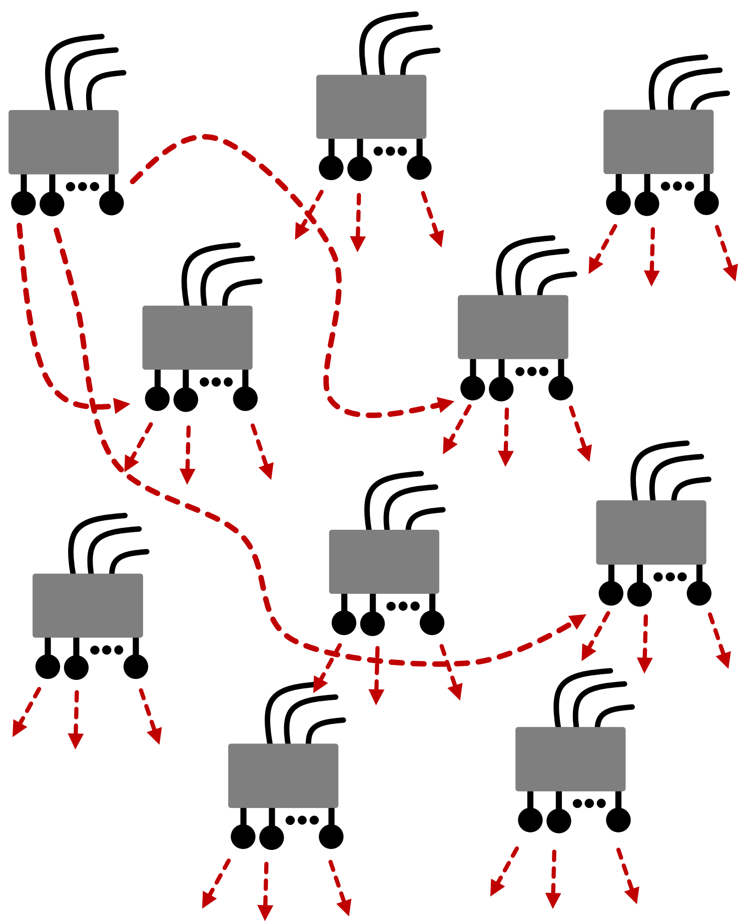
∃ Path of length 2 between two routers

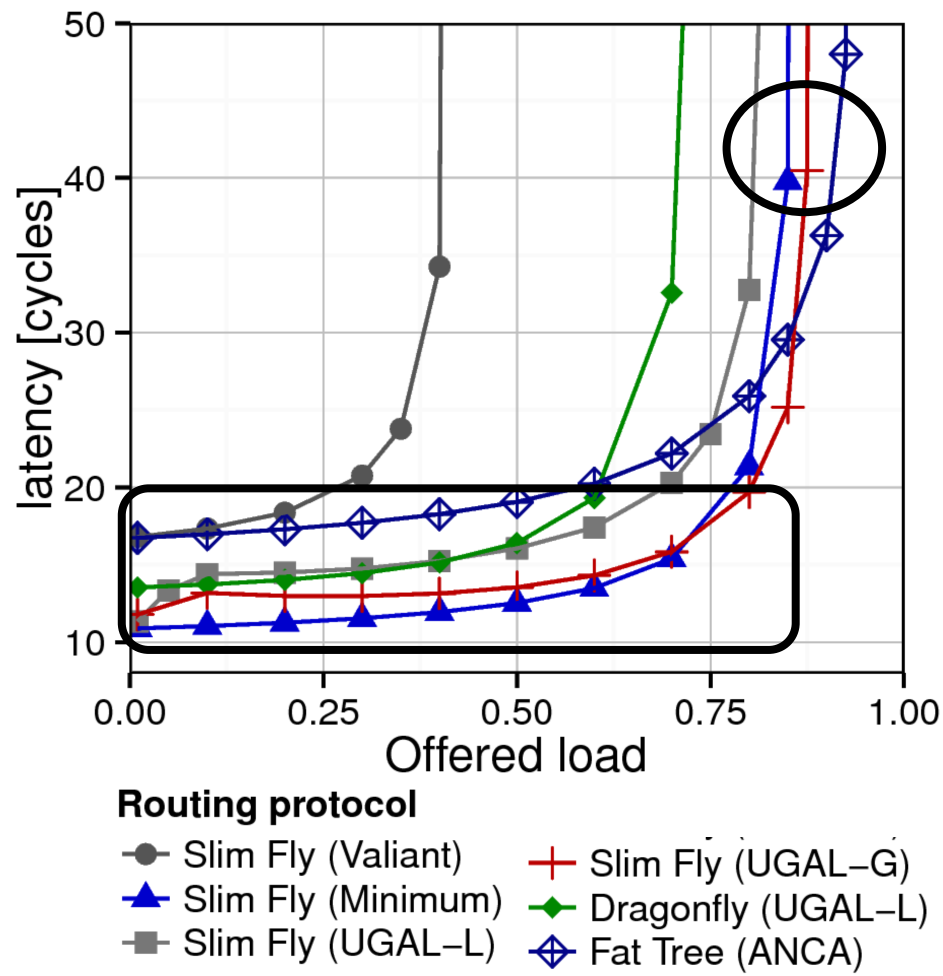


PERFORMANCE & ROUTING

RANDOM UNIFORM TRAFFIC

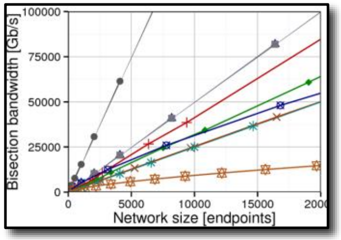




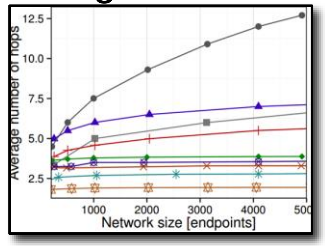


OTHER RESULTS

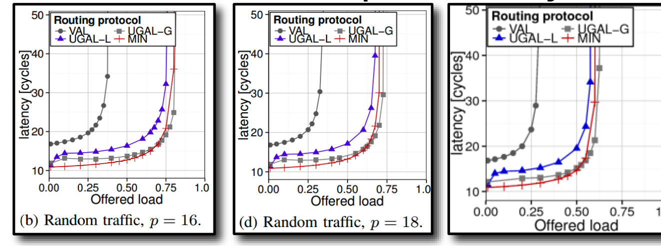
Bisection b.



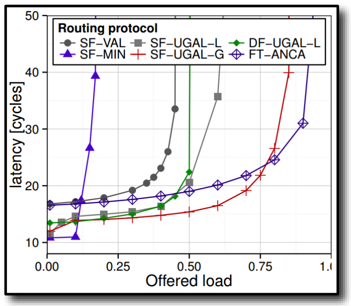
Avg. distance



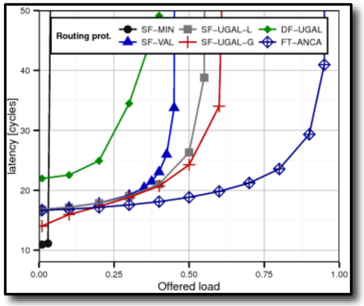
Oversubscription analysis



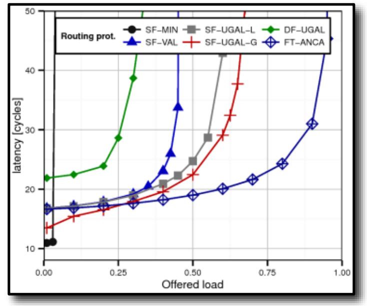
Bit reverse



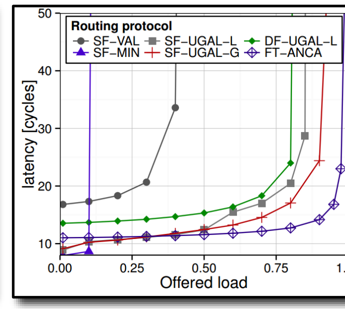
Bit complement



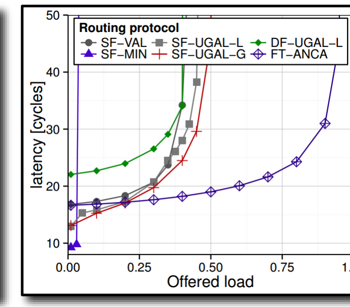
Shuffle



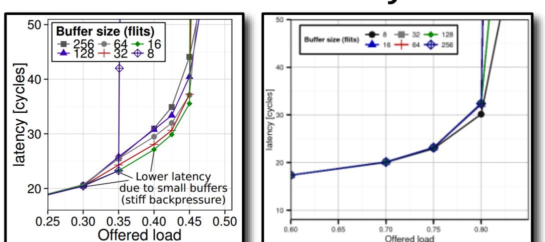
Shift



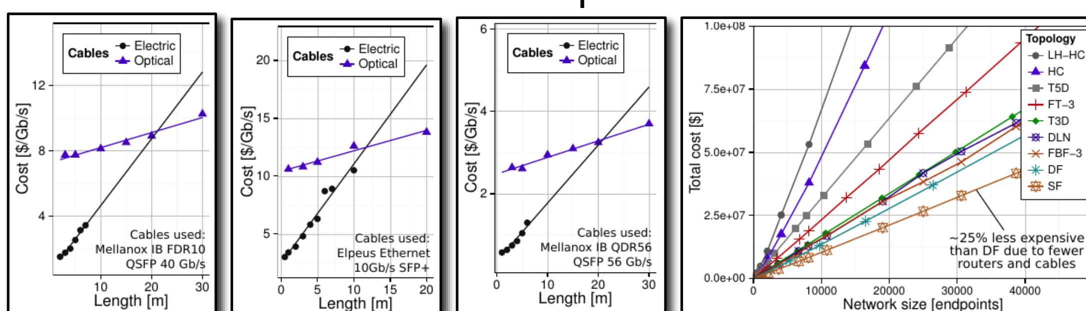
Adversarial

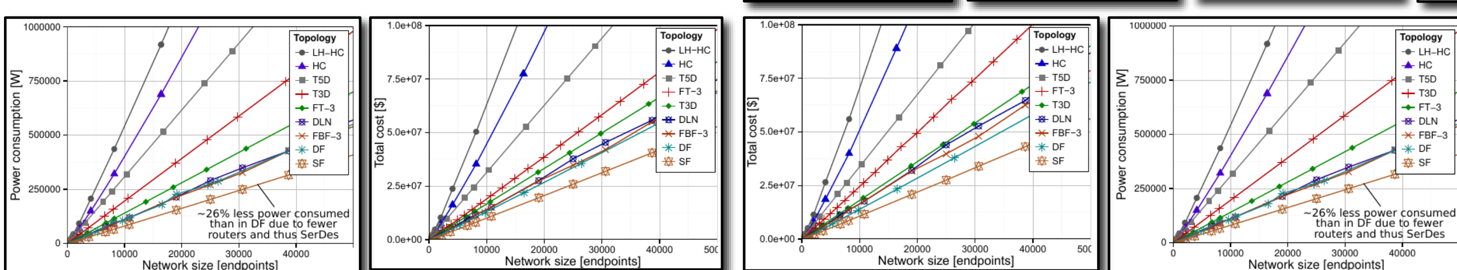


Buffer size analysis



Other cost & power results



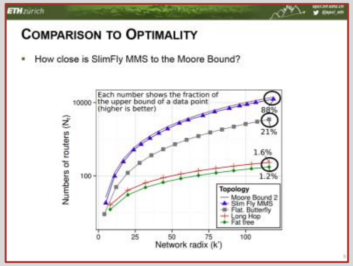


Topology	Dragonfly	Slim Fly
Endpoints (N)	10,890	**10,830**
Routers (N_r)	990	**722**
Radix (k)	43	**43**
Electric cables	6,885	**6,669**
Fiber cables	1,012	**6,869**
Cost per node [$]	1,365	**1,033**
Power per node [W]	10.9	**8.02**

Conclusions

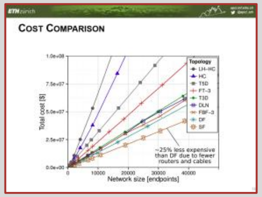
Topology design

Optimizing towards the Moore Bound reduces expensive network resources

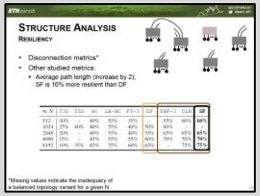


Advantages of SlimFly

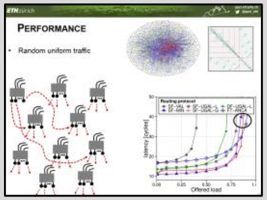
Cost & power



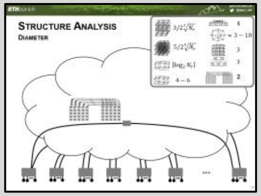
Resilience



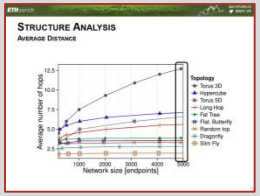
Performance



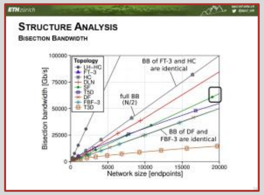
Diameter



Avg. distance

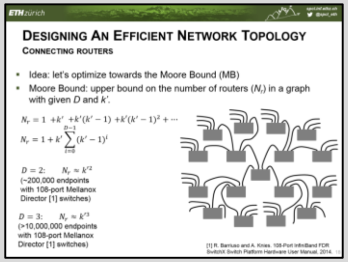


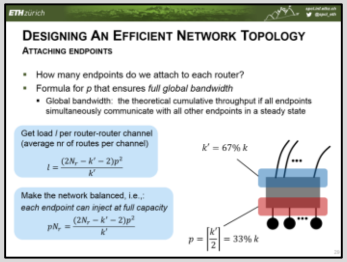
Bandwidth



Optimization approach

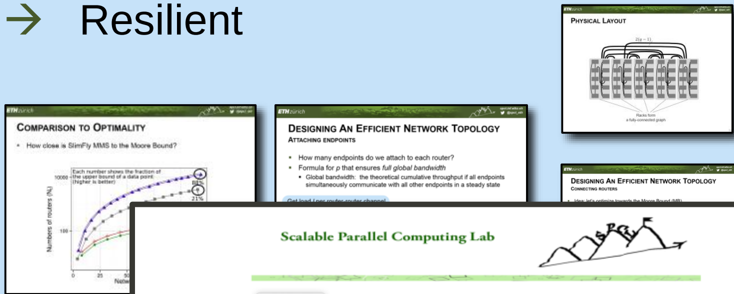
Combining mathematical optimization and current technology trends effectively tackles challenges in networking

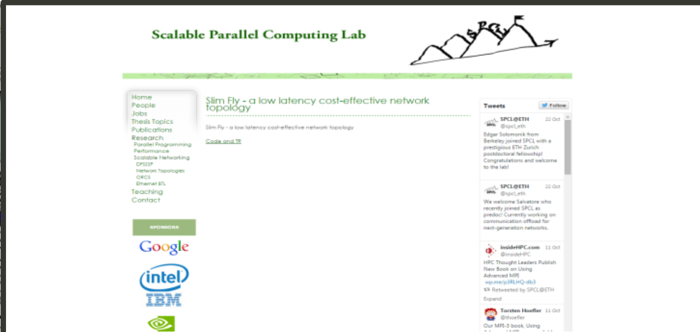




A lowest-diameter topology

→ Viable set of configurations
→ Resilient

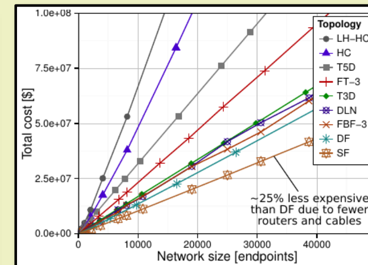


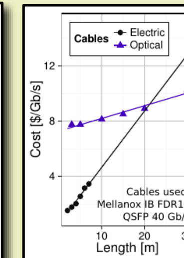


http://spcl.inf.ethz.ch/Research/Scalable_Networking/SlimFly

A cost & power effective topology

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



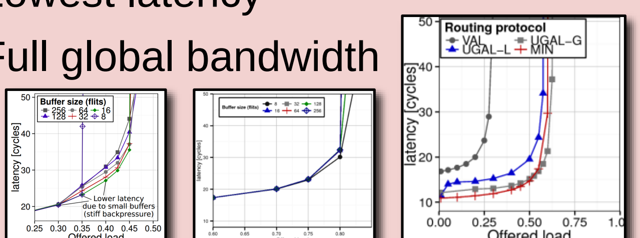


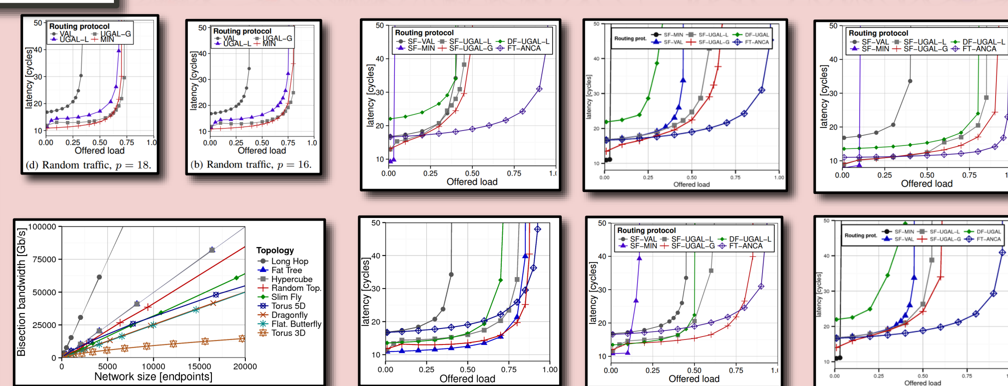
AWAY M

Thank you for your attention

A high-performance topology

→ Lowest latency
→ Full global bandwidth

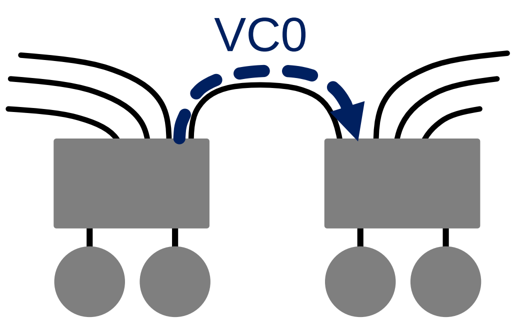


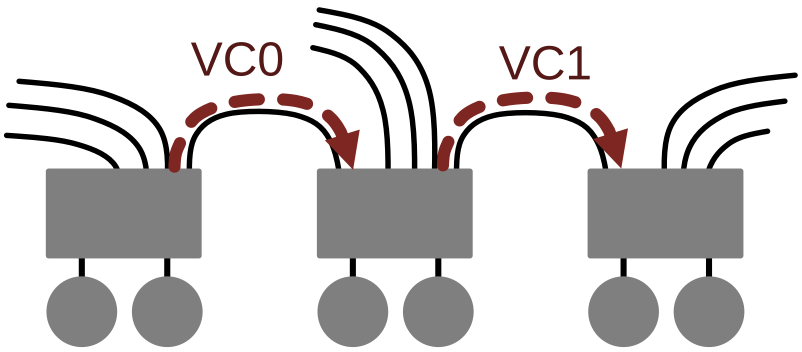


Deadlock Freedom

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]



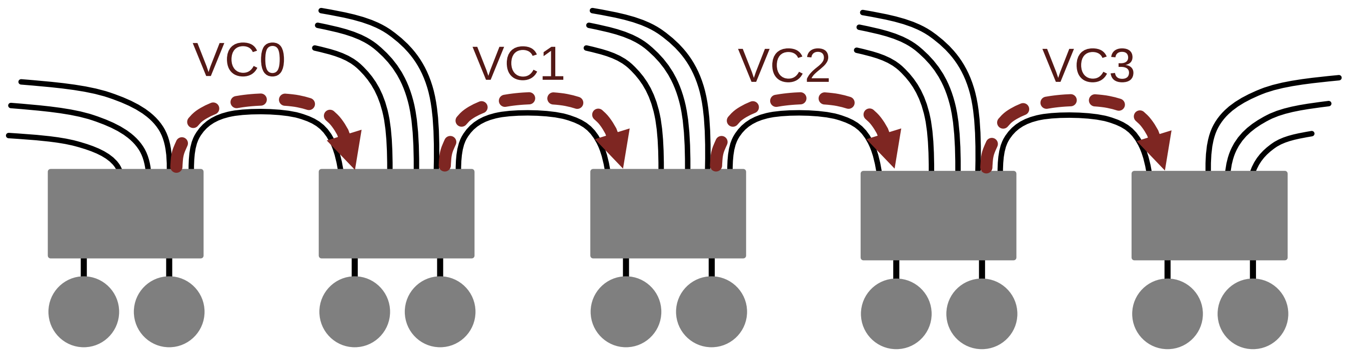


[1] J. Domke, T. Hoefler, and W. Nagel. Deadlock-Free Oblivious Routing for Arbitrary Topologies. IPDPS'11

Deadlock Freedom

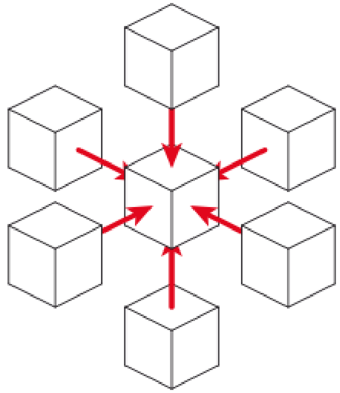
Adaptive Routing

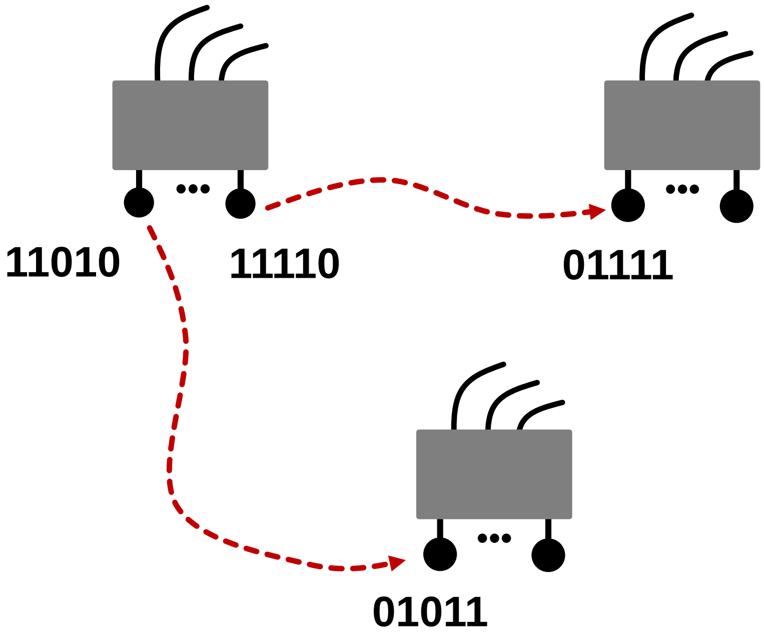
- Simple generalization of the previous scheme
- Assign four virtual channels (VC0 – VC3) to each link
- For hop *k* path use VC*k*, 0 <= k <= 3

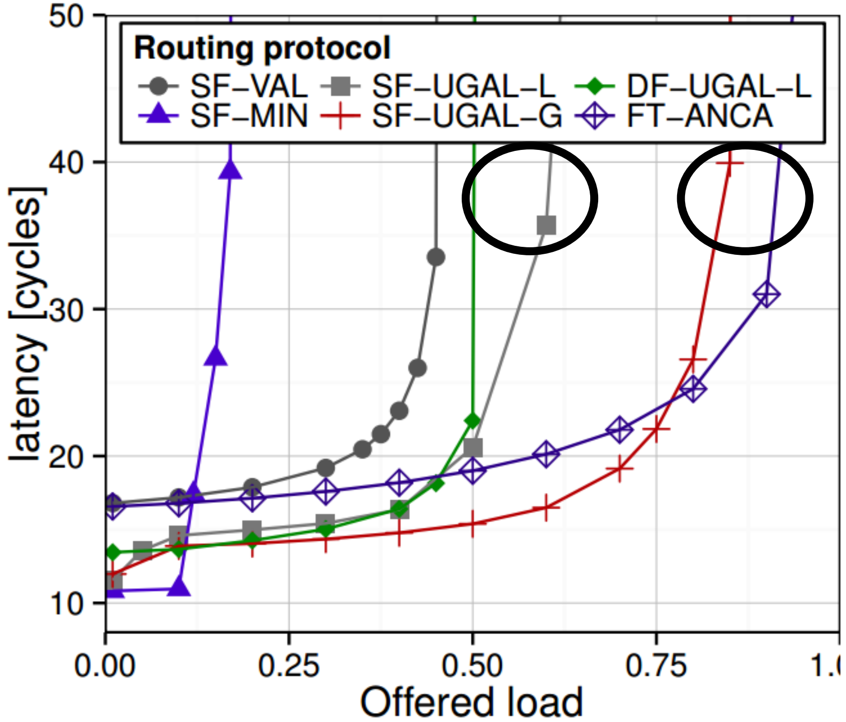


Performance

- Bit permutation traffic







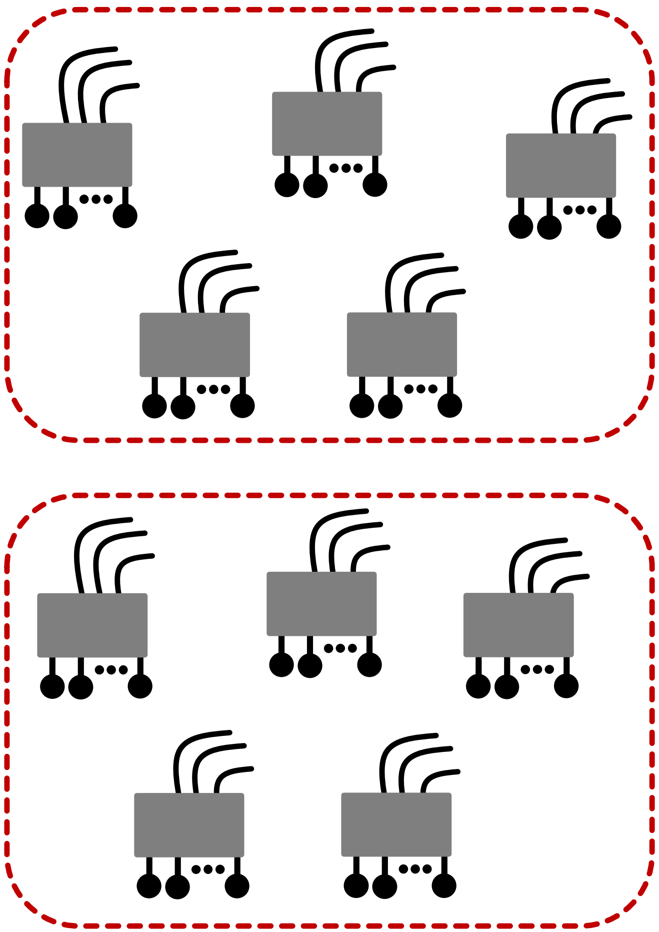
PERFORMANCE

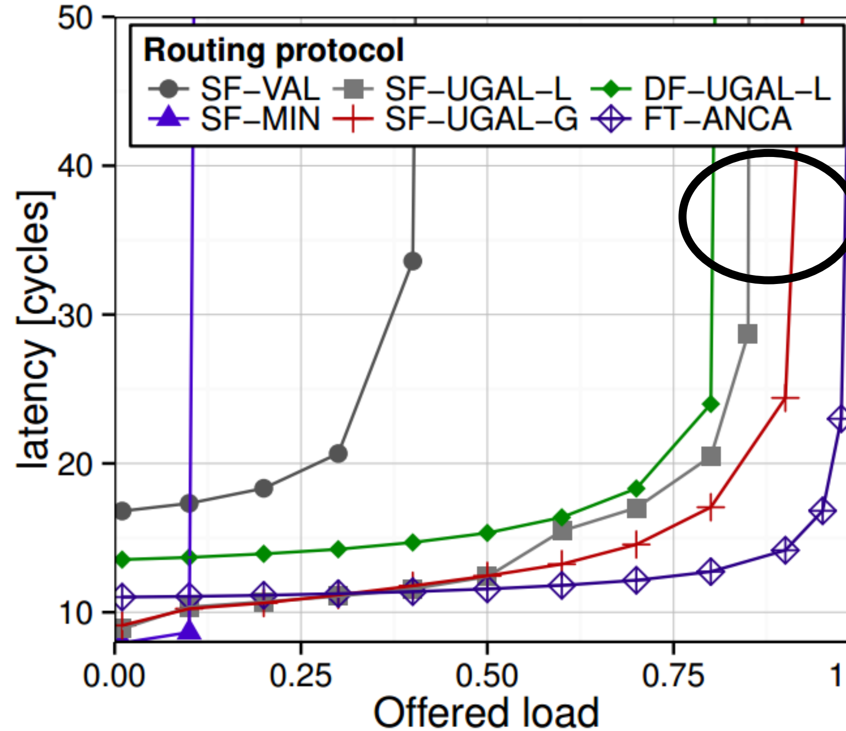
- Shift traffic

dest id — source id

$$d = \left(s \bmod \frac{N}{2}\right) + \frac{N}{2}$$

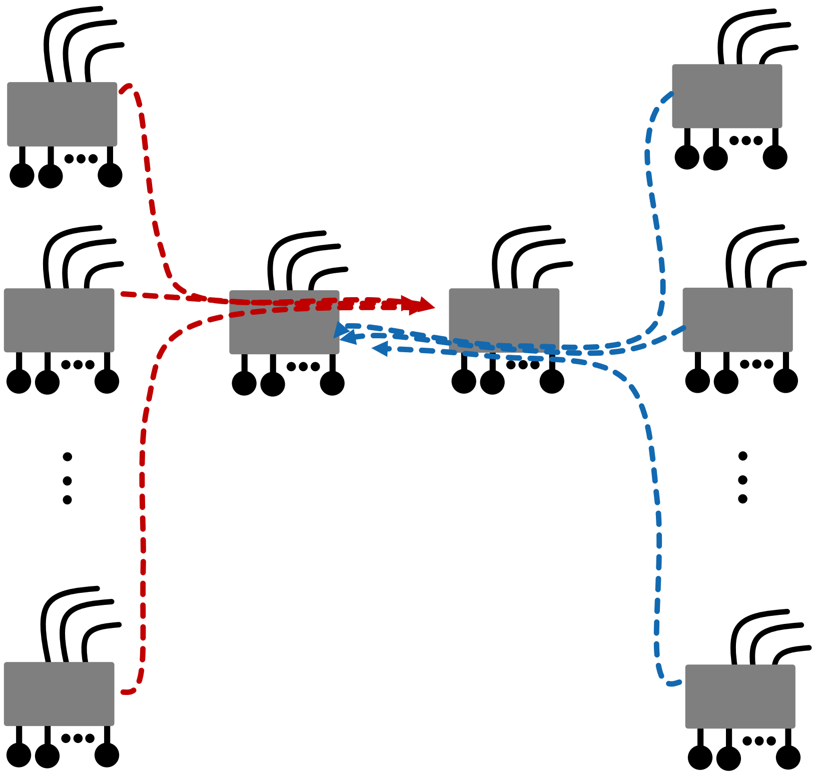
$$d = s \bmod \frac{N}{2}$$

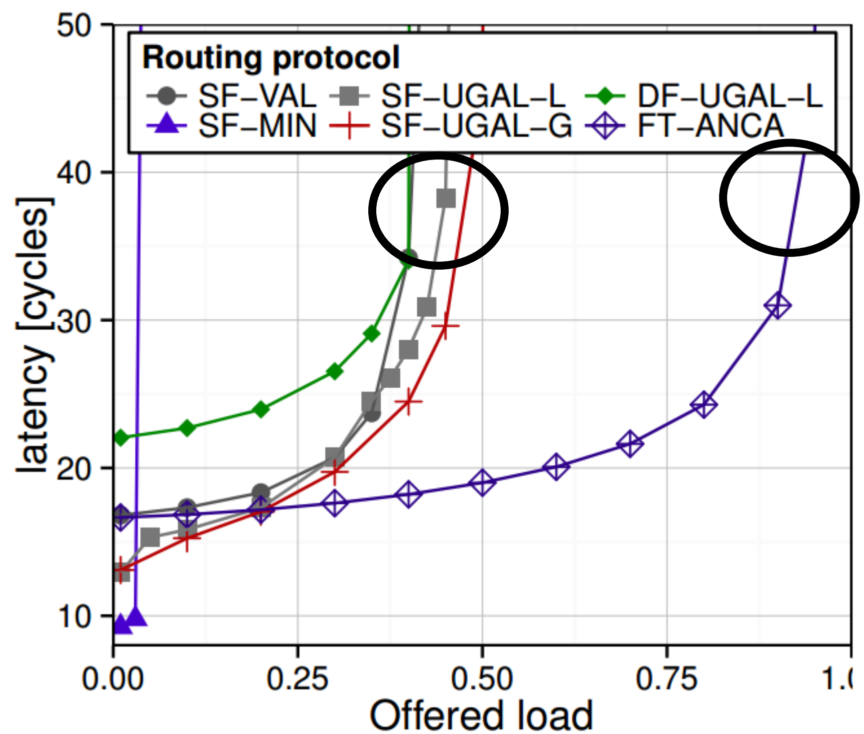




Performance

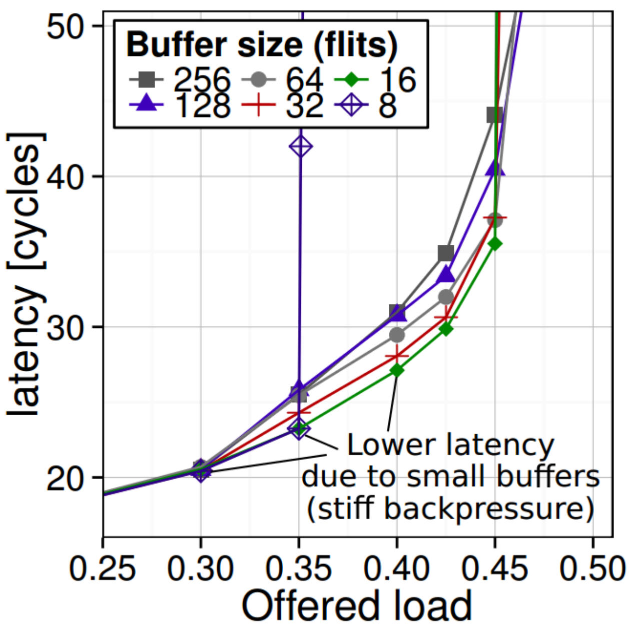
- Worst-case traffic



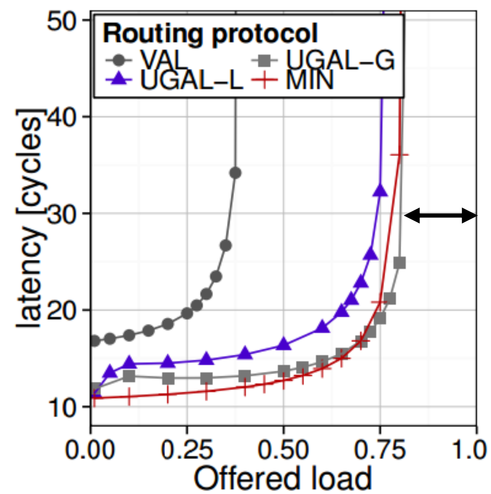


PERFORMANCE

- Buffer sizes (UGAL-L, worst-case traffic)
- Oversubscription (64 flits)

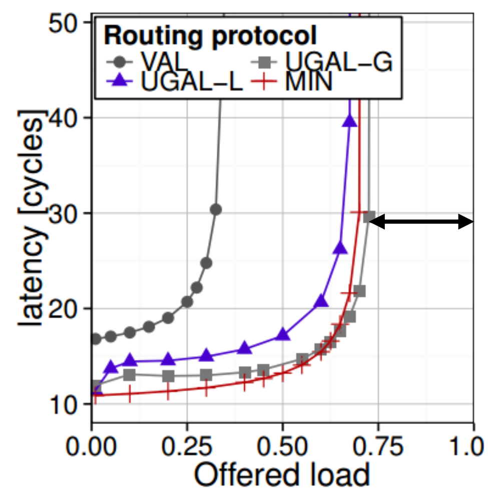


p=16



(b) Random traffic, $p = 16$.

p=18

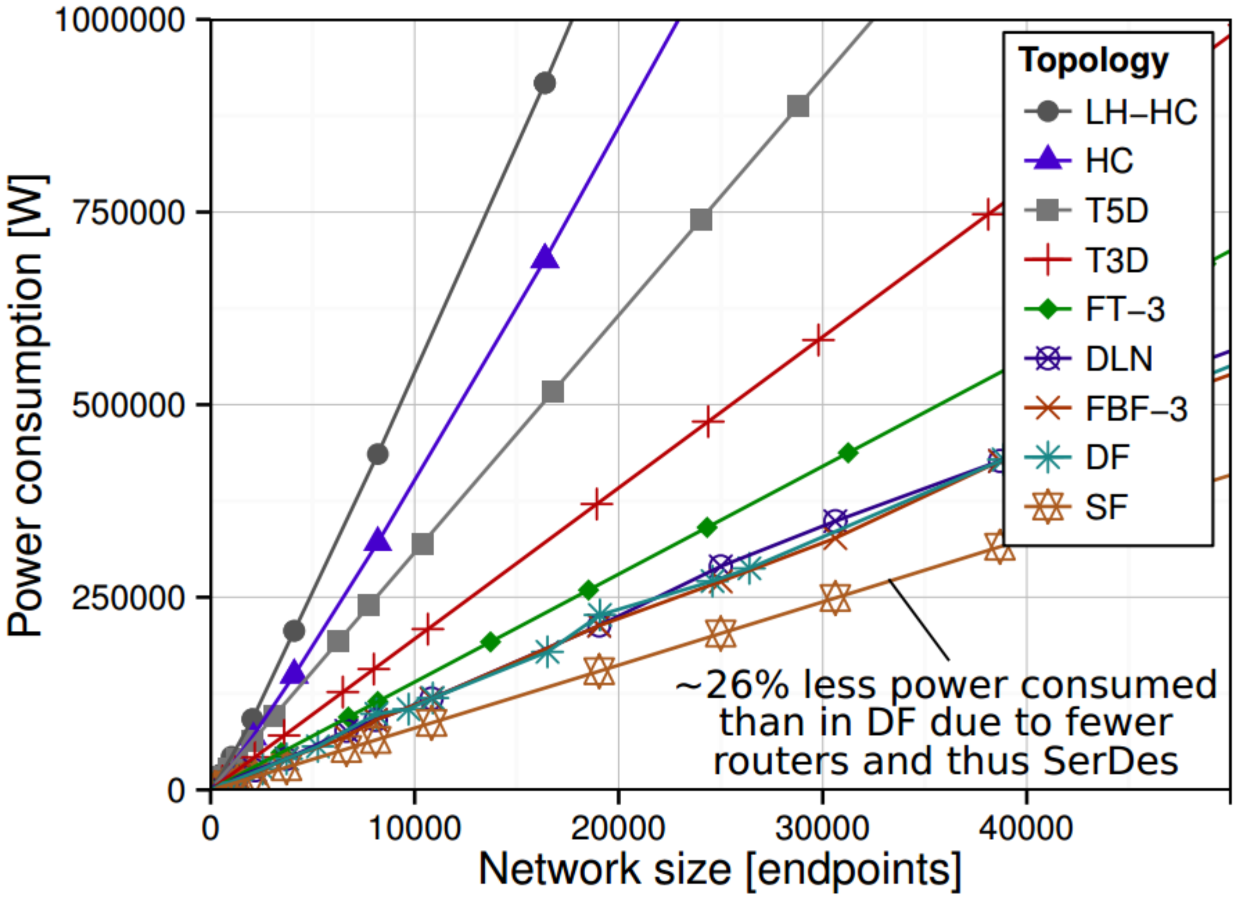


(d) Random traffic, $p = 18$.

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



[1] D. Abts et al. Energy Proportional Datacenter Networks. ISCA'10

Cost & Power Comparison

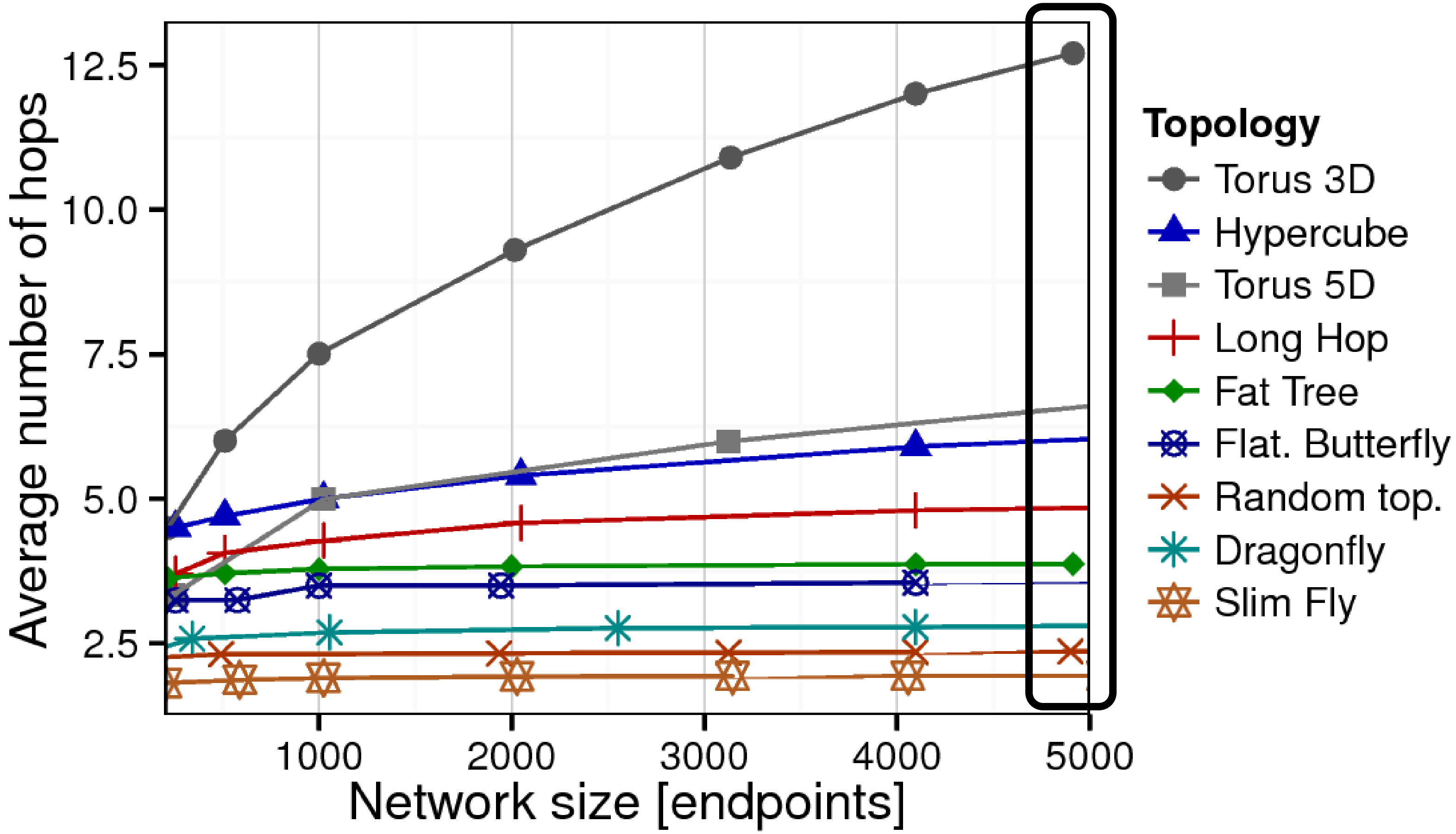
Detailed Case-Study: high-radix topologies

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**

Structure Analysis

Average Distance

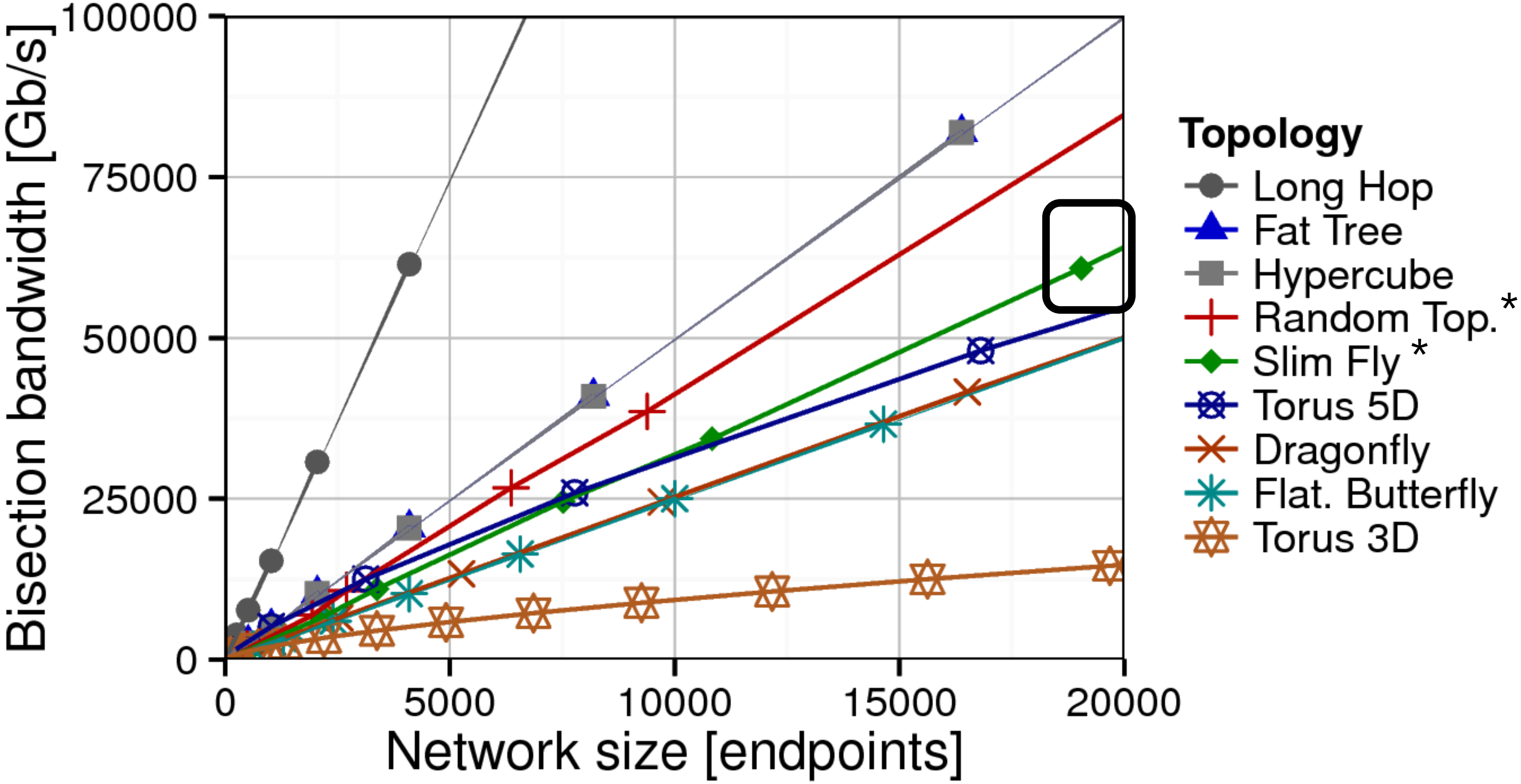
Random uniform traffic
using minimum path routing



STRUCTURE ANALYSIS

BISECTION BANDWIDTH (BB)

*BB approximated with the Metis partitioner [1]



[1] G. Karypis, V. Kumar. A fast and high quality multilevel scheme for partitioning irregular graphs. ICPP'95

Designing an Efficient Network Topology

Connecting routers: Diameter 2

6 *Intra-group connections*

Router $(0, x, y) \leftrightarrow (0, x, y')$
iff $y - y' \in X$

Router $(0, m, c) \leftrightarrow (0, m, c')$
iff $c - c' \in X'$

E *Example:* $q = 5$

$X = \{1,4\}$

Take Routers $(0,0,.)$

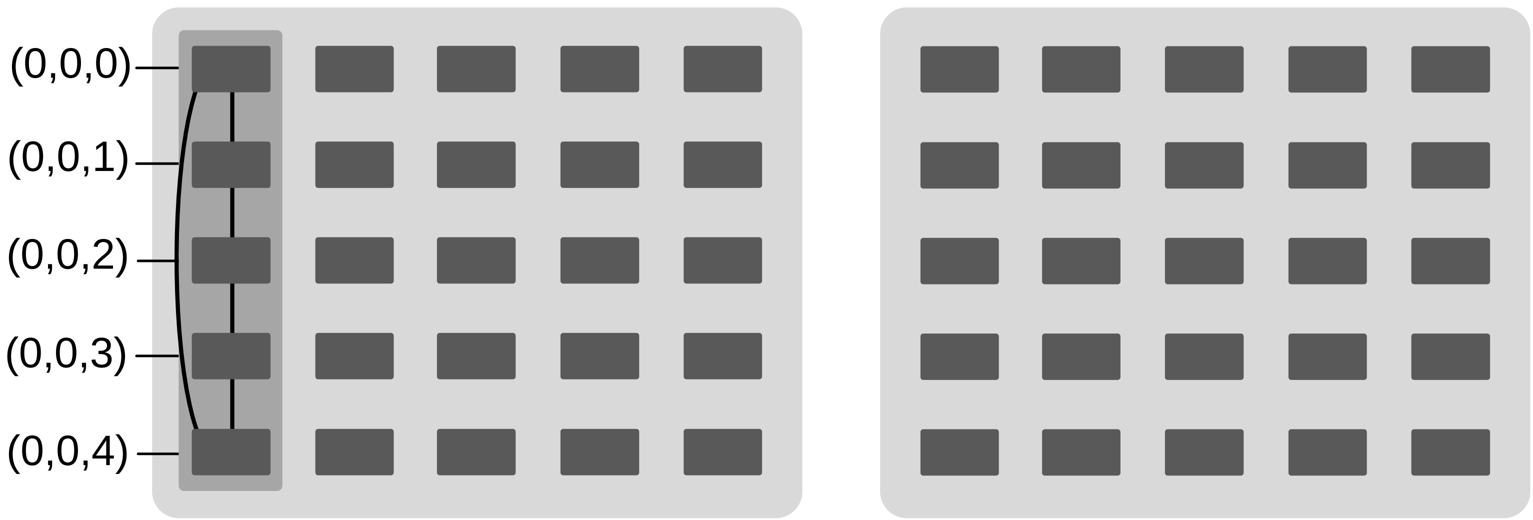
(0,0,0), (0,0,1): $y - y' = 1 \in X$ ✓

(0,0,0), (0,0,2): $y - y' = 2 \notin X$ ✗

(0,0,1), (0,0,2): $y - y' = 1 \in X$ ✓

…

(0,0,0), (0,0,4): $y - y' = 4 \in X$ ✓



Designing An Efficient Network Topology

Connecting routers: Diameter 2

6 *Intra-group connections*

Router $(0, x, y) \leftrightarrow (0, x, y')$
iff $y - y' \in X$

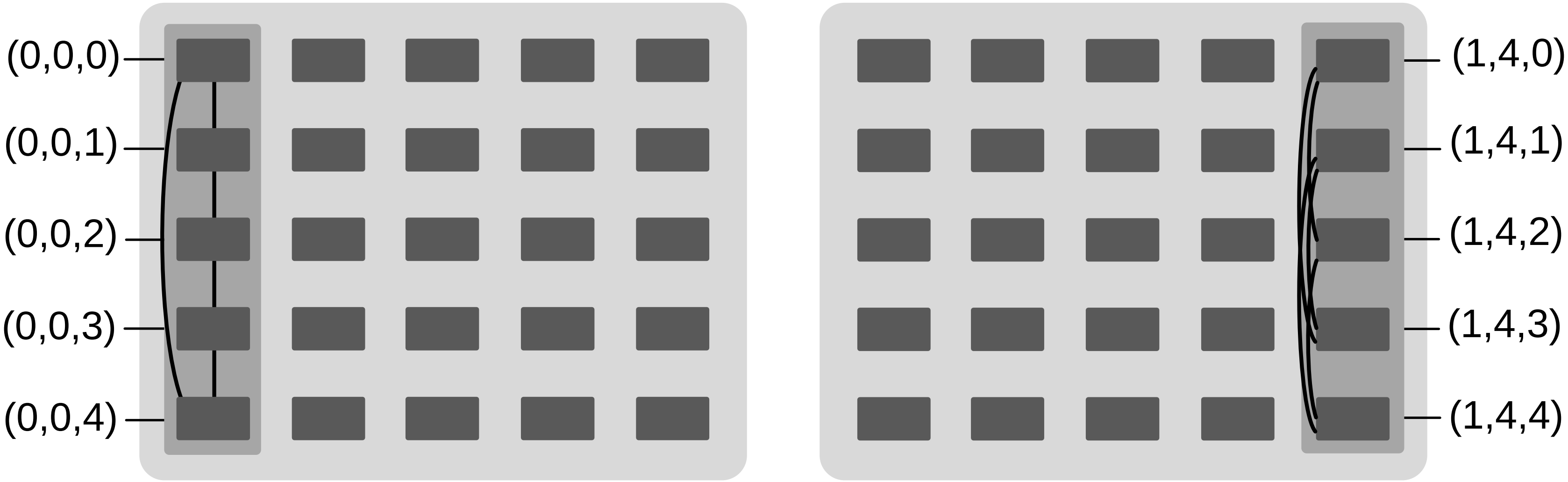
Router $(0, m, c) \leftrightarrow (0, m, c')$
iff $c - c' \in X'$

E *Example:* $q = 5$

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

Take Routers $(1,4,.)$

(1,4,0), (1,4,1): $y - y' = 1 \notin X'$ ✗
(0,0,0), (0,0,2): $y - y' = 2 \in X'$ ✓
(0,0,1), (0,0,4): $y - y' = 3 \in X'$ ✓
...
(0,0,0), (0,0,4): $y - y' = 4 \notin X'$ ✗



Designing An Efficient Network Topology

Connecting routers: Diameter 2

7 *Inter-group connections*

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

iff $\quad y = mx + c$

E *Example:* $q = 5$

Take Router $(1,1,0)$

(0,0,0): $y = 0 \quad mx + c = 0$ ✓

(0,1,1): $y = 1 \quad mx + c = 1$ ✓

(0,2,2): $y = 2 \quad mx + c = 2$ ✓

...

(0,4,4): $y = 4 \quad mx + c = 4$ ✓