HOW TO BUILD A BETTER TESTBED

LESSONS FROM A DECADE OF NETWORK EXPERIMENTS ON EMULAB

Fabien Hermenier

Université Nice Sophia Antipolis

Robert Ricci

THE UNIVERSITY OF UTAH

TridentCom' 2012

Network testbeds

- support for experiments in networked systems
- highly customizable networked environments
- raw access to a variety of specific hardware
- a physical design and features to match the needs

Testbed physical design

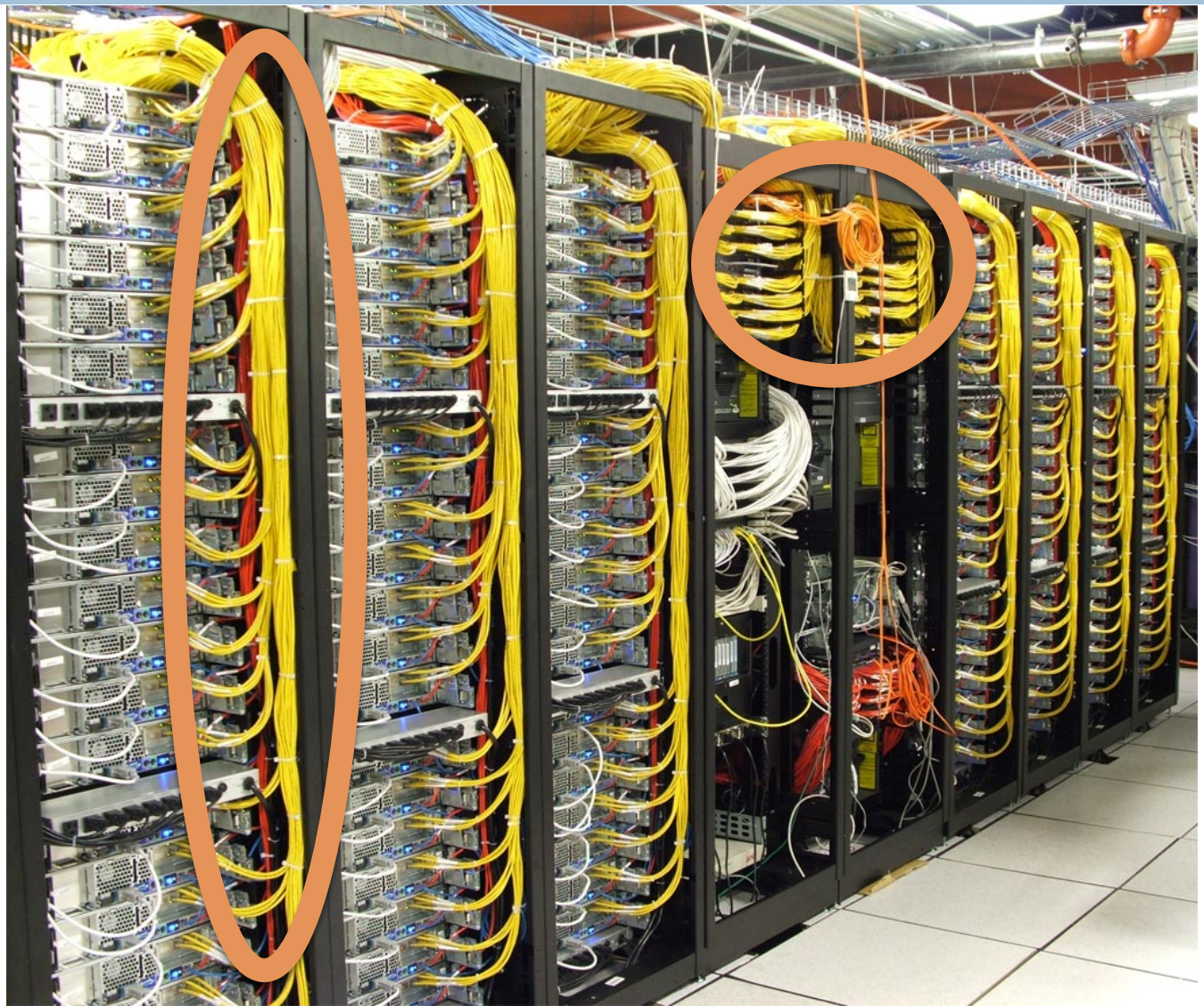
- no extensive studies for effective physical testbed design
  - assumption-based
  - budget-constrained
- lack of data from real experimenters

- bad design decisions have consequences
  - prevent support for certain experiments
  - over-commitment on un-needed hardware

How to build better testbeds ?

- careful analysis of the Utah Emulab facility usage
  - one of the largest testbeds
  - used in production since 2001
  - > 4 dozen testbeds worldwide with a similar designs
- several alternative testbed designs
- evaluation of new designs using real workload

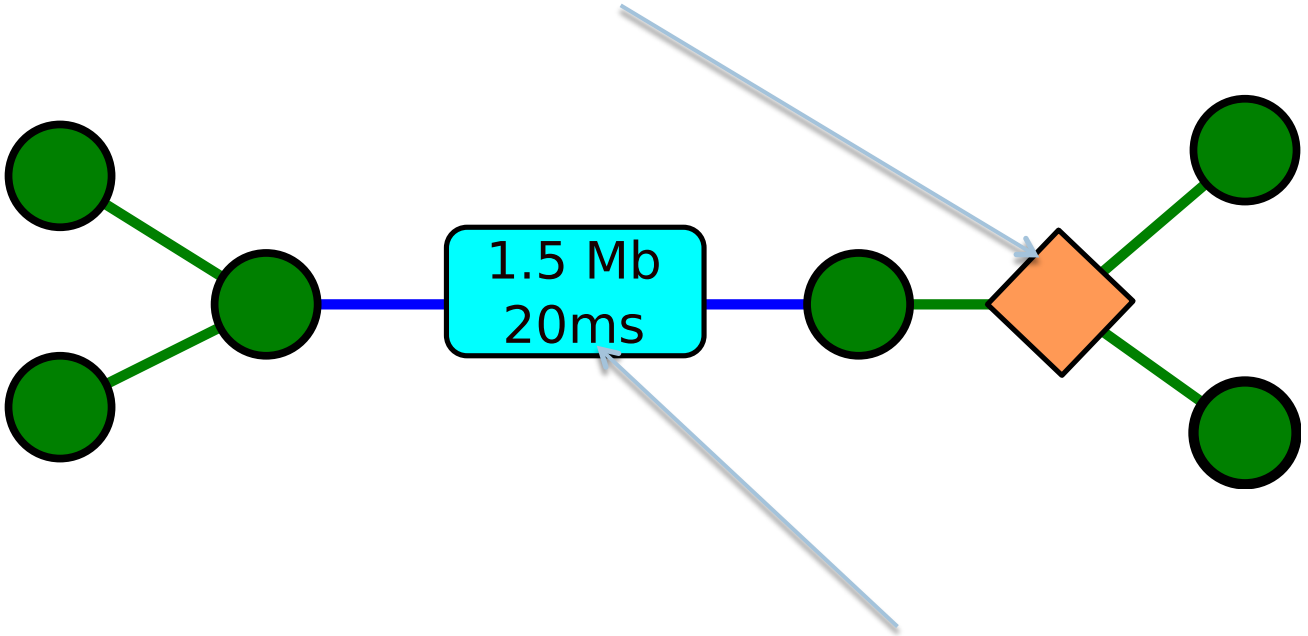
The Utah Emulab facility: not just another pretty cluster



Virtual topology in Emulab

Lan node

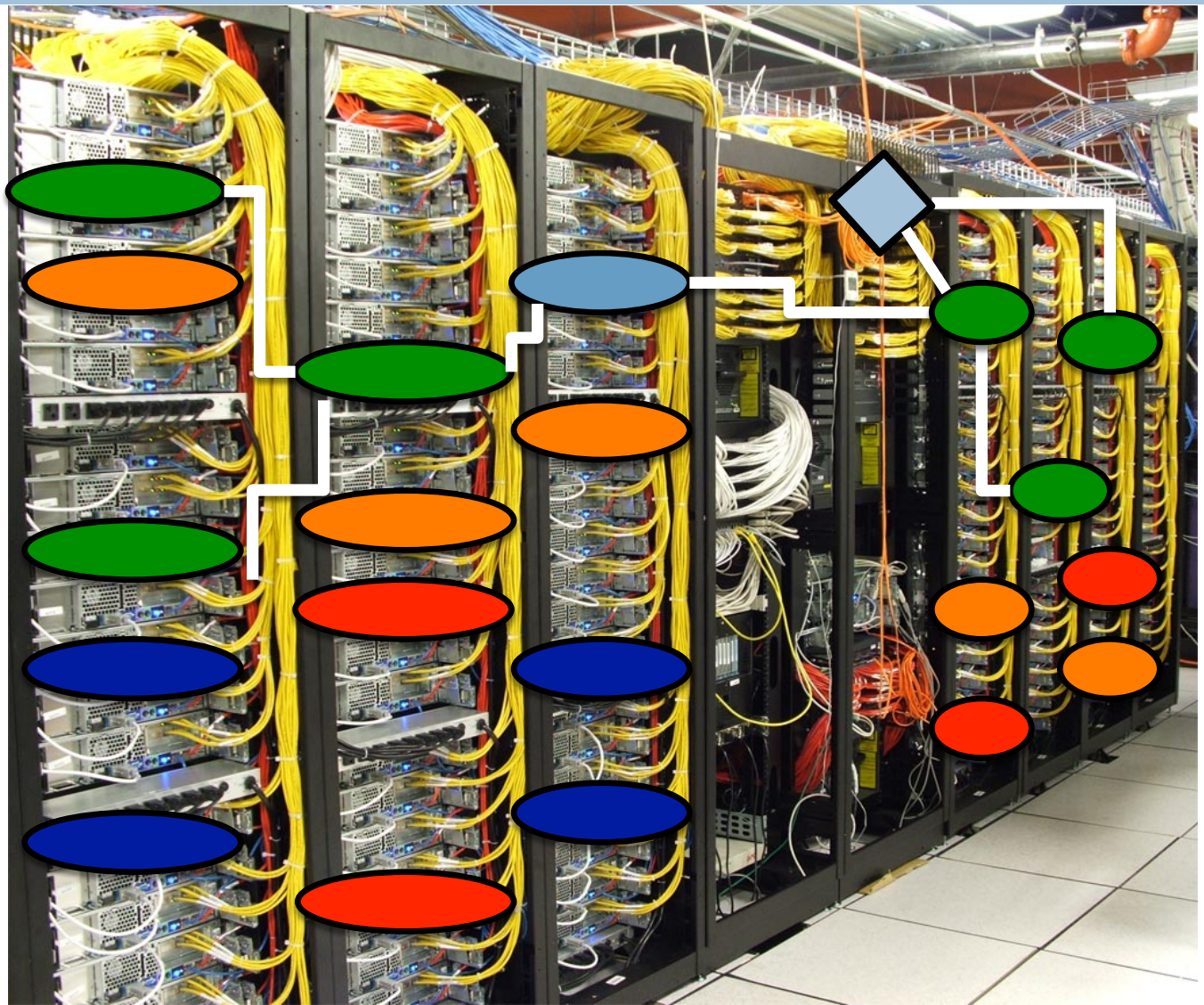
Full bisection bandwidth between members

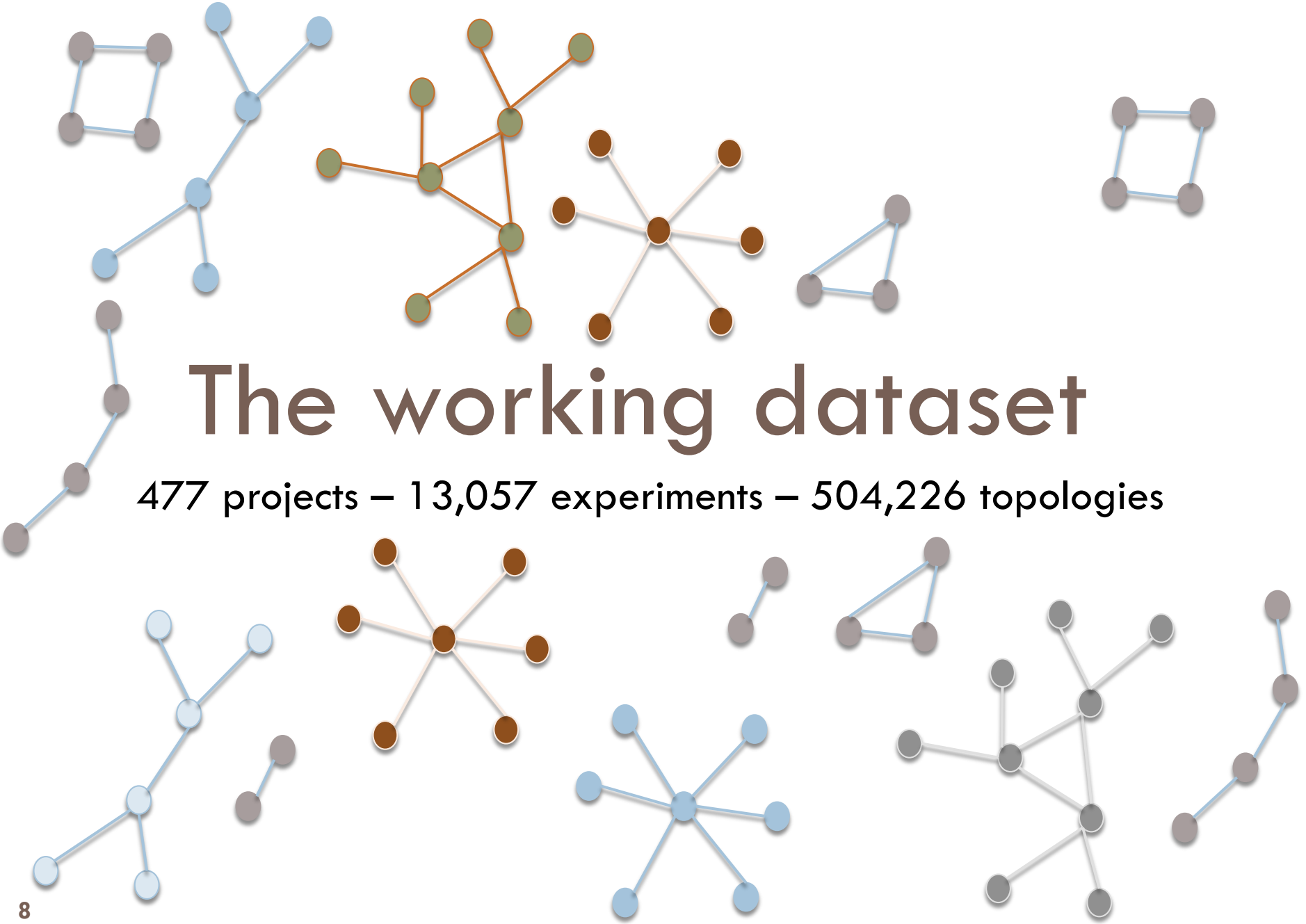


Traffic shaping

Implemented with a PC “delay node”

Making the virtual into reality

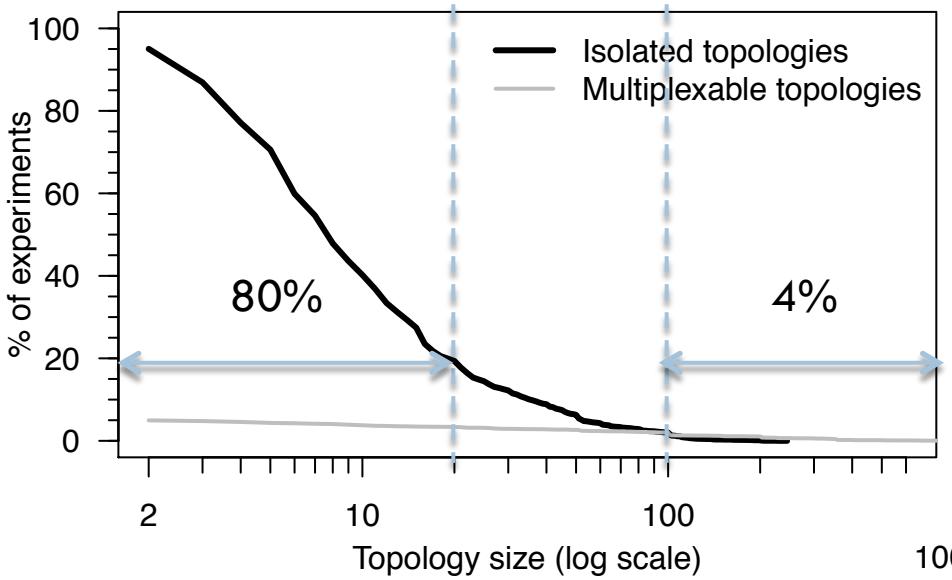




The working dataset

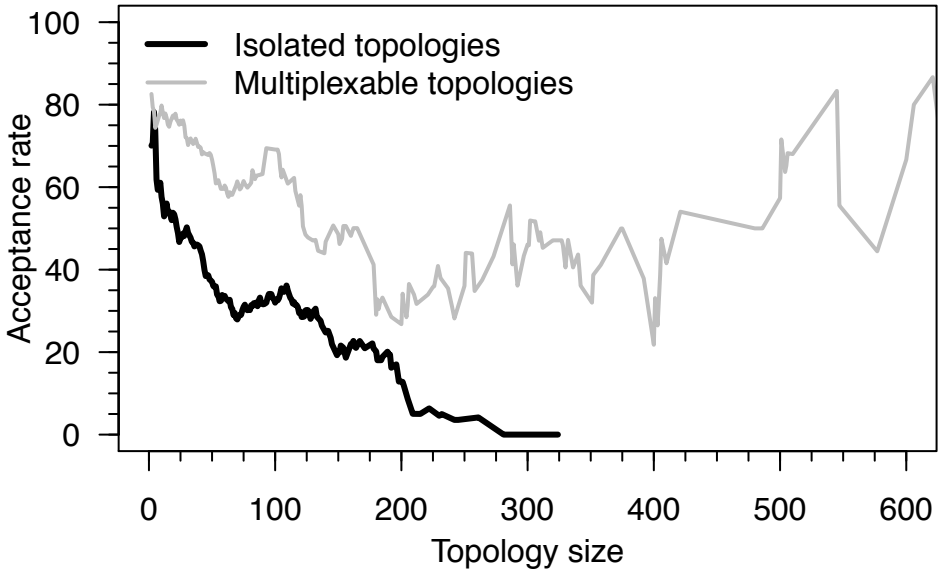
477 projects – 13,057 experiments – 504,226 topologies

Most experiments are small

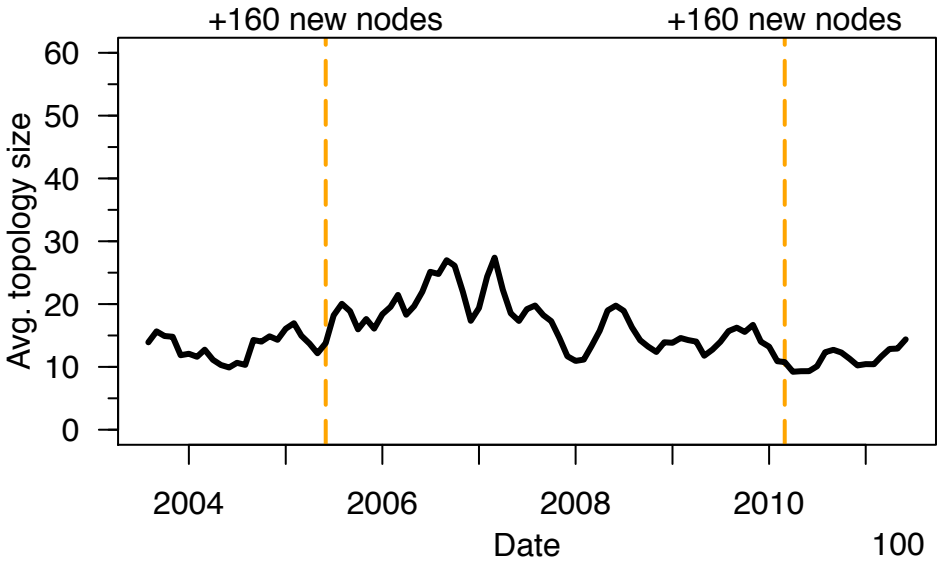


- prepare with small topologies
- evaluate with larger

- multiplexable topologies
  - to increase acceptance rate
  - to deploy larger experiments

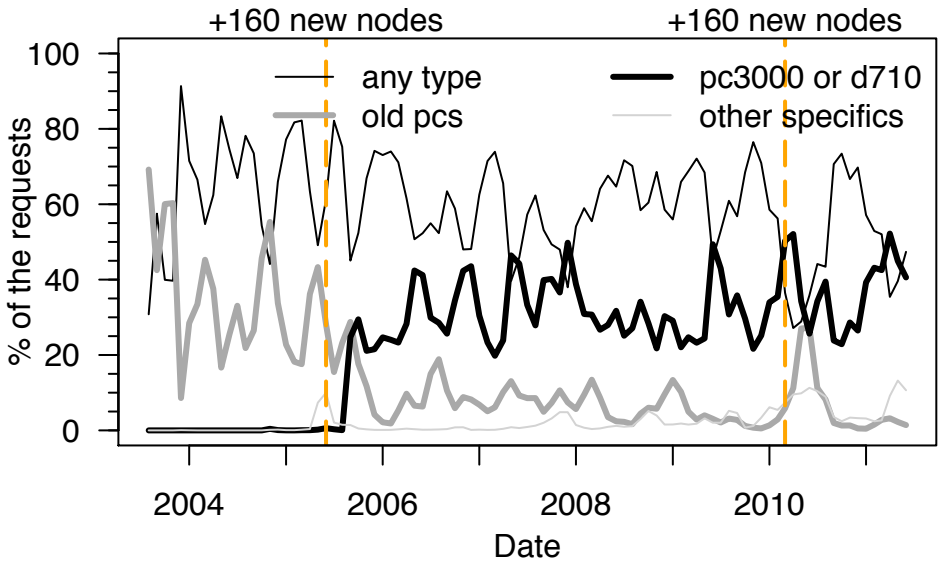


Attractive nodes are the bottleneck



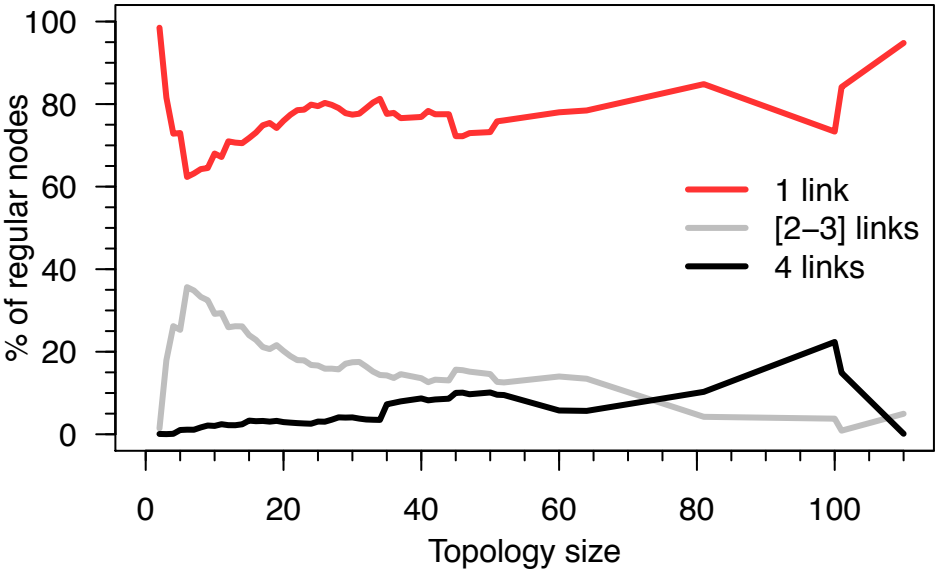
- upgrading is a necessity
  - to meet the demand
  - to support larger experiments

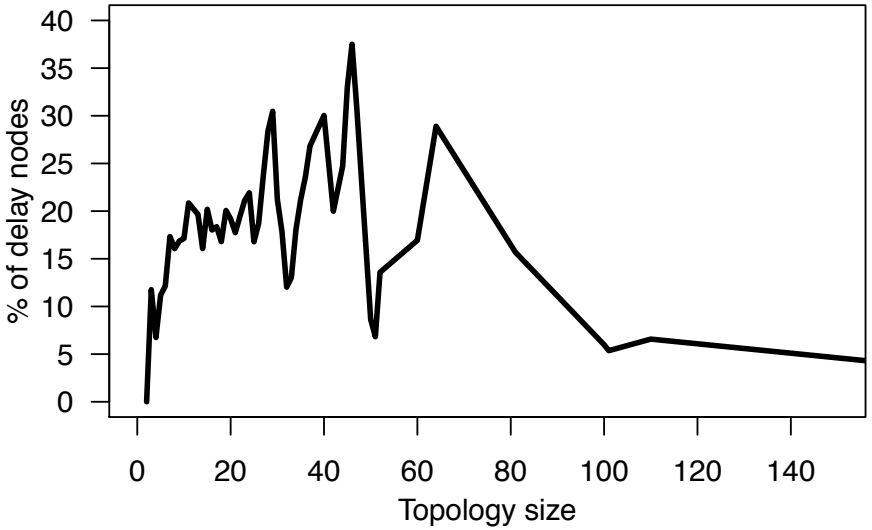
- beauty is ephemeral
- repeatability is forever



Most requests use few interfaces

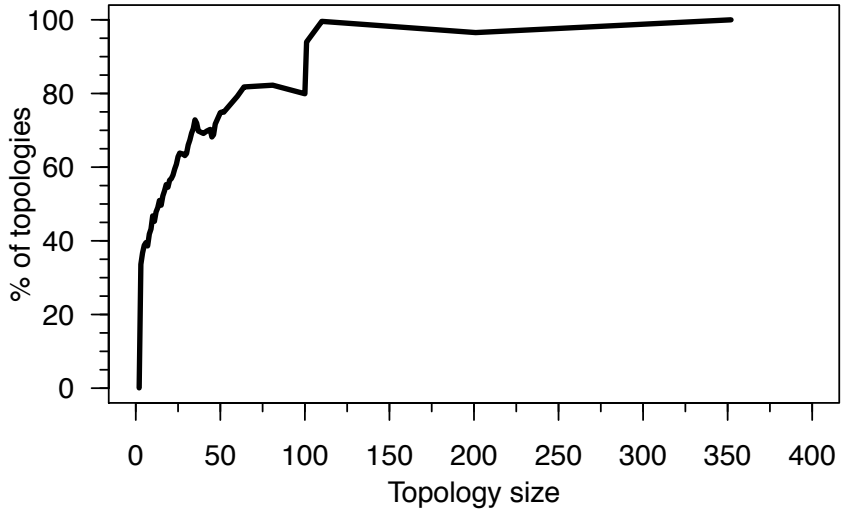
- network interfaces are wasted
- homogeneous connectivity does not reflect the reality of usage





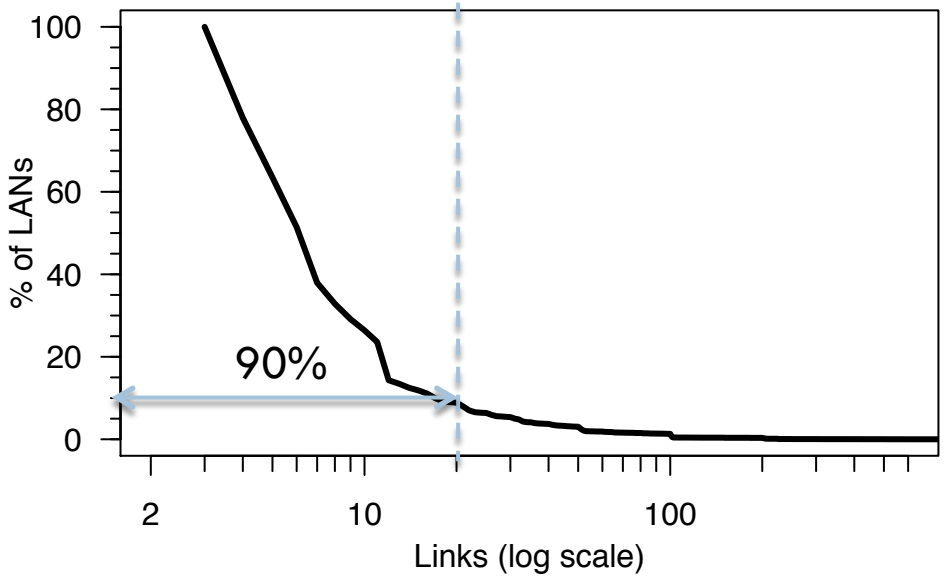
- traffic shaping is a prime feature
- multiple interfaces are required

LANs are common, but most are small



- another prime feature
- LANs dominate large experiments

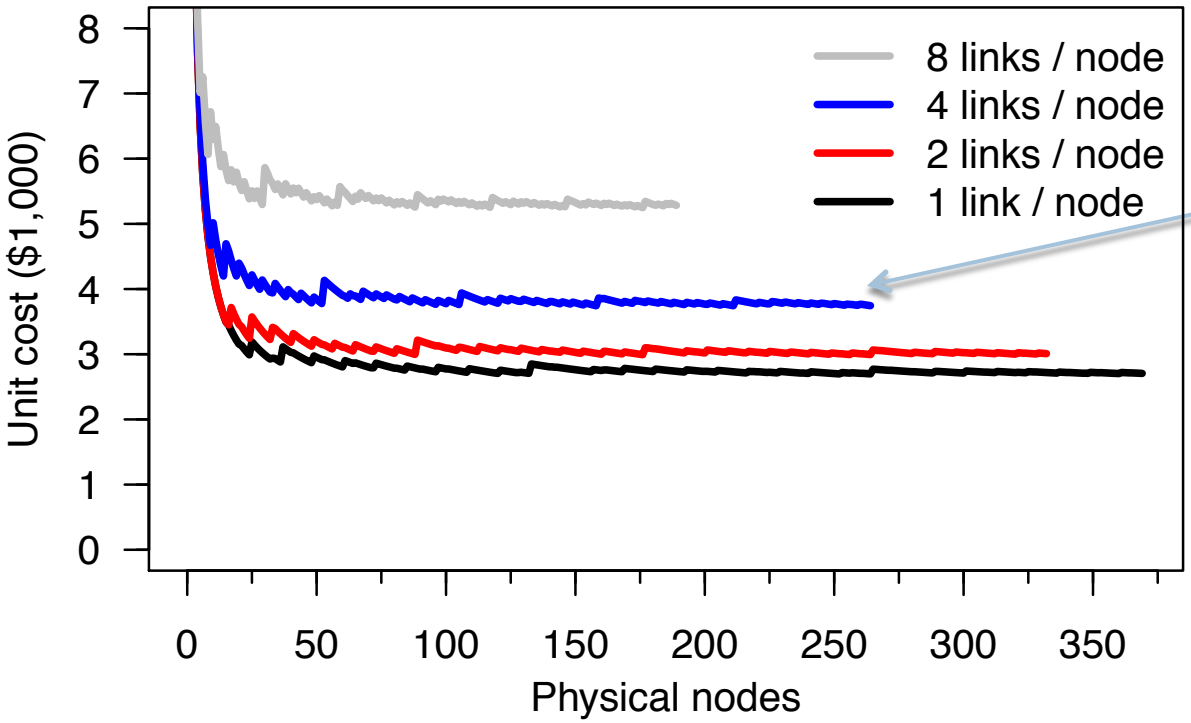
- small LANs
- a moderated usage of the interswitch bandwidth ?



Facts from experimenters data

- the testbed size limits
  - its acceptance rate
  - experiments size
- connectivity is overprovisioned
- improved designs must provide
  - some nodes with multiple interfaces
  - non-blocking bandwidth between a few nodes

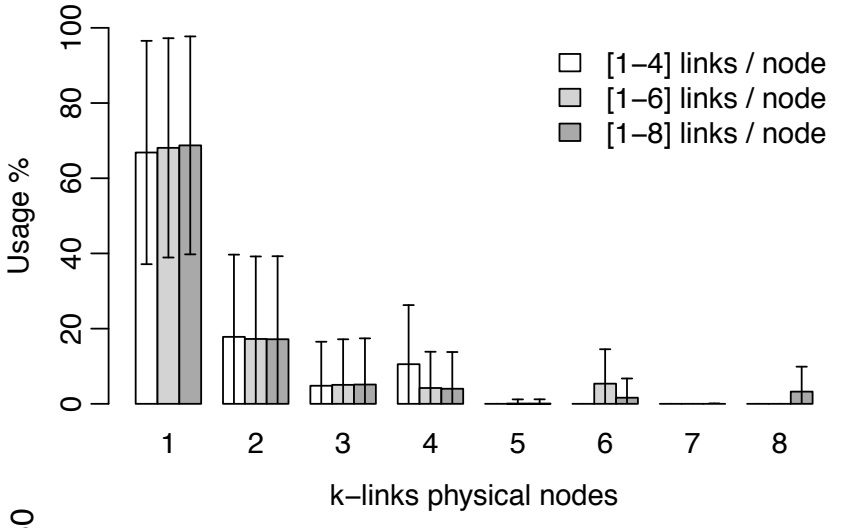
A cost model for testbeds

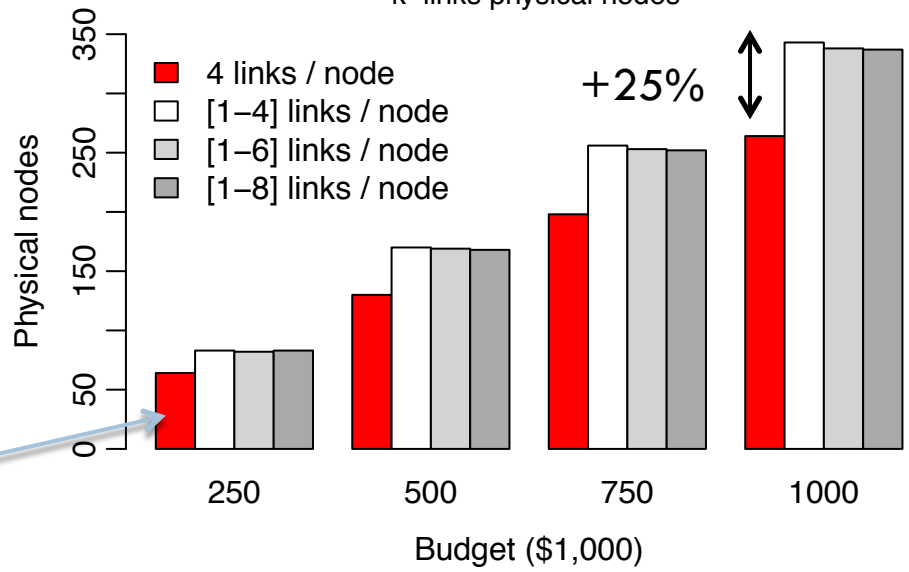


- the impact of high connectivity is significant at scale
  - $100,000: 34 2-link nodes or 27 4-link nodes ?
  - $1,000,000: 370 2-link nodes or 270 4-link nodes ?

Heterogeneous node connectivity

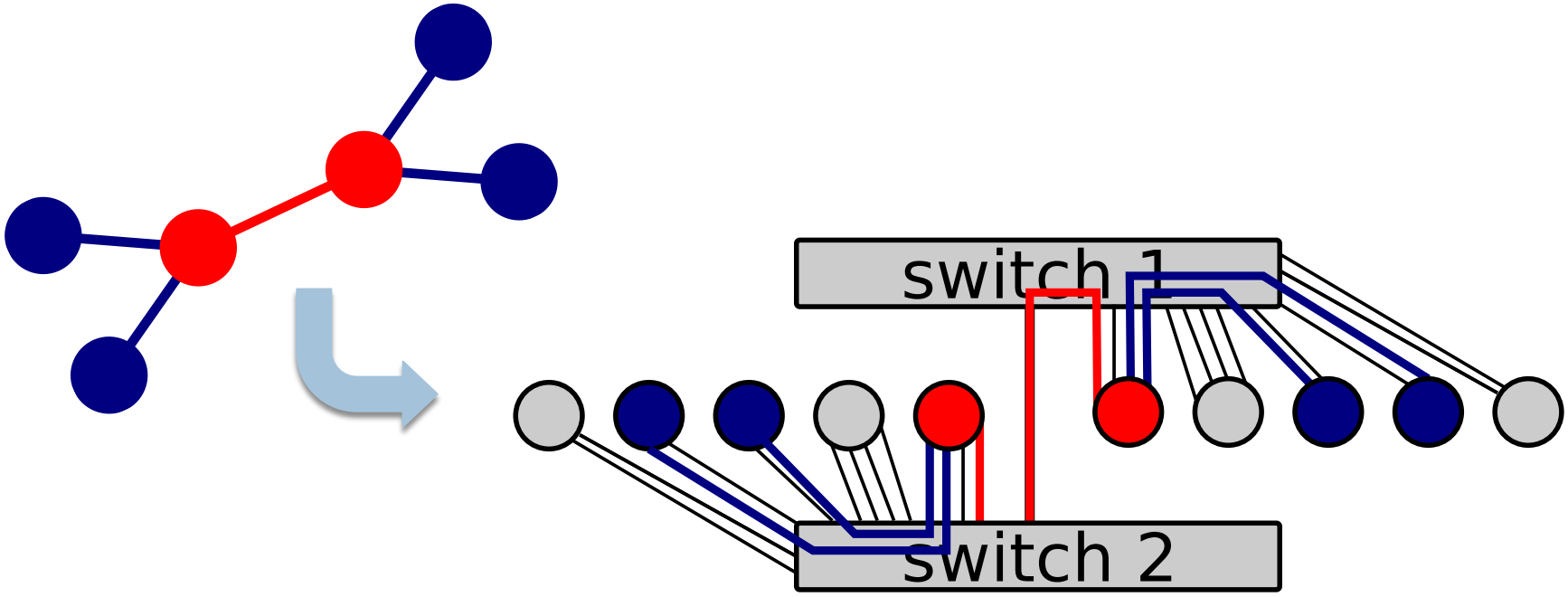
- the best nodes for a topology do not have extra links
- from an ideal node connectivity distribution…
- … to a testbed that minimizes link waste





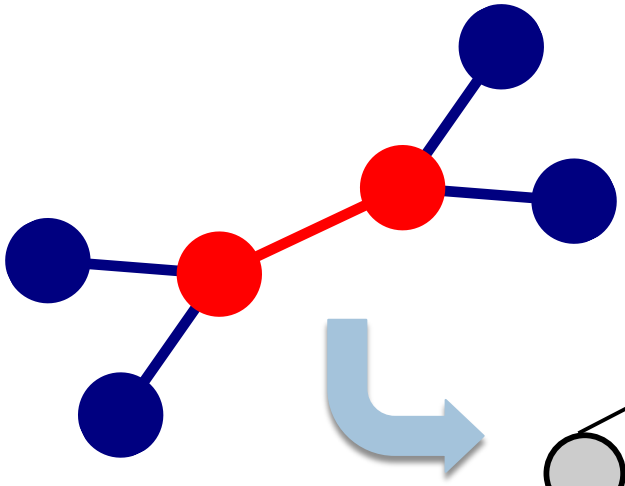
Alternative for switch connectivity

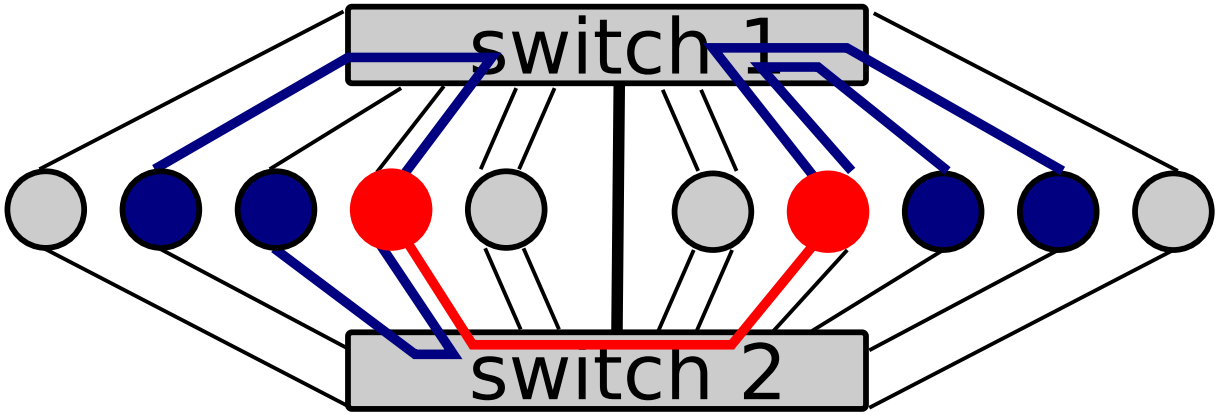
- interswitch bandwidth
  - limit LANs, experiments among switches
- faster interconnect are expensive
- link concentration limits direct communication



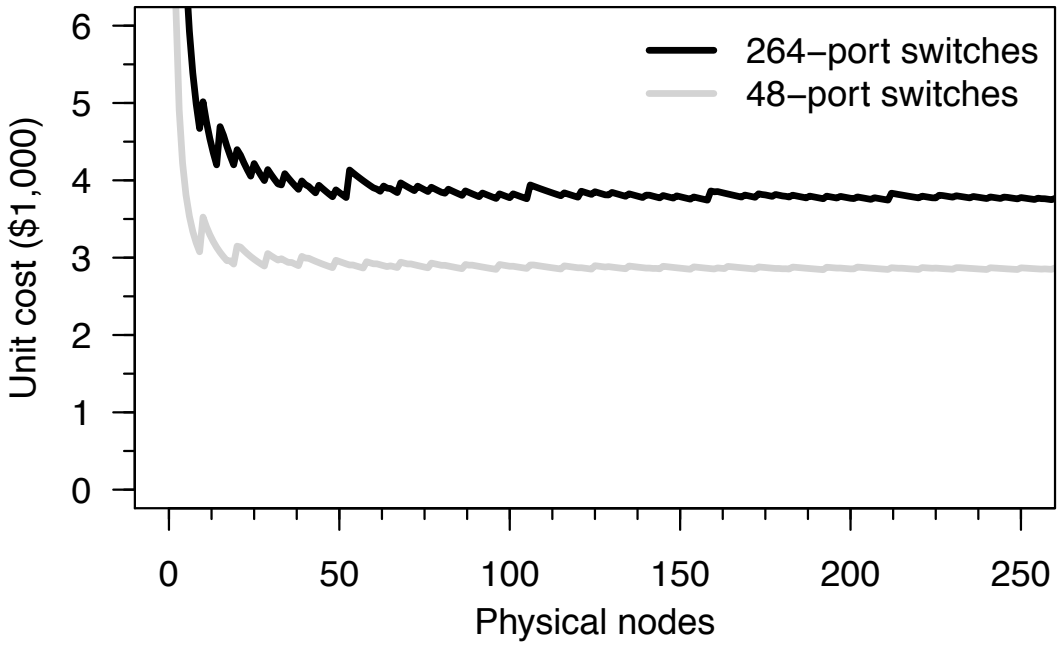
Striping links

- direct communication between nodes
- scalability limited by switch size
- hard to mix with heterogeneous connectivity





Big switches, small testbeds

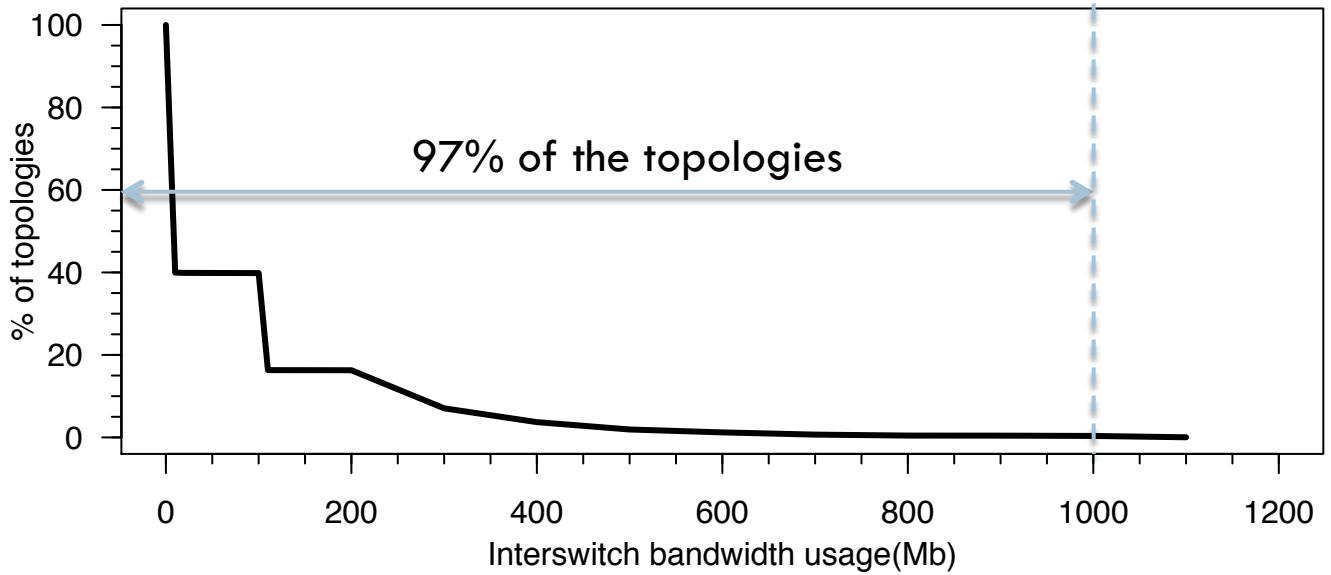


- small switches lead to a 30% bigger testbed …
- … but the reduced bisection bandwidth between the nodes limits maximum LAN size

The interswitch bandwidth myth

BUSTED

- a simulation replayed the user requests for the pc3000 nodes



- interswitch links are overprovisioned
- an opportunity for smaller and cheaper switches

Evaluating new designs

- impact of relevant testbed designs on
  - completion time
  - rejection rate

- the cost model as a testbed generator
- the 15k topologies focusing pc3000 nodes as workload
- a simulator to replay the mapping
  - FIFO scheduling policy
  - each topology runs for a day

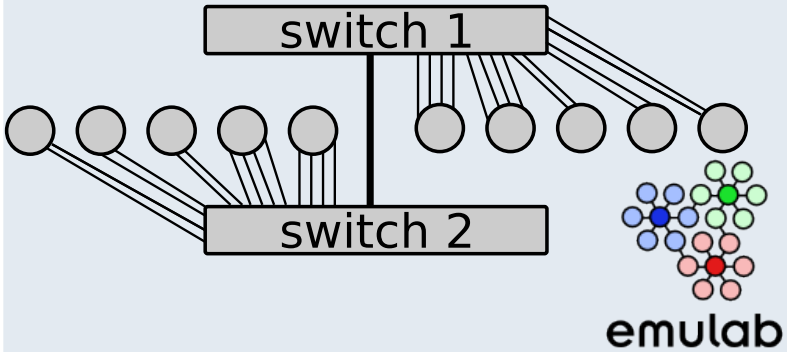
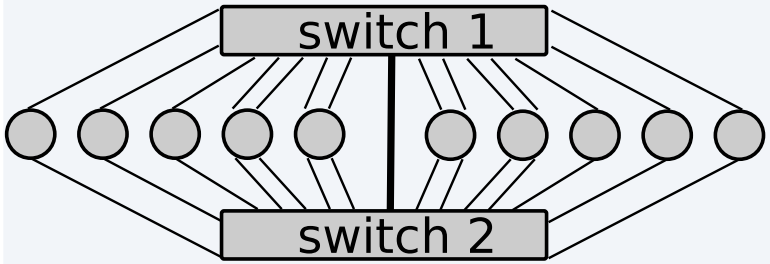
Connectivity for nodes: good trade

- testbeds
  - $500,000 as funds
  - nodes with 2 or 4 links

2-link nodes	Nodes	Bandwidth	Rejections	Time (days)
0% 	130	1.46 Gb	0	1564
60%	148	1.11 Gb	0	1394
90%	161	30 Mb	33 (0.2%)	1487

Striping annihilates bandwidth requirements

- testbeds
  - 148 nodes; 60% with 2-links
  - 2 switches

Configuration	Bandwidth	Rejections	Time (days)
	1.11 Gb	0	1394
	85 Mb	0	1392

Small switches, big testbed

Configuration	Cost	Nodes	Bandwidth	Rejections	Time (days)
264-ports 	$498,796	148	1.11 Gb	0	1394
48-ports	$498,354	186	1.06 Gb	138 (0.9%)	996
48-ports	$390,268	148	883 Mb	142 (0.9%)	1314

- we are the 99%
- using large switches, support the 0.9% “hard” topologies
  - with 40% more time
  - with $108,000. $761 per “hard” topology vs. $33.5

Conclusions

- the testbed size is the bottleneck, not the network
- facts lead to new design suggestions
  - lower connectivity
  - smaller switches
  - link striping
- cost models and replays for a good insight into the testbed's effectiveness
- what to give up to support the outliers?

Read the paper !

How to Build a Better Testbed

Lessons from a decade of network experiments on Emulab

Datacenter vs. network testbed physical design

- datacenters
  - node centric
  - network as a support to maximize performance
  - non-explicit communications
- network testbeds
  - network centric
  - explicit communication to reproduce
  - conservative allocation