A **Throughput-Centric** View of the Performance of Datacenter Topologies

Pooria Namyar (USC)
Sucha Supittayapornpong (VISTEC)
Mingyang Zhang (USC)
Minlan Yu (Harvard University)
Ramesh Govindan (USC)

USC Viterbi School of Engineering
VISTEC VIDYASIRIMEDHI INSTITUTE OF SCIENCE AND TECHNOLOGY
HARVARD UNIVERSITY

When experts design a network, they try to provision the network to handle expected traffic demands…

When cloud providers design a datacenter network, they try to provision the network to handle ***any possible traffic demand***.

* To a first approximation. We discuss oversubscription in the paper.

Why any possible traffic demand

Datacenters are long-lived

Why any possible traffic demand

Datacenters are long-lived

Traffic can change significantly

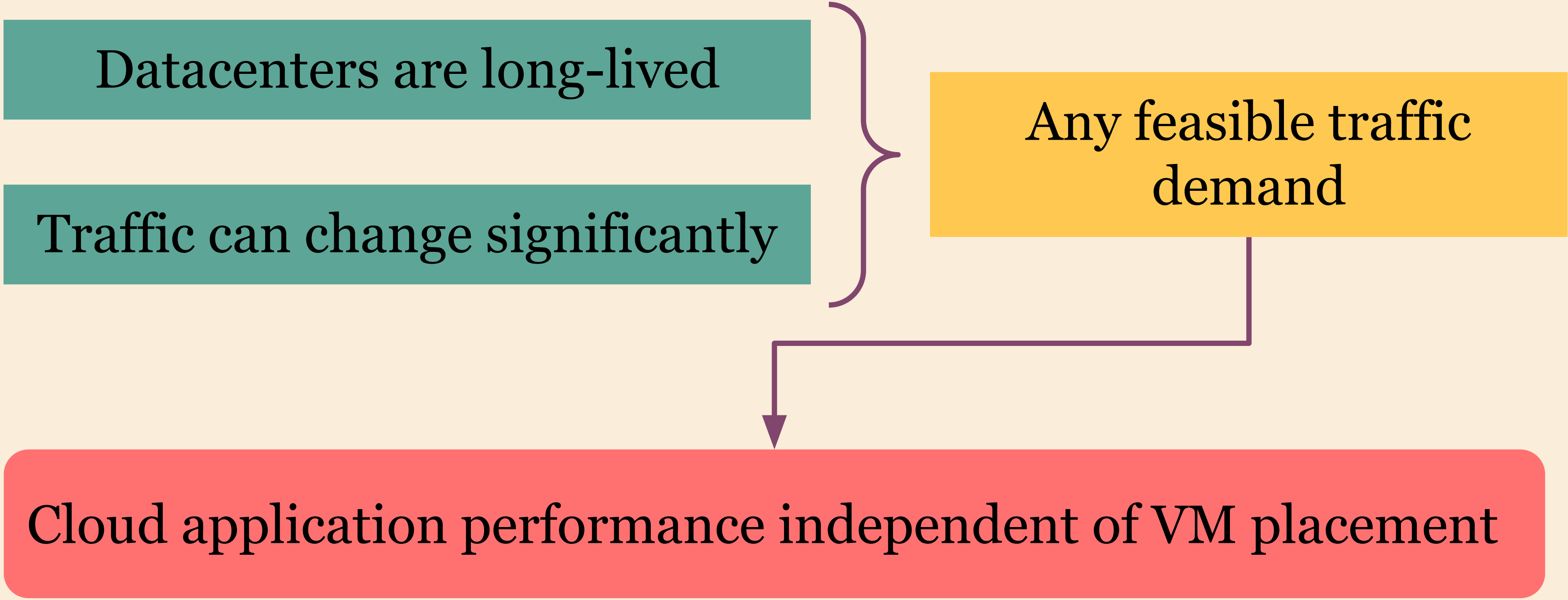
Why any possible traffic demand

Datacenters are long-lived

Traffic can change significantly

Any feasible traffic demand

Why any possible traffic demand



Why any possible traffic demand

Dat

Traffi

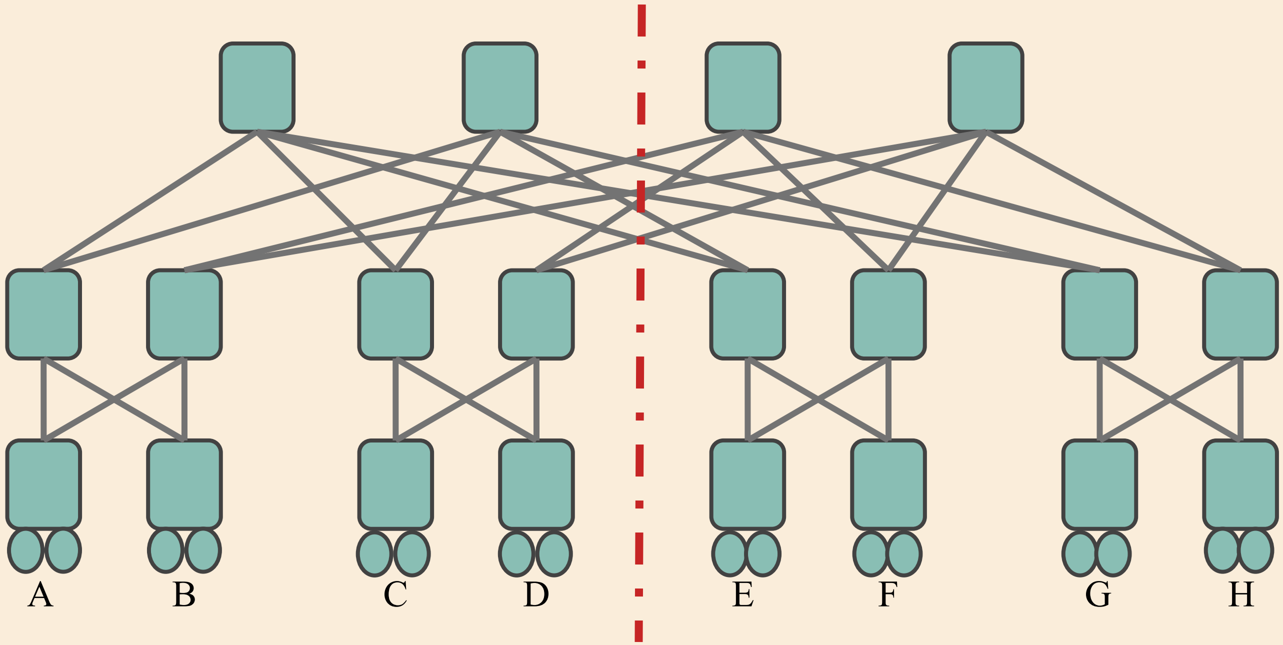
raffic

Non-blocking Topology;
A topology that does not block any traffic demand

Cloud application performance independent of VM placement

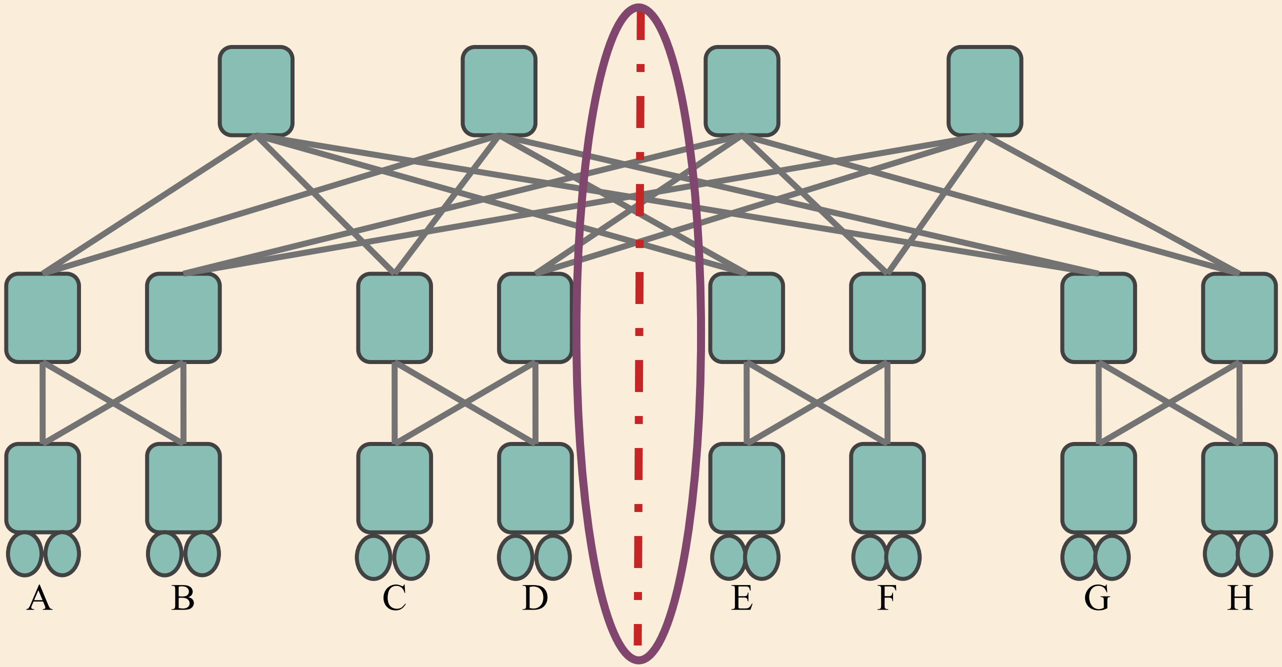
How to assess whether a datacenter topology is non-blocking?

Early Work uses Bisection Bandwidth



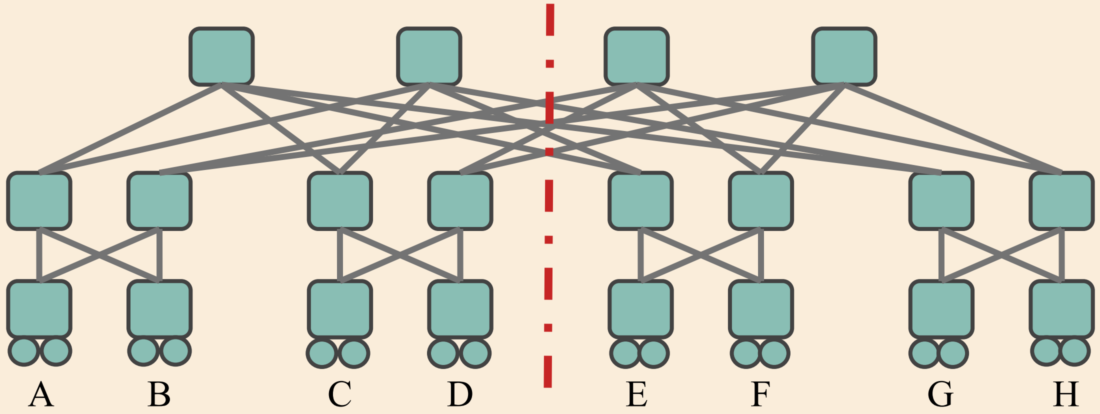
Bisection Bandwidth

Early Work uses Bisection Bandwidth



Bisection Bandwidth

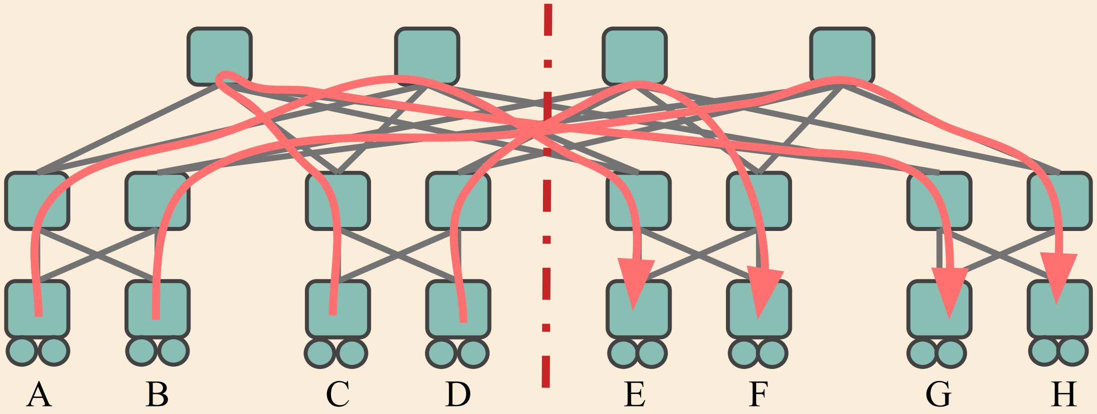
Early Work uses Bisection Bandwidth



Full Bisection Bandwidth

Bisection Bandwidth ≥ #servers/2

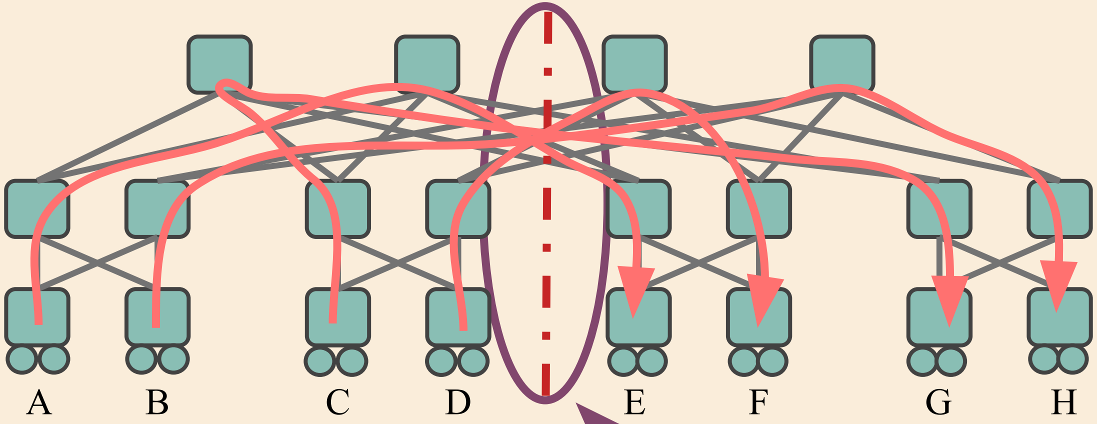
Early Work uses Bisection Bandwidth



Full Bisection Bandwidth

Bisection Bandwidth ≥ #servers/2

Early Work uses Bisection Bandwidth



Full Bisection Bandwidth

Bisection Bandwidth ≥ #servers/2

Early Work uses Bisection Bandwidth

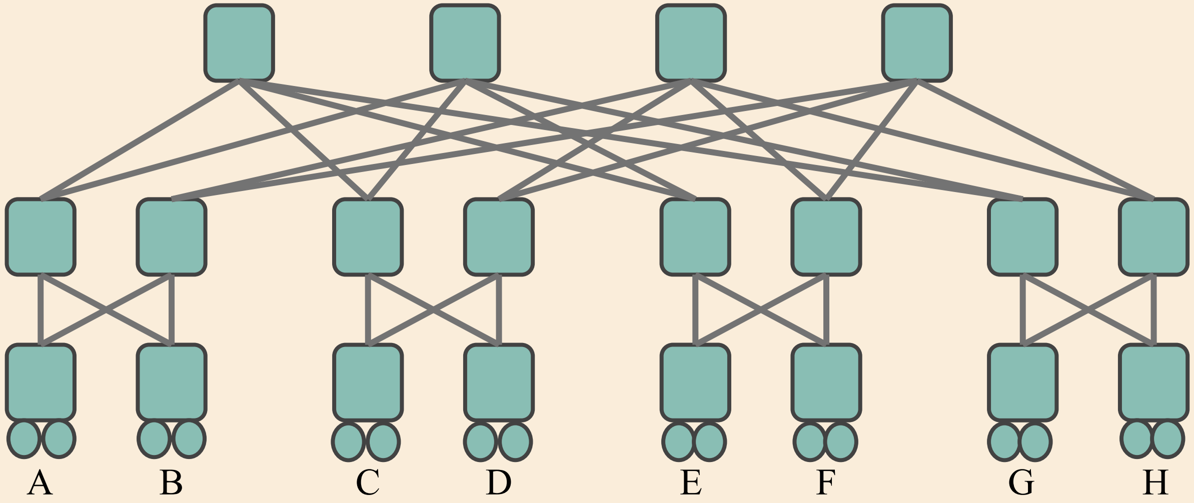
Full Bisection Bandwidth → Non-blocking Topology

Early Work uses Bisection Bandwidth

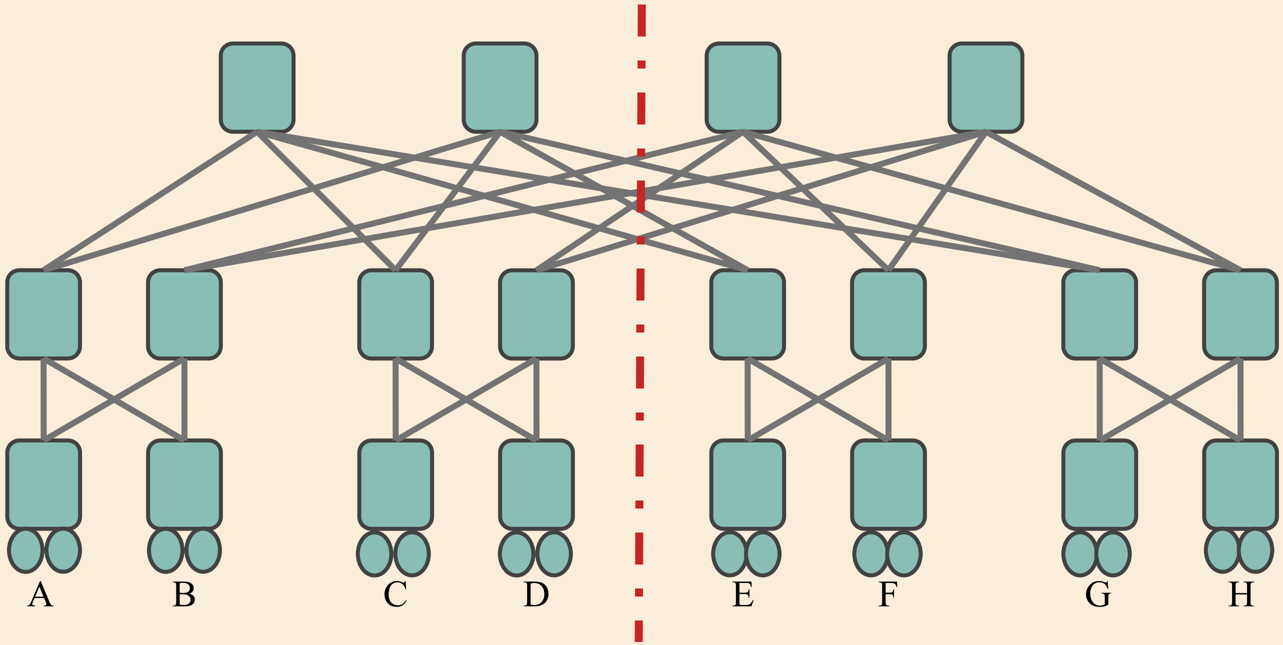
Full Bisection Bandwidth → Non-blocking Topology

This holds for a specific topology family called **Clos**.

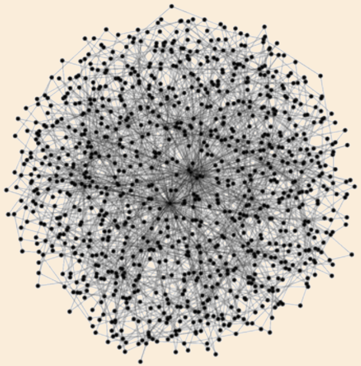
Most Commercial Datacenters are Clos



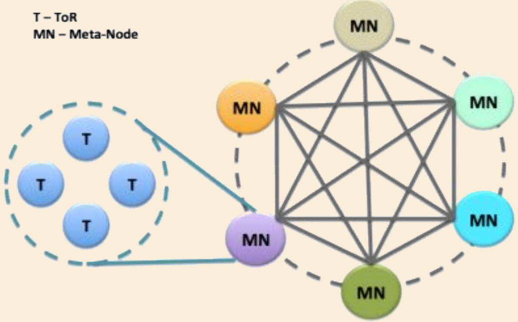
But Clos is Expensive



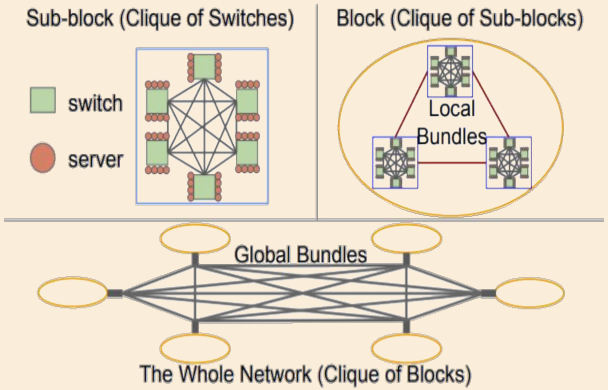
Recently Proposed Topologies: Expanders



Jellyfish
[NSDI'12]



Xpander
[CoNEXT'16]



FatClique
[NSDI'19]

Recently Proposed Topologies: Expanders

Lower Cost (#Switches, #Links, #Racks, ...)

Recently Proposed Topologies: Expanders

Lower Cost (#Switches, #Links, #Racks, ...)

Better Management Complexity (Expansion, Wiring, ...)

Recently Proposed Topologies: Expanders

Lower Cost (#Switches, #Links, #Racks, ...)

Better Management Complexity (Expansion, Wiring, ...)

Better Failure Resiliency (Random Failure, ...)

For expanders, can bisection bandwidth help assess whether topology is non-blocking?

* It is for Clos → proof in the paper.

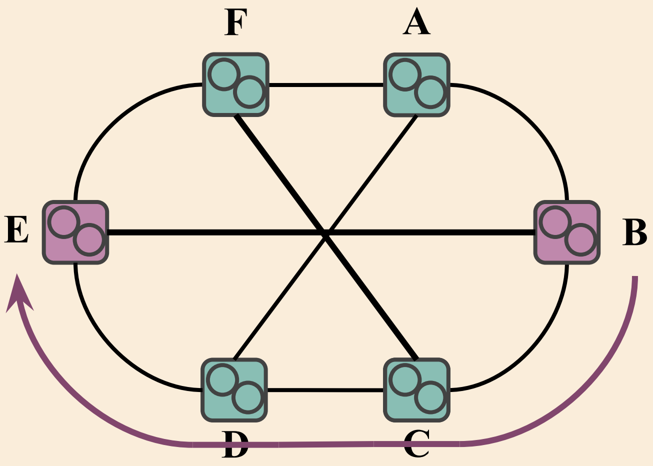
Prior Work Has Proposed Another Metric

Throughput of the topology for a given *traffic matrix* measures the fraction of demand that network can sustain

Prior Work Has Proposed Another Metric

Throughput of the topology for a given *traffic matrix* measures the fraction of demand that network can sustain

Demand from B to E =2.0

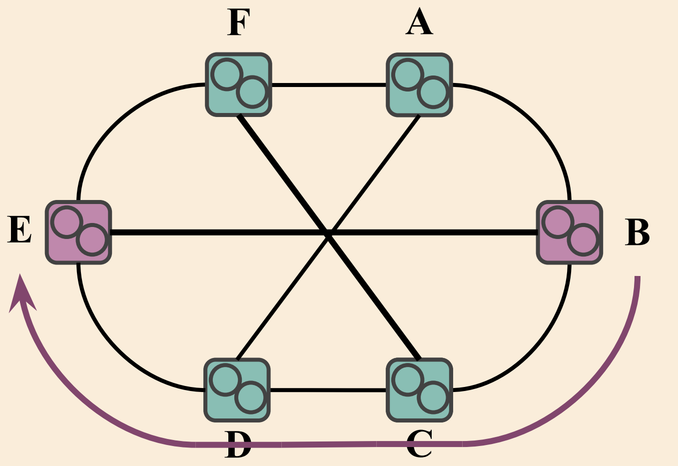


Prior Work Has Proposed Another Metric

Throughput of the topology for a given *traffic matrix* measures the fraction of demand that network can sustain

Demand from B to E =2.0

Network can sustain =1.5



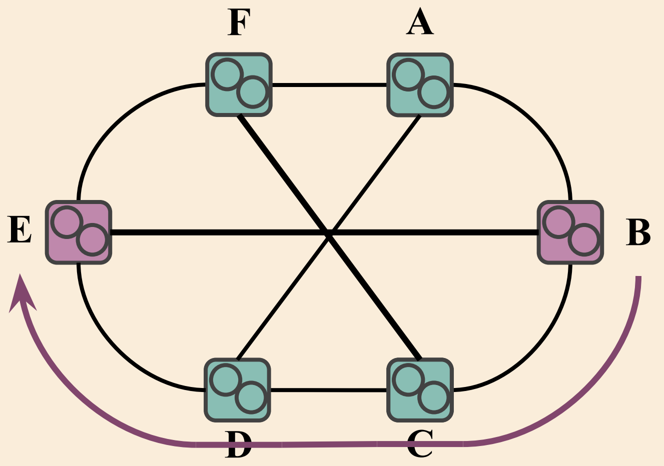
Prior Work Has Proposed Another Metric

Throughput of the topology for a given *traffic matrix* measures the fraction of demand that network can sustain

Demand from B to E =2.0

Network can sustain =1.5

Throughput = 0.75



Prior Work Has Proposed Another Metric

Throughput of the topology for a given *traffic matrix* measures the fraction of demand that network can sustain

Throughput of 1 means network can support the traffic matrix

Prior Work Has Proposed Another Metric

Throughput of the topology for a given *traffic matrix* measures the fraction of demand that network can sustain

Throughput of topology is the **smallest throughput** across all possible traffic matrices

Prior Work Has Proposed Another Metric

Throughput of the topology for a given *traffic matrix* measures the fraction of demand that network can sustain

Throughput of topology is the **smallest throughput** across all possible traffic matrices

↓

Throughput of 1 means network is non-blocking

Prior Work Has Proposed Another Metric

Throughput of the topology for a given *traffic matrix* measures the fraction of demand that network can sustain

Throughput of topology is the smallest throughput across all possible traffic matrices

Throughput is expensive to compute

For expanders, is bisection bandwidth equivalent to throughput?

Findings

1

A full bisection bandwidth Expander may not have full throughput.

Findings

1

A full bisection bandwidth Expander may not have full throughput.

Theory

There are always exist a size beyond which no full throughput Expander topology exists.

Practice

Even Expanders with 10-15K servers may not have full throughput even if they have full bisection bandwidth

Findings

1

A full bisection bandwidth Expander may not have full throughput.

2

Cost, manageability, and failure resilience comparisons affected significantly when throughput is used at large-scale.

But Computing Throughput is Expensive

An accurate upper bound for throughput of Expanders and Clos topologies that scales well.

Outline

1

A full bisection bandwidth Expander may not have full throughput.

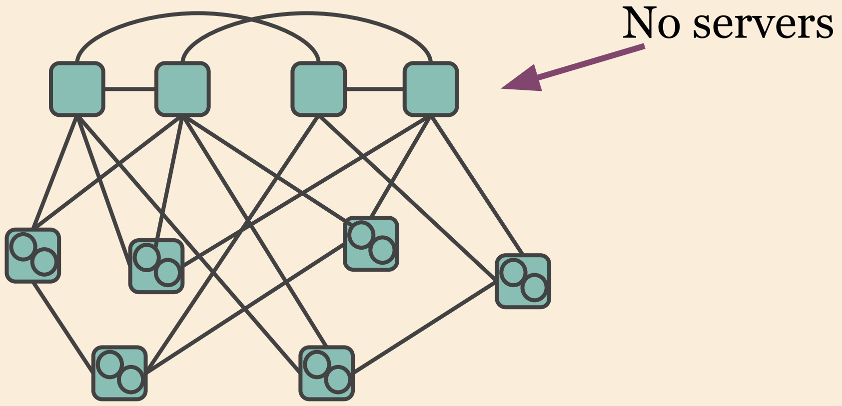
2

Cost, manageability, and failure resilience comparisons affected significantly when throughput is used at large-scale.

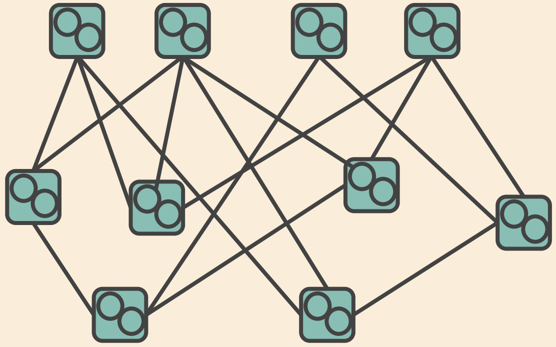
3

An accurate upper bound for throughput of Expanders and Clos topologies that scales well.

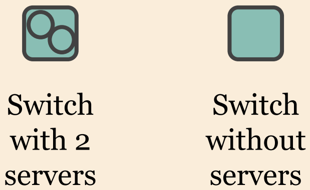
Clos vs Expanders



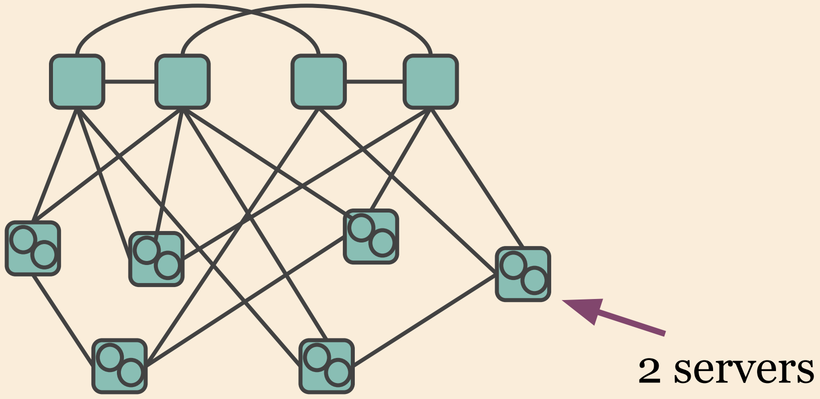
Clos



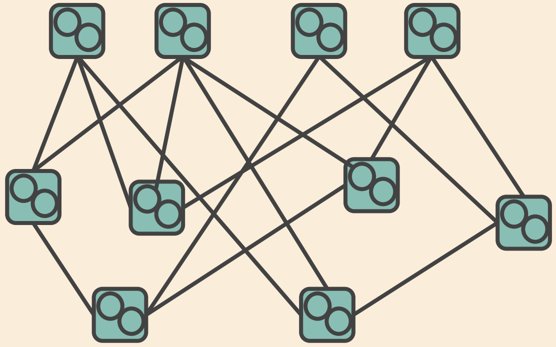
Expanders



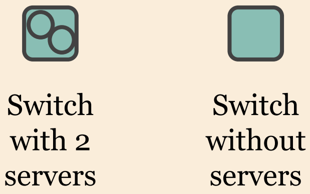
Clos vs Expanders



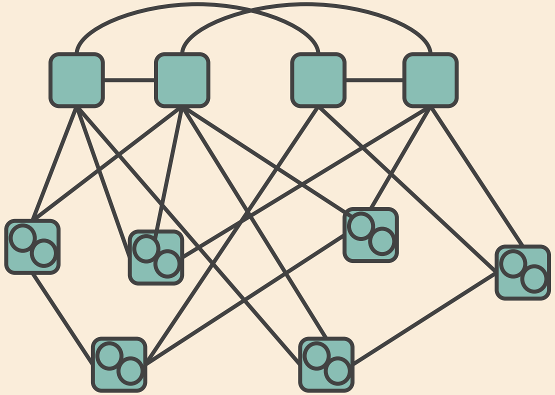
Clos



Expanders

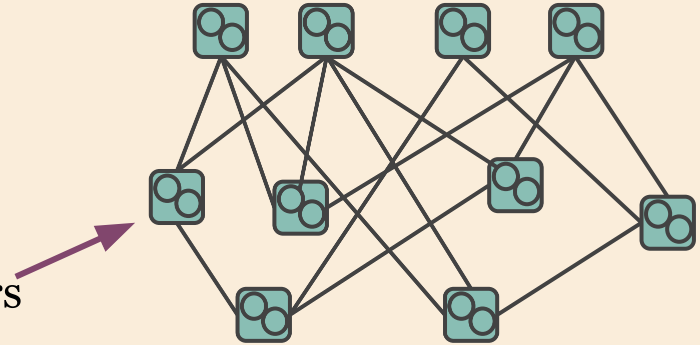


Clos vs Expanders



Clos

2 servers



Expanders

Switch with 2 servers

Switch without servers

Scaling Limitations (Expanders)

Servers Per Switch

	8	7	6
Full-Throughput	111K	256K	3.97M
Full-Bisection Bandwidth	>20M	>20M	>20M

Scaling Limitations (Expanders)

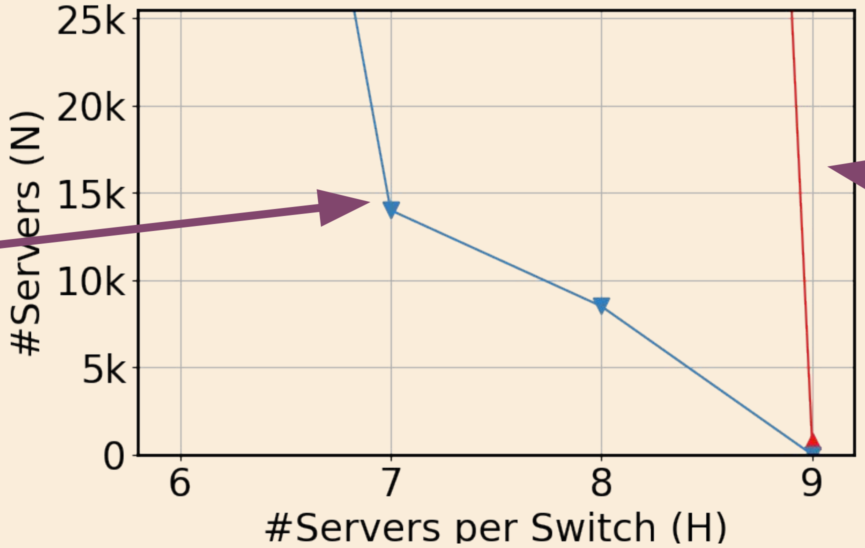
Servers Per Switch

	8	7	6
Full-Throughput	111K	256K	3.97M
Full-Bisection Bandwidth	>20M	>20M	>20M

If a designer wants a **non-blocking expander**, the size of the datacenters is **limited** (not so for Clos)

Scaling Limitations: Frontier Curve

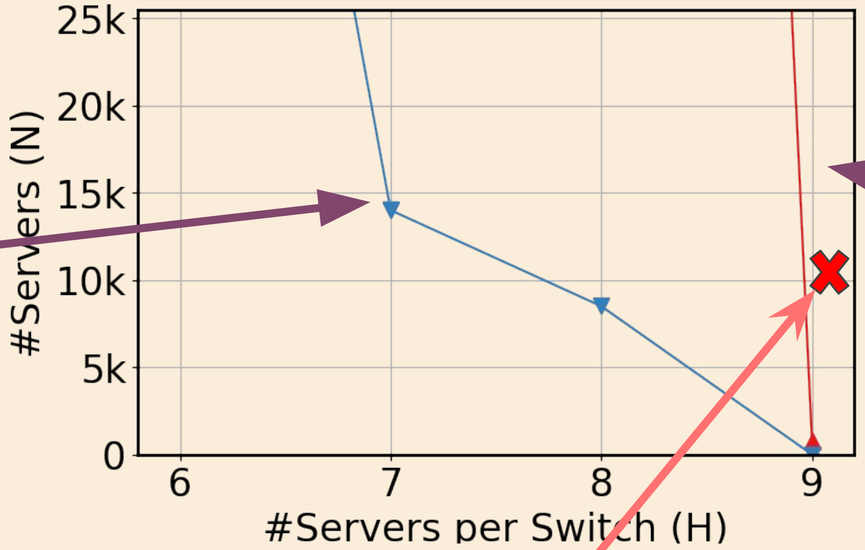
Throughput Frontier



Bisection Bandwidth Frontier

Scaling Limitations: Frontier Curve

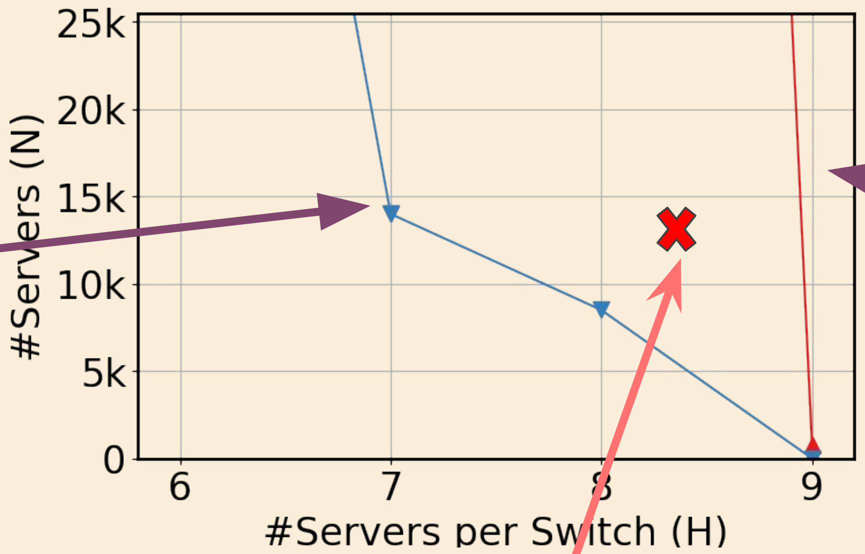
Throughput Frontier



Bisection Bandwidth Frontier

Not Full Bisection Bandwidth
Not Full Throughput

Scaling Limitations: Frontier Curve

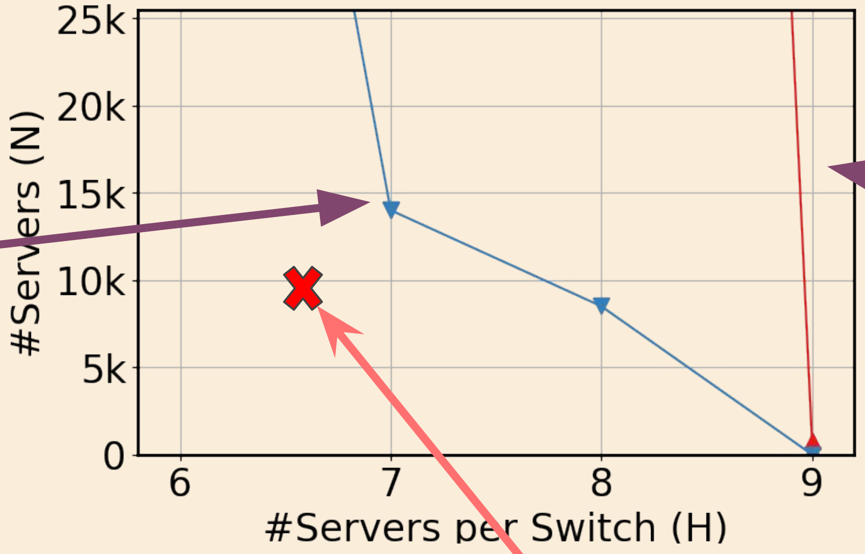


Throughput Frontier

Bisection Bandwidth Frontier

Full Bisection Bandwidth
Not Full Throughput

Scaling Limitations: Frontier Curve

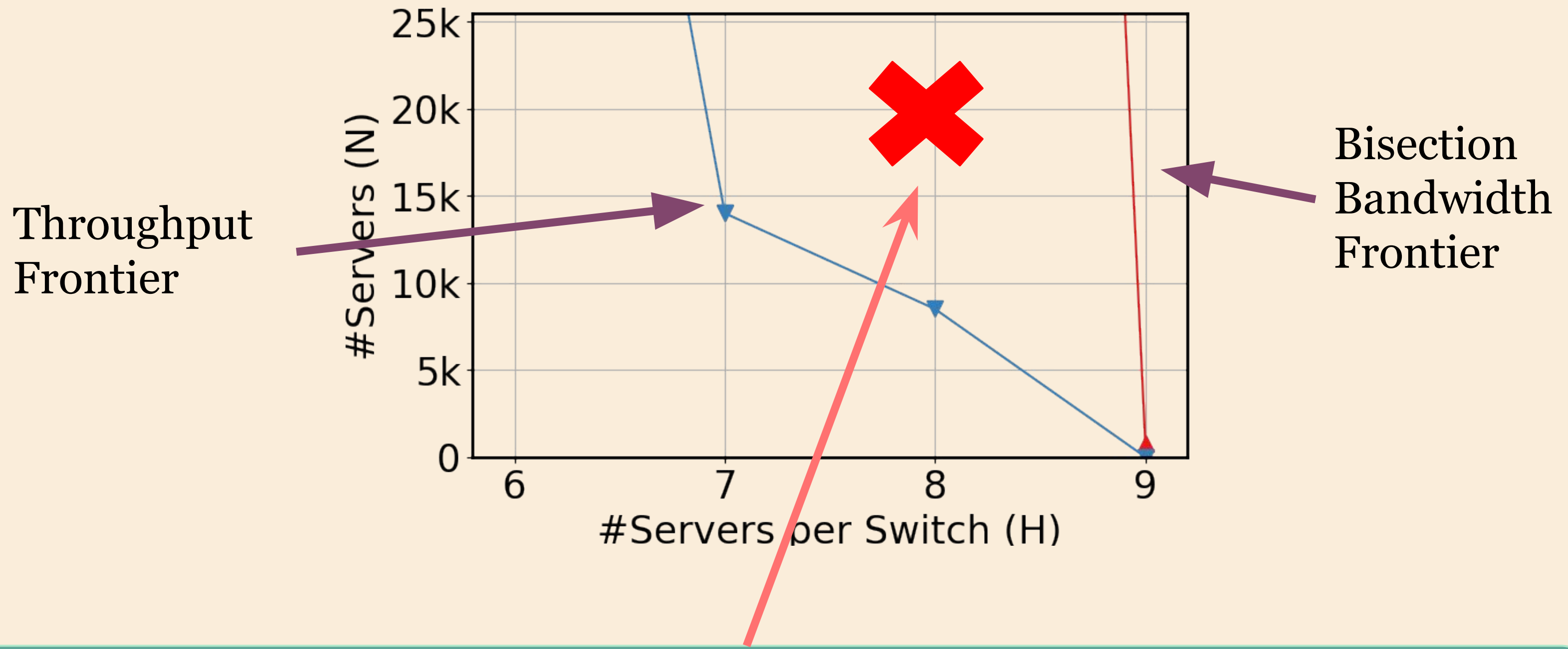


Throughput Frontier

Bisection Bandwidth Frontier

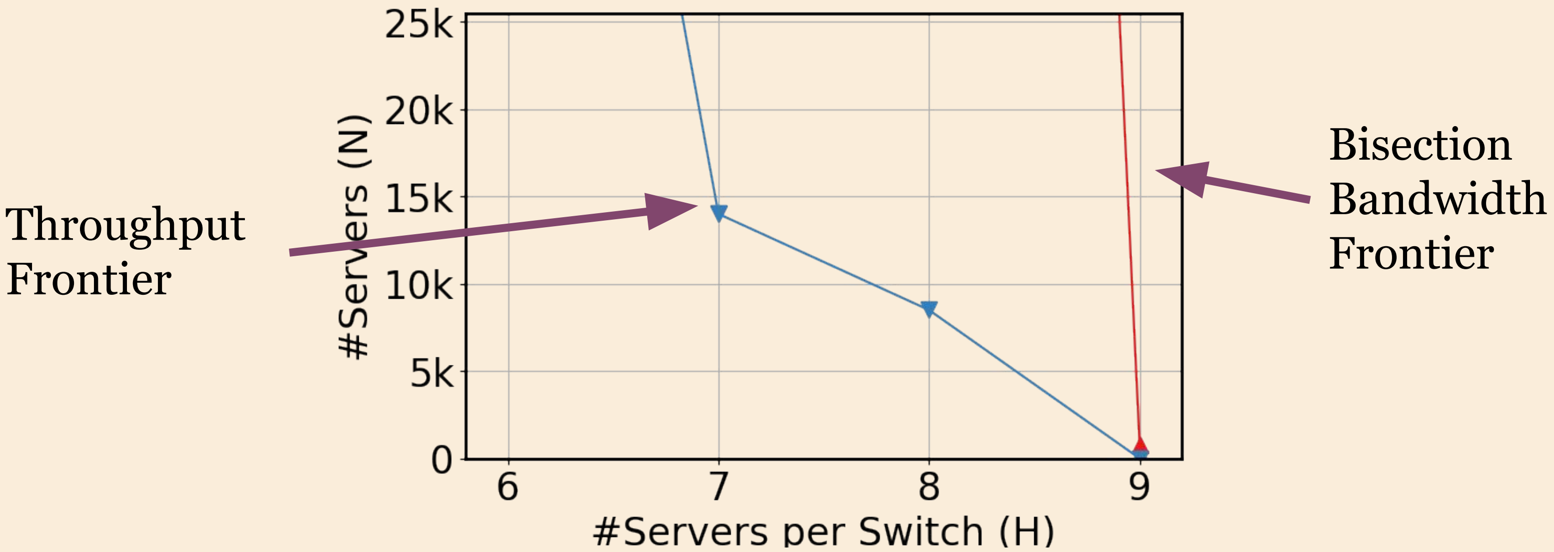
Full Bisection Bandwidth
Full Throughput

Scaling Limitations: Frontier Curve



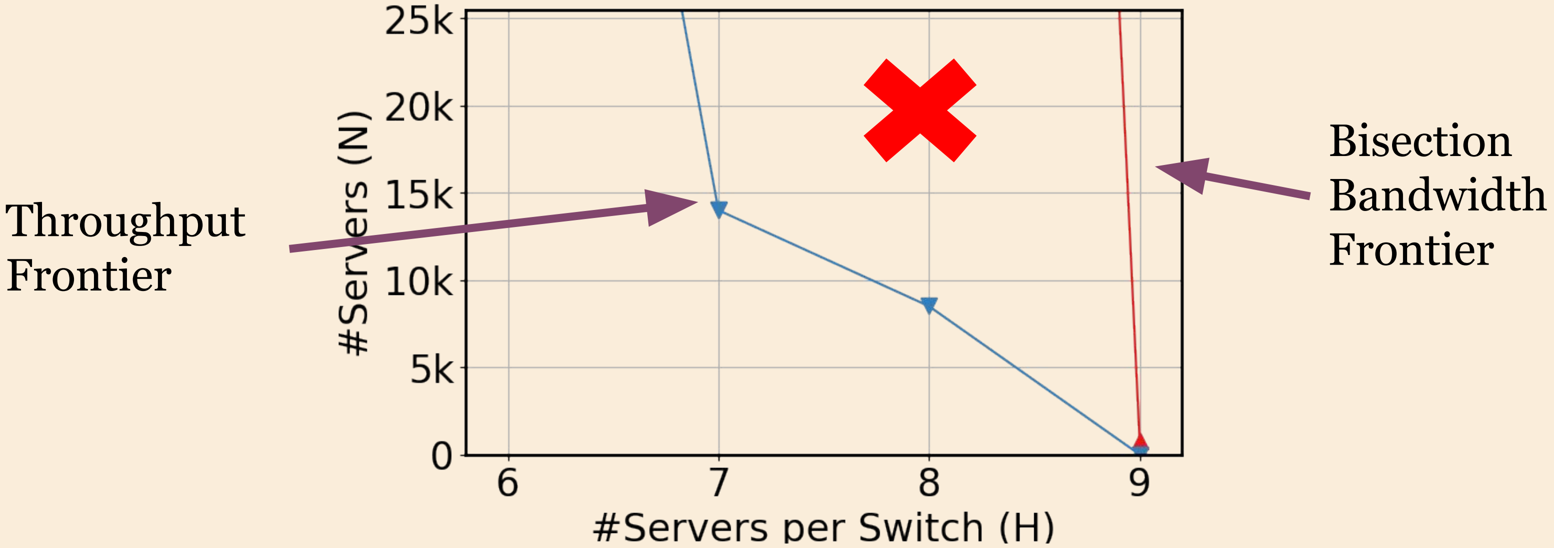
Full bisection bandwidth expanders may not be non-blocking (not so for Clos)

Scaling Limitations: Frontier Curve



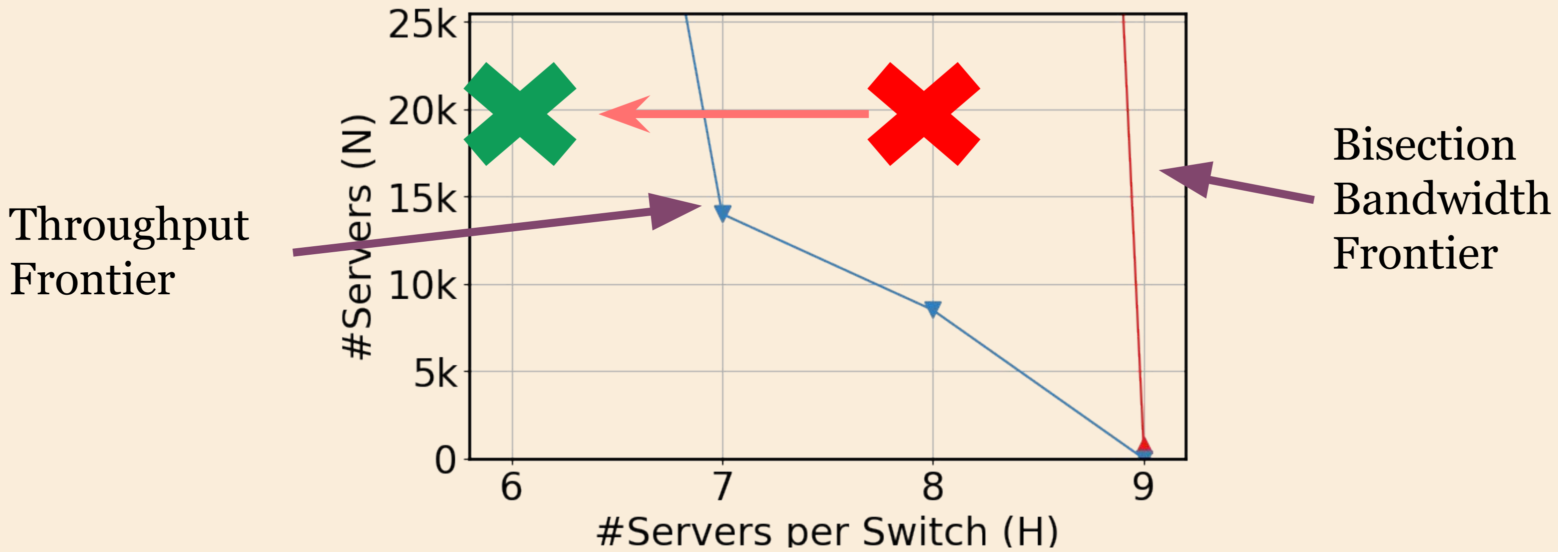
A designer may need to pick topology parameters carefully: even a small-scale expander may not be non-blocking

Scaling Limitations: Frontier Curve



A designer may need to pick topology parameters carefully: even a small-scale expander may not be non-blocking

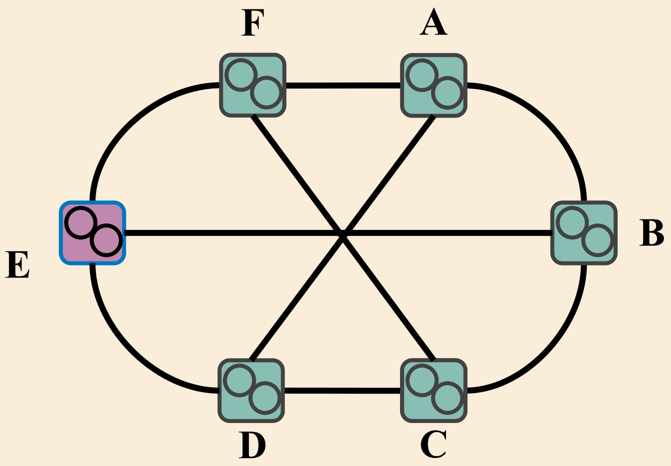
Scaling Limitations: Frontier Curve



A designer may need to pick topology parameters carefully: even a small-scale expander may not be non-blocking

Why Expanders have scaling limitations?

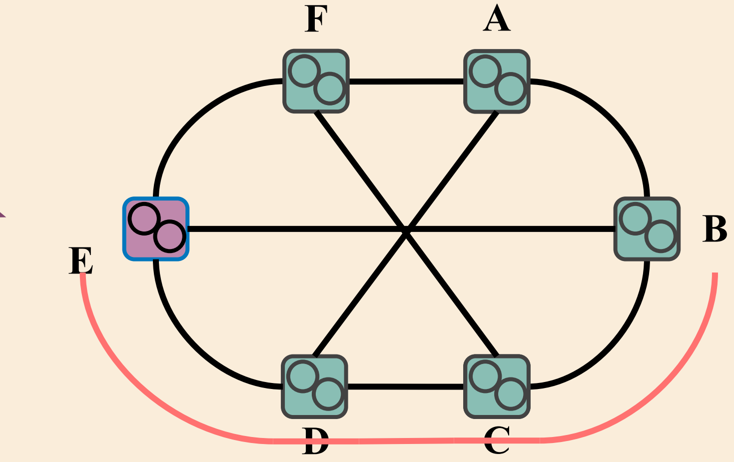
Two types of traffic in datacenter: Transit Traffic, Traffic originated/destined to connected server



Why Expanders have scaling limitations?

Two types of traffic in datacenter: Transit Traffic, Traffic originated/destined to connected server

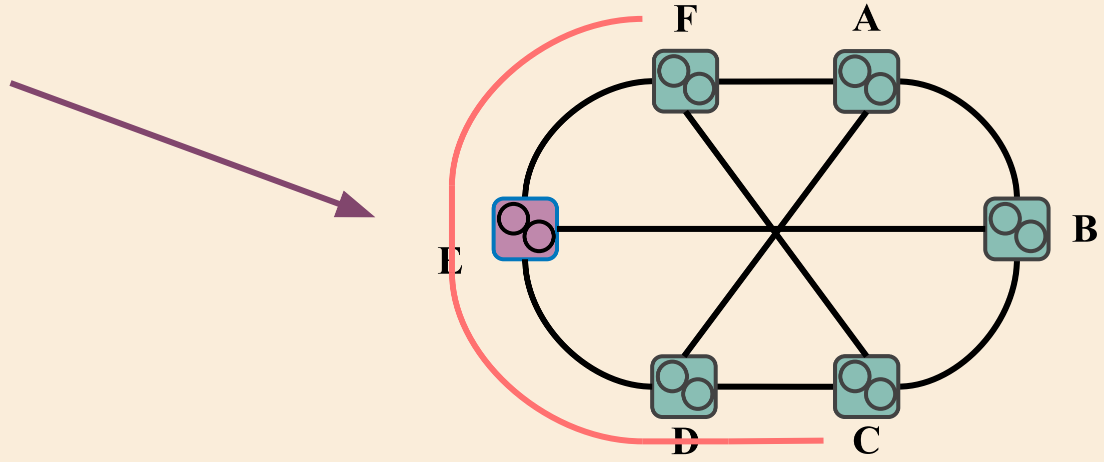
Traffic from/to the servers



Why Expanders have scaling limitations?

Two types of traffic in datacenter: Transit Traffic, Traffic originated/destined to connected server

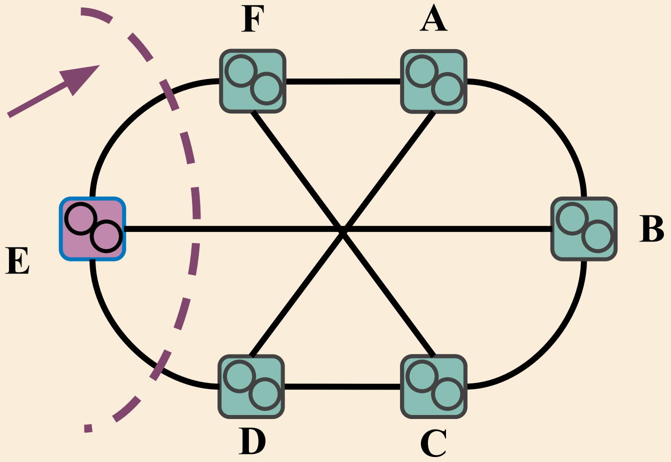
Transit Traffic



Why Expanders have scaling limitations?

Each switch has limited up-facing capacity.

Each Switch has 3 up-facing capacity

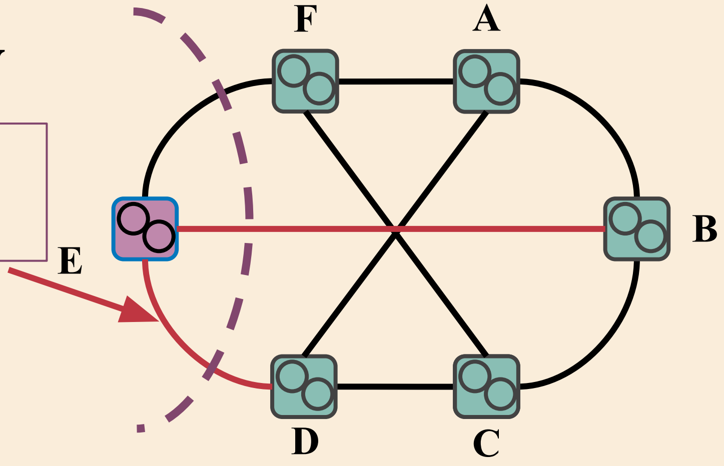


Why Expanders have scaling limitations?

In Expander, each switch has a fixed number of servers

Each Switch has 3 up-facing capacity

Each Switch connected to 2 Servers



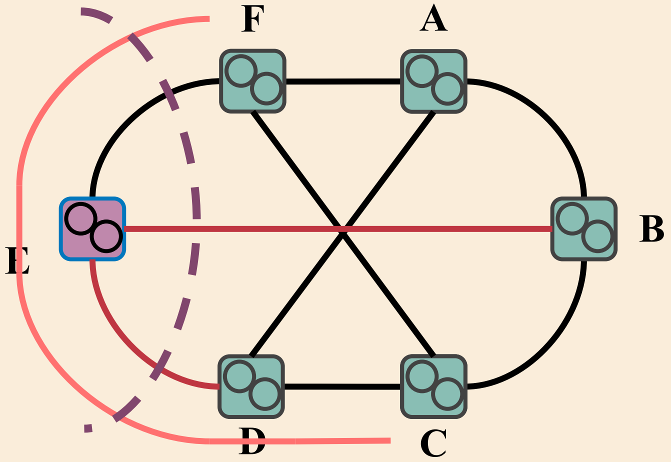
Why Expanders have scaling limitations?

In Expanders, each switch has limited capacity to handle transit traffic.

Each Switch has 3 up-facing capacity

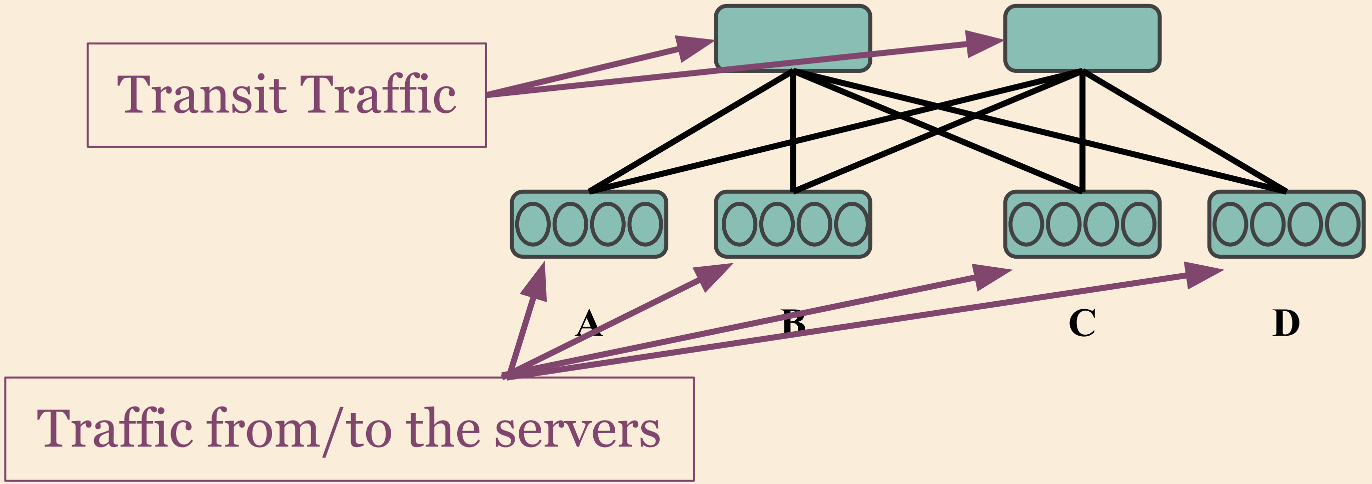
Each Switch connected to 2 Servers

1 up-facing capacity left for transit traffic



Why Expanders have scaling limitations?

In Clos, each switch either handles transit traffic or routes the traffic from/to their servers.



Why Expanders have scaling limitations?

In Clos, each switch either handles transit traffic or routes the traffic from/to their servers.

In Expanders, each switch handles both.

In Expander, number of servers per switch should be reduced so that each switch has more capacity left for transit traffic.

Outline

1

A full bisection bandwidth Expander may not have full throughput.

2

Cost, manageability, and failure resilience comparisons affected significantly when throughput is used at large-scale.

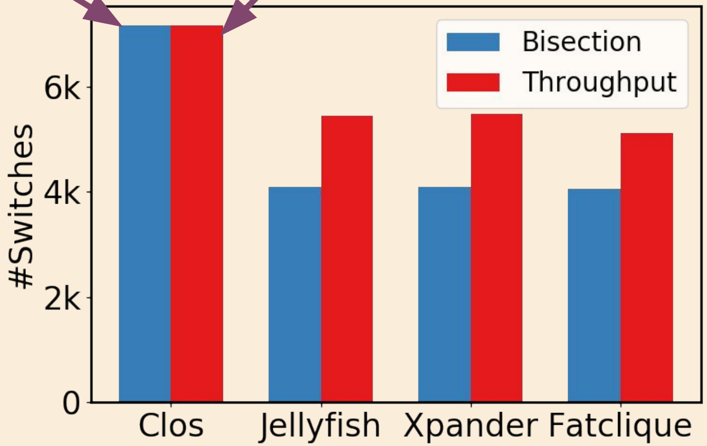
3

An accurate upper bound for throughput of Expanders and Clos topologies that scales well.

Cost Comparison

Full Bisection Bandwidth Clos

Full Throughput Clos

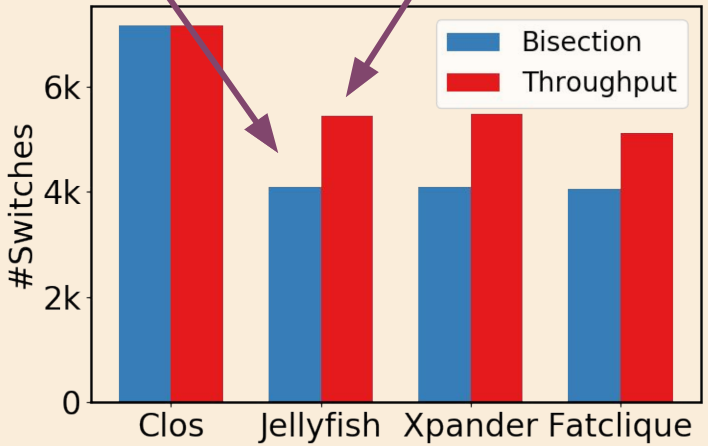


N = 32K, R=32

Cost Comparison

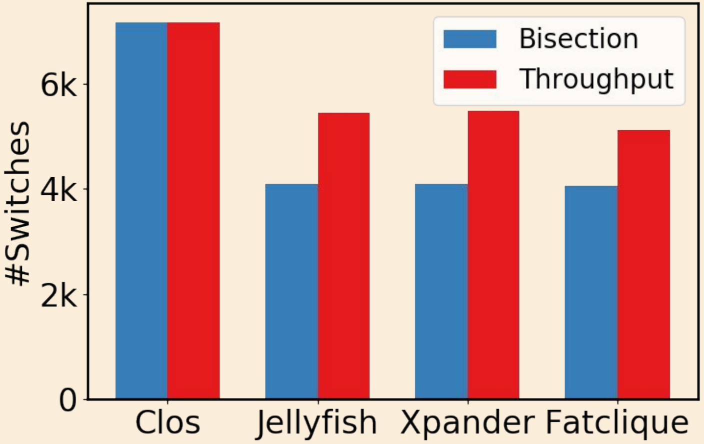
Full Bisection Bandwidth Jellyfish

Full Throughput Jellyfish

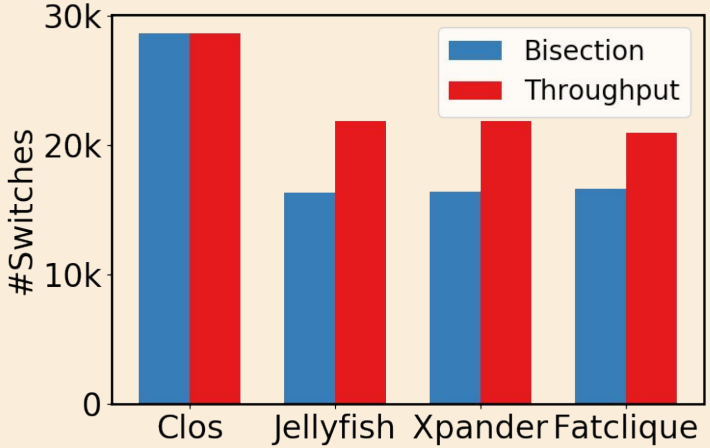


N = 32K, R=32

Cost Comparison

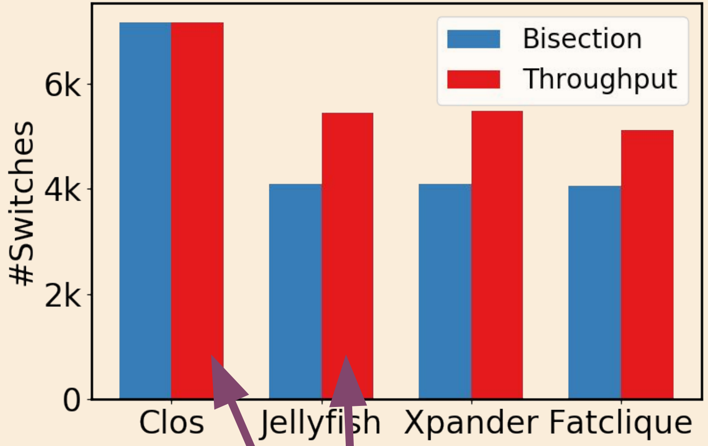


N = 32K, R=32

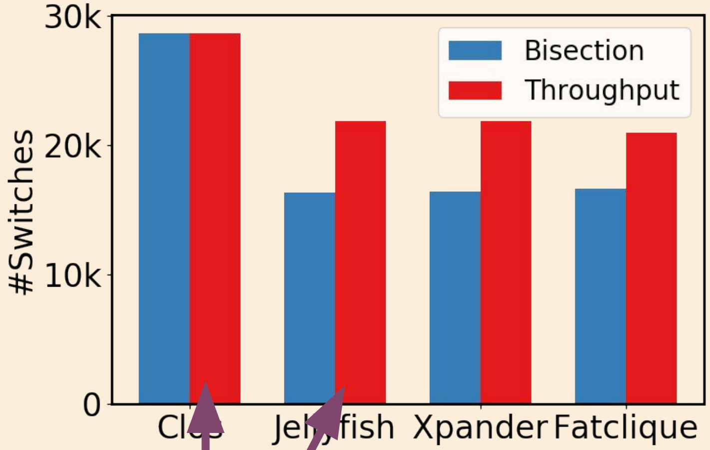


N = 131K, R=32

Cost Comparison



N = 32K, R=32



N = 131K, R=32

Expanders are less attractive from cost perspective! Their cost advantage over Clos drops by 2x when throughput is used.

Other Results

Expansion of Expanders requires advanced planning, otherwise it might cause throughput degradation.

Throughput measures the oversubscription ratio better than bisection bandwidth.

Expanders can deviate from perfect resiliency by up to 20%.

Outline

1

A full bisection bandwidth Expander may not have full throughput.

2

Cost, manageability, and failure resilience comparisons affected significantly when throughput is used at large-scale.

3

An accurate upper bound for throughput of Expanders and Clos topologies that scales well.

Throughput Upper Bound

Goal: Estimate throughput of a network

- Efficiently
- Accurately

Throughput of a topology

Minimum throughput over all the feasible traffic demands

Throughput of a Traffic Demand

Maximum scaling factor to make the traffic demand satisfiable.

Throughput of a Traffic Matrix

Maximum scaling factor to make the traffic matrix satisfiable.

	A	B	C
A	0	0	2
B	2	0	0
C	0	2	0

Throughput of a Traffic Matrix

Maximum scaling factor to make the traffic matrix satisfiable.

	A	B	C
A	0	0	2
B	2	0	0
C	0	2	0

	A	B	C
A	0	0	1.5
B	1.5	0	0
C	0	1.5	0

Throughput of a Traffic Matrix

Maximum scaling factor to make the traffic matrix satisfiable.

	A	B	C
A	0	0	2
B	2	0	0
C	0	2	0

Throughput = 0.75 →

	A	B	C
A	0	0	1.5
B	1.5	0	0
C	0	1.5	0

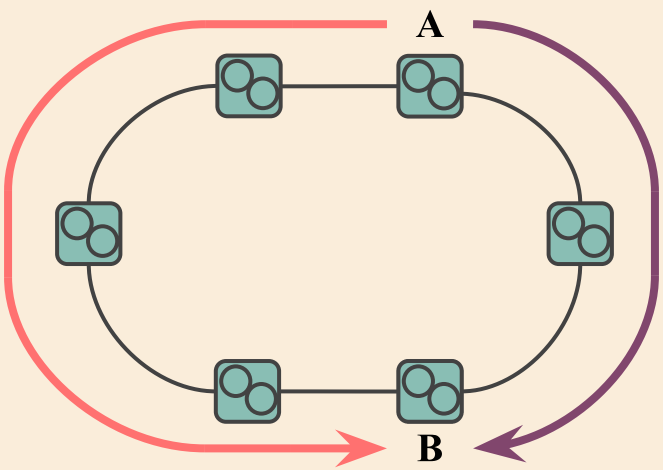
Hard to Compute Throughput of a Traffic Matrix

Throughput of a traffic matrix = Maximum scaling factor to make the traffic matrix satisfiable.

- LP Optimization → Does not scale to size of commercial datacenters

We Estimate an Upper Bound on Throughput

Routing each flow through the shortest path consumes the minimum capacity



We Estimate an Upper Bound on Throughput

Routing each flow through the shortest path consumes the minimum capacity

Assuming shortest paths provide enough diversity to handle all the flows

↓

Upper bound on throughput of a traffic demand

Hard to Compute Throughput of Network

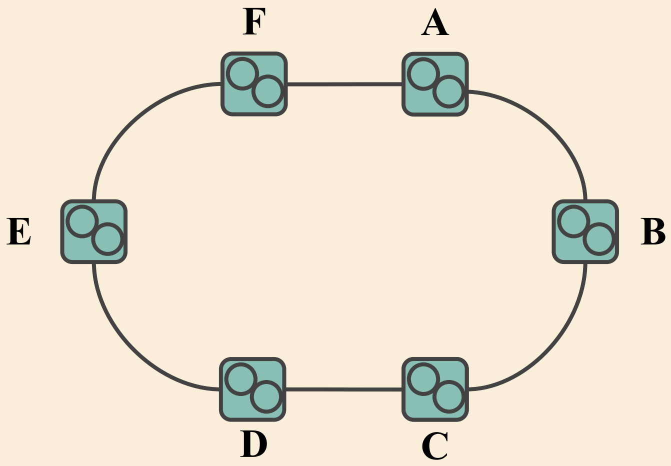
Throughput of a topology = Minimum throughput over all the feasible traffic matrices

- Infinite number of feasible traffic matrices

Our Approach: Focus on Permutation Traffic Matrices

Permutation Traffic

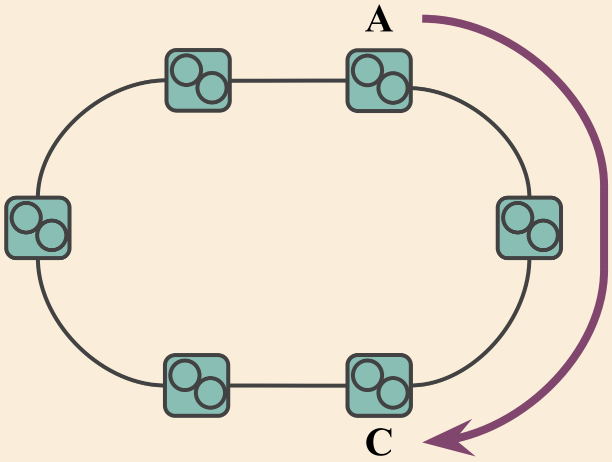
Each ToR sends/receives traffic to/from only one other ToR



Our Approach: Focus on Permutation Traffic Matrices

Permutation Traffic

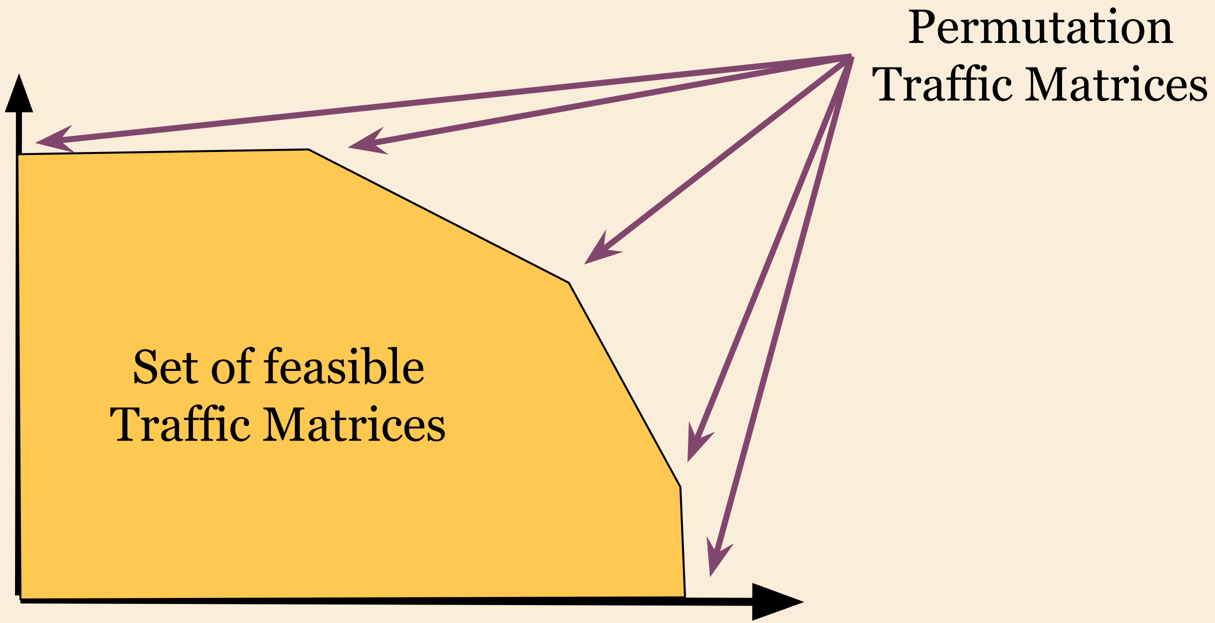
Each ToR sends/receives traffic to/from only one other ToR



	A	B	C	D	E	F
A	0	0	2	0	0	0

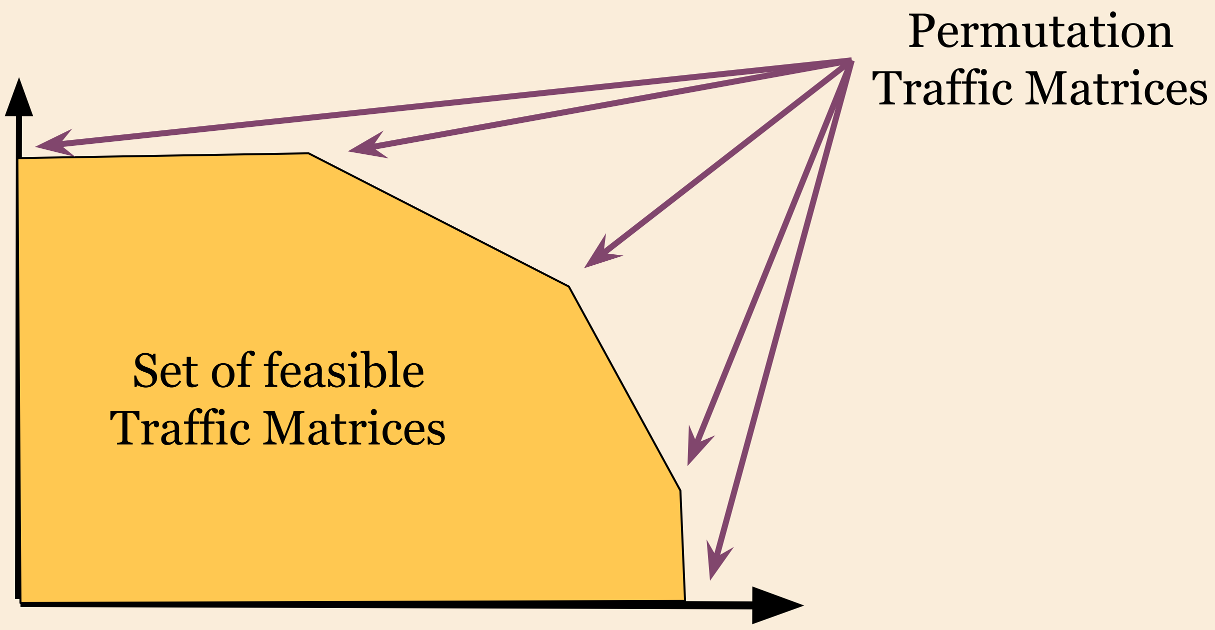
Our Approach: Focus on Permutation Traffic Matrices

Every Traffic Demand is a convex combination of Permutation Traffic Matrices.



Our Approach: Focus on Permutation Traffic Matrices

Permutation Traffic Matrices are sufficient to find the Minimum Throughput.



A Maximal Permutation Matrix has Lowest Throughput

Still Infeasible to Enumerate all the Permutation Traffic Matrices!!!

A Maximal Permutation Matrix has Lowest Throughput

Still Infeasible to Enumerate all the Permutation Traffic Matrices!!!

Assuming shortest paths provide enough diversity → Upper bound

A Maximal Permutation Matrix has Lowest Throughput

Still Infeasible to Enumerate all the Permutation Traffic Matrices!!!

Assuming shortest paths provide enough diversity → Upper bound

Maximal Permutation Traffic

Permutation Traffic with longest total shortest path length

Algorithm for Throughput Upper Bound (TUB)

1

Compute all pairs shortest path lengths

Algorithm for Throughput Upper Bound (TUB)

1

Compute all pairs shortest path lengths

2

Find Maximal Permutation Matrix using
Maximum Weight Matching in Full Bipartite Graph

S. A. Jyothi et. al. "Measuring and Understanding Throughput of Network Topologies" SC '16

Algorithm for Throughput Upper Bound (TUB)

1

Compute all pairs shortest path lengths

2

Find Maximal Permutation Matrix using
Maximum Weight Matching in Full Bipartite Graph

S. A. Jyothi et. al. "Measuring and Understanding Throughput of Network Topologies" SC '16

3

Compute Upper bound on Throughput of Maximal Permutation

Accuracy & Scalability

Evaluation Set up

Baseline

- K-shortest path MCF with high enough K on Maximal Permutation TM (KSP-MCF)

Throughput Gap

- Absolute difference from KSP-MCF

Comparison Alternatives

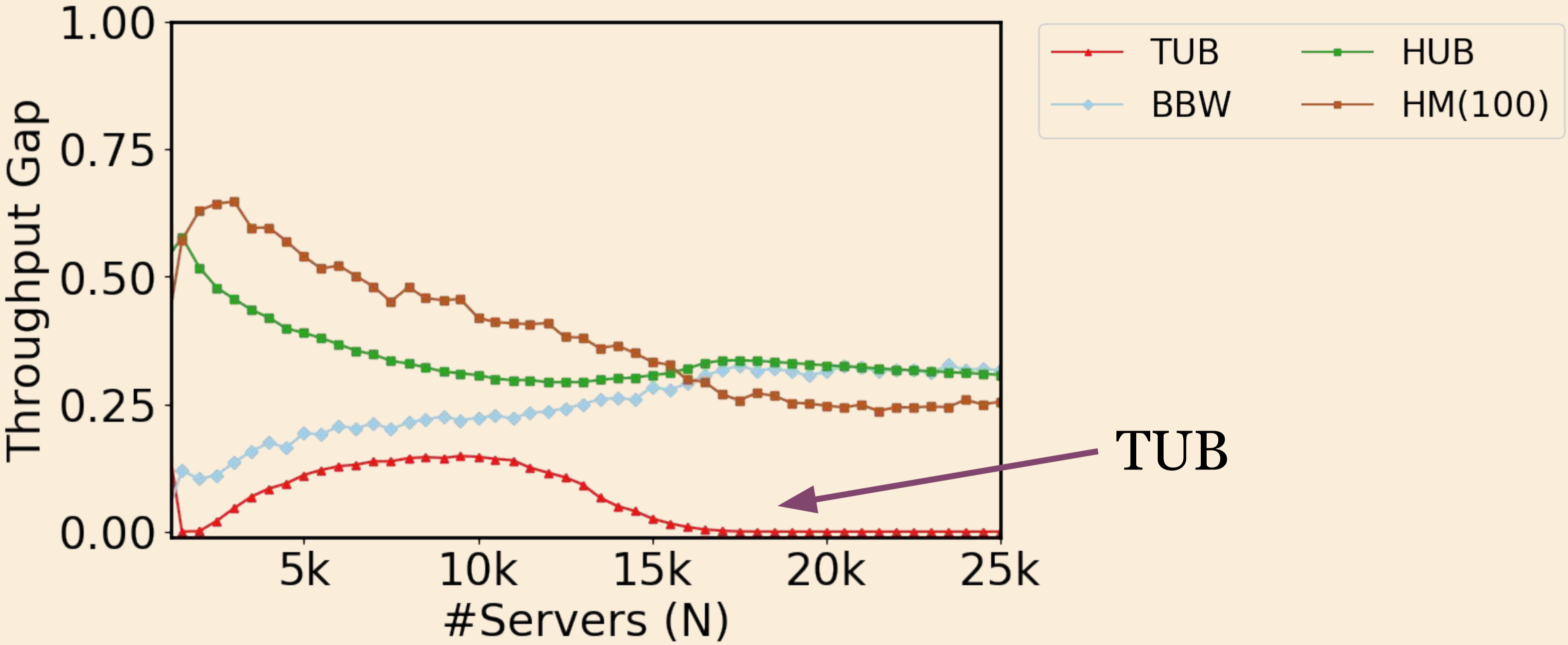
1) Bisection Bandwidth (BBW)

2) Upper-bound in (HUB)
   - A. Singla et. al. “High Throughput Data Center Topology Design” NSDI’14

3) Hoefler’s method (HM)
   - T. Hoefler et. al. “Multistage switches are not crossbars: Effects of static routing in high-performance networks”, 2008 IEEE International Conference on Cluster Computing
   - X. Yuan et. al. “ A New Routing Scheme for Jellyfish and Its Performance with HPC Workloads“ SC’13

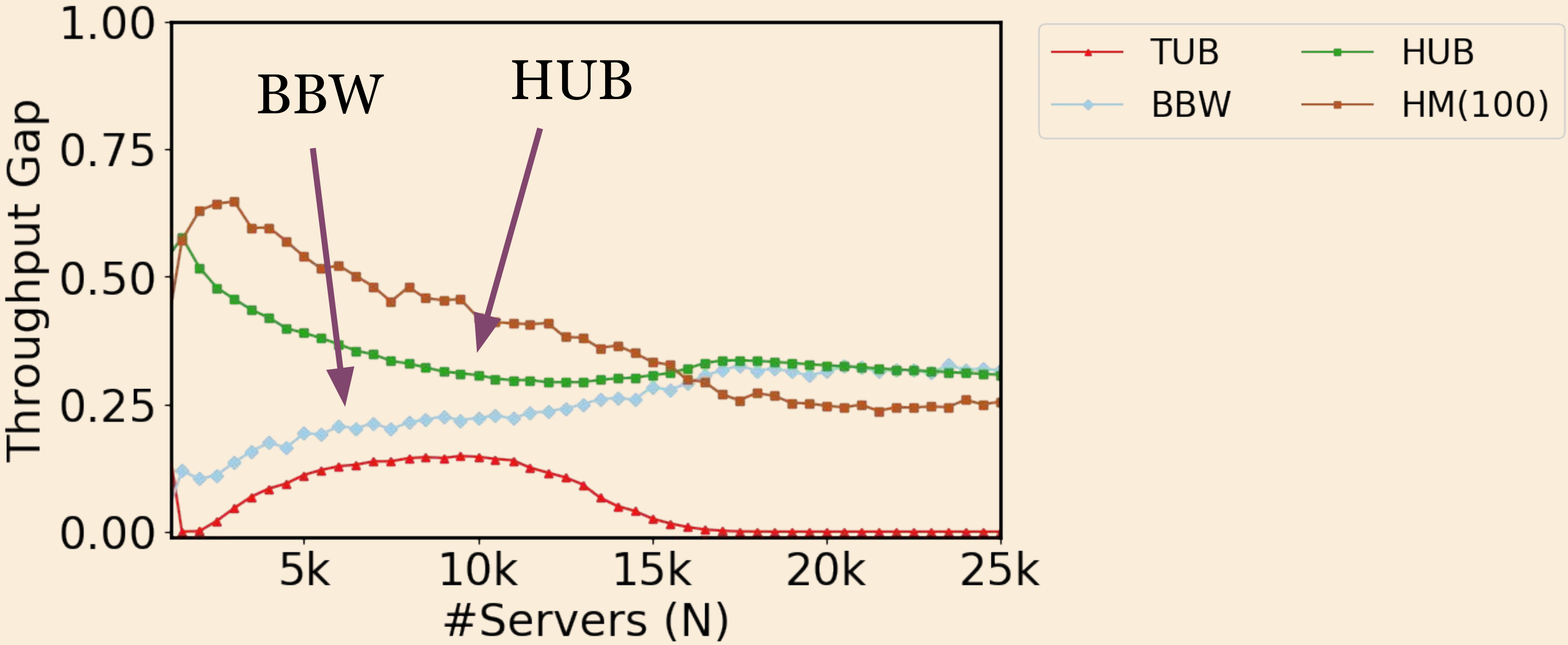
Comparison

Our Upper bound (TUB) is more accurate than other alternatives.



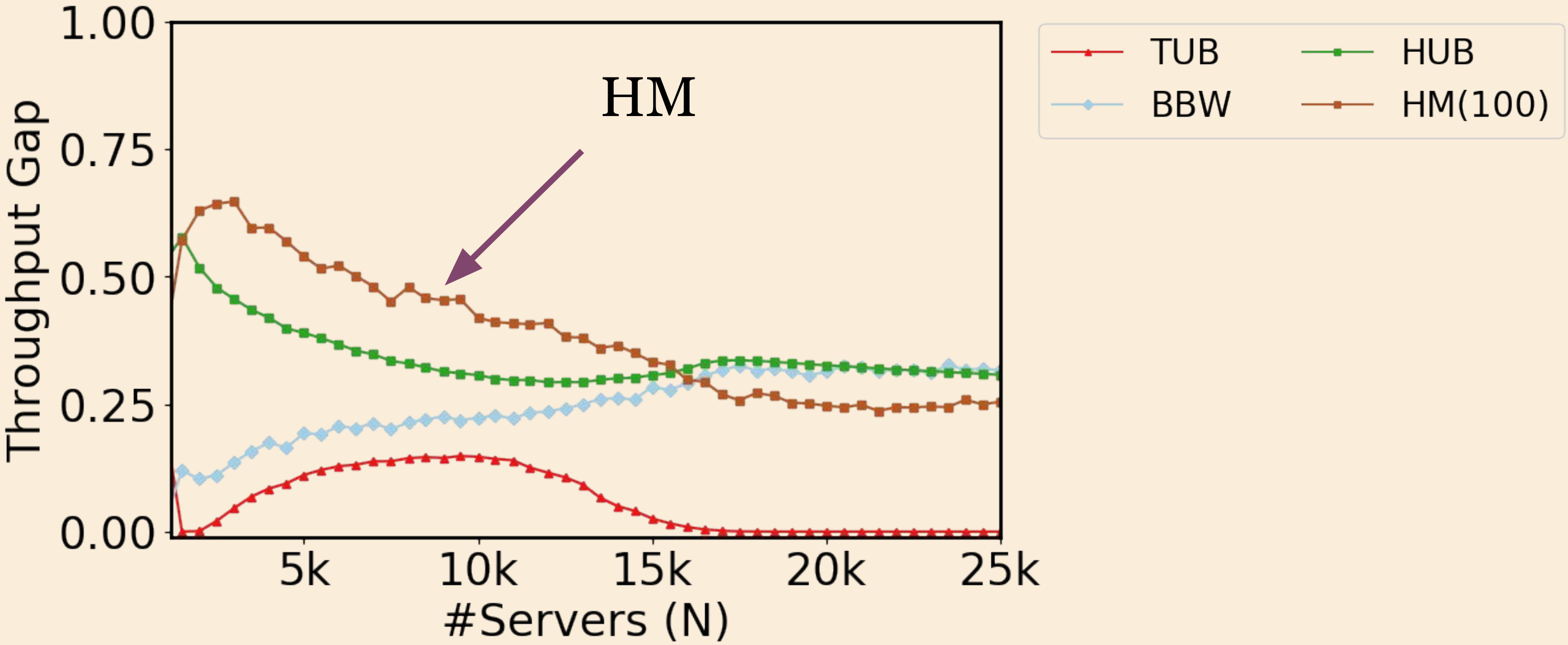
Comparison

Our Upper bound (TUB) is more accurate than other alternatives.



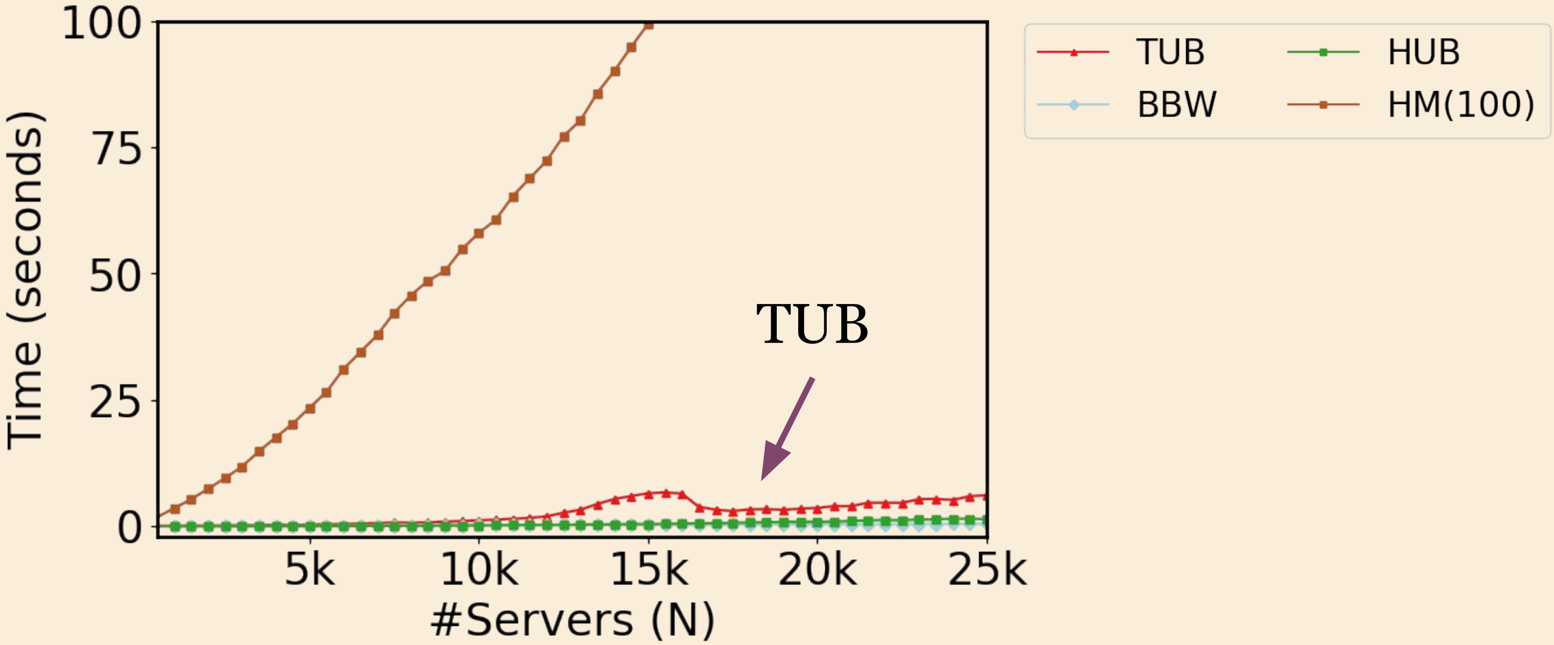
Comparison

Our Upper bound (TUB) is more accurate than other alternatives.



Comparison

Our Upper bound (TUB) scales scales as well or better than alternatives.



Conclusion

1. A full bisection bandwidth Expander may not have full throughput.

2. Cost, manageability, and failure resilience comparisons affected significantly when throughput is used at large-scale.

3. An accurate upper bound for throughput of Expanders and Clos topologies that scales well.

Future Work

- Practical routing evaluation
- Parallel Throughput upper bound computation
- Further Improvement of accuracy

Thank you!

Email: namyar@usc.edu
Twitter: @PooriaNamyar