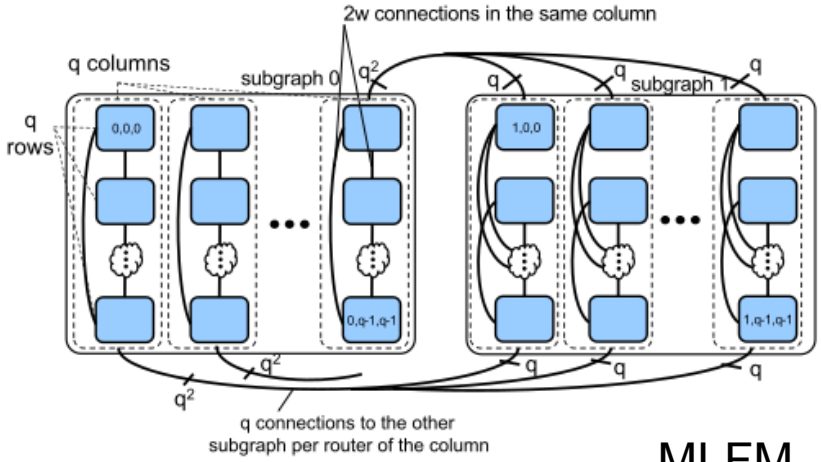
TORSTEN HOEFLER

HiPINEB'16 - Panel

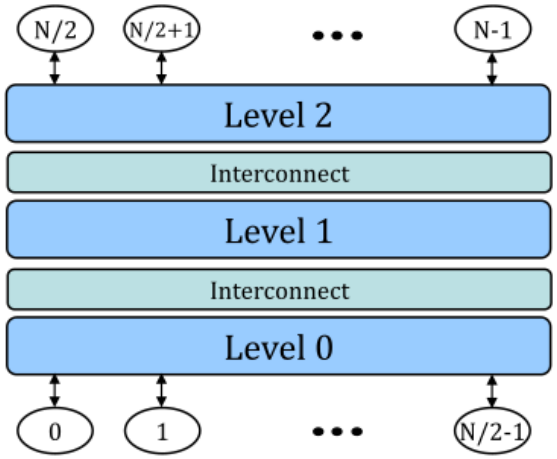
What's your view regarding topologies and routings suitable for very-large interconnection networks? - Topologies

- **Two main criteria: (1) cost (energy and $s) and (2) throughput**
  (1) & (2) Will force us to lowest diameter 2

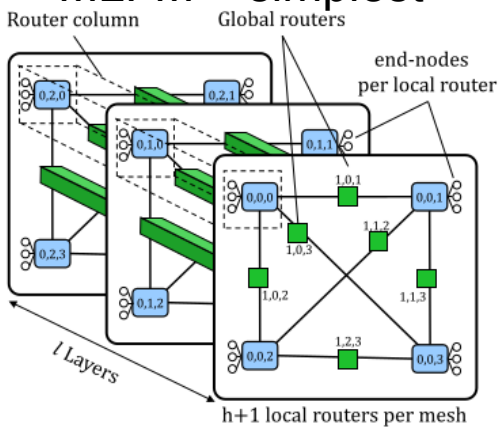
Slim Fly - best direct



OFT - best indirect

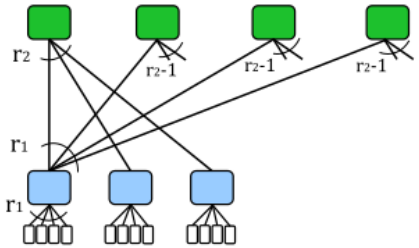


MLFM – simplest

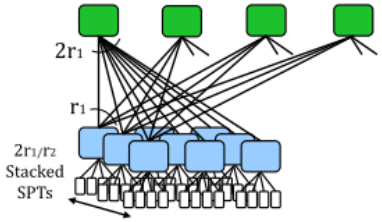


What's your view regarding topologies and routings suitable for very-large interconnection networks? - Topologies

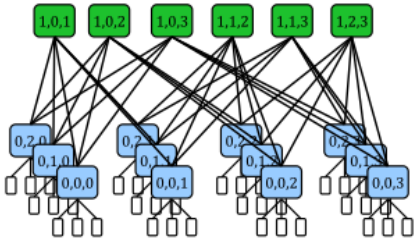
- **In fact, all diameter-2 topologies can be drawn as trees [1]**



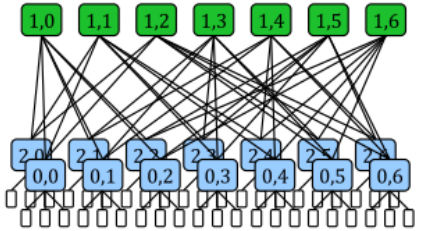
(a) SPT.



(b) SSPT.



(c) 3,3,3-MLFM.



(d) 3-OFT.

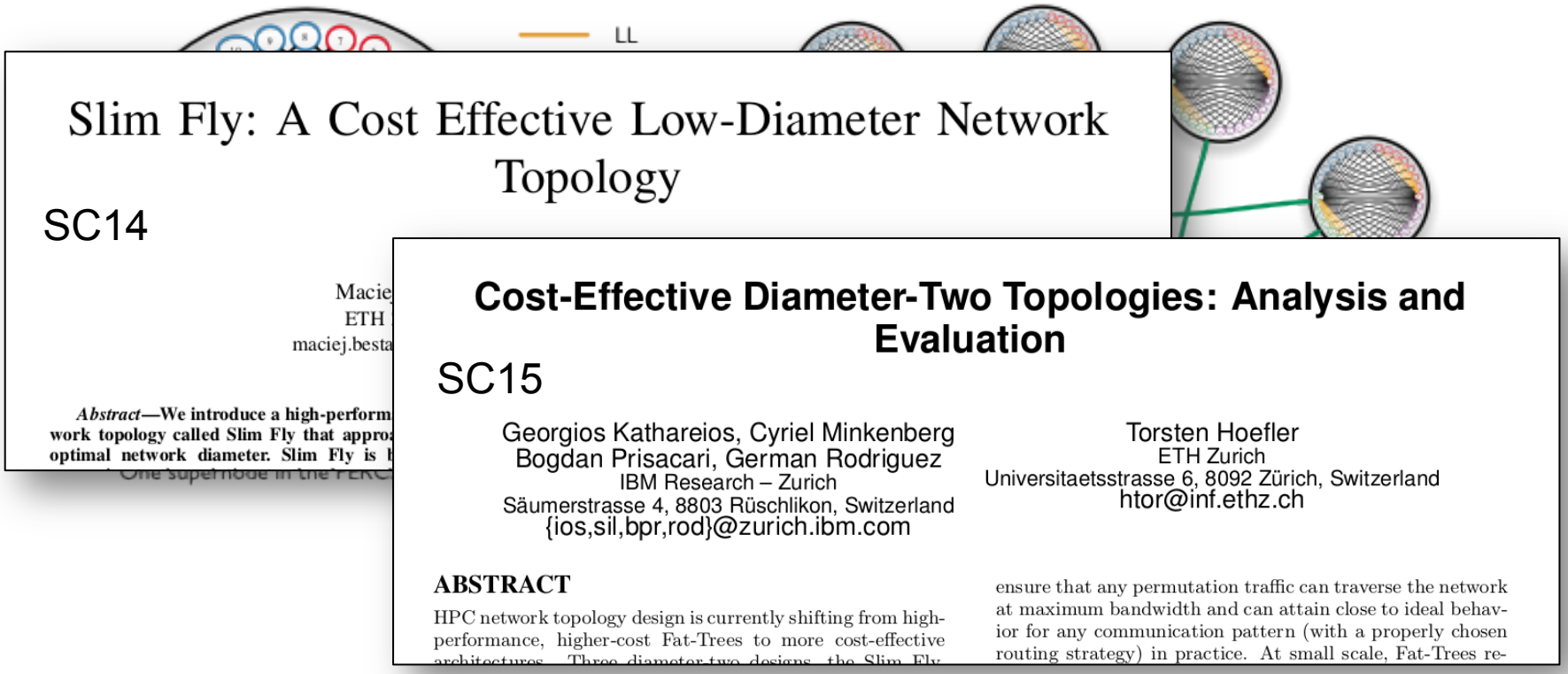
	Topology	Diam.	Scale	$\frac{N_l}{N}$	$\frac{N_p}{N}$
Direct	2D HyperX	2	$\approx r^3/27$	2	3
	Slim Fly (SF)	2	$\approx r^3/8$	2*	3*
Indirect	2-lvl Fat-Tree	2	$r^2/2$	2	3
	3-lvl Fat-Tree	4	$r^3/4$	3	5
	MLFM	2	$\approx r^3/8$	2	3
	OFT	2	$\approx r^3/4$	2	3

[1] Kathareios et al.: Cost-Effective Diameter-Two Topologies: Analysis and Evaluation, SC15

What's your view regarding topologies and routings suitable for very-large interconnection networks? - Topologies

- **With more money, go diameter 3 (where we are today)**

  Less (re-)engineering overhead



Arimilli et al.: "The PERCS High-Performance Interconnect" HOTI'10

What's your view regarding topologies and routings suitable for very-large interconnection networks? - Topologies

- **Btw., failures are NOT a big problem – Fail in Place**
  - *The problems are in the routing and system software*
  - *Seen some nice talks here ☺*

Fail-in-Place Network Design: Interaction between Topology, Routing Algorithm and Failures

SC14

Jens Domke
Global Scientific Information and Computing Center
Tokyo Institute of Technology, Japan
Email: domke.j.aa@m.titech.ac.jp

Torsten Hoefler
Computer Science Department
ETH Zurich, Switzerland
Email: htor@inf.ethz.ch

Satoshi Matsuoka
Global Scientific Information and Computing Center
Tokyo Institute of Technology, Japan
Email: matsu@is.titech.ac.jp

***Abstract*—The growing system size of high performance computers results in a steady decrease of the mean time between failures. Exchanging network components often requires whole system downtime which increases the cost of failures. In this work, we study a fail-in-place strategy where broken network elements remain untouched. We show, that a fail-in-place strategy**

scale data centers with millions of hard drives. For example, Microsoft owned approximately one million servers in 2013, i.e., even an optimistic failure rate of 1% per year and two hypothetically hard drives per server would result in a mean time between failure of 26 minutes. Instead of replacing the hard drives of a server, the storage system, such as IBM's

Domke et al.: "Fail-in-Place Network Design: Interaction between Topology, Routing Algorithm and Failures", SC14

What's your view regarding topologies and routings suitable for very-large interconnection networks? - Routing

- **The structure of low diameter networks forces adaptive routing**
  Small number of shortest paths – need to take longer ones (sometime)

TARA

Myricom

Probing Adaptive Routing [1]

UGAL [2]

[1]: Geoffray, TH: “Adaptive Routing Strategies for Modern High Performance Networks “, HOTI’08
[2]: Singh: “Load-balanced routing in interconnection networks”, PhD Thesis Stanford, 2005

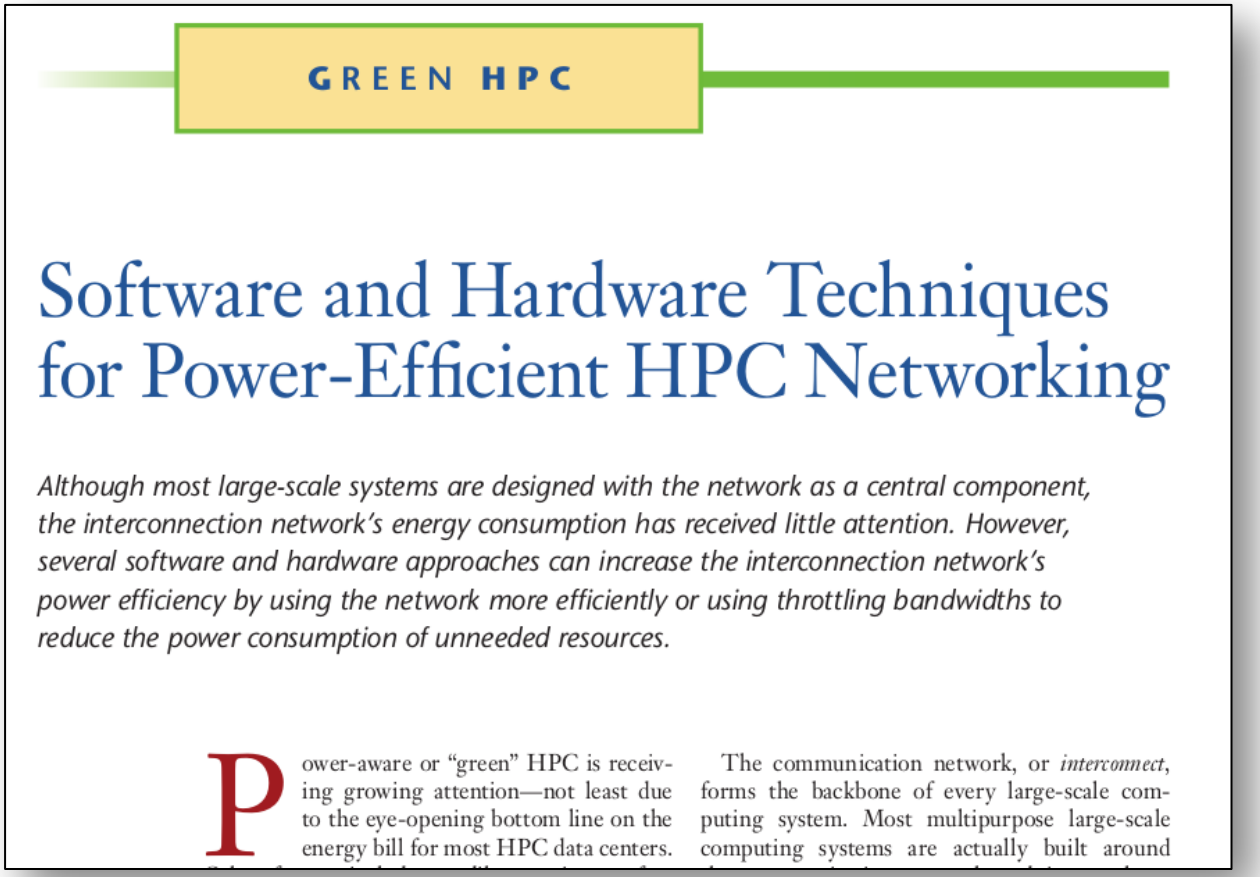
How could congestion be managed in huge Exascale Supercomputers or Big-Data systems?

- **Reactive (TCP, IB) or static (maybe IB, research)**
  Very hard topic – underlying problem (MCF) is challenging
- **Maybe need new approaches? Interesting research topic!**

Will power management techniques become mandatory in huge Exascale or Big-Data systems?

- **YES – but less clear if it's needed at the interconnect level ☺**

  Power-proportional networking anyone?



GREEN HPC

Software and Hardware Techniques for Power-Efficient HPC Networking

Although most large-scale systems are designed with the network as a central component, the interconnection network's energy consumption has received little attention. However, several software and hardware approaches can increase the interconnection network's power efficiency by using the network more efficiently or using throttling bandwidths to reduce the power consumption of unneeded resources.

Power-aware or "green" HPC is receiving growing attention—not least due to the eye-opening bottom line on the energy bill for most HPC data centers.

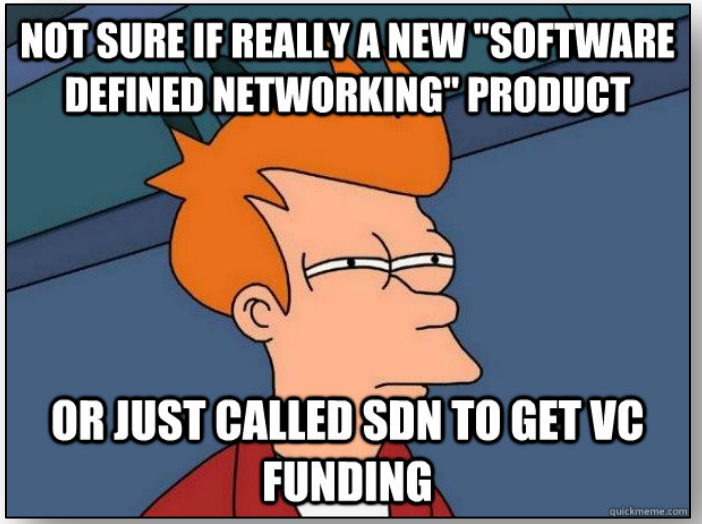
The communication network, or *interconnect*, forms the backbone of every large-scale computing system. Most multipurpose large-scale computing systems are actually built around

What's your opinion about the use of either SDN or locally-adaptive policies in networks of Exascale or Big-Data systems, especially considering Network Scalability?

- **SDN = Ethernet (OMG!)**
  - Definition fuzzy: something central, reactive routing to flows

    Somewhere between IB (TARA) and locally adaptive (decentral)

- **What is a "flow" anyway???**
  - HPC, Big Data operates with messages or remote accesses



- **Remember the topologies …**
  - Diameter-2 opens offers many research angles …

    E.g., look-ahead adaptive [1]

- **I bet on a mix of both skewed towards locally adaptive**

  Many open research topics if you're a grad student!

[1]: Geoffray, TH: "Adaptive Routing Strategies for Modern High Performance Networks ", HOTI'08