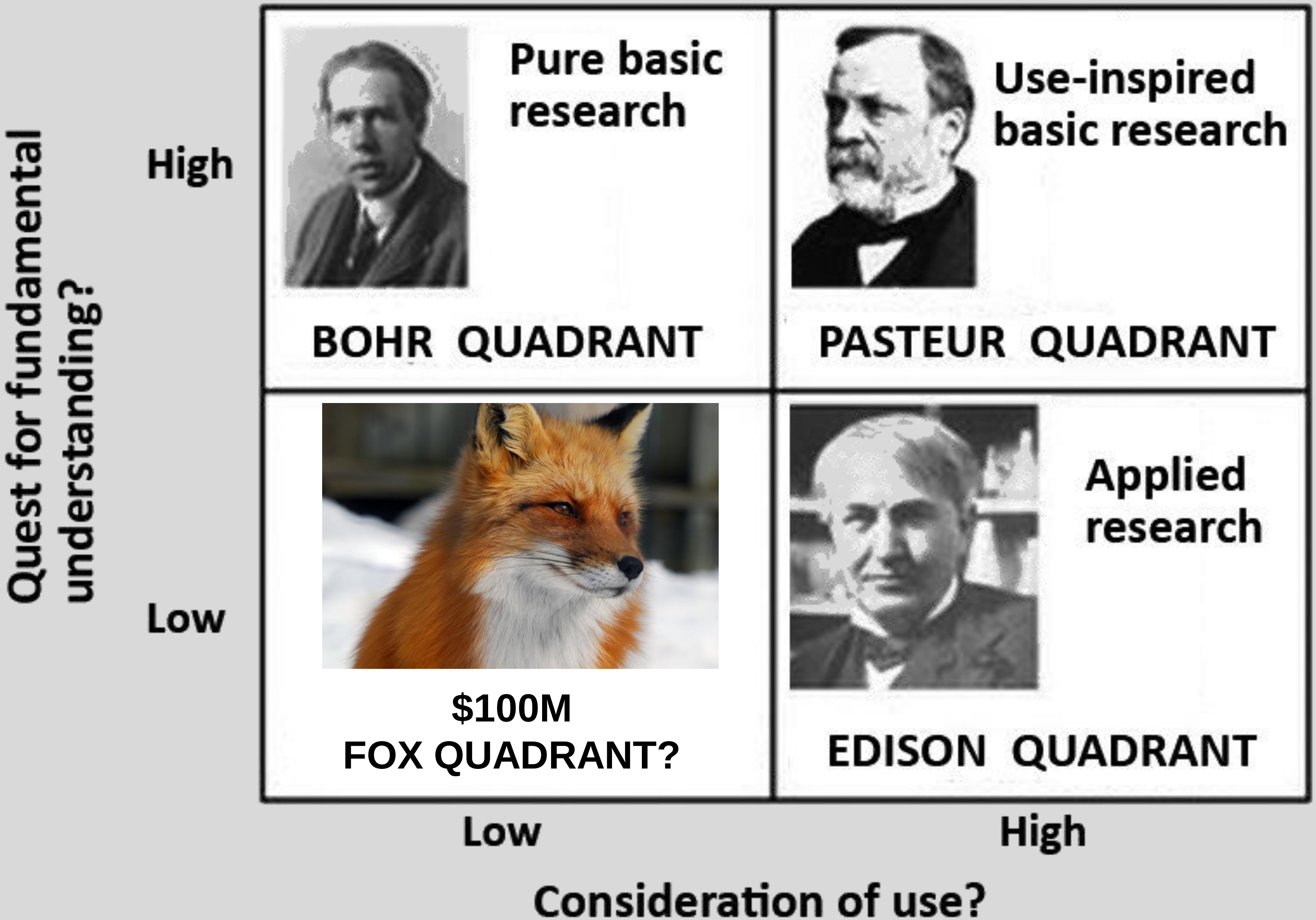
SlimFly: A Cost Effective Low-Diameter Network Topology

TORSTEN HOEFLER

WITH MACIEJ BESTA

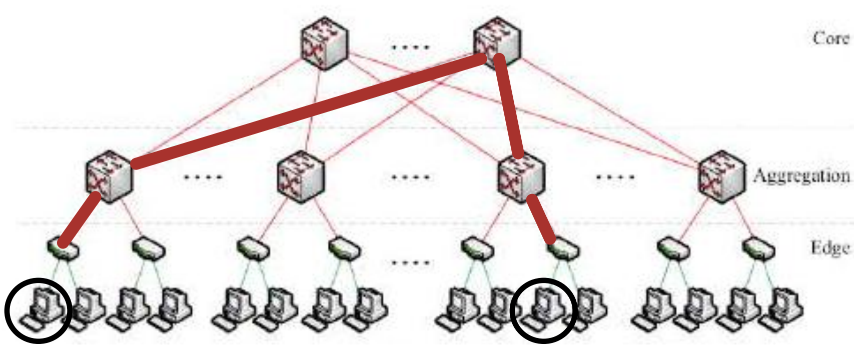
Edison’s vs. Pasteur’s quadrant



Designing an Efficient Network Topology

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

Fat tree: Diameter == 4

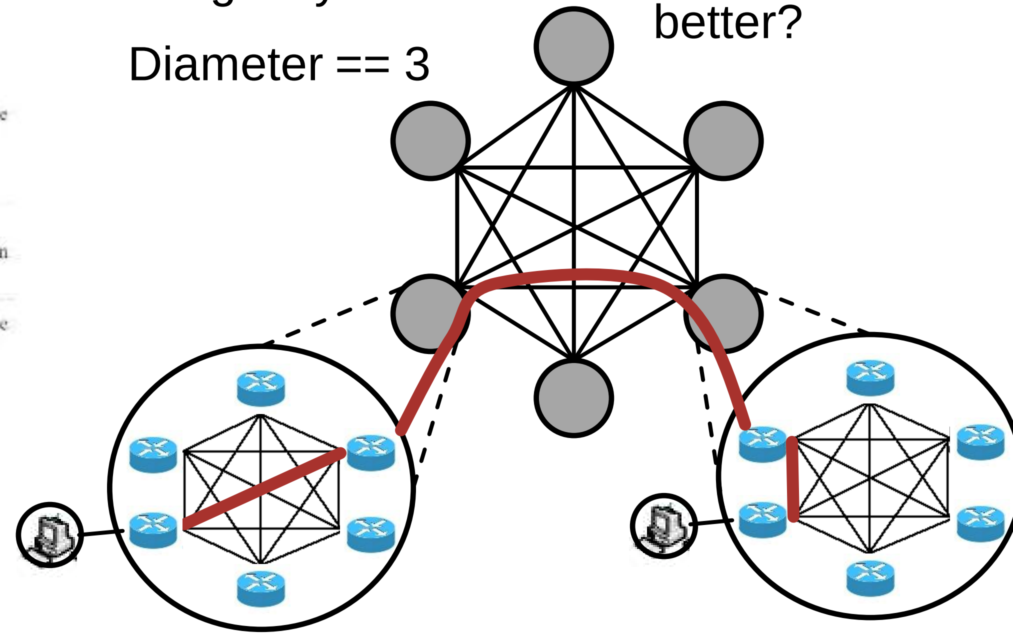


...still high ☹

Dragonfly:

Diameter == 3

...can we do better?



Designing An Efficient Network Topology

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

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

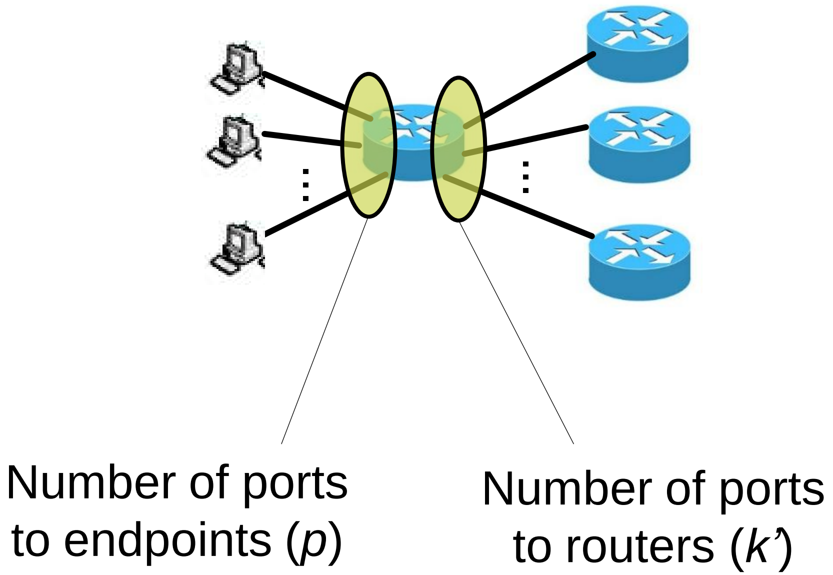
$$N_r = 1 + k' \sum_{i=0}^{D-1} (k' - 1)^i$$

$$D = 2: \qquad N_r \approx k'^2$$

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

$$D = 3: \qquad N_r \approx k'^3$$

(>10,000,000 endpoints
with 108-port switches)



Designing an Efficient Network Topology

- Degree-Diameter problem

Graph with the maximum N_r for a given D and k

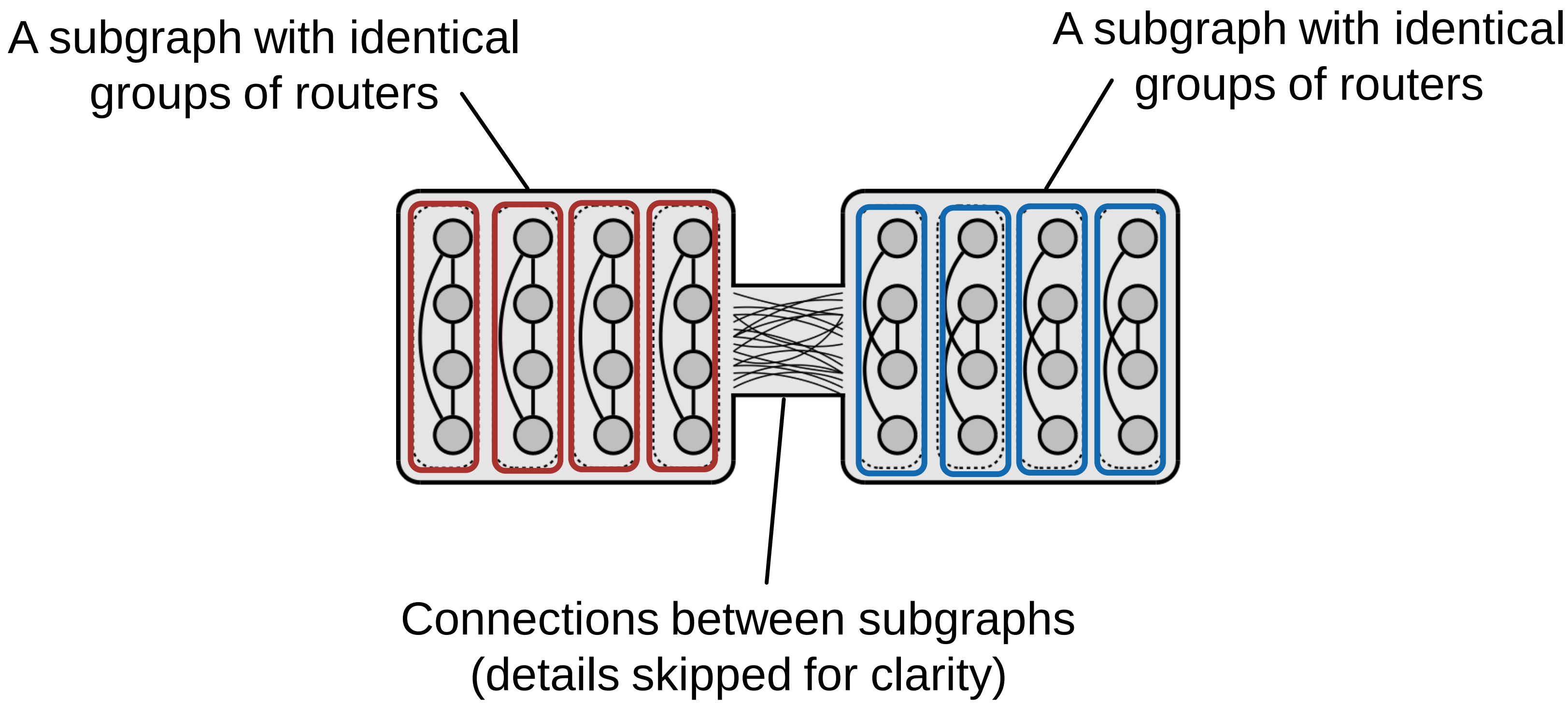
Diameter (D) →

Degree (k) ↓

	2	3	4	5	6	7	8	9	10
3	10 100%	20 100%	38 100%	92 76.08%	188 70.21%	380 51.57%	764 43.97%	1532 39.16%	3068 40.74%
4	15 100%	52 78.84%	160 61.25%	484 75.20%	1456 50.82%	4372 30.19%	13120 24.71%	39364 19.24%	118096 14.99%
5	24 100%	104 69.23%	424 50%	1704 36.61%	6824 40.62%	27304 20.20%	109224 15.59%	436904 13.23%	1747624 10.70%
6	32 100%	186 59.67%	936 41.66%	4686 29.96%	23436 33.78%	117186 16.54%	585936 13.04%	2929686 10.50%	14648436 8.55%
7	50 100%	301 55.81%	1813 37.06%	10885 25.31%	65317 18.35%	391909 13.46%	2351461 10.61%	14108773 8.66%	84652645 7.09%
8	63 90.47%	456 55.48%	3200 34.37%	22408 22.58%	156864 25.29%	1098056 11.94%	7686400 9.56%	53804808 7.88%	376633664 6.61%
9	80 92.50%	657 89.04%	5265 29.43%	42129 19.46%	337041 22.51%	2696337 10.37%	21570705 7.81%	172565649 7.02%	1380525201 4.77%
10	99 91.91%	910 71.42%	8200 27.87%	73810 17.80%	664300 20.27%	5978710 9.75%	53808400 7.97%	484275610 5.78%	4358480500 4.61%

Designing An Efficient Network Topology

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

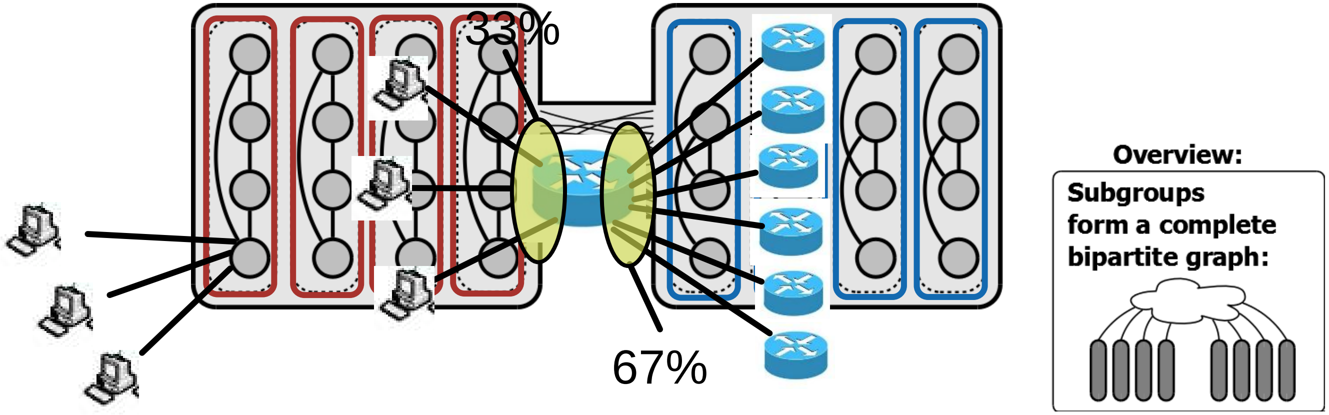


[1] B. D. McKay, M. Miller, and J. Siran. A note on large graphs of diameter two and given maximum degree. Journal of Combinatorial Theory, Series B, 74(1):110 – 118, 1998

Attaching Endpoints

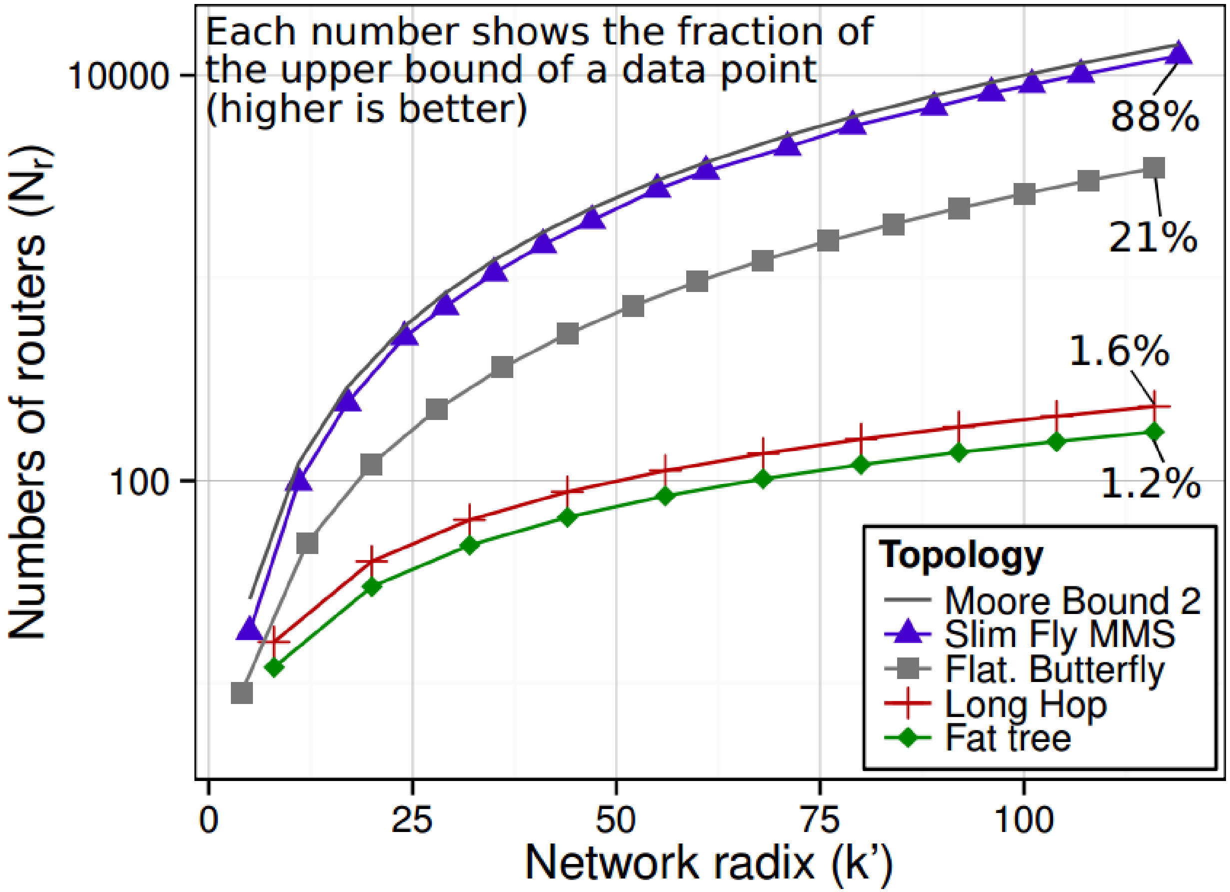
- How many endpoints do we attach to each router?
- Maximize for *p* while maintaining full global bandwidth
  - Global bandwidth: the theoretical cumulative throughput if all processes simultaneously communicate with all other processes in a steady state
  - Result:

$$p = \left\lceil \frac{k'}{2} \right\rceil$$



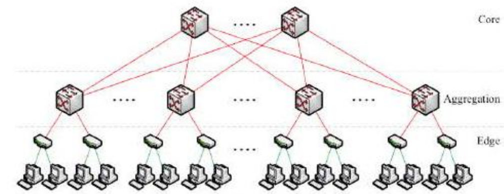
Comparison to Optimality

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

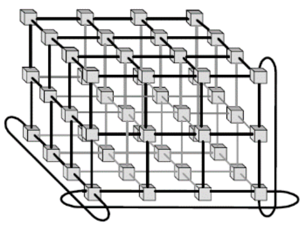


STRUCTURE ANALYSIS

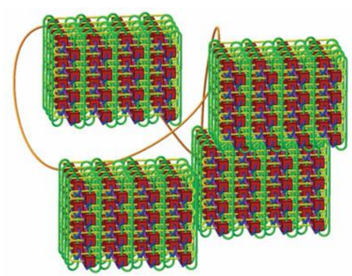
COMPARISON TARGETS



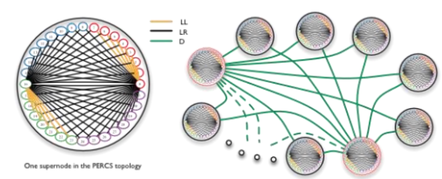
Fat tree



Torus 3D

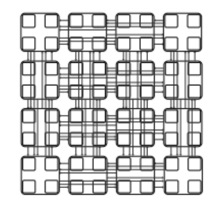


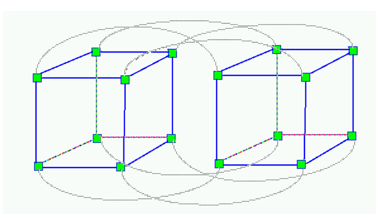
Torus 5D



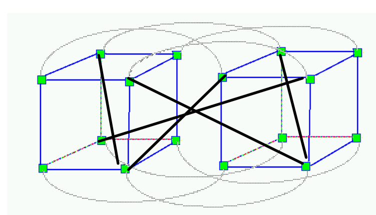
Dragonfly

Flattened Butterfly

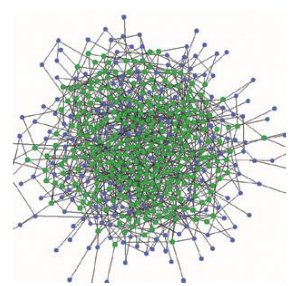




Hypercube



Long Hop [1]



Random networks

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

STRUCTURE ANALYSIS

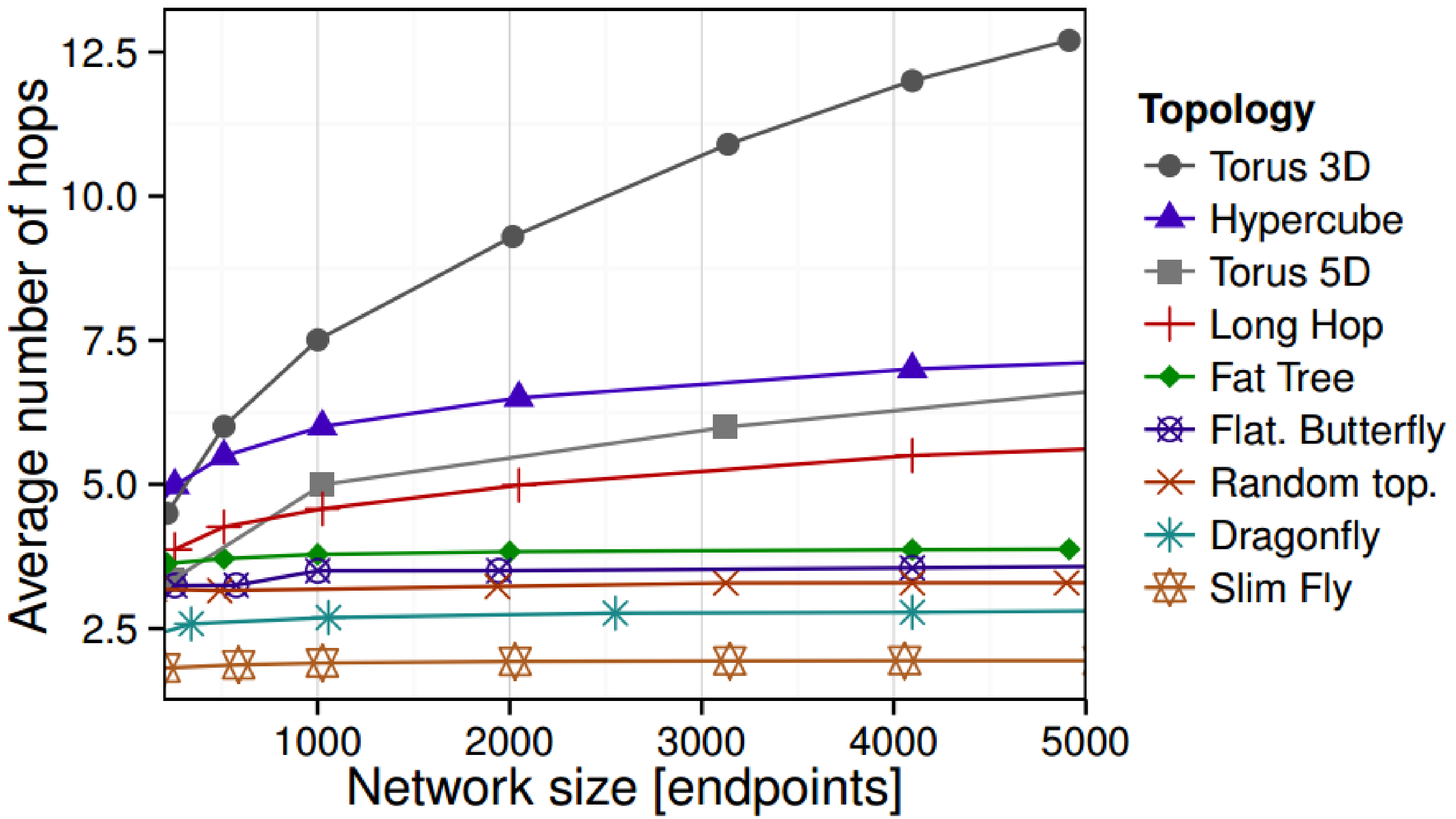
DIAMETER

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

Structure Analysis

Average Distance

Random uniform traffic
using minimum path routing



STRUCTURE ANALYSIS

RESILIENCY

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

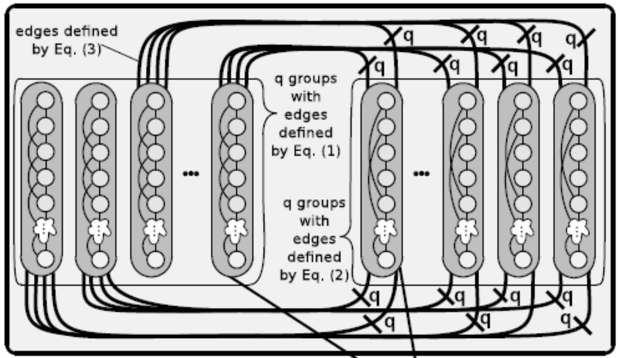
- Disconnection metrics*
- Other studied metrics (N≈8192):
  - Diameter (increase by 2) [1]; SF: 40%, DF: 25%, DLN: 60%
  - Average path length (increase by 2); SF: 55%, DF: 45%, DLN: 60%

$\approx N$	T3D	T5D	HC	LH-HC	FT-3	DF	FBF-3	DLN	**SF**
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%**

[1] M. Koibuchi et al. A case for random shortcut topologies for HPC interconnects. ISCA'12

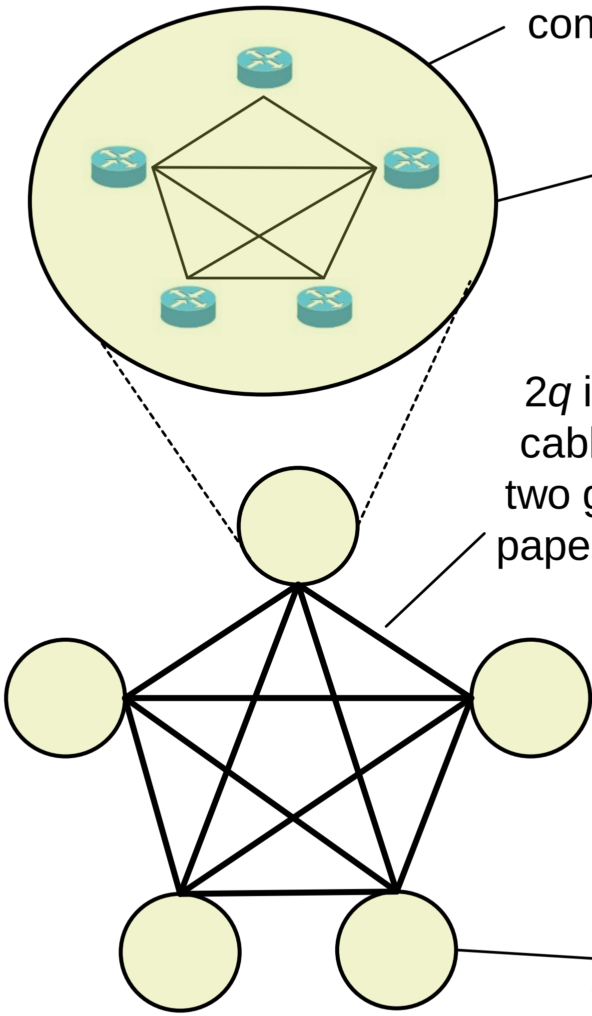
PHYSICAL LAYOUT

Step 1: generate an MMS graph

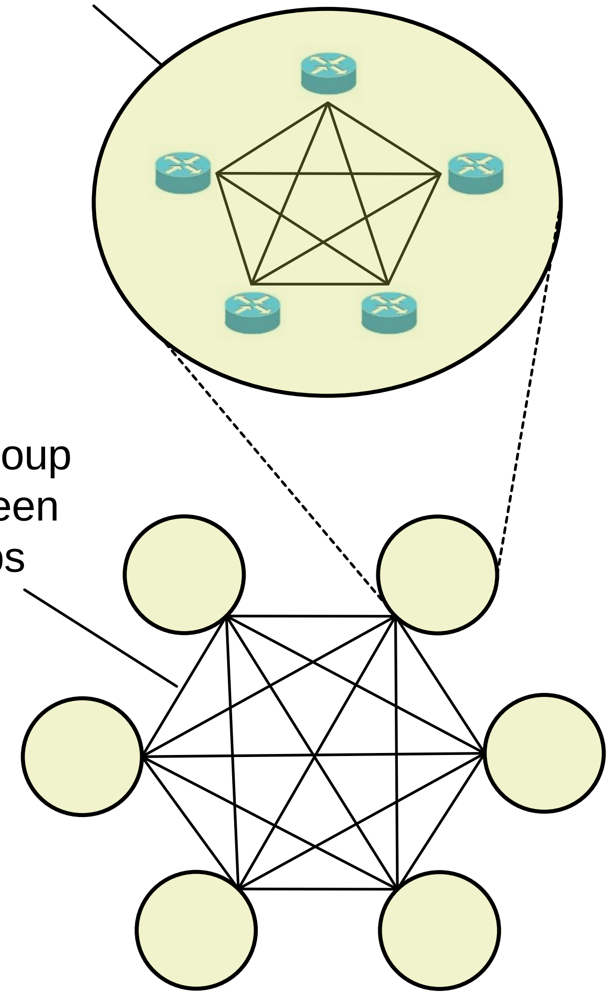


COMPARISON TO DRAGONFLY

SlimFly:



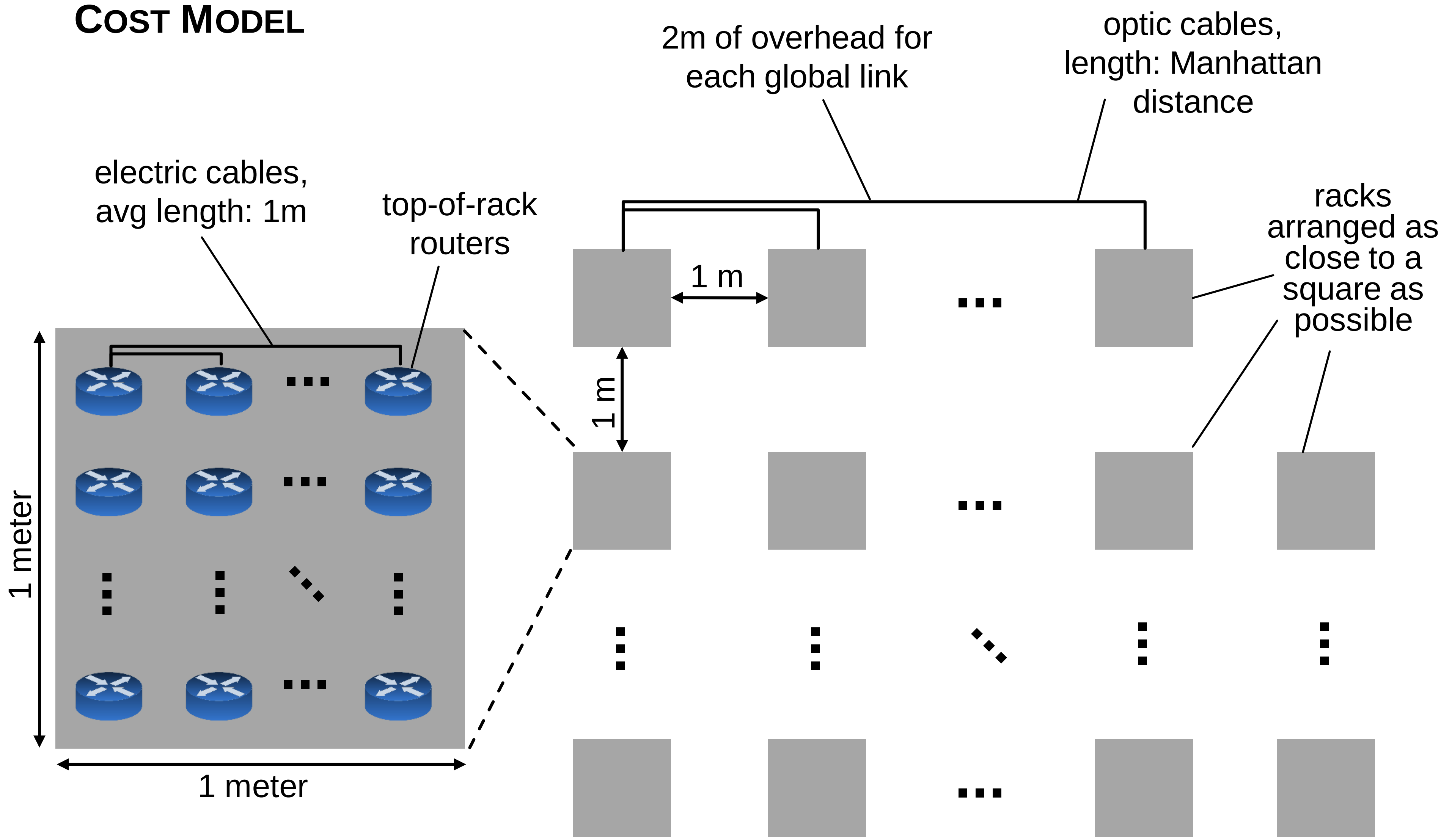
Dragonfly:



Cost Comparison

Cost Model

*Most cables skipped for clarity

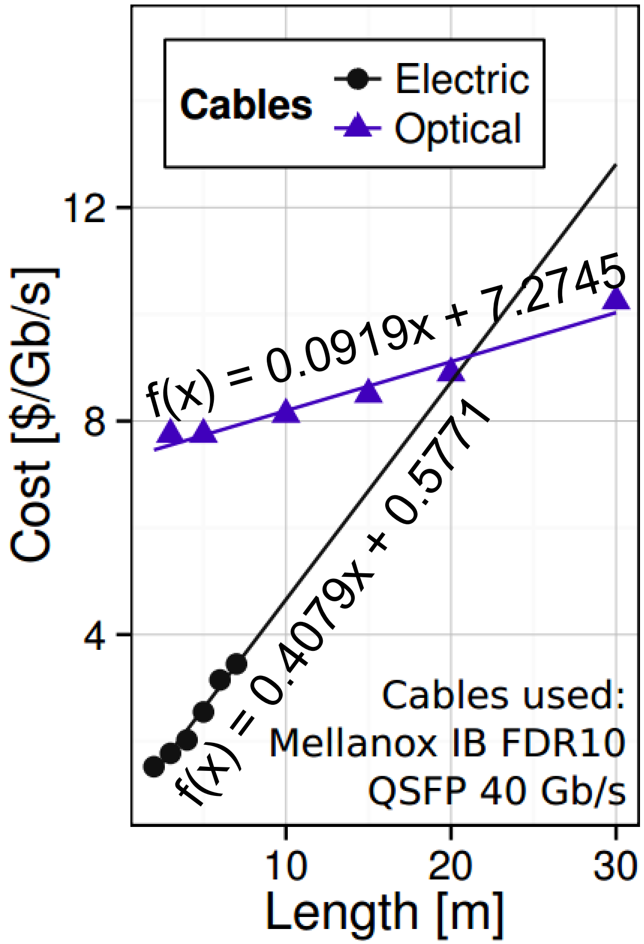


J. Kim, W. J. Dally, and D. Abts. Flattened Butterfly: A Cost-efficient Topology for High-radix Networks. ISCA '07

Cost Comparison

Cable Cost Model

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

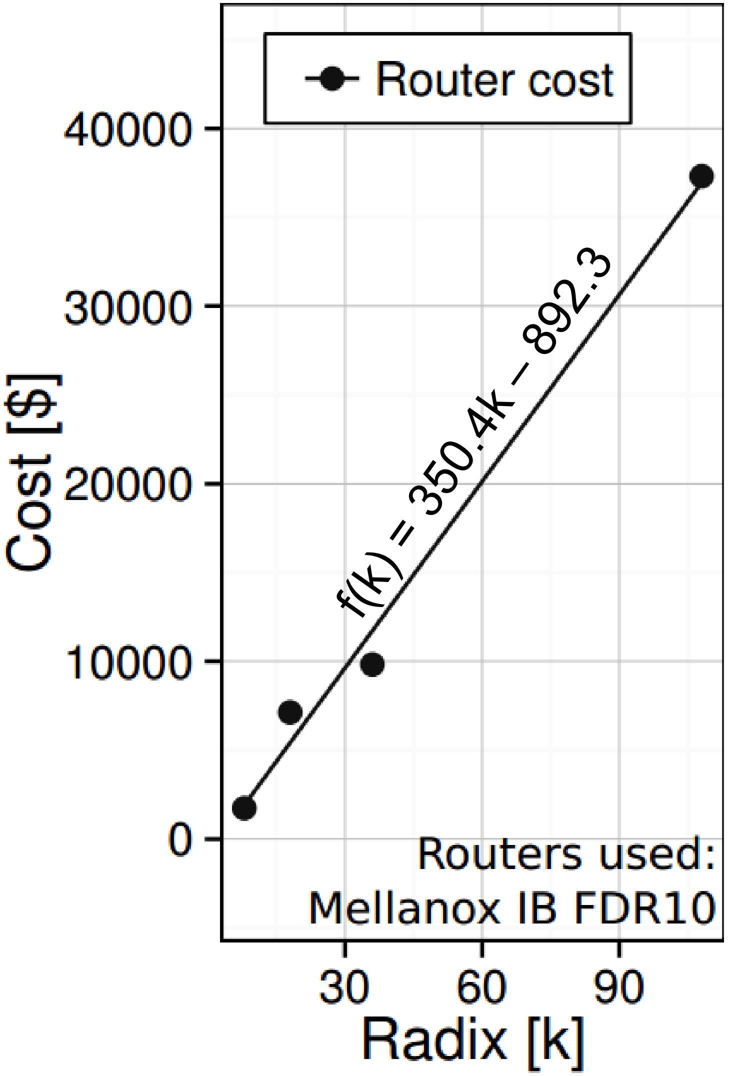


*Prices based on ColfaxDirect, June 2014

Cost Comparison

Router Cost Model

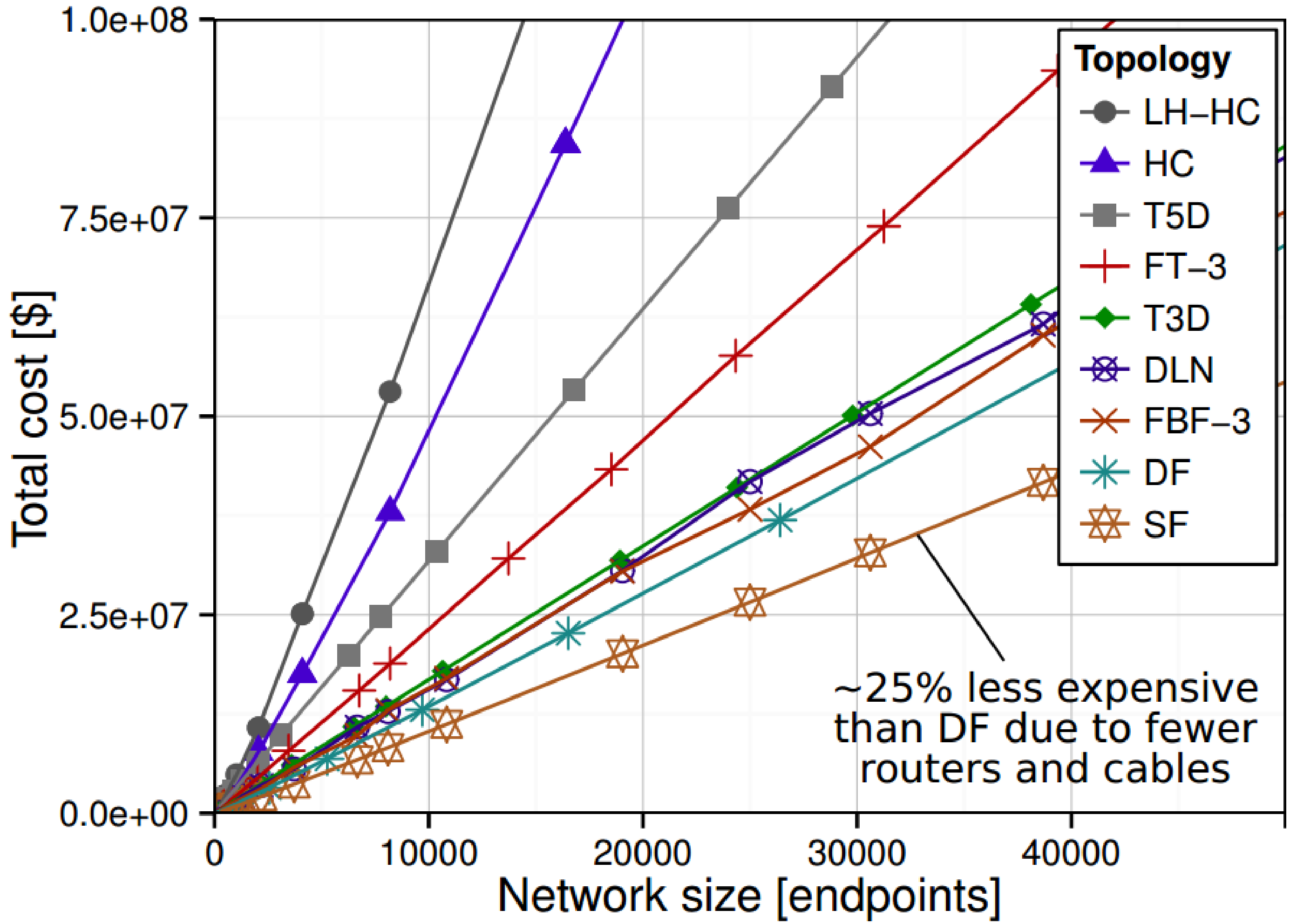
- Router cost as a function of radix
  - The function obtained using linear regression*
- Used routers:
  - Mellanox Ethernet 10/40Gb



*Prices based on ColfaxDirect, June 2014

Cost Comparison

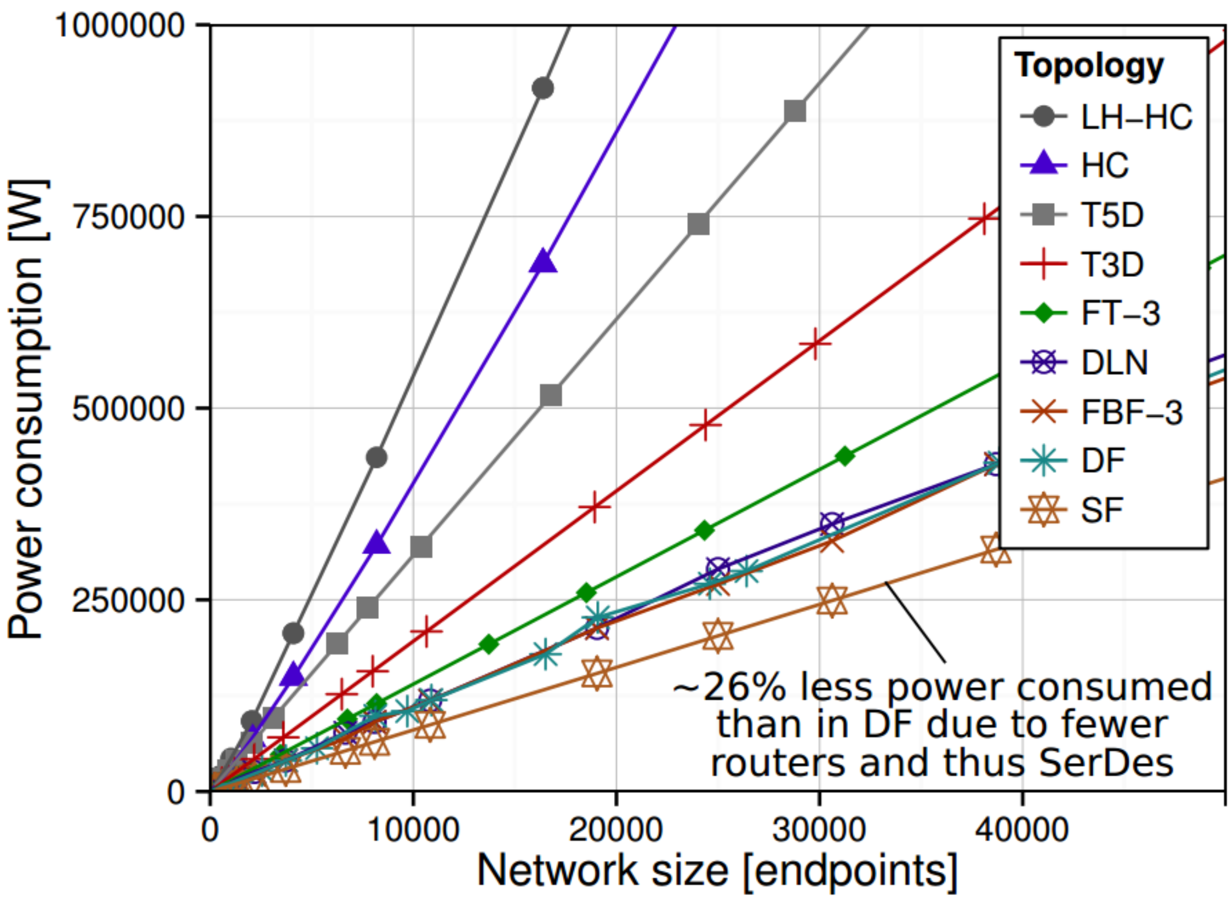
Results



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

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

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

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]

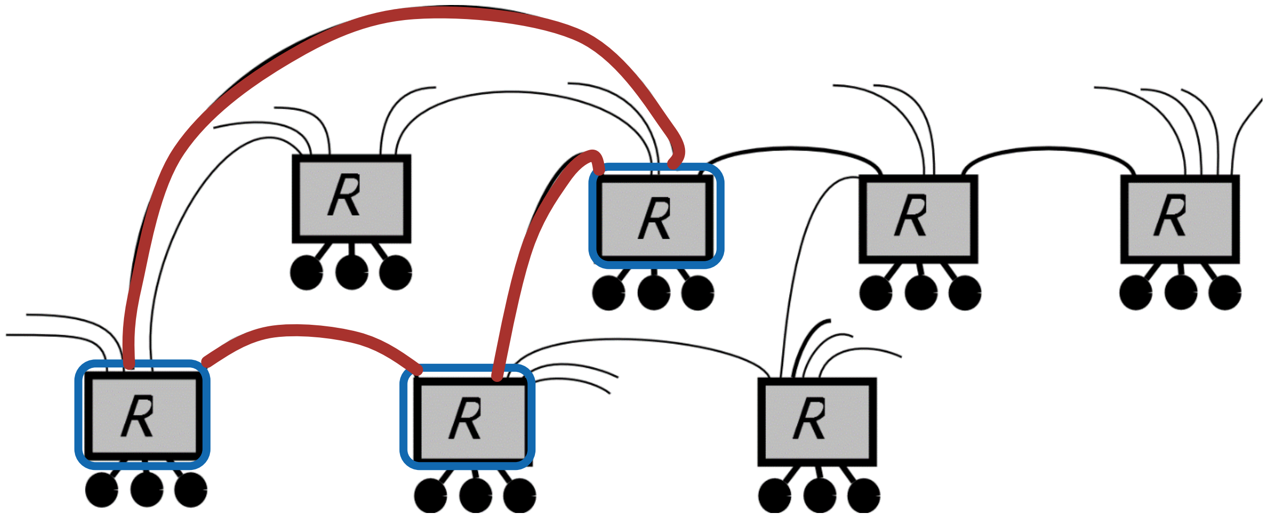
[1] J. Domke, T. Hoefler, and W. Nagel. Deadlock-Free Oblivious Routing for Arbitrary Topologies. IPDPS'11

PERFORMANCE

- Cycle-based flit-level simulations (Booksim)
- Routing protocols:
  - Minimum static routing
  - Valiant's random routing
  - Universal Globally-Adaptive Load-Balancing routing

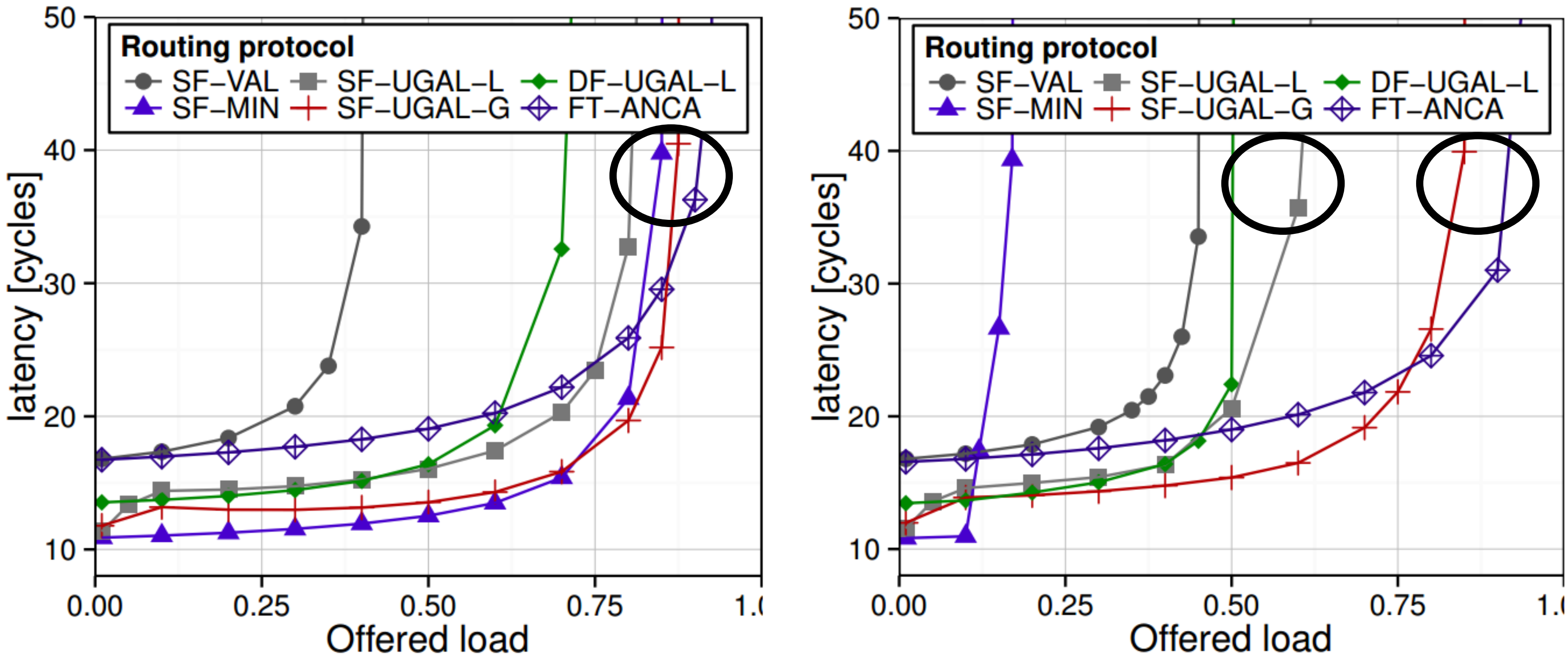
    UGAL-L: each router has access to its local output queues

    UGAL-G: each router has access to the sizes of all router queues in the network



PERFORMANCE

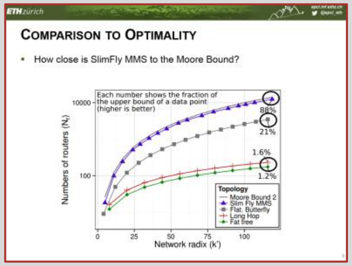
- Random uniform traffic
- Bit permutation (reverse) traffic



Conclusions

Topology design

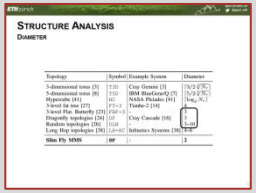
Optimizing towards the Moore Bound reduces expensive network resources



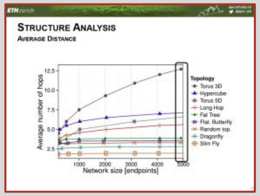
Google PhD fellowship for parallel computing

Advantages of SlimFly

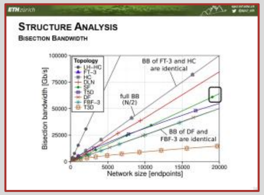
Diameter



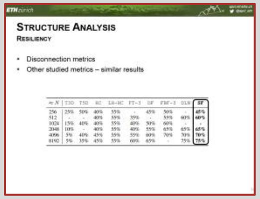
Avg. distance



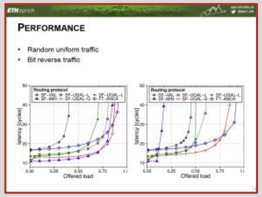
Bisection bandwidth



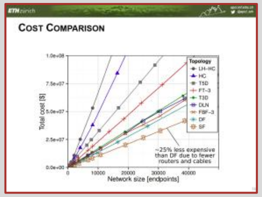
Resilience



Performance



Cost & power



Optimization approach

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