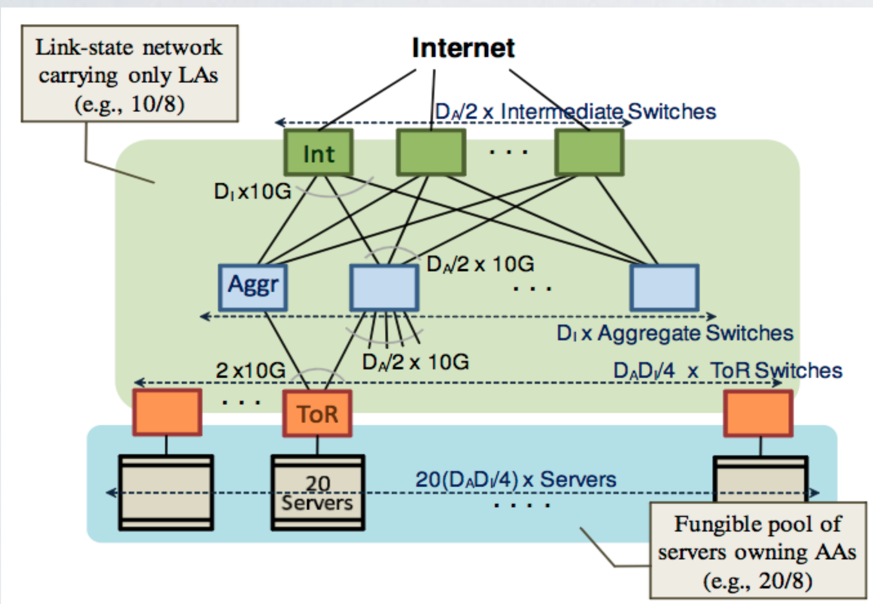
XPANDER: TOWARDS OPTIMAL-PERFORMANCE DATACENTERS

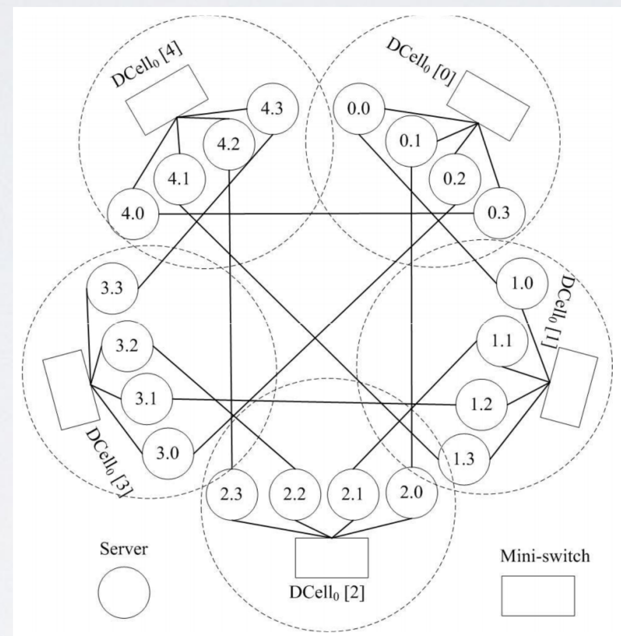
Asaf Valadarsky (Hebrew University)
Gal Shahaf (Hebrew University)
Michael Dinitz (Johns Hopkins University)
Michael Schapira (Hebrew University)

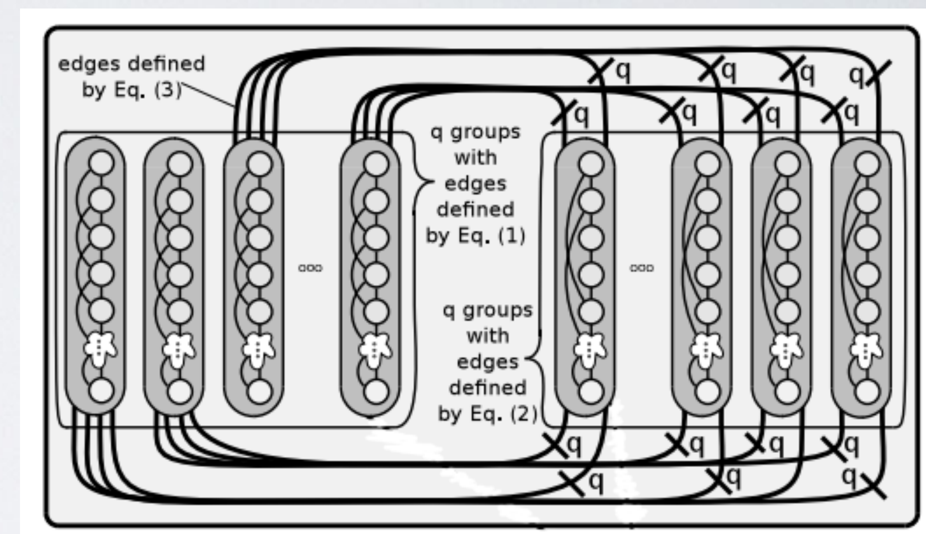
JOHNS HOPKINS UNIVERSITY

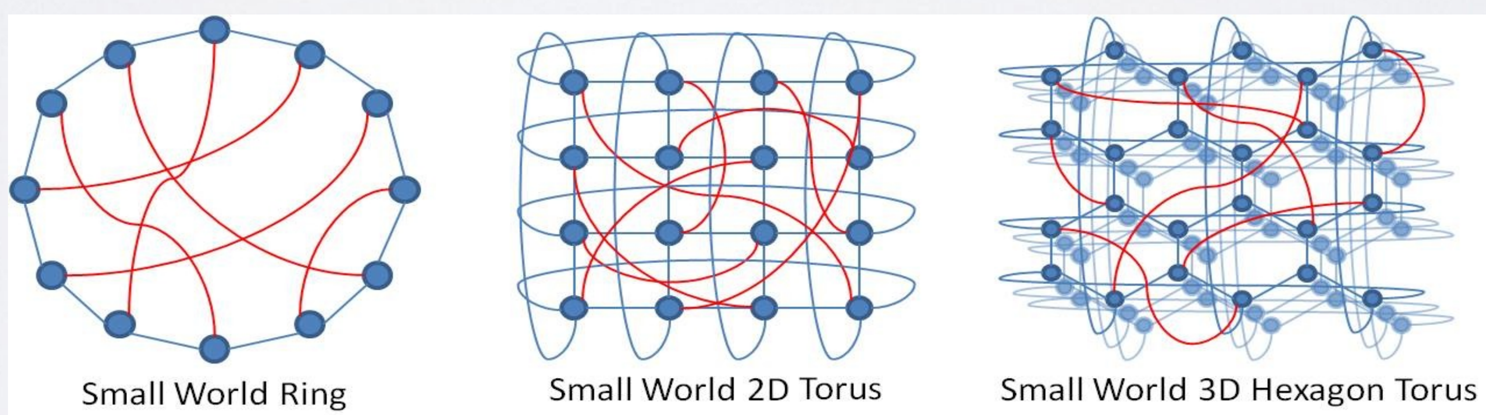
האוניברסיטה העברית בירושלים
THE HEBREW UNIVERSITY OF JERUSALEM

DESIGNING A DATACENTER ARCHITECTURE









Network Topology? Routing? Congestion Control?

DESIGNING A DATACENTER ARCHITECTURE

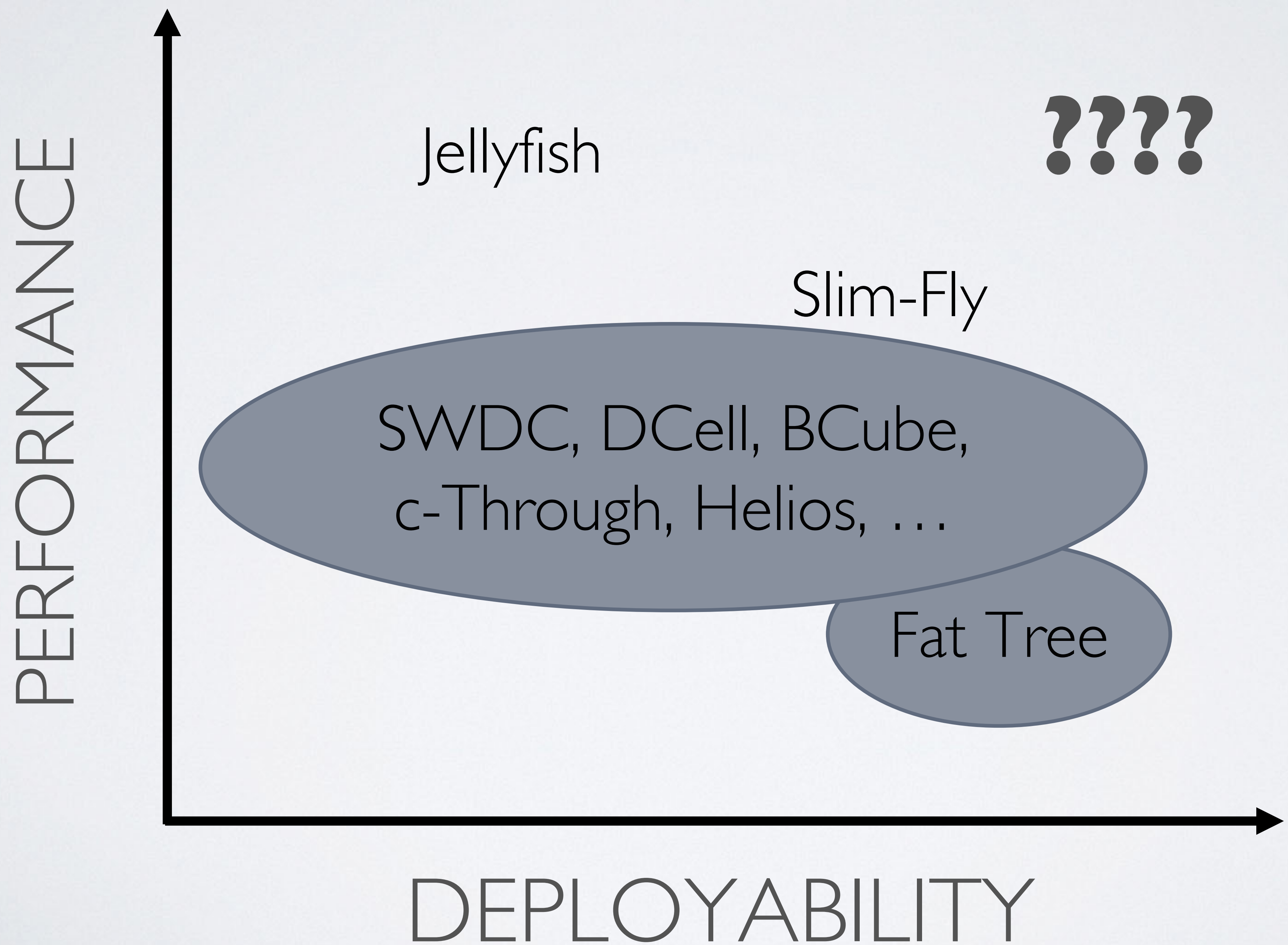
Performance

- ➡Throughput
- ➡Resiliency to failures
- ➡Path diversity
- ➡…

Deployability

- ➡Cabling complexity
- ➡Operations cost
- ➡Equipment costs
- ➡…

WHAT IS THE "RIGHT" DATACENTER ARCHITECTURE?



AGENDA

- Reaching that upper-right corner entails designing "expander datacenters"
- **Xpander**: a tangible and near-optimal datacenter design

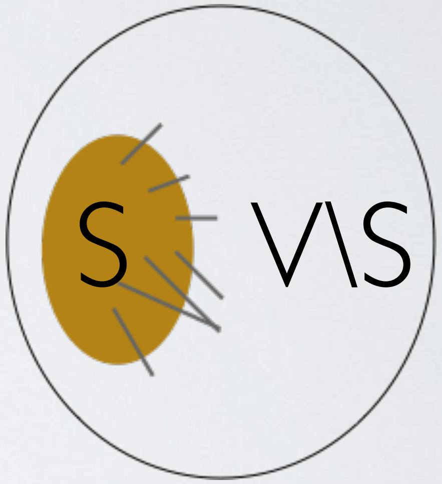
EXPANDER DATACENTERS

- An expander datacenter architecture:
  - ➡ Utilizes an expander graph as its network topology ***(see next slide)***
  - ➡ Employs (multi-path) routing and congestion control to exploit path diversity

EXPANDER GRAPHS: INTUITION

- A graph is called an "expander graph" if it has "good" edge expansion

$$\min_{S \subset V, 0 < |S| \leq \frac{n}{2}} \frac{EdgesBetween(S, V \setminus S)}{|S|}$$



- **Intuition:** In an expander graph, the capacity traversing each cut is "large"
  - ➡ Traffic is never bottlenecked at small set of links
  - ➡ High path diversity

CONSTRUCTING EXPANDERS

- Constructing expanders is a prominent research area in mathematics and computer science
- Applications in networking, computational complexity, coding, and beyond

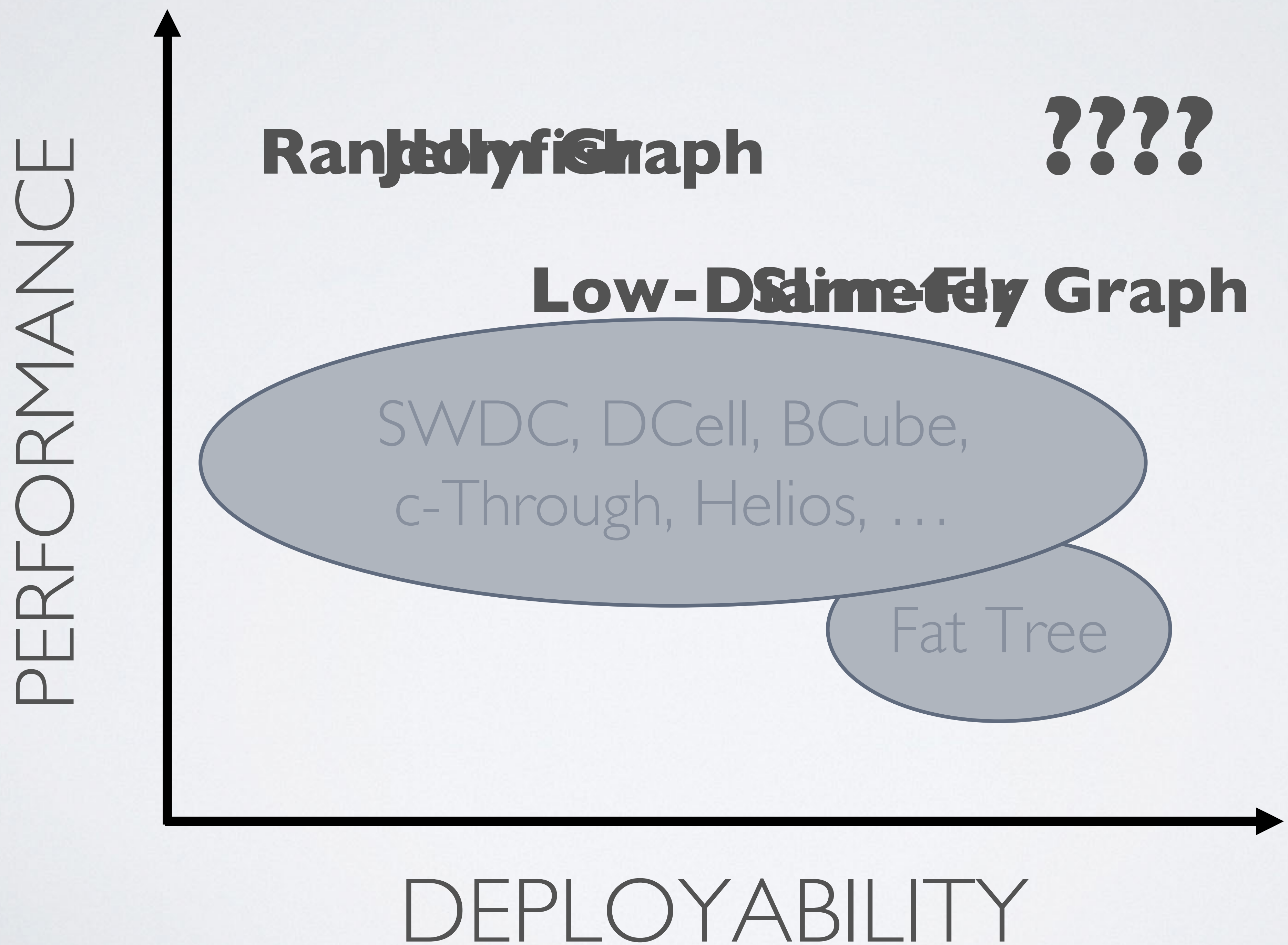
EXPANDER DATACENTERS ACHIEVE NEAR-OPTIMAL PERFORMANCE

- ➡ Support higher traffic loads
- ➡ More resilient to failures
- ➡ Support more servers with less network devices
- ➡ Multiple short-paths between hosts
- ➡ Incrementally expandable

OUR EVALUATION

- ➡ Theoretical analyses
- ➡ Flow- and packet-level simulations
- ➡ Experiments on network emulator
- ➡ Experiments on an SDN-capable network

EXPANDER DATACENTERS **<u>ARE</u>** THE STATE-OF-THE-ART



CAN WE HAVE IT ALL?

A well structured design + Near optimal performance

YES! :)

XPANDER DATACENTER ARCHITECTURE

Near-Optimal Performance

- ➡ Throughput
- ➡ F… lures
- ➡ …
- ➡ …

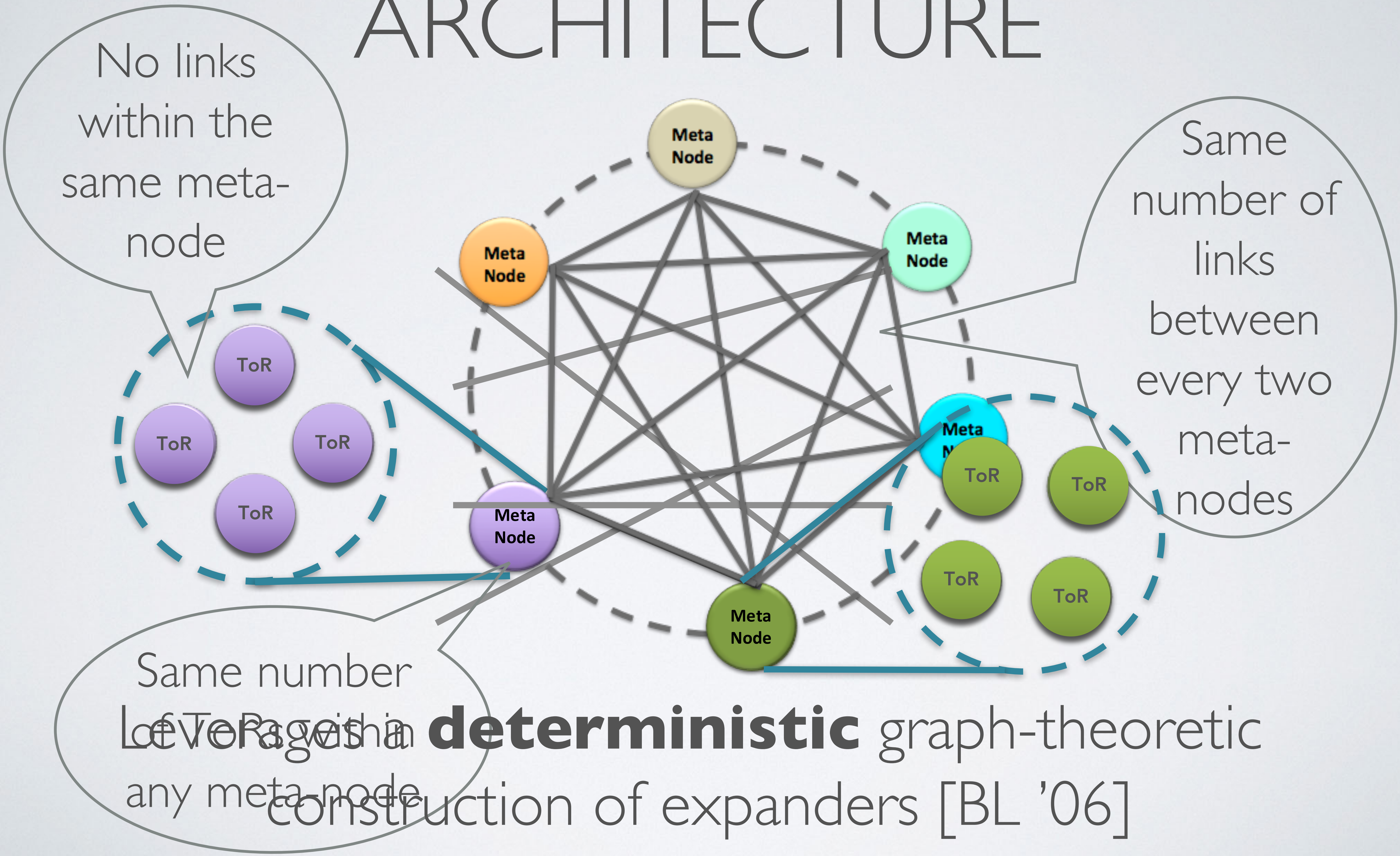
Expander Datacenter

Deployable

- ➡ Cabling … lexity
- ➡ O…
- ➡ Eq…
- ➡ …

Deployment-Oriented Construction

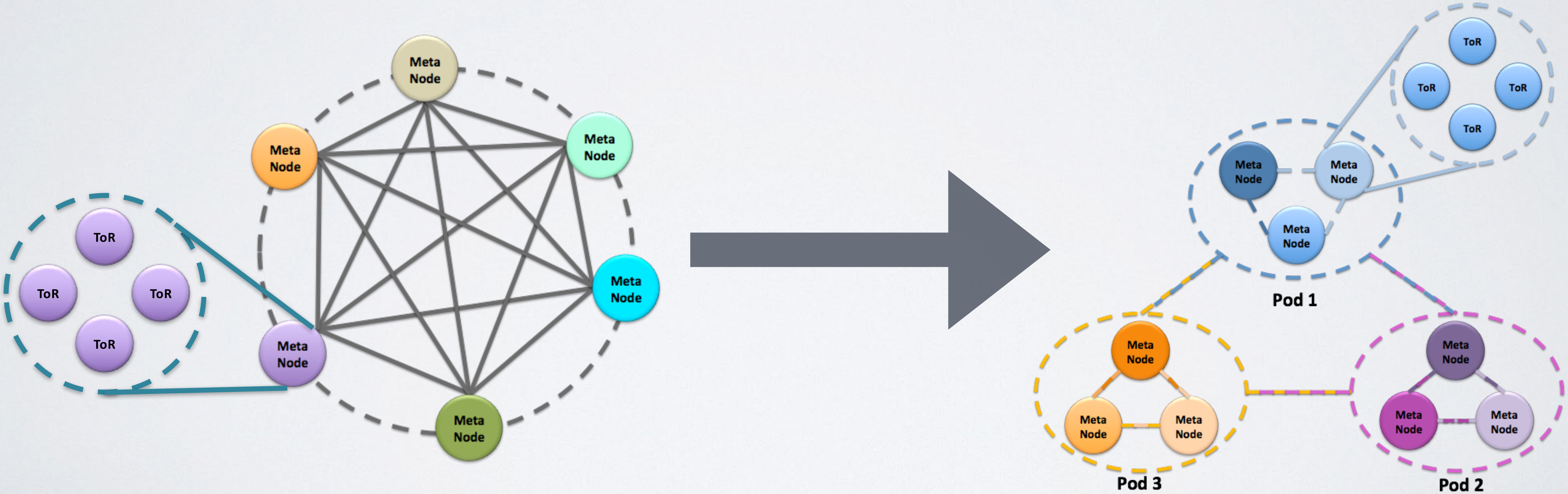
XPANDER DATACENTER ARCHITECTURE



Leverages a **deterministic** graph-theoretic construction of expanders [BL '06]

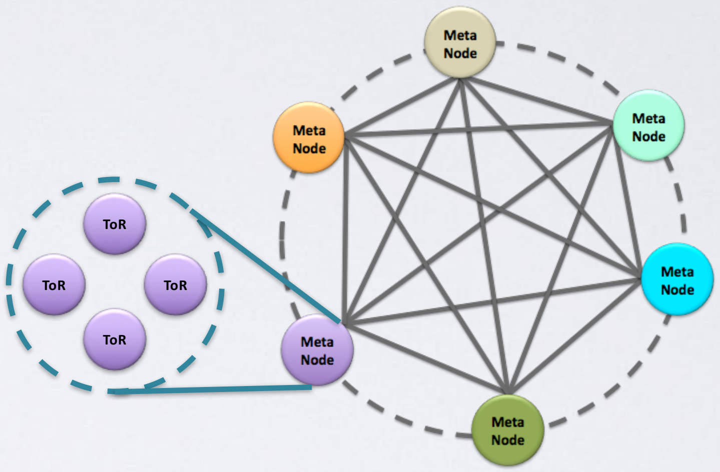
WHERE ARE MY PODS?

An Xpander can be divided into smaller "Xpander pods"



XPANDER DATACENTER ARCHITECTURE

Topology



Routing

Multipath Routing (K-Shortest Paths)

Congestion Control

Multipath Congestion Control (Multipath-TCP)

EXPANDER DATACENTERS ACHIEVE NEAR-OPTIMAL PERFORMANCE

➡ **Support higher traffic loads**

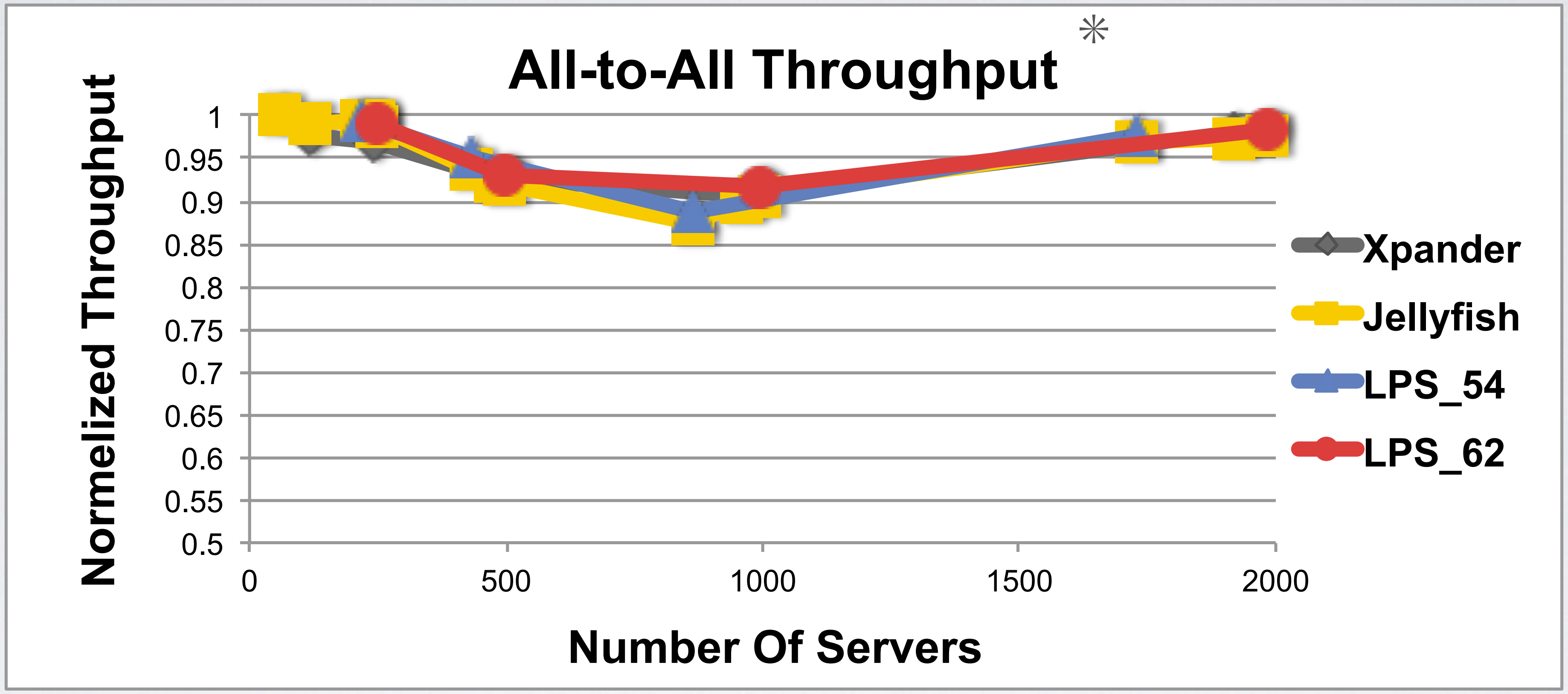
➡ **More resilient to failures**

➡ **Support more servers with less network devices**

➡ Multiple short-paths between hosts

➡ Incrementally expandable

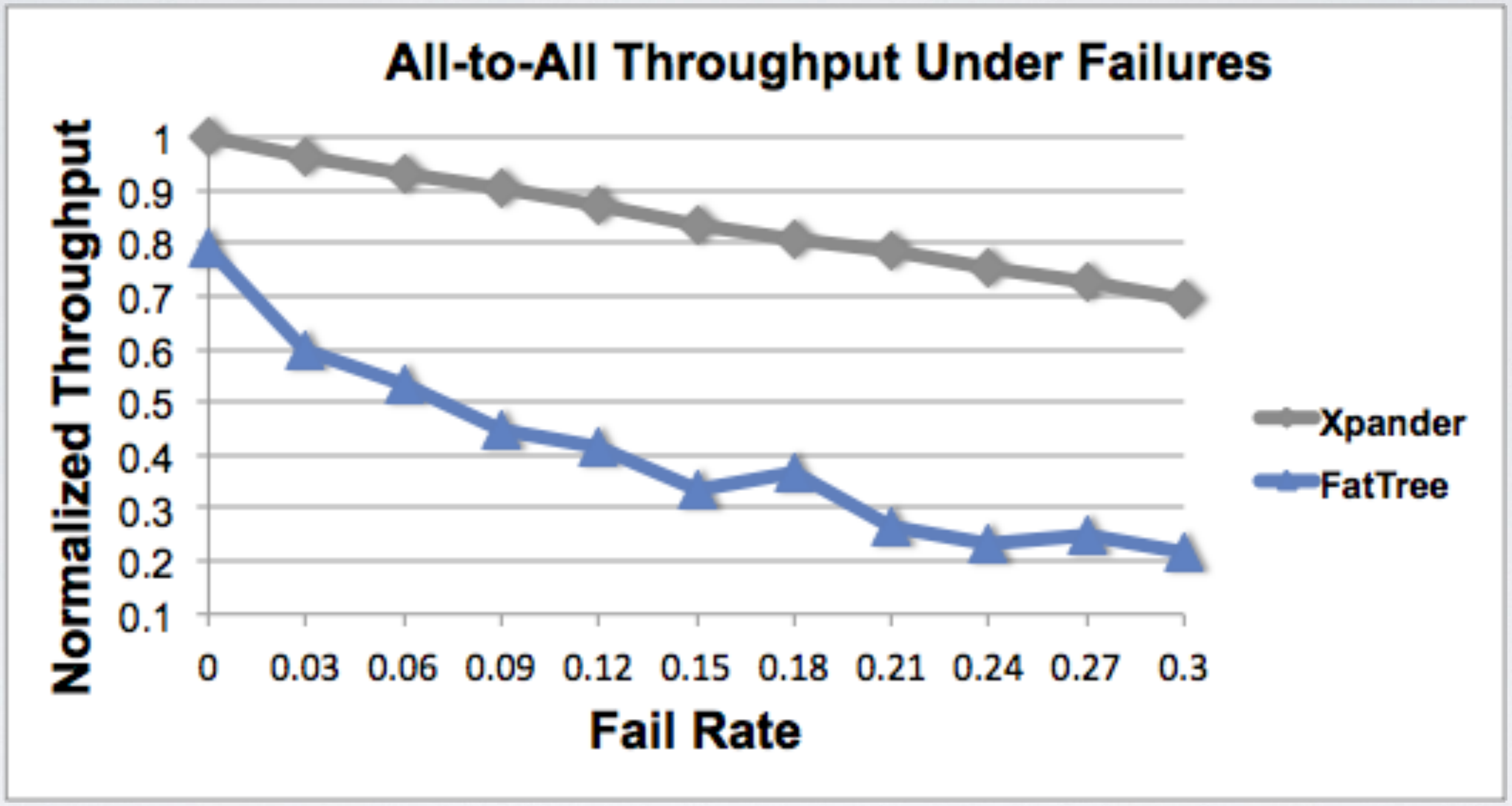
NEAR OPTIMAL ALL-TO-ALL THROUGHPUT



* 18-port switches

Theorem: In the all-to-all setting, the throughout of any d-regular expander G on n vertices is within a factor of O(logd) of that of the throughput-optimal d-regular graph on n vertices

RESILIENCE TO FAILURES



Theorem: In any d-regular expander, any two vertices are connected by exactly d edge-disjoint paths.

NEAR-OPTIMAL THROUGHPUT UNDER SKEWED TRAFFIC MATRICES

- Expander datacenters empirically attain near-optimal throughput under skewed TMs (mice and elephants)
- We prove that expander datacenters are **optimal** with respect to **adversarial** traffic conditions

COST EFFICIENCY: XPANDER VS. FAT-TREE

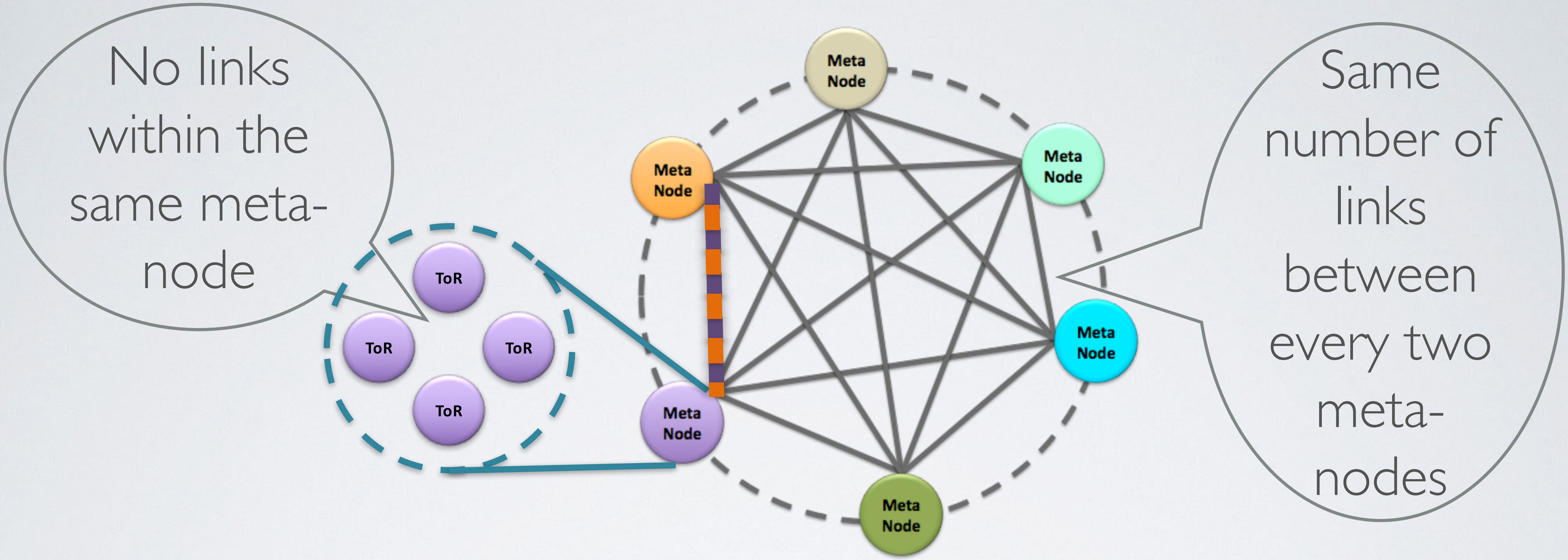
Switch Degree	#Switches	All-to-All Throughput
8*	80%	121%
10	100%	157%
24	80%	111%

*Validated using Mininet experiments

SEE PAPER FOR

- Analysis of shortest-paths and diameter
- Physical layout and costs
- Incremental expansion of expander datacenters
- Results for skewed traffic matrices
- Results for Xpander vs. Jellyfish
- Results for Xpander vs. Slim-Fly
- Additional results for Xpander vs. Fat Tree
- Experiments with the Mininet network emulator
- Experiments on the OCEAN SDN-capable network testbed
- …

DEPLOYING XPANDER



- ➡ Place ToRs of each meta-node in close proximity
- ➡ Bundle cables between two meta-nodes
- ➡ Use color-coding to distinguish between different meta-nodes and bundles of cables

DEPLOYING XPANDER

- Analysed physical layout, cabling complexity, #cables and cable length for both large-scale and "container" datacenters

Switch Ports	#Switches	#Servers	#Cables	Cable Length	Throughput
32	42 vs. 48 (87.5%)	504 vs. 512 (98.44%)	420 vs. 512 (82%)	4.2 km vs 5.12km (82%)	109%
48	66 vs. 72 (92%)	1056 vs. 1152 (92%)	1056 vs. 1152 (92%)	10.5 km vs 11.5km (92%)	142%

CONCLUSION

- We show that expander datacenters outperform traditional datacenters
  - ✓ Sheds light on past results about random and low-diameter graphs based datacenters
- We present **Xpander**, a novel datacenter architecture
  - ✓ Suggests a **tangible** alternative to today's datacenter architectures
  - ✓ Achieves **near-optimal** performance

QUESTIONS?

THANK YOU!

See project webpage at:
https://husant.github.io/Xpander/