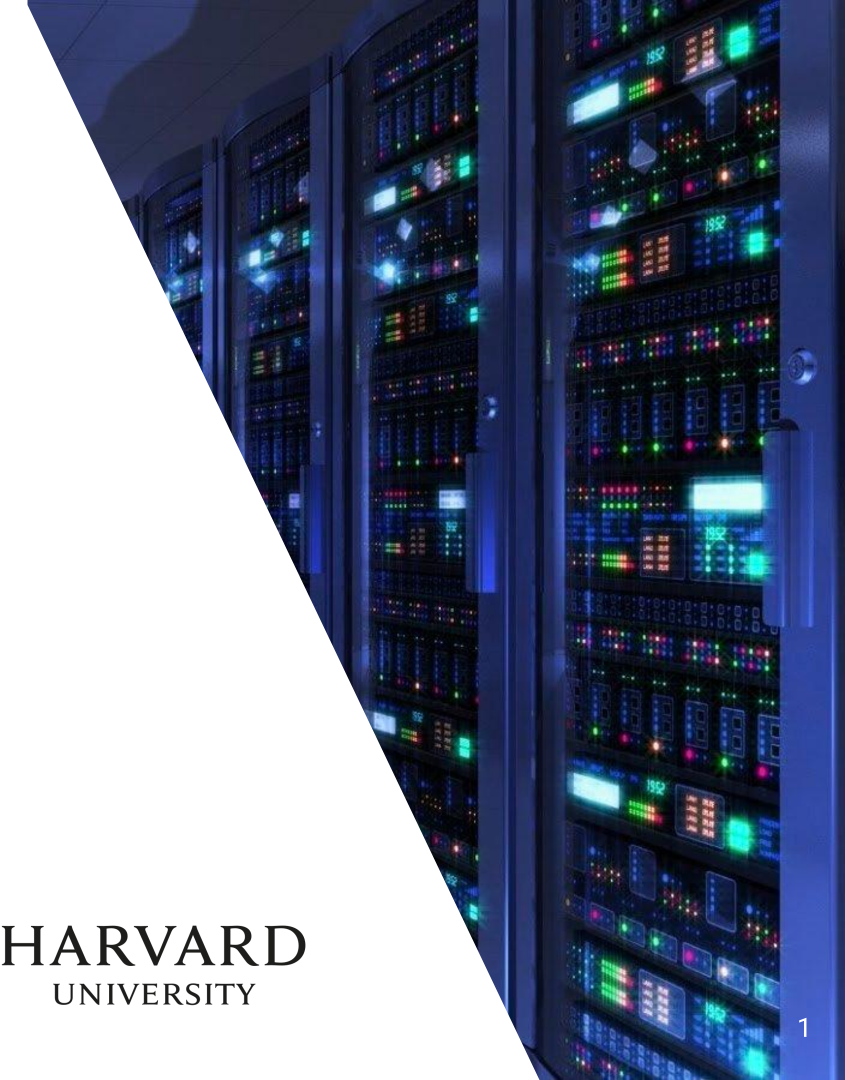
A **Throughput-Centric** View of the Performance of Datacenter Topologies

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



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

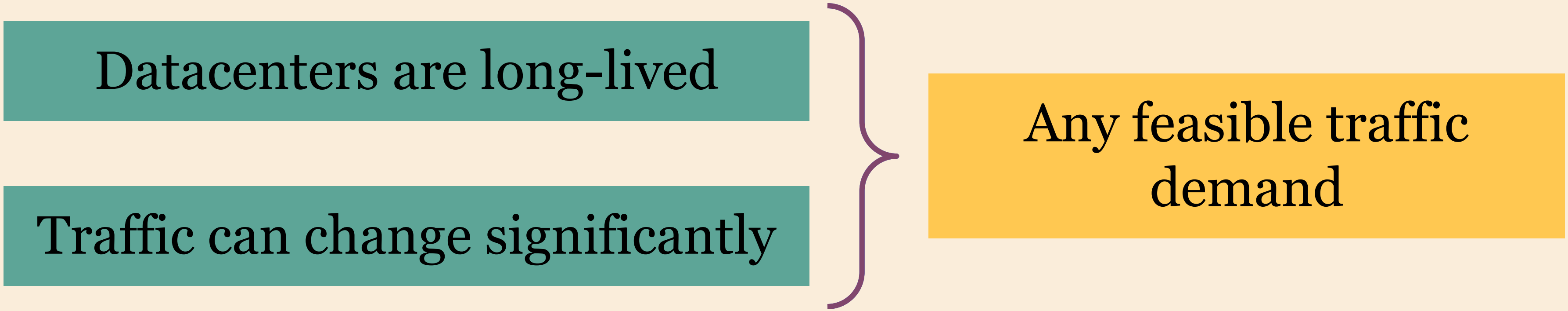
Datacenters are long-lived

Why any possible traffic demand

Datacenters are long-lived

Traffic can change significantly

Why any possible traffic demand

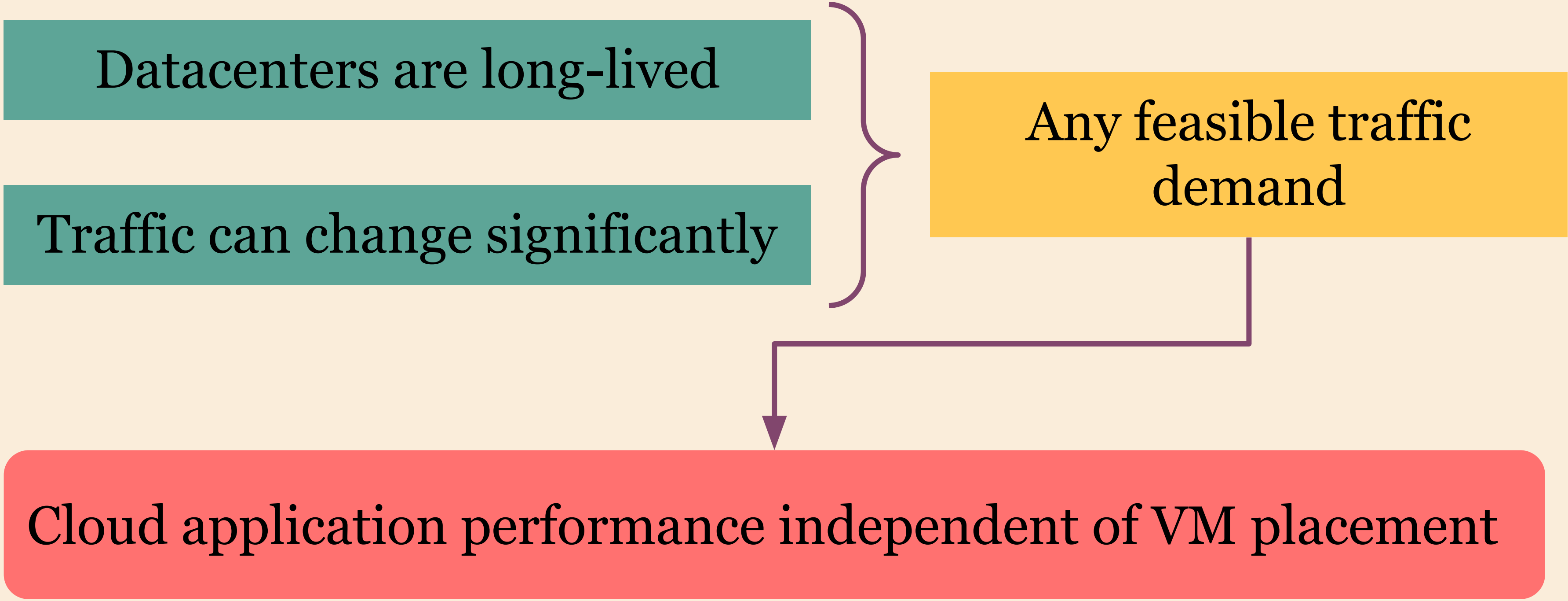


Datacenters are long-lived

Traffic can change significantly

Any feasible traffic demand

Why any possible traffic demand



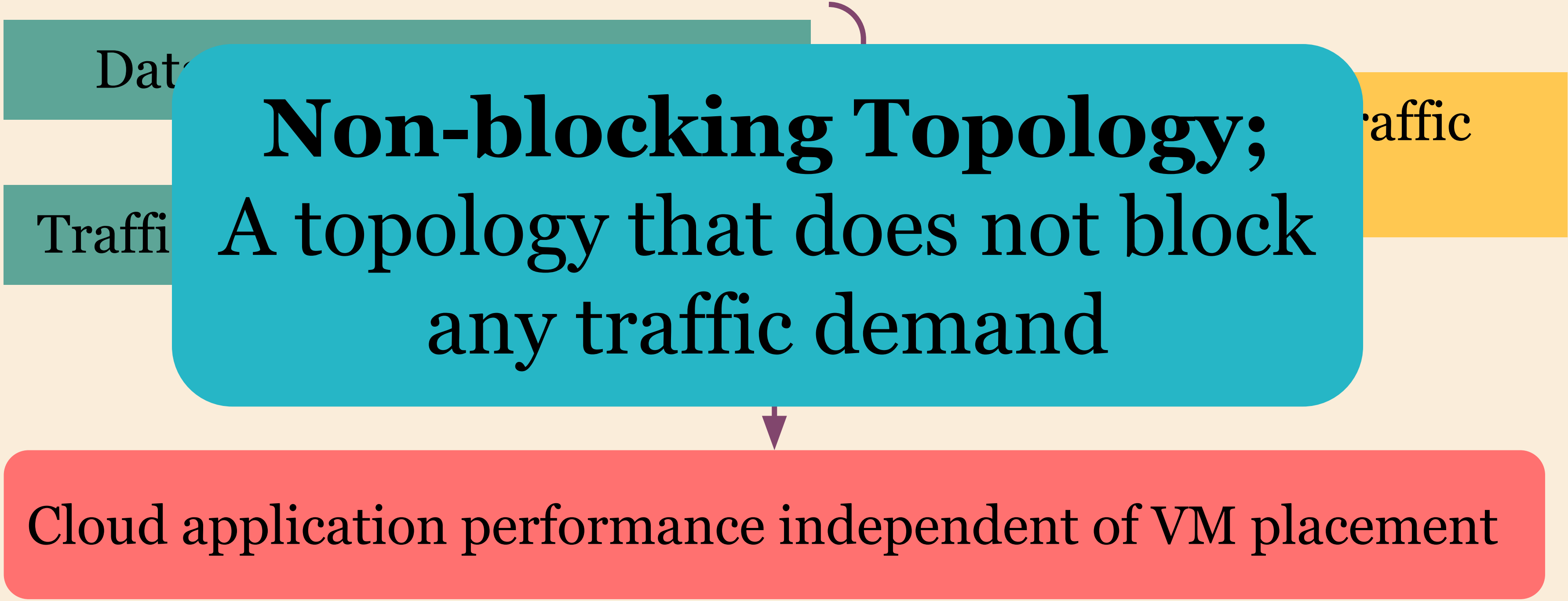
Datacenters are long-lived

Traffic can change significantly

Any feasible traffic demand

Cloud application performance independent of VM placement

Why any possible traffic demand



Dat

Traffi

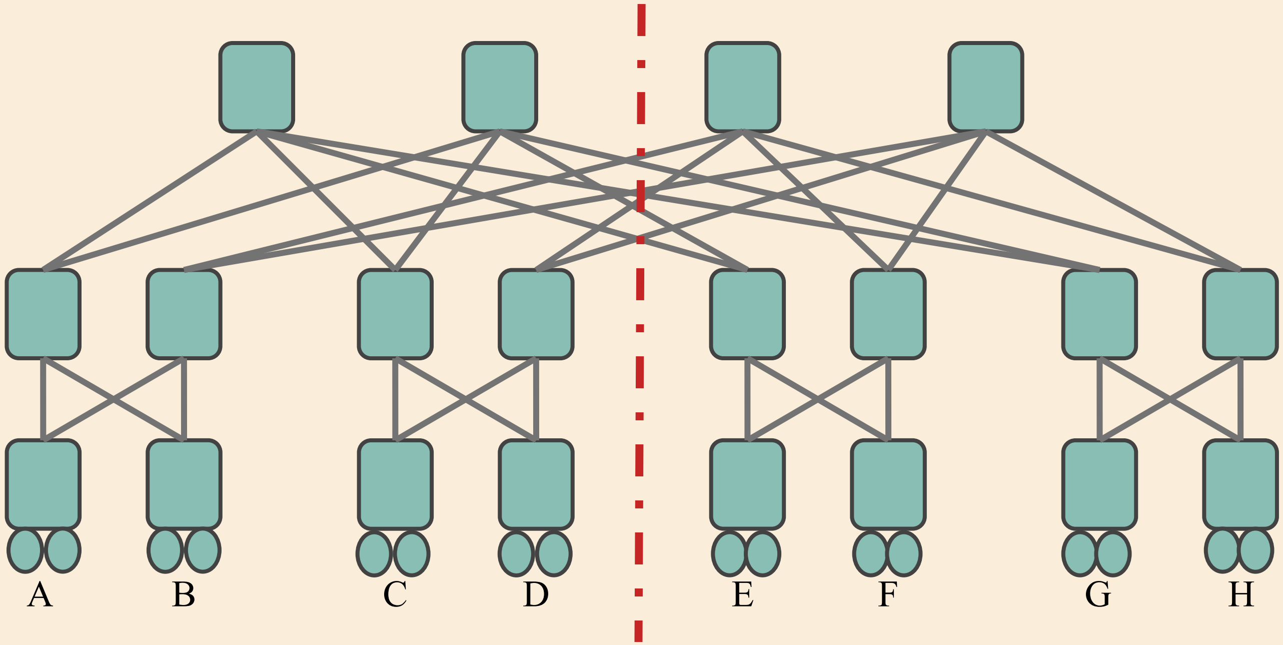
affic

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

Cloud application performance independent of VM placement

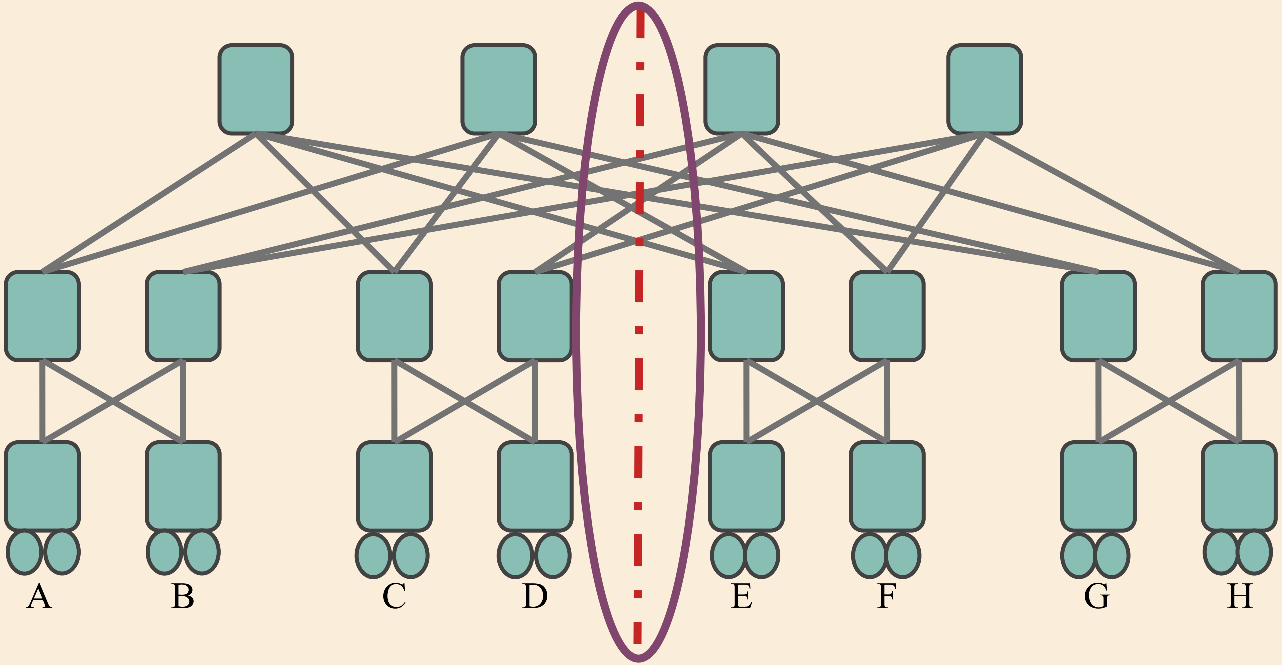
How to assess whether a datacenter topology is non-blocking?

Early Work uses Bisection Bandwidth



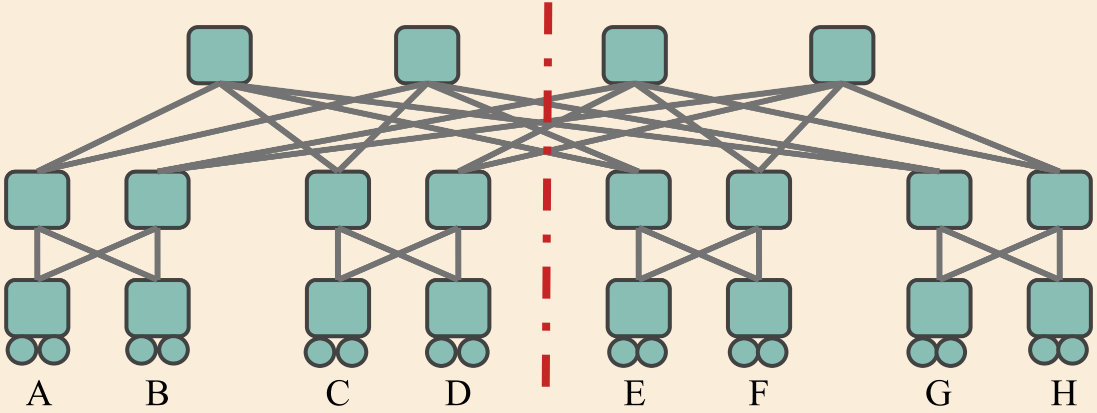
Bisection Bandwidth

Early Work uses Bisection Bandwidth



Bisection Bandwidth

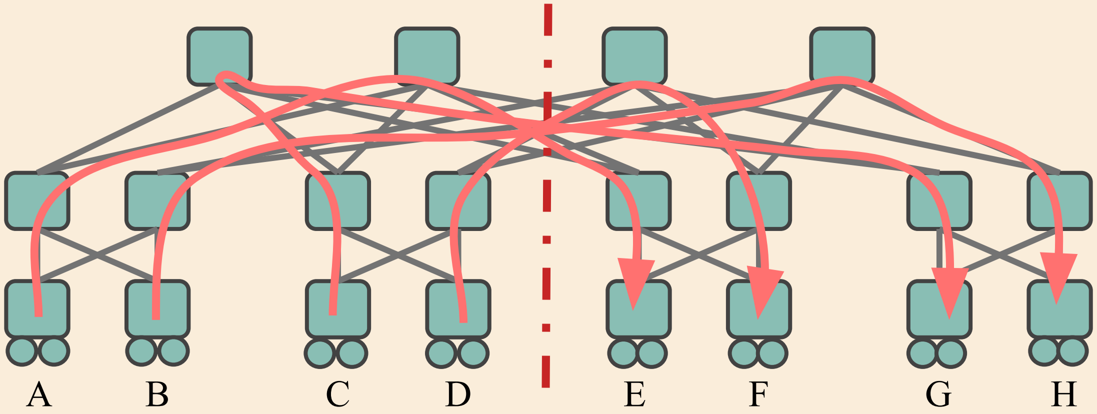
Early Work uses Bisection Bandwidth



Full Bisection Bandwidth

Bisection Bandwidth ≥ #servers/2

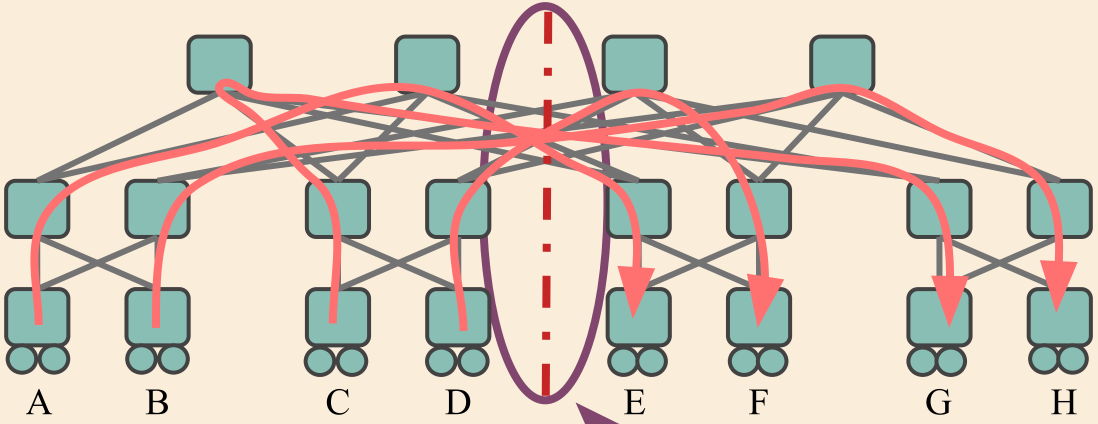
Early Work uses Bisection Bandwidth



Full Bisection Bandwidth

Bisection Bandwidth ≥ #servers/2

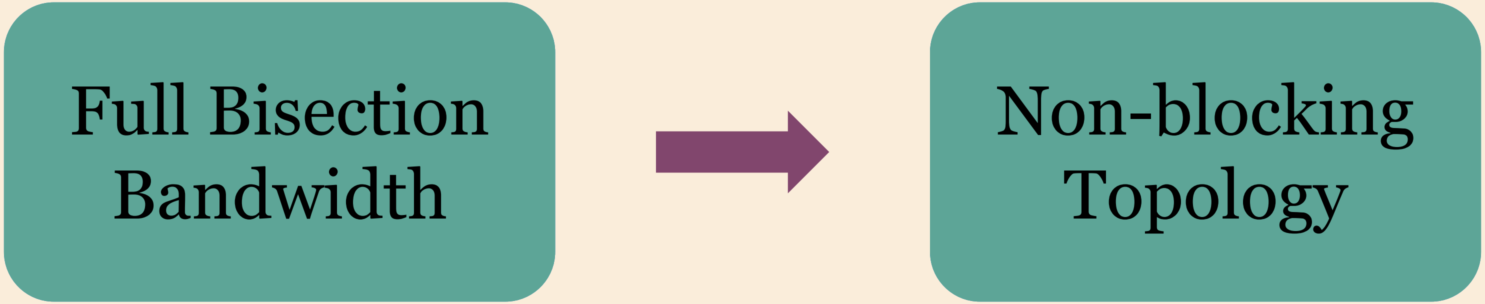
Early Work uses Bisection Bandwidth



Full Bisection Bandwidth

Bisection Bandwidth ≥ #servers/2

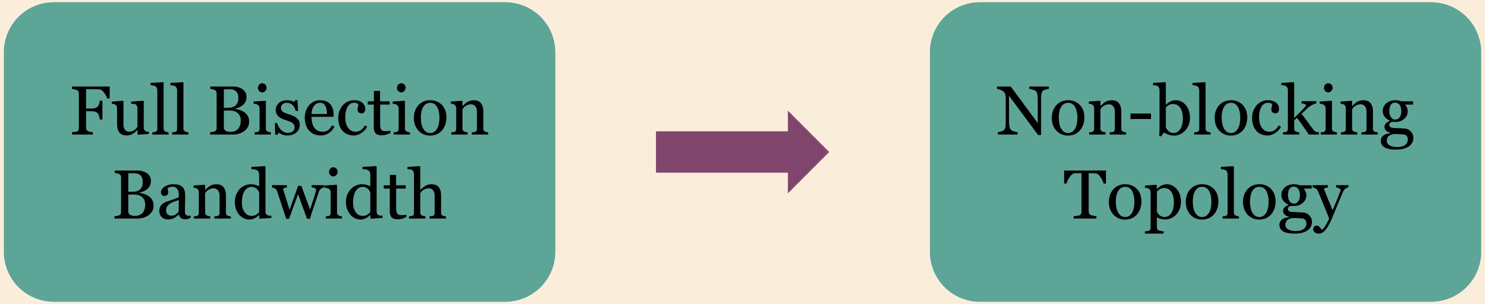
Early Work uses Bisection Bandwidth



Full Bisection Bandwidth

Non-blocking Topology

Early Work uses Bisection Bandwidth



Full Bisection Bandwidth

Non-blocking Topology

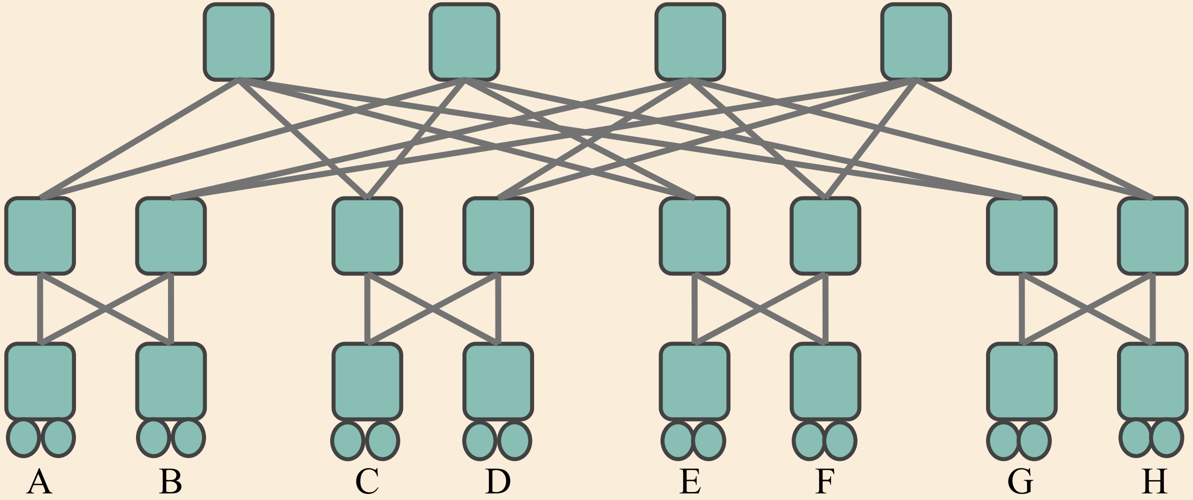
This holds for a specific topology family called **Clos**.

Most Commercial Datacenters are Clos

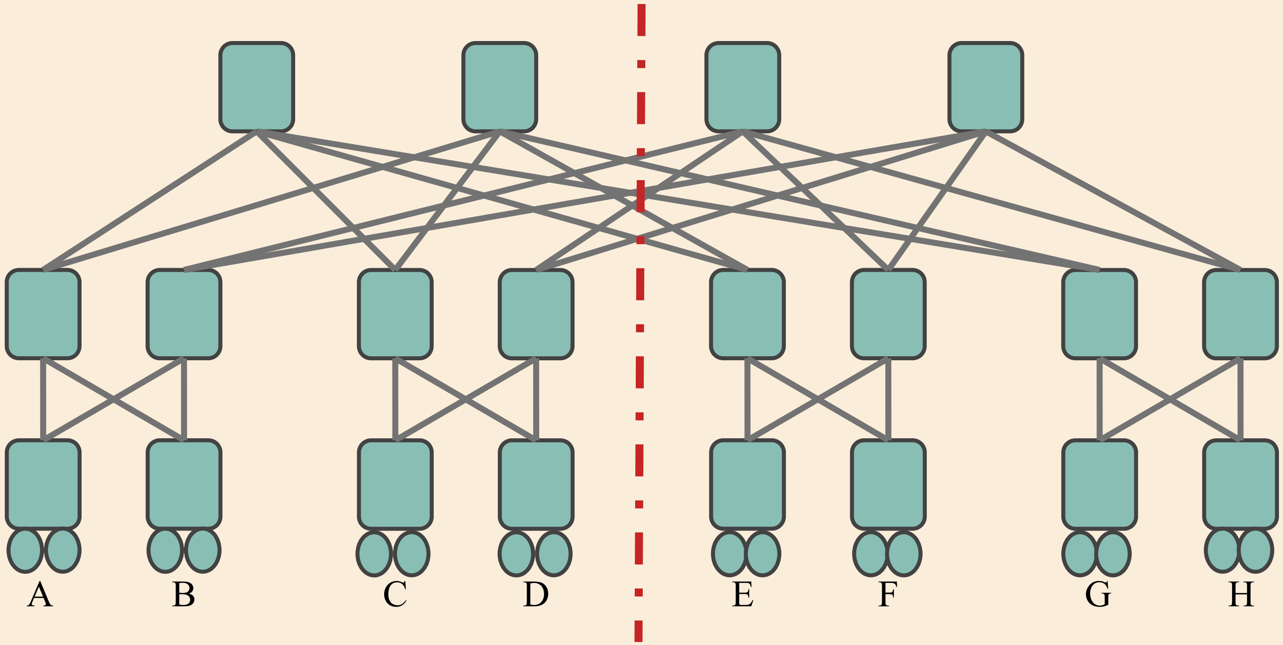
Microsoft

Google

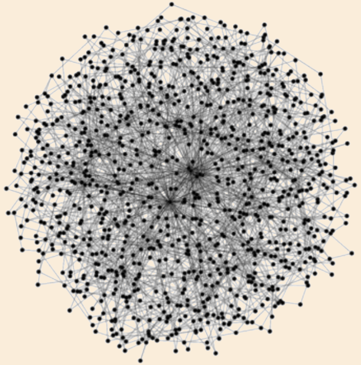
facebook



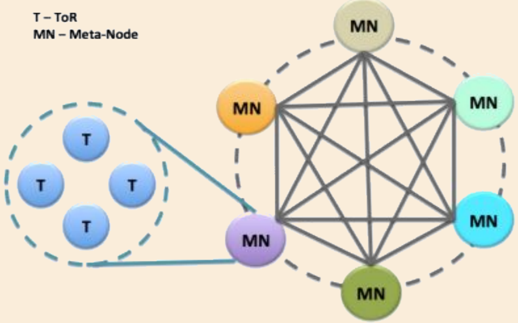
But Clos is Expensive



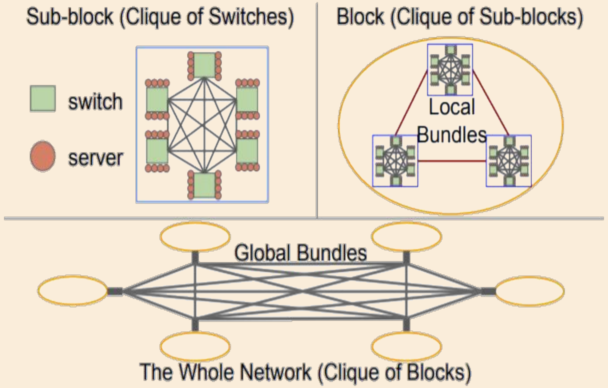
Recently Proposed Topologies: Expanders



Jellyfish
[NSDI'12]



Xpander
[CoNEXT'16]


Sub-block (Clique of Switches)
Block (Clique of Sub-blocks)
switch
server
Local Bundles
Global Bundles
The Whole Network (Clique of Blocks)

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.

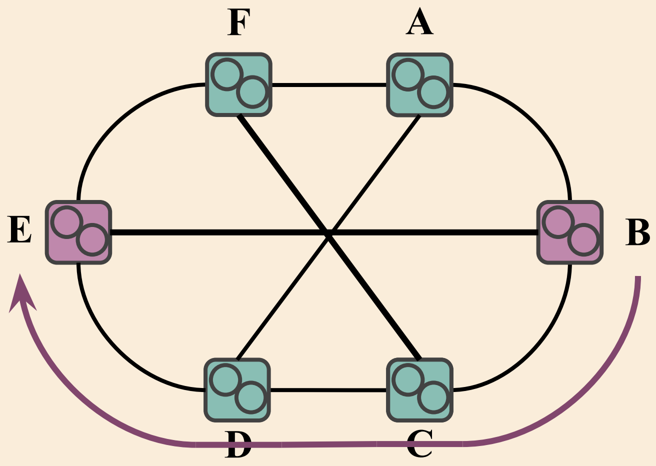
Prior Work Has Proposed Another Metric

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

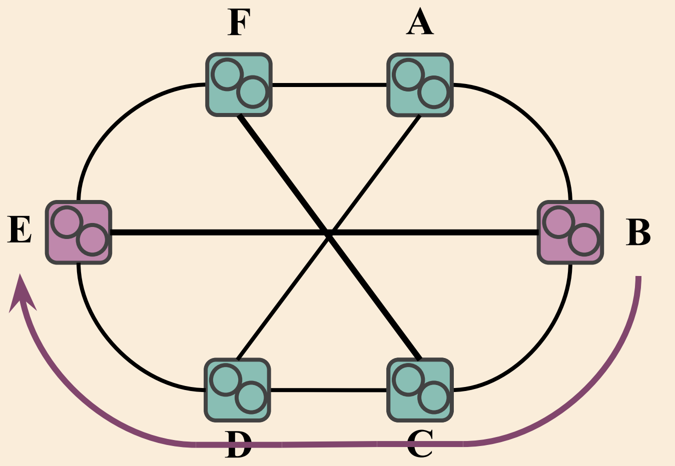


Prior Work Has Proposed Another Metric

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



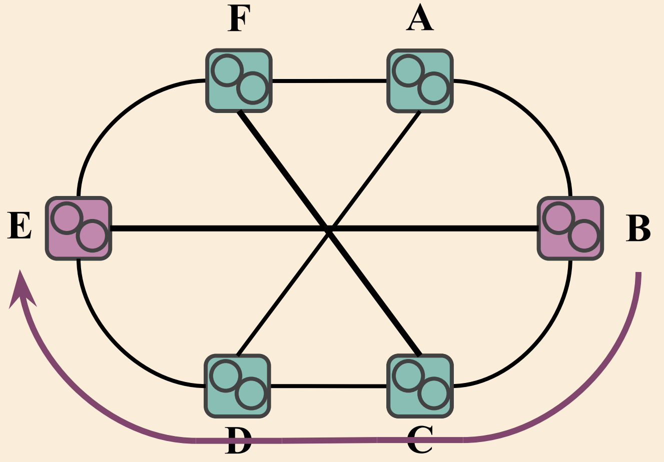
Prior Work Has Proposed Another Metric

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.

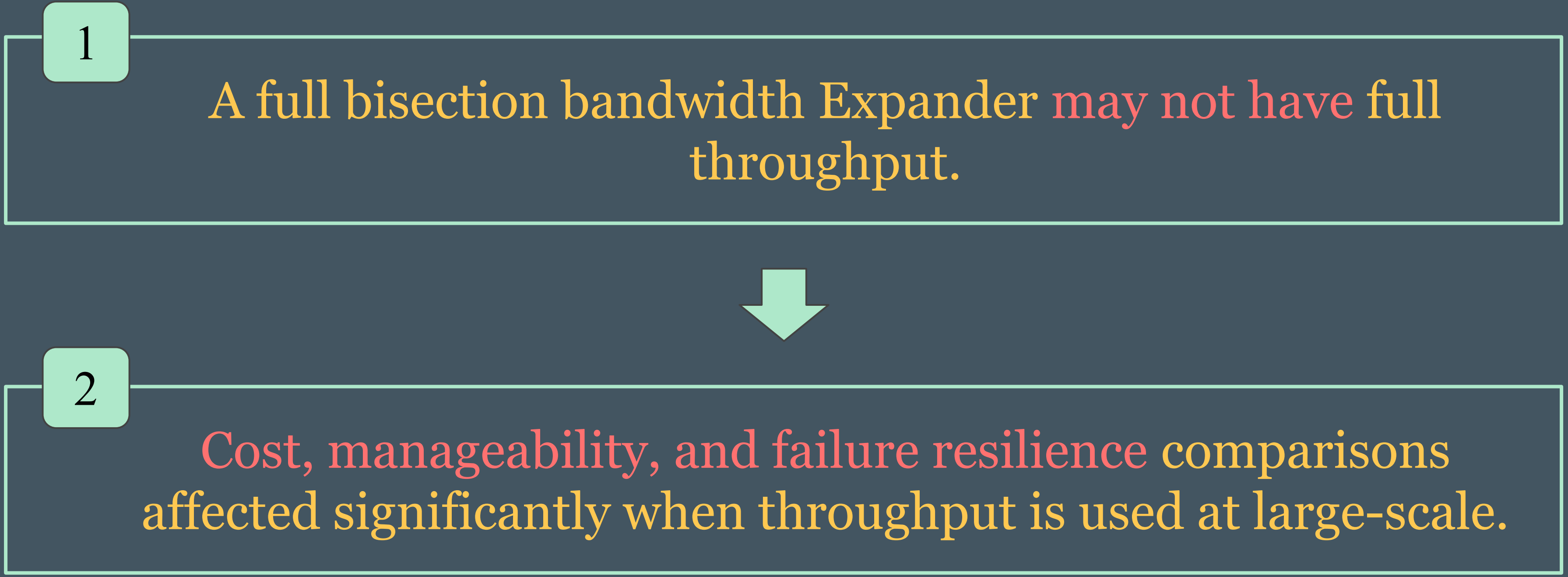
Theory

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.

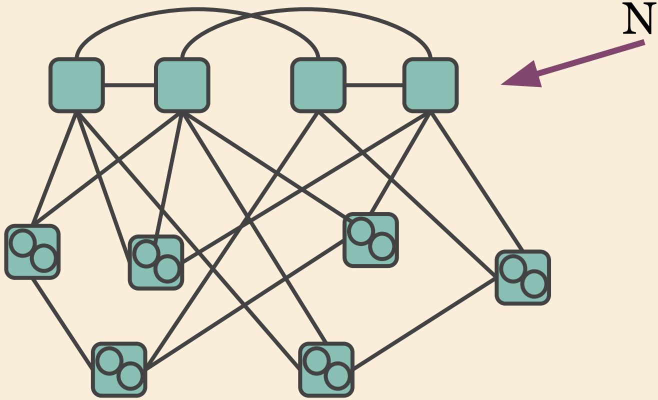
2

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

3

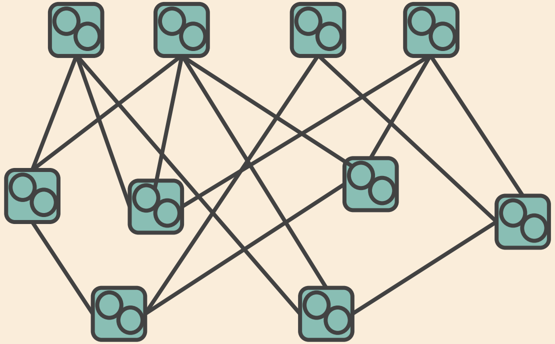
An accurate upper bound for throughput of Expanders and Clos topologies that scales well.

Clos vs Expanders



No servers

Clos

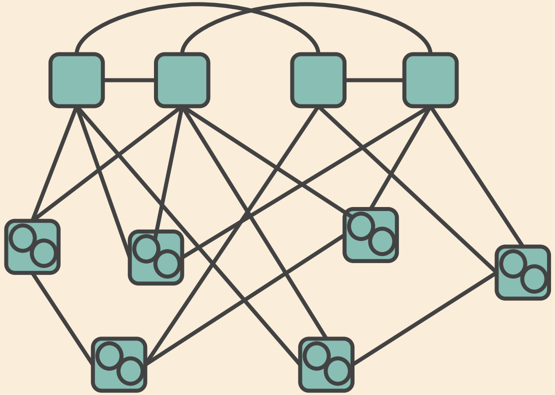


Expanders

Switch with 2 servers

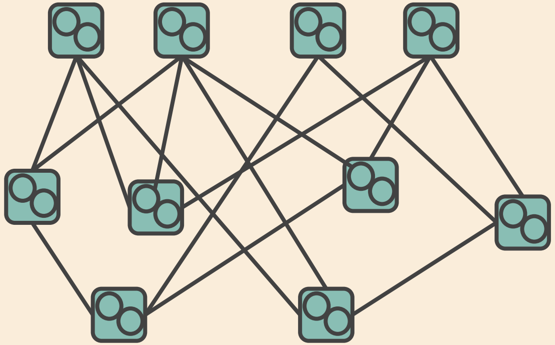
Switch without servers

Clos vs Expanders



2 servers

Clos

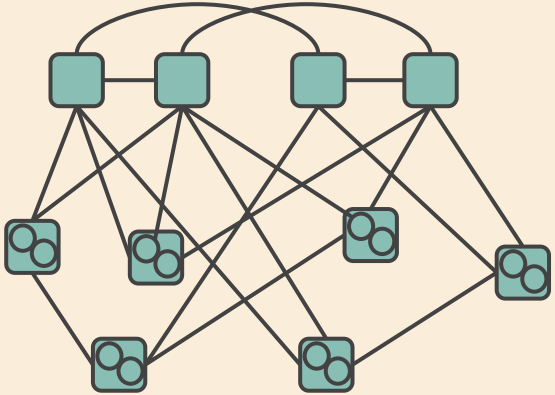


Expanders

Switch with 2 servers

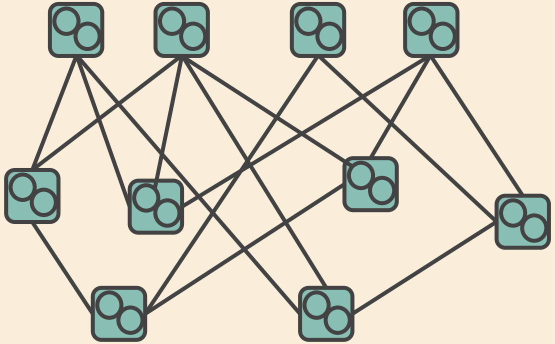
Switch without servers

Clos vs Expanders



Clos

2 servers

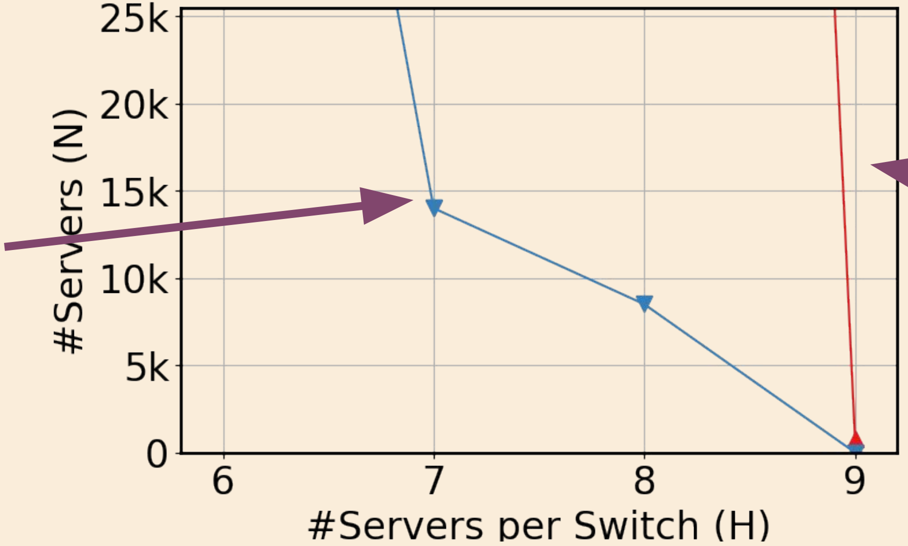


Expanders

Switch with 2 servers

Switch without servers

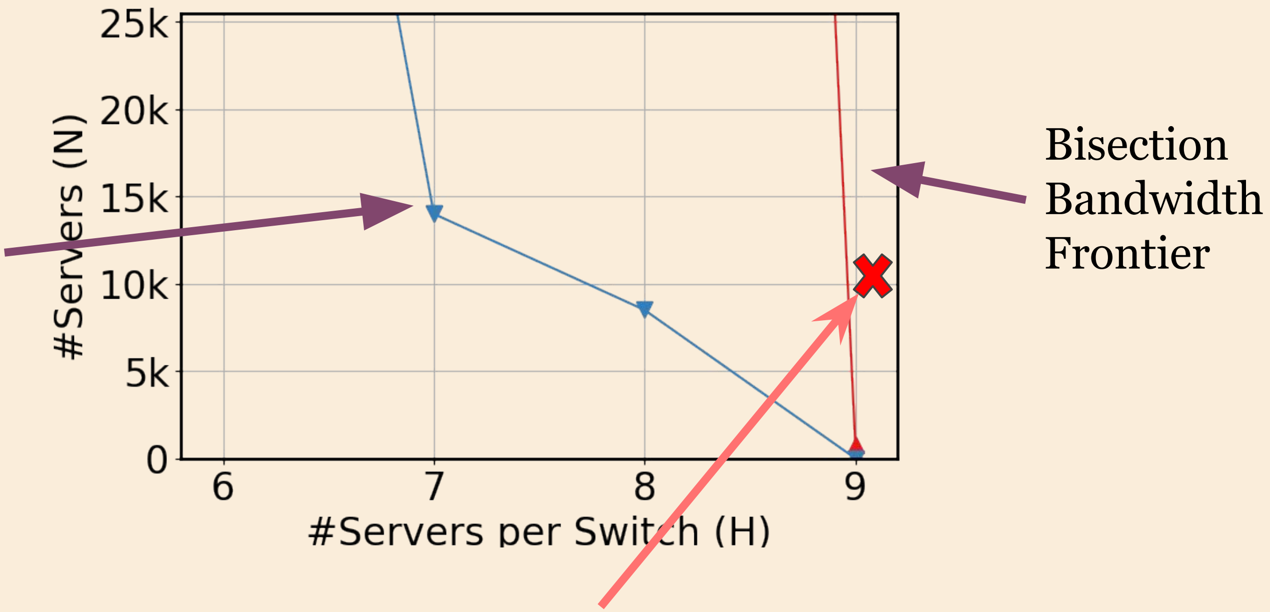
Scaling Limitations: Frontier Curve



Throughput Frontier

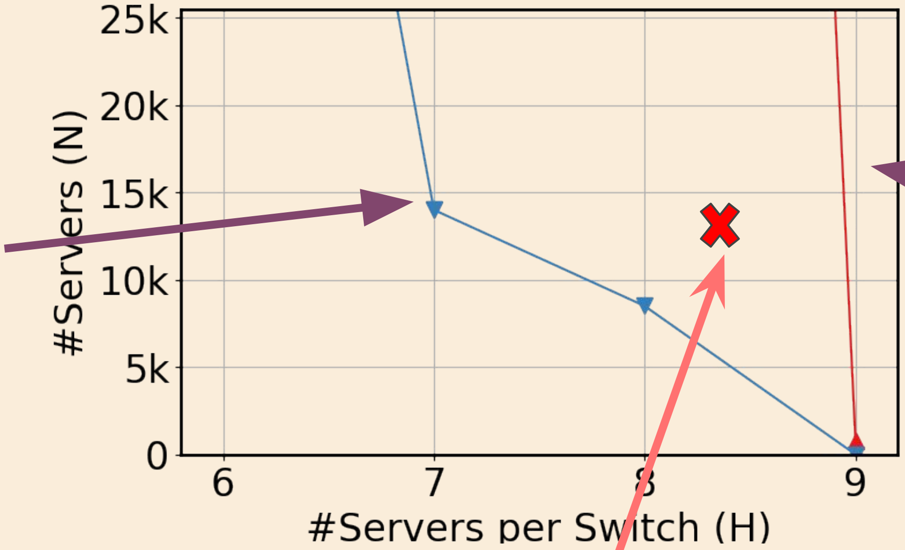
Bisection Bandwidth Frontier

Scaling Limitations: Frontier Curve



Not Full Bisection Bandwidth
Not Full Throughput

Scaling Limitations: Frontier Curve

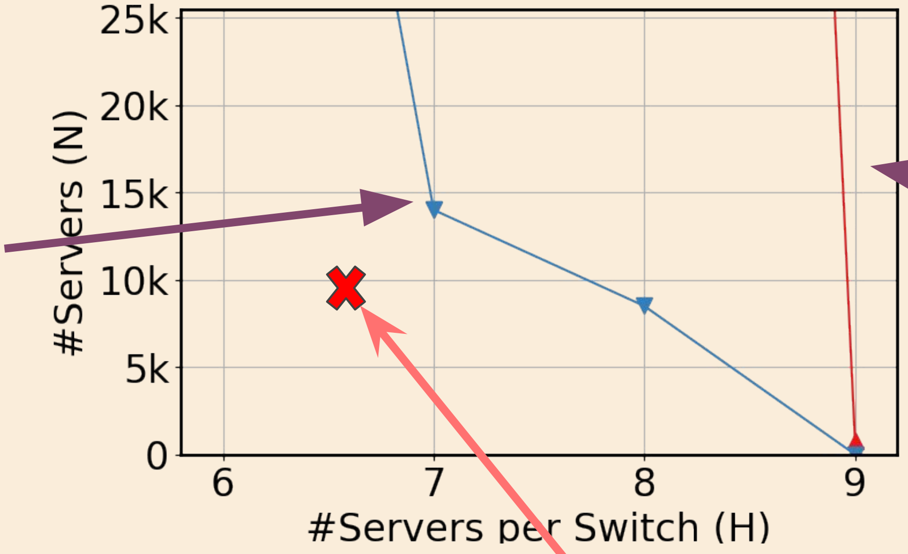


Throughput Frontier

Bisection Bandwidth Frontier

Full Bisection Bandwidth
Not Full Throughput

Scaling Limitations: Frontier Curve

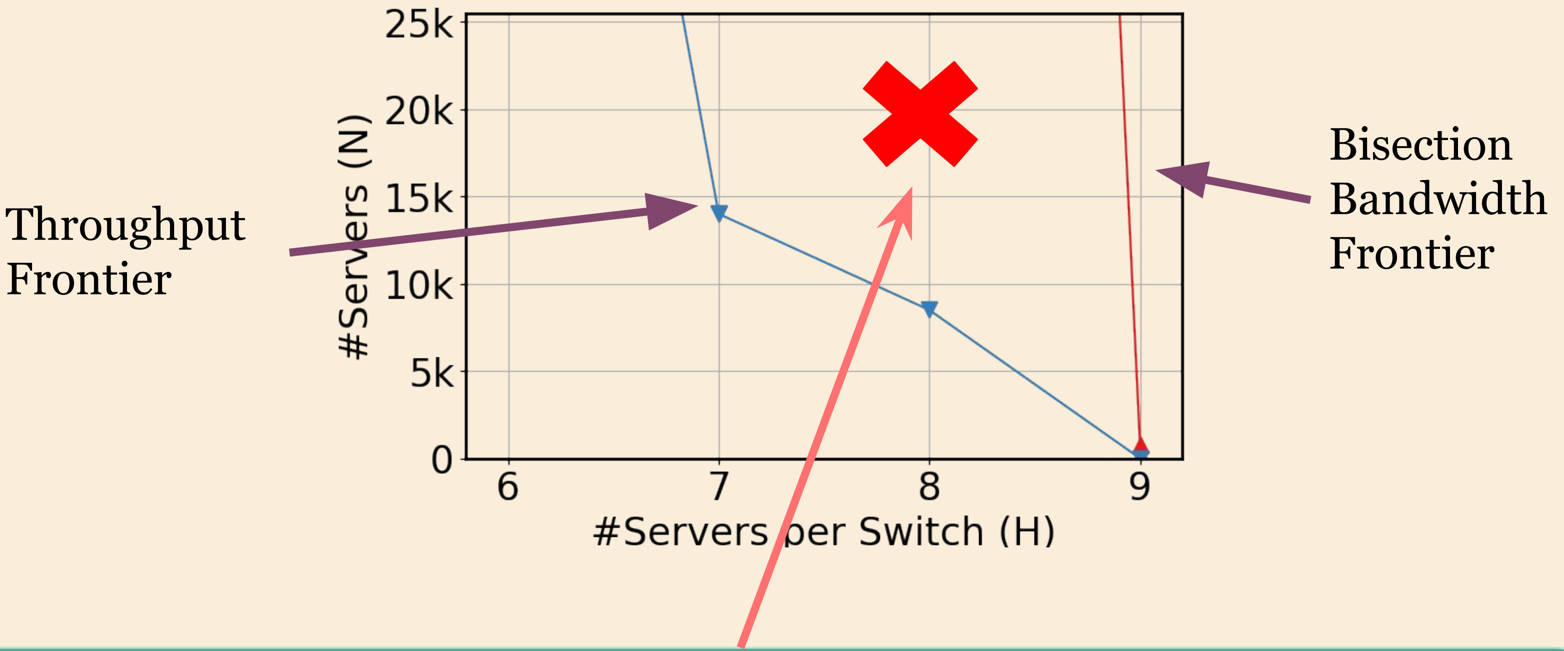


Throughput Frontier

Bisection Bandwidth Frontier

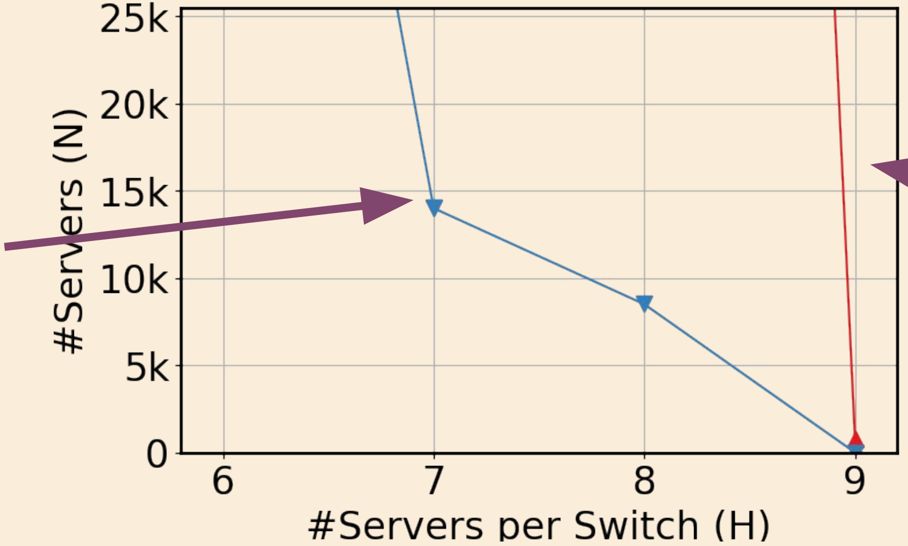
Full Bisection Bandwidth
Full Throughput

Scaling Limitations: Frontier Curve



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

Scaling Limitations: Frontier Curve

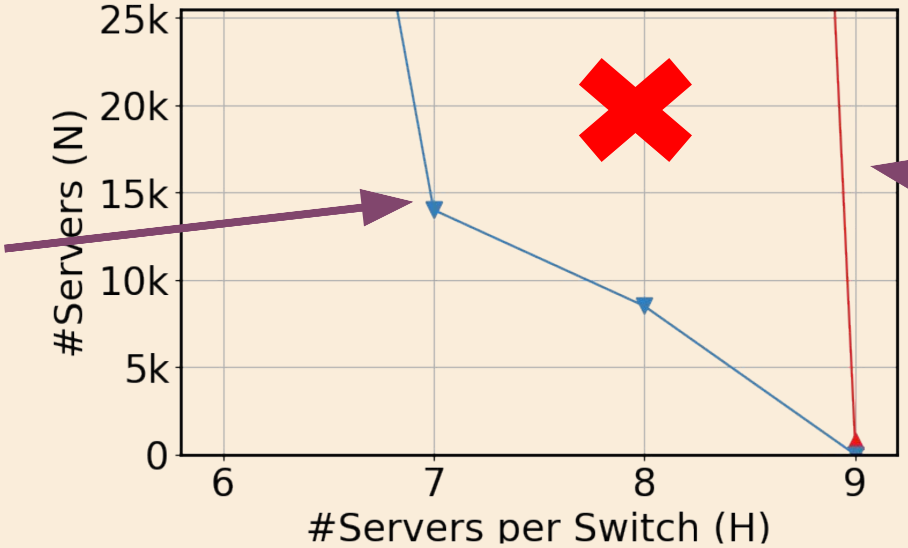


Throughput Frontier

Bisection Bandwidth Frontier

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

Scaling Limitations: Frontier Curve

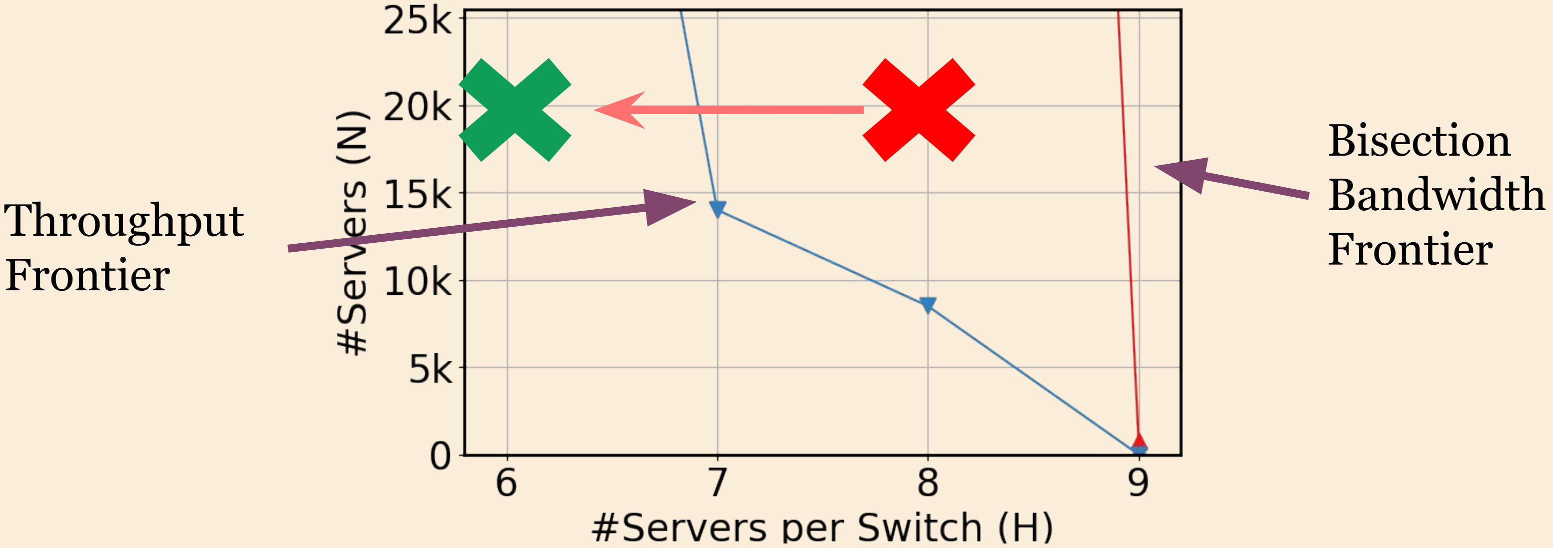


Throughput Frontier

Bisection Bandwidth Frontier

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

Scaling Limitations: Frontier Curve



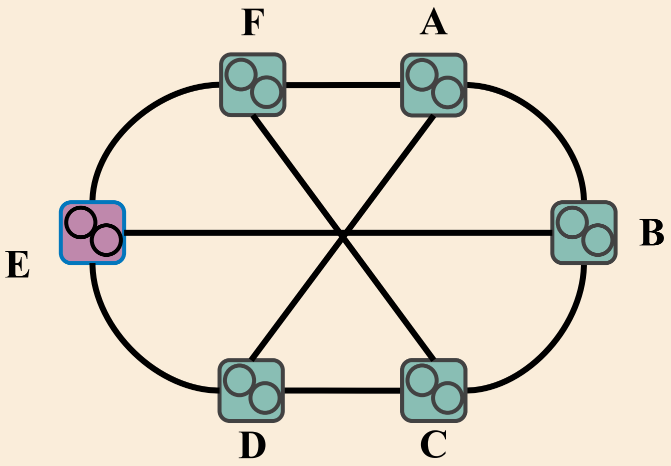
Throughput Frontier

Bisection Bandwidth Frontier

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

Why Expanders have scaling limitations?

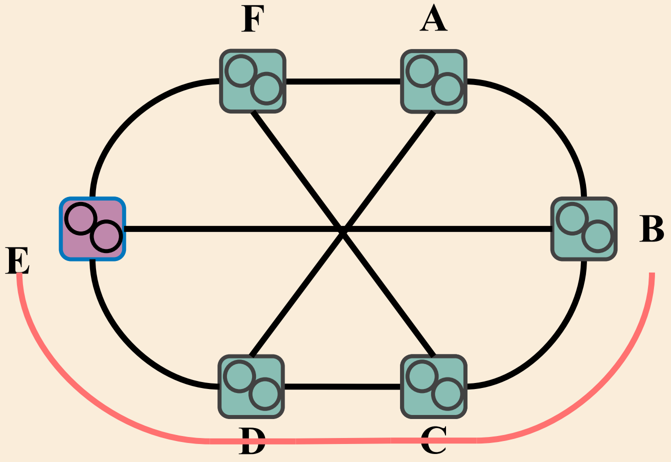
Two types of traffic in datacenter: Transit Traffic, Traffic originated/destined to connected server



Why Expanders have scaling limitations?

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

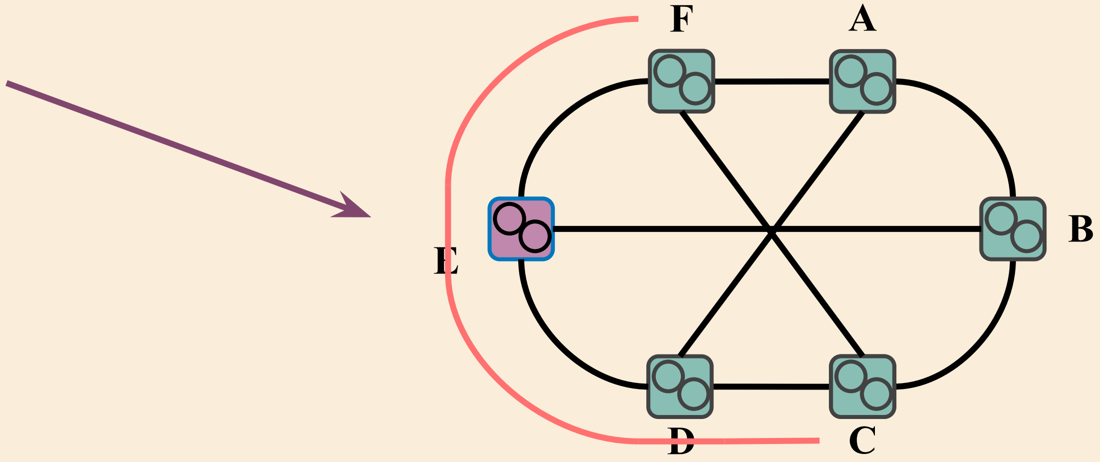
Traffic from/to the servers



Why Expanders have scaling limitations?

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

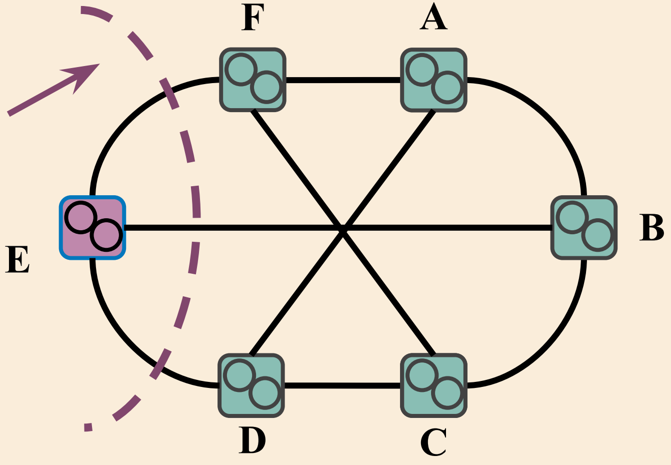
Transit Traffic



Why Expanders have scaling limitations?

Each switch has limited up-facing capacity.

Each Switch has 3 up-facing capacity

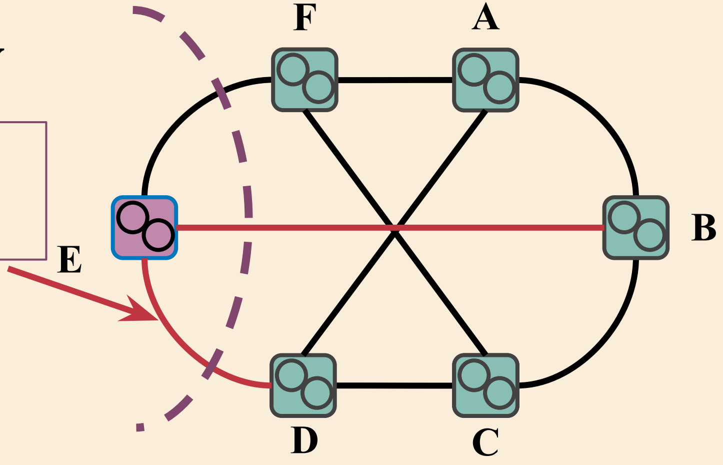


Why Expanders have scaling limitations?

In Expander, each switch has a fixed number of servers

Each Switch has 3 up-facing capacity

Each Switch connected to 2 Servers



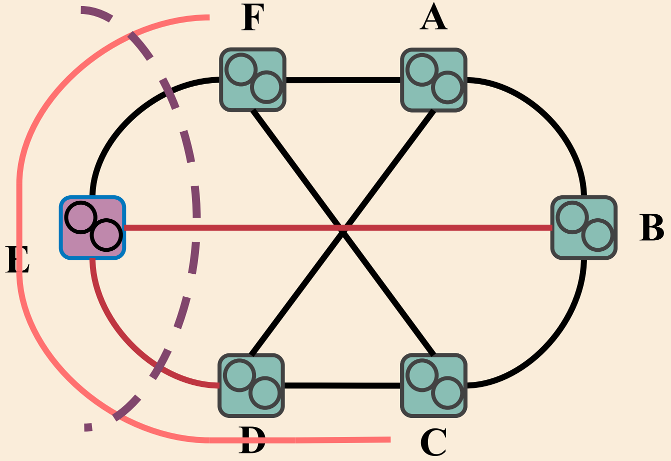
Why Expanders have scaling limitations?

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

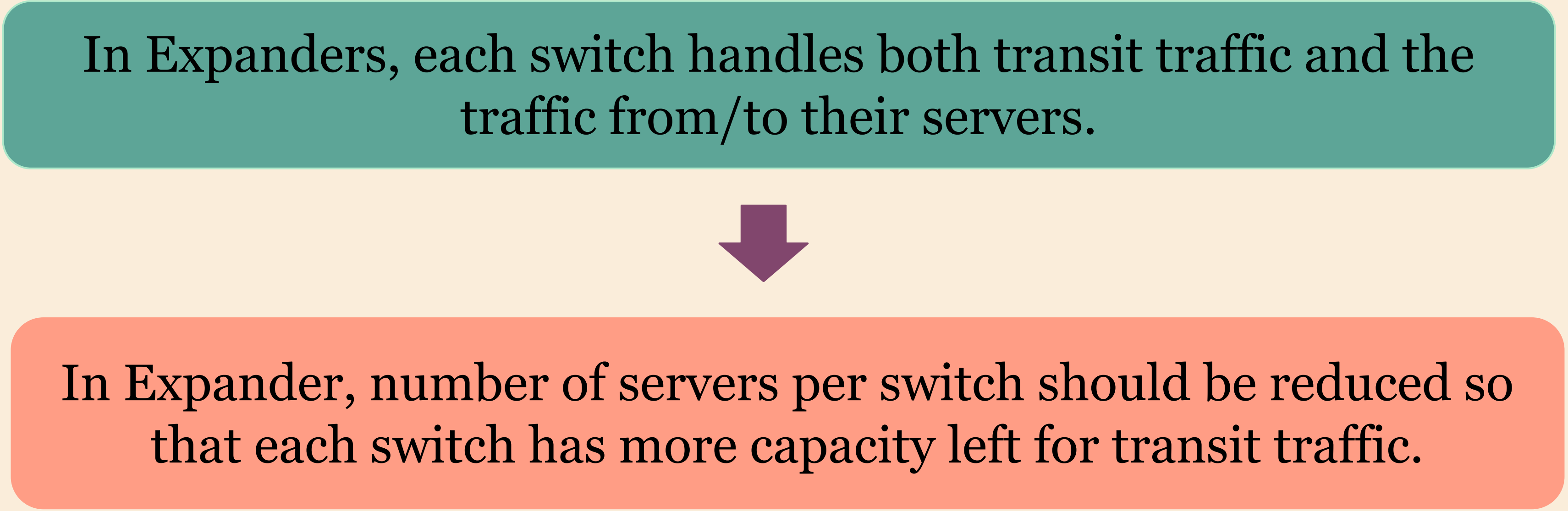
Each Switch has 3 up-facing capacity

Each Switch connected to 2 Servers

1 up-facing capacity left for transit traffic



Why Expanders have scaling limitations?



In Expanders, each switch handles both transit traffic and the traffic from/to their servers.

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

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