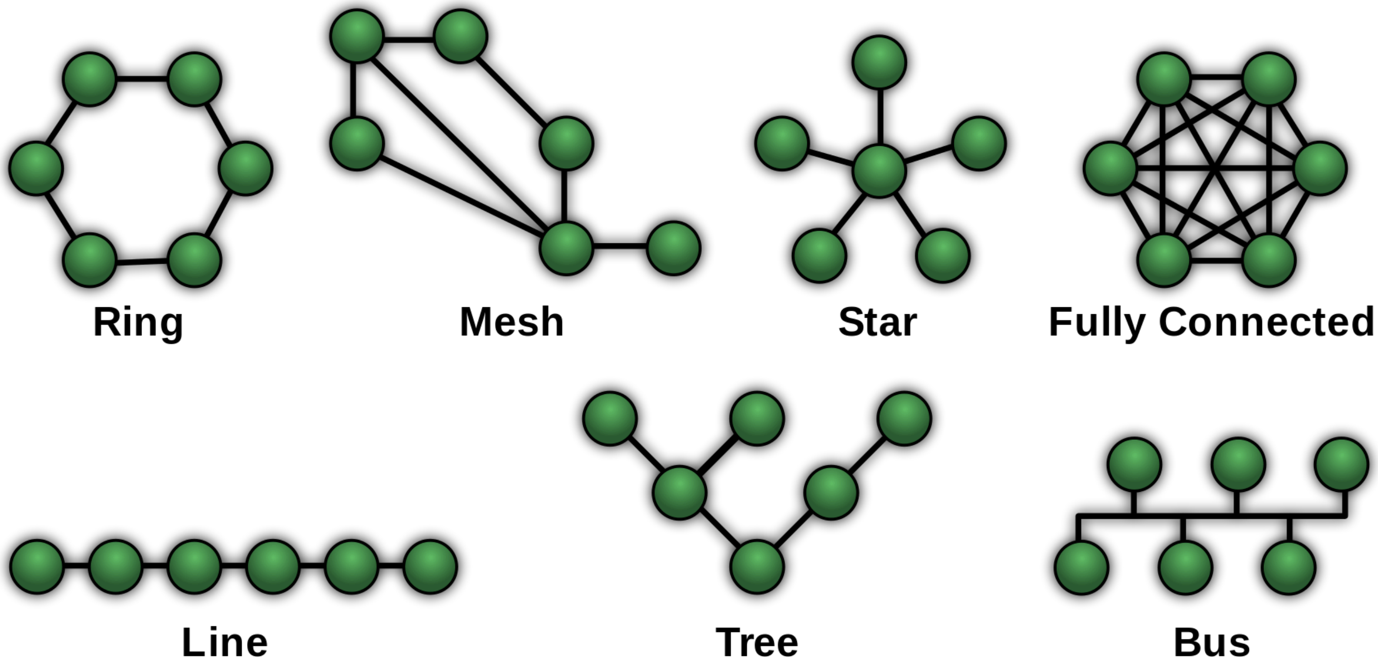
Improve Operations of Data Center Networks with Physical-Layer Programmability

Yiting Xia

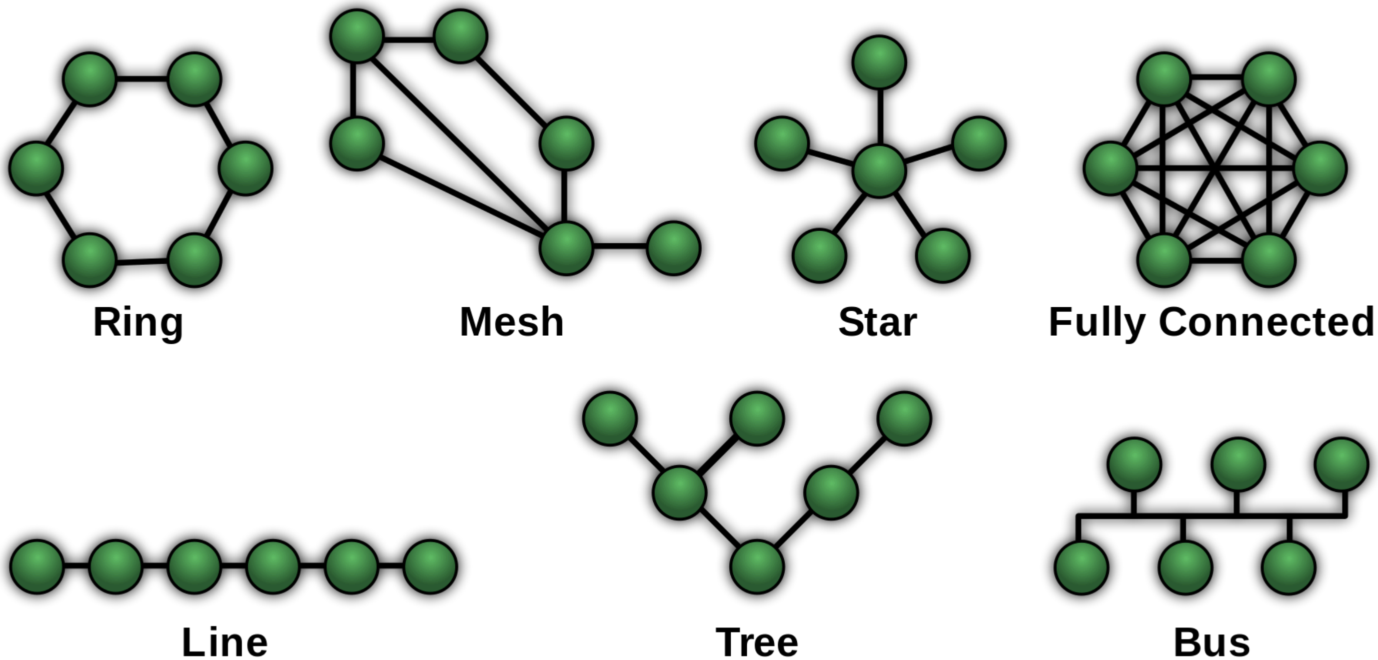
Research Scientist, Facebook

Network is a static graph



Common Network Topologies

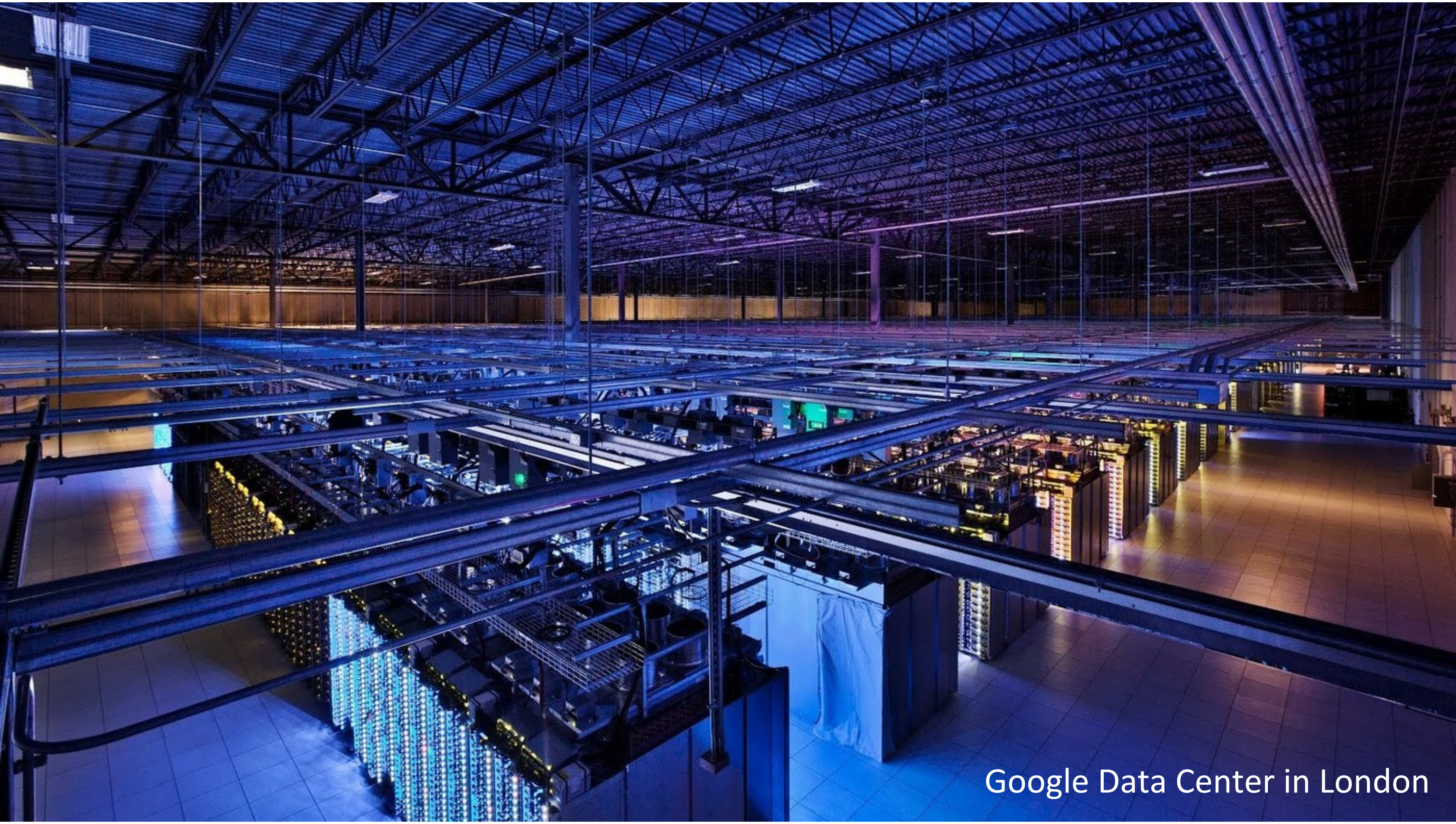
Network is a static graph



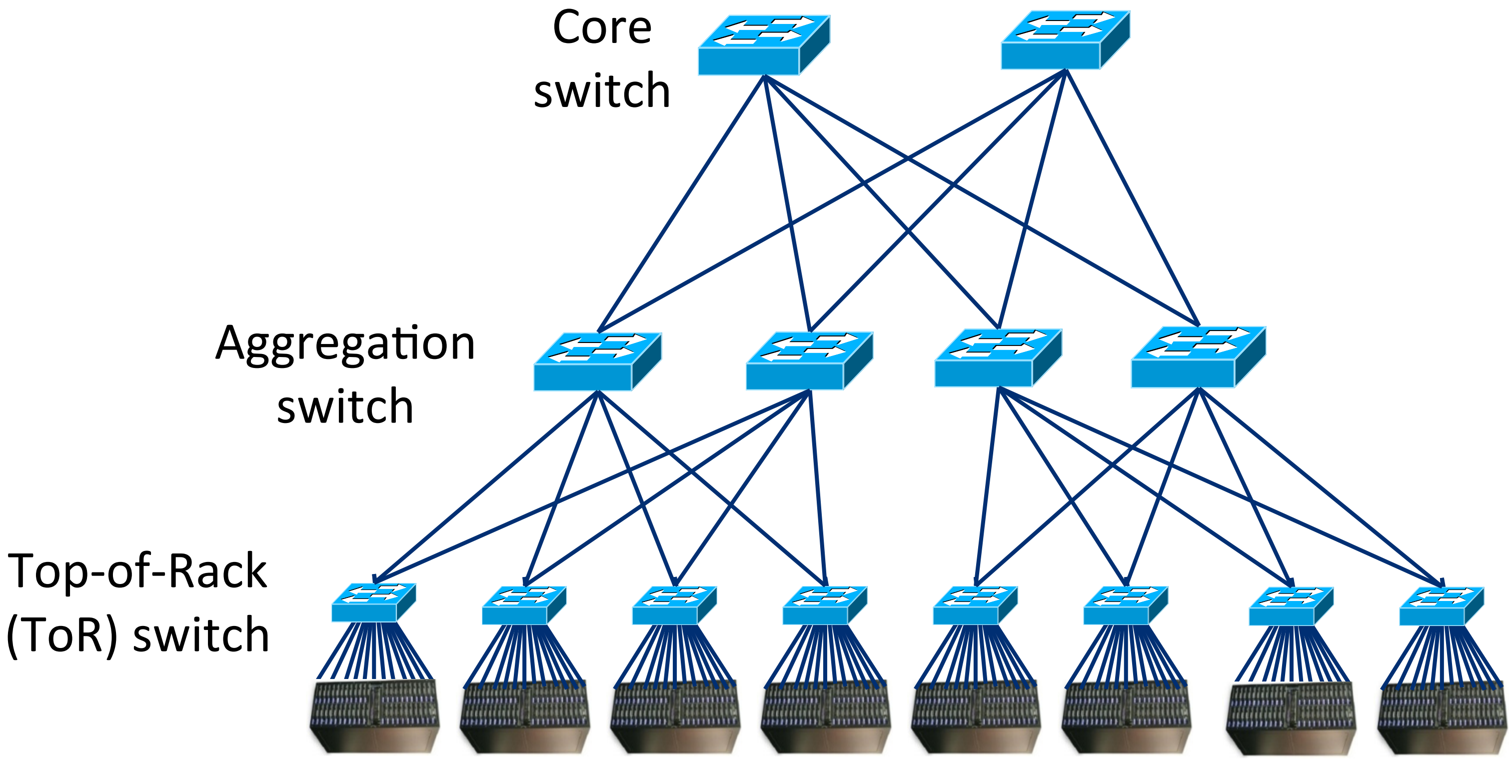
Common Network Topologies

Basic assumption of the networking world

Cloud Data Center Network



Topology of Data Center Network



Clos (Multi-Rooted Tree) Topology

Network operation is hard

Network operation is hard

- Failures

Network operation is hard

- Failures

Google Cloud Outage Triggered By Networking Issue

Google's Tuesday afternoon outage brought down popular services, including Spotify and Snapchat.

By Gina Narcisi — July 17, 2018, 04:50 PM EDT

Google Cloud suffered an outage that slowed down or stopped several popular services on Tuesday afternoon, including Spotify and Snapchat.

Google confirmed via its cloud status dashboard that it became aware of a networking issue impacting its load balancers just after noon PT on Tuesday.

RELATED STORIES

News Cloud
Google Partners Embrace Tech Giant's Enterprise Cloud Mission

Network operation is hard

- Failures

Google Cloud Outage Trigge

Google's Tuesday afternoon outage brought

By Gina Narcisi

Google Cloud suffered an outage that slowe
popular services on Tuesday afternoon, inclu

Google confirmed via its cloud status dashb
networking issue impacting its load balancer

PM EDT

ech
sion

The New York Times

A Failure Here, Damaged Fiber There and a Day of Internet Glitches

Cloudflare and Google dealt with issues that affected countless sites and users on Tuesday.

By David Yaffe-Bellany

July 2, 2019

When a website won't load, many internet users turn to DownDetector, a site that keeps track of online disruptions, providing frequent updates on the status of the world's digital infrastructure.

But the site, which calls itself the "weatherman of the digital world," was no help on Tuesday when thousands of major websites showed the same so-called 502 error message for part of the morning. In a twist, DownDetector had also gone down.

Network operation is hard

The New York Times

- Failures

Google Cloud Outage Trigge

Google's Tuesday afternoon

By Gina Narcisi

Google Cloud suffered an ou
popular services on Tuesday

Google confirmed via its clou
networking issue impacting it

A Failure Here, Damaged Fiber There and a Day of Internet Glitches

FB **Facebook**
@Facebook

We're aware that some people are currently having trouble accessing the Facebook family of apps. We're working to resolve the issue as soon as possible.

47K 9:49 AM - Mar 13, 2019

29.4K people are talking about this

showed the same so-called 502 error message for part of the morning. In a twist, DownDetector had also gone down.

Network operation is hard

- Failures

The New York Times

A Failure Here, Damaged Fiber There and a Day of Internet Glitches

Google Cloud Outage Trigge

Google's Tuesday afternoon

By Gina Narcisi

Facebook
@Facebook

Microsoft

Even Microsoft faced its share of cloud outages this year affecting Azure, Microsoft 365, Dynamics, and DevOps. In May, Microsoft had to face an outage that lasted for more than an hour showing network connectivity errors in Microsoft Azure that deeply affected its cloud services including Office 365, Microsoft Teams, Xbox Live, and several others which are widely used by Microsoft's commercial customers. Engineers identified the root cause to be an incorrect name server delegation issue that affected DNS resolution, network connectivity, and downstream impact. While the services were recovered, no customer DNS records were impacted during this incident.

Network operation is hard

- Failures

The New York Times

A Failure Here, Damaged Fiber There and a Day of Internet Glitches

Google Cloud Outage Trigge

Google's Tuesday afternoon

By Gina Narcisi

Facebook
@Facebook

Microsoft

Even Microsoft faced its share of cloud outages this year affecting Azure, Microsoft 365, Dynamics, and DevOps. In May, Microsoft had to face an outage that lasted for more than an hour showing network connectivity errors in Microsoft Azure that deeply affected its cloud services including Office 365, Microsoft Teams, Xbox Live, and several others which are widely used by Microsoft's commercial customers. Engineers identified the root cause to be an incorrect name server delegation issue that affected DNS resolution, network connectivity, and downstream impact. While the services were recovered, no customer DNS records were impacted during this incident.

Network operation is hard

- Failures

The New York Times

A Failure Here, Damaged Fiber There and a Day of Internet Glitches

Google Cloud Outage Trigge

Google's Tuesday afternoon

By Gina Narcisi

FB **Facebook**
@Facebook

Microsoft

Even Microsoft faced its share of cloud outages this year affecting Azure, Microsoft 365, Dynamics, and DevOps. In May, Microsoft had to face an outage that lasted for more than an hour showing network connectivity errors in Microsoft Azure that deeply affected its cloud services including Office 365, Microsoft Teams, Xbox Live, and several others which are widely used by Microsoft's commercial customers. Engineers identified the root cause to be an incorrect name server delegation issue that affected DNS resolution, network connectivity, and downstream impact. While the services were recovered, no customer DNS records were impacted during this incident.

- Median case of failures: 10% less traffic delivered
- Worst 20% cases of failures: 40% less traffic delivered

Gill et al., Understanding Network Failures in Data Centers: Measurement, Analysis, and Implications, SIGCOMM 2011

Network operation is hard

- Failures

The New York Times

A Failure Here, Damaged Fiber There and a Day of Internet Glitches

Google Cloud Outage Trigge

Google's Tuesday afternoon

By Gina Narcisi

Facebook
@Facebook

Microsoft

Even Microsoft faced its share of cloud outages this year affecting Azure, Microsoft 365, Dynamics, and DevOps. In May, Microsoft had to face an outage that lasted for more than an

- Failures are disruptive
- Fixed topology: have to live with a crippled network

- Median case of failures: 10% less traffic delivered
- Worst 20% cases of failures: 40% less traffic delivered

Gill et al., Understanding Network Failures in Data Centers: Measurement, Analysis, and Implications, SIGCOMM 2011

Network operation is hard

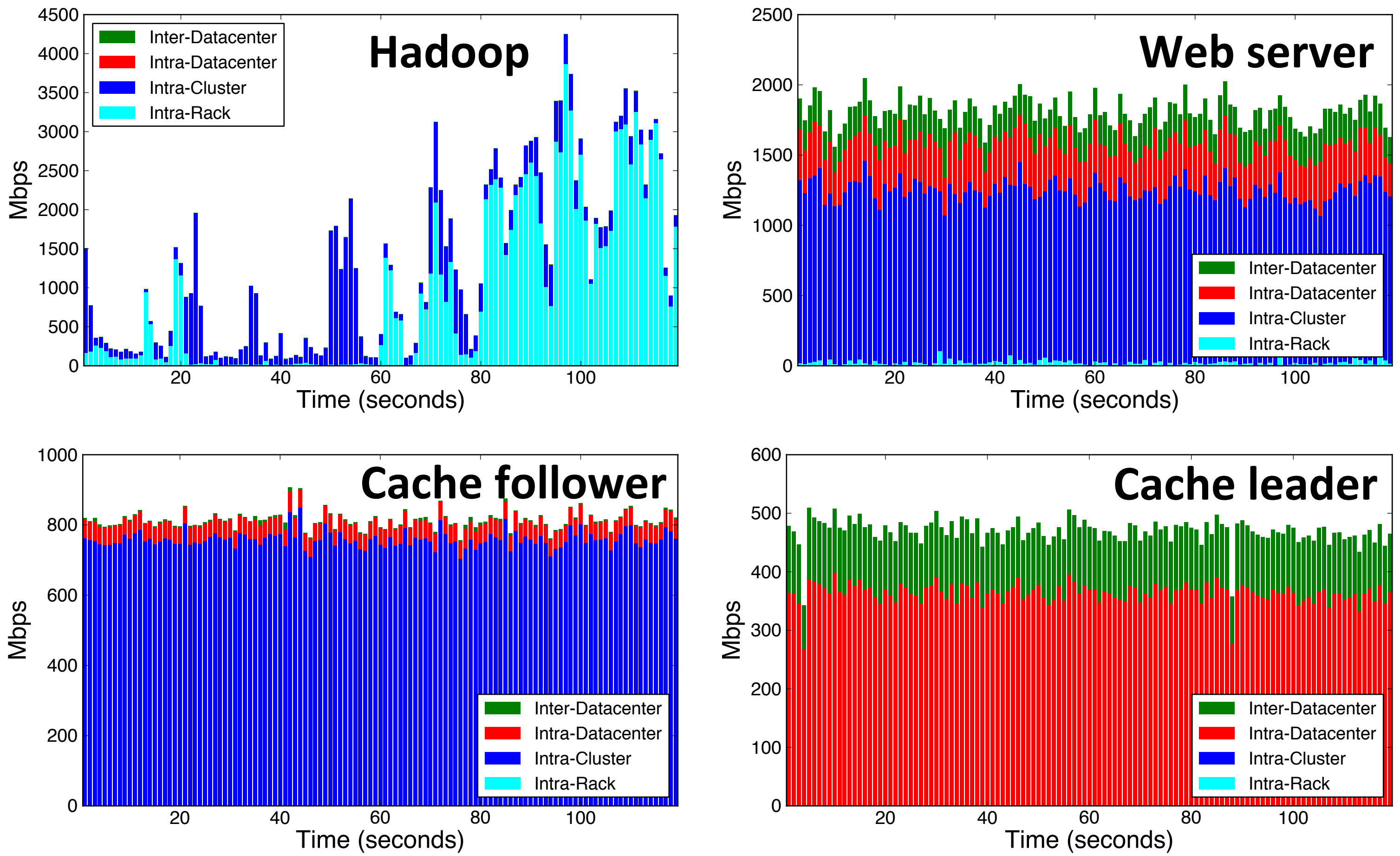
- Service provisioning

Network operation is hard

- Service provisioning
  - *Public cloud: VM clusters*
  - *Private cloud: sub-systems supporting the service*

Network operation is hard

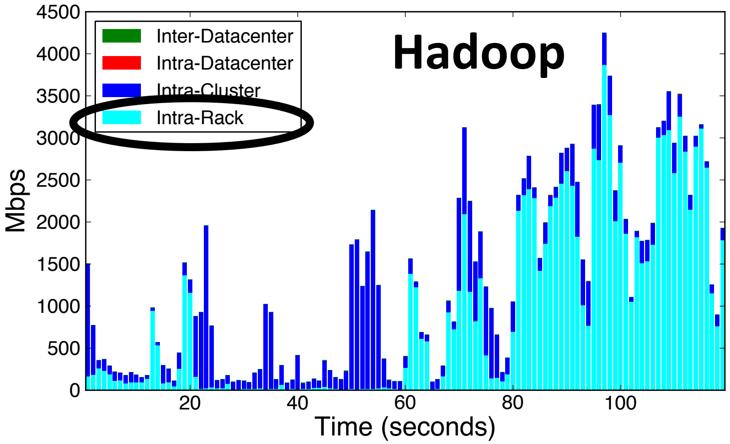
- Service provisioning

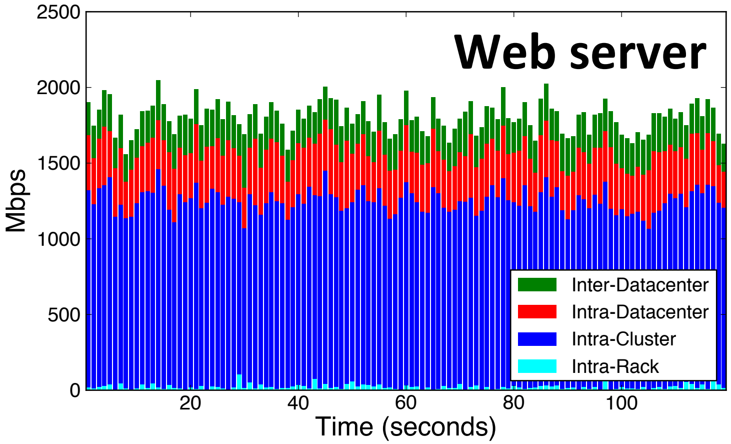


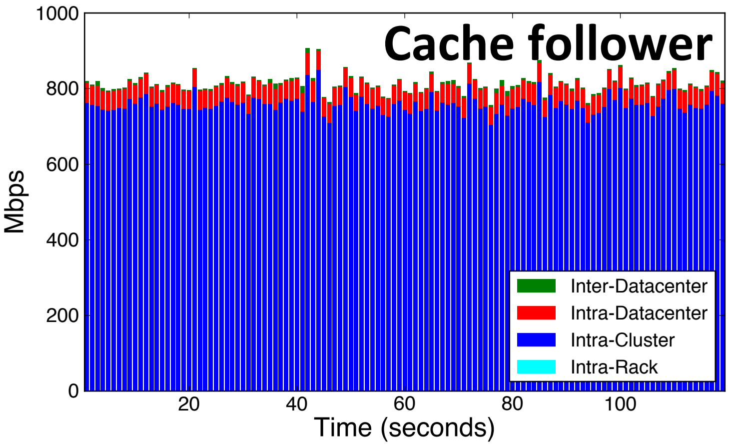
Roy et al., Inside the Social Network's (Datacenter) Network, SIGCOMM 2015

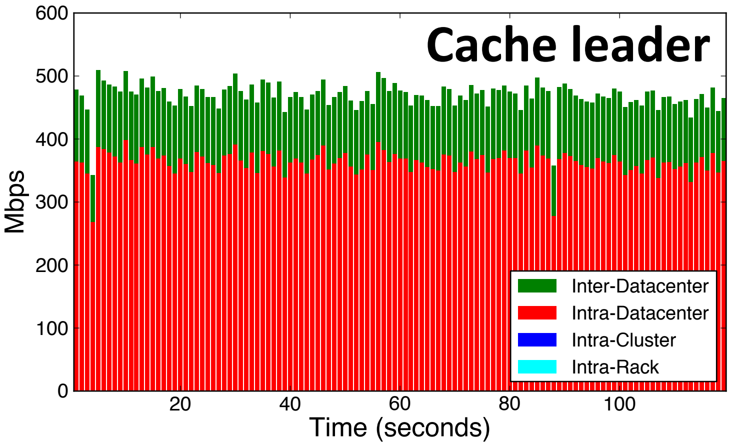
Network operation is hard

- Service provisioning





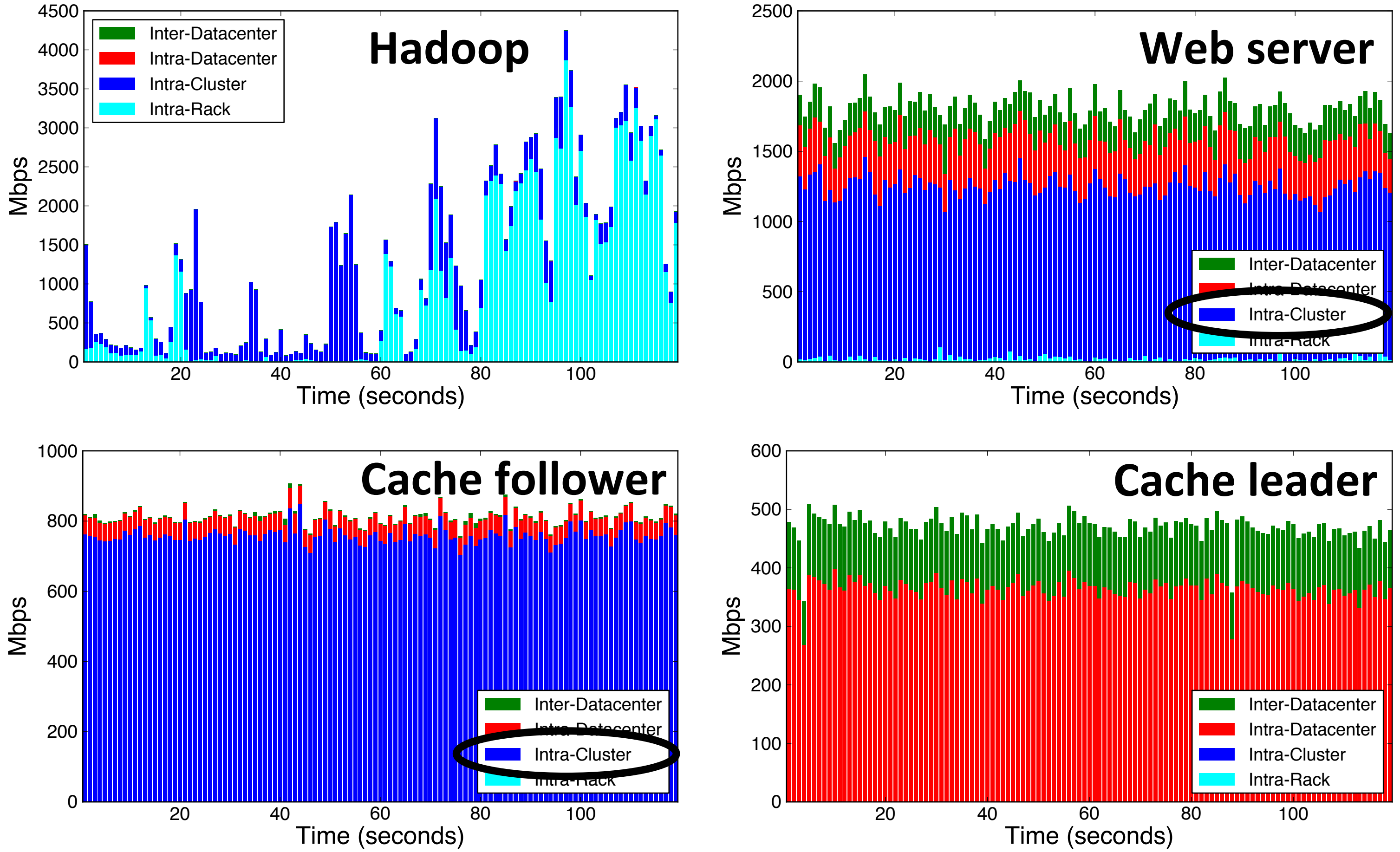




Roy et al., Inside the Social Network's (Datacenter) Network, SIGCOMM 2015

Network operation is hard

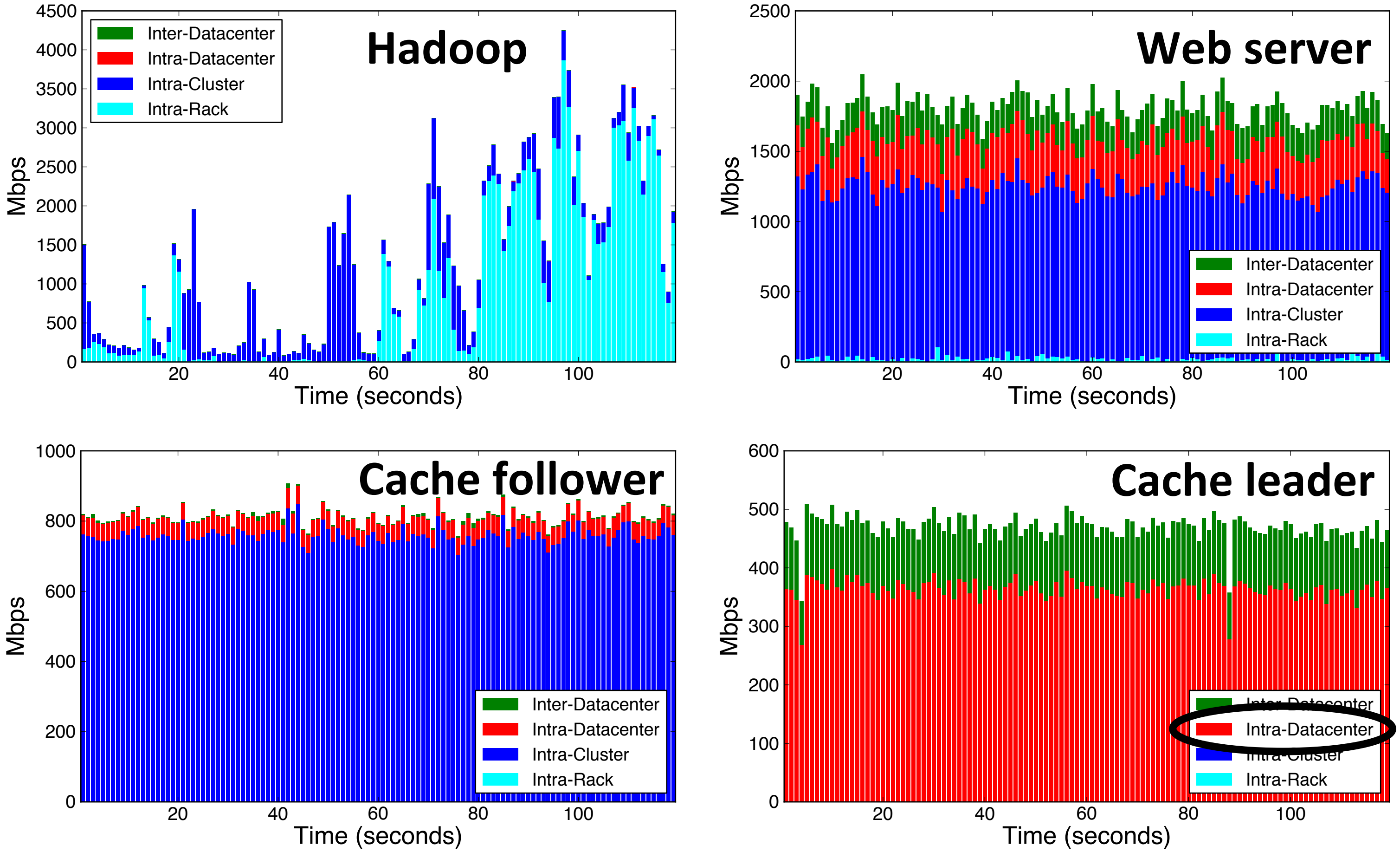
- Service provisioning



Roy et al., Inside the Social Network's (Datacenter) Network, SIGCOMM 2015

Network operation is hard

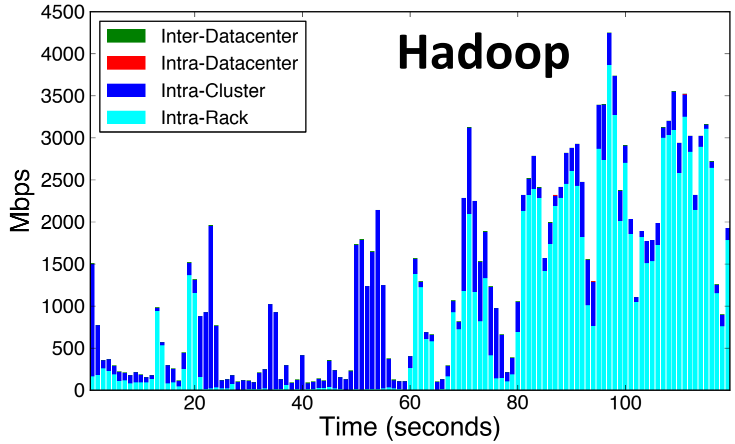
- Service provisioning

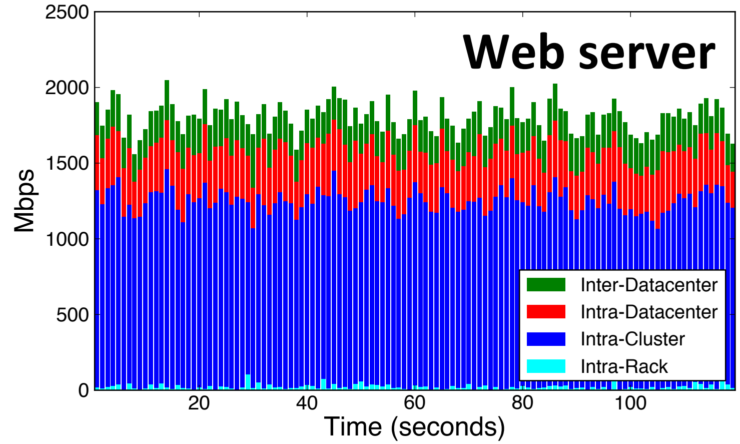


Roy et al., Inside the Social Network's (Datacenter) Network, SIGCOMM 2015

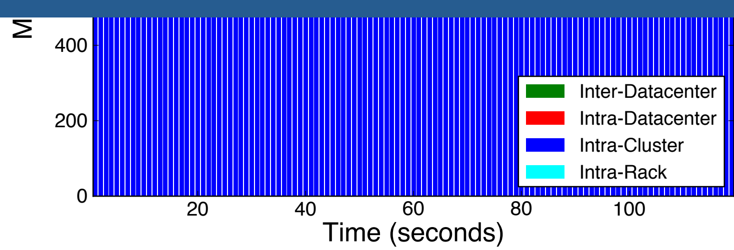
Network operation is hard

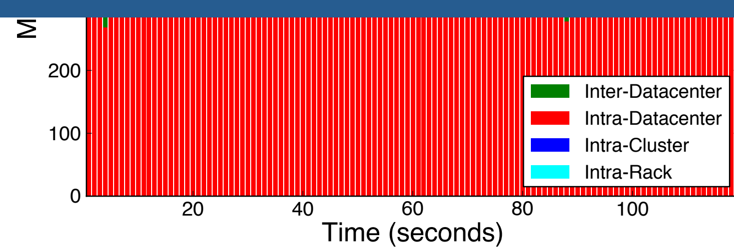
- Service provisioning





- Different clusters have different traffic localities
- Hard to fit into the same network topology





Roy et al., Inside the Social Network's (Datacenter) Network, SIGCOMM 2015

Network operation is hard

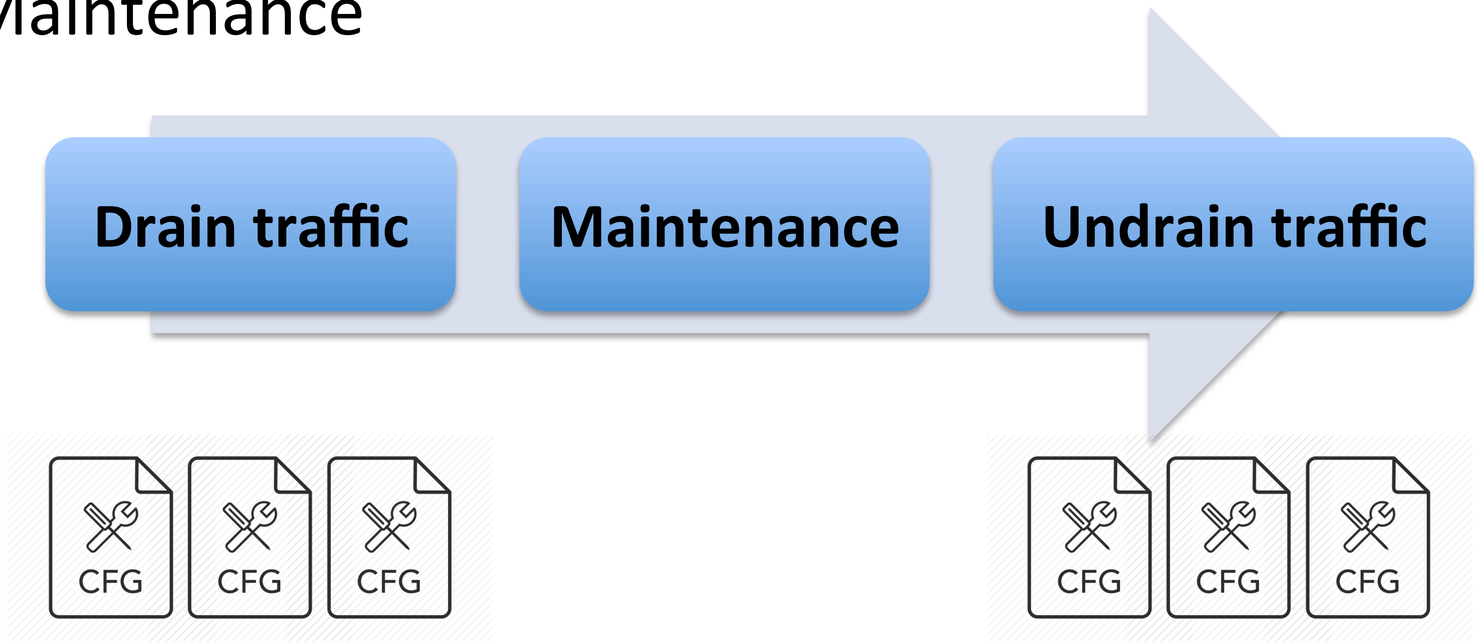
- Maintenance

Network operation is hard

- Maintenance

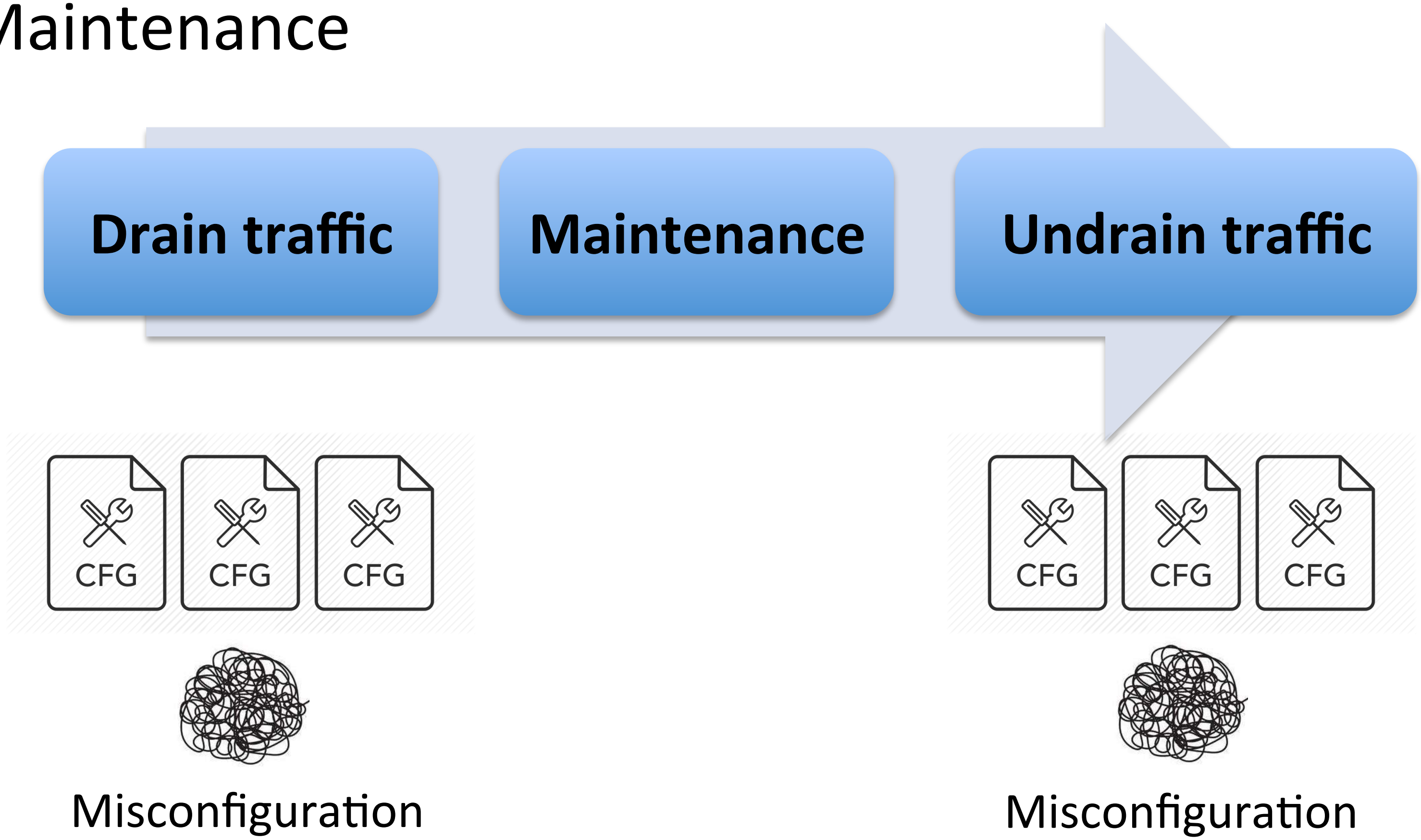
Network operation is hard

- Maintenance



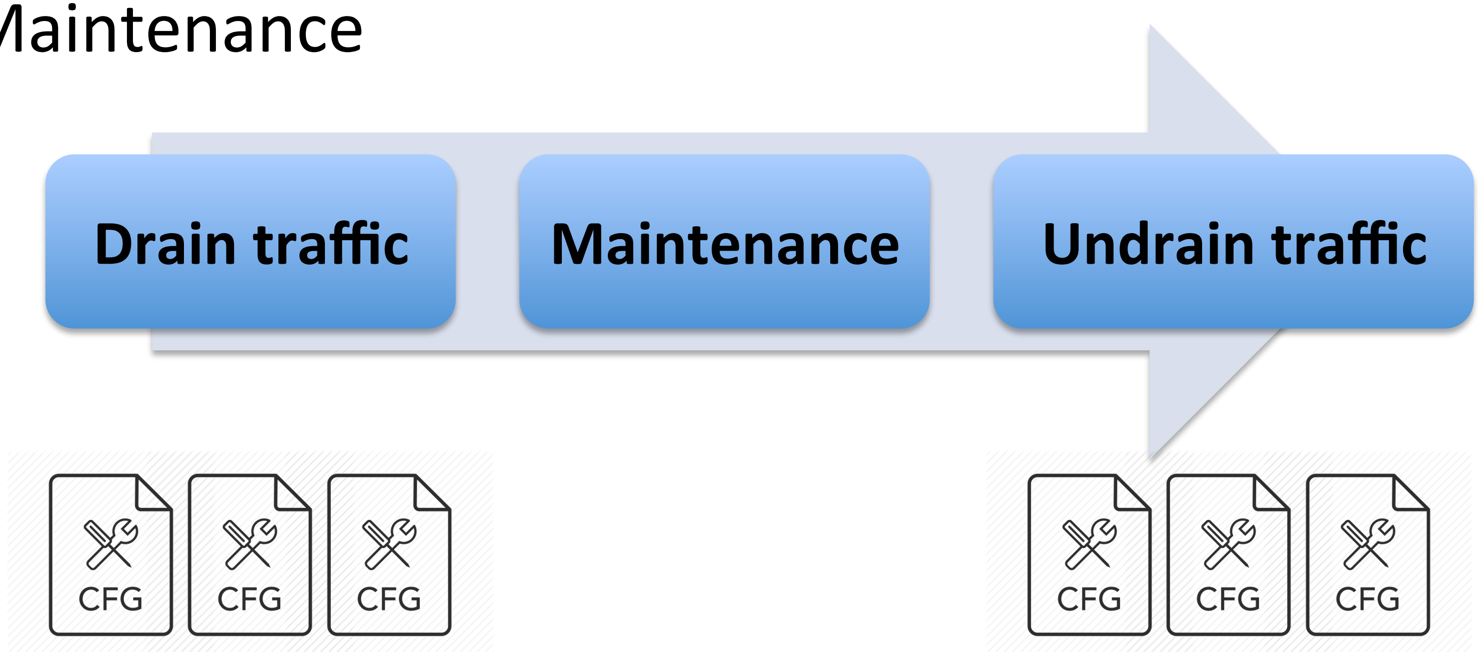
Network operation is hard

- Maintenance



Network operation is hard

- Maintenance



- Important source of oncall problems
- Change network states to “fake” loss of connectivity

Network operation is hard

- Wiring

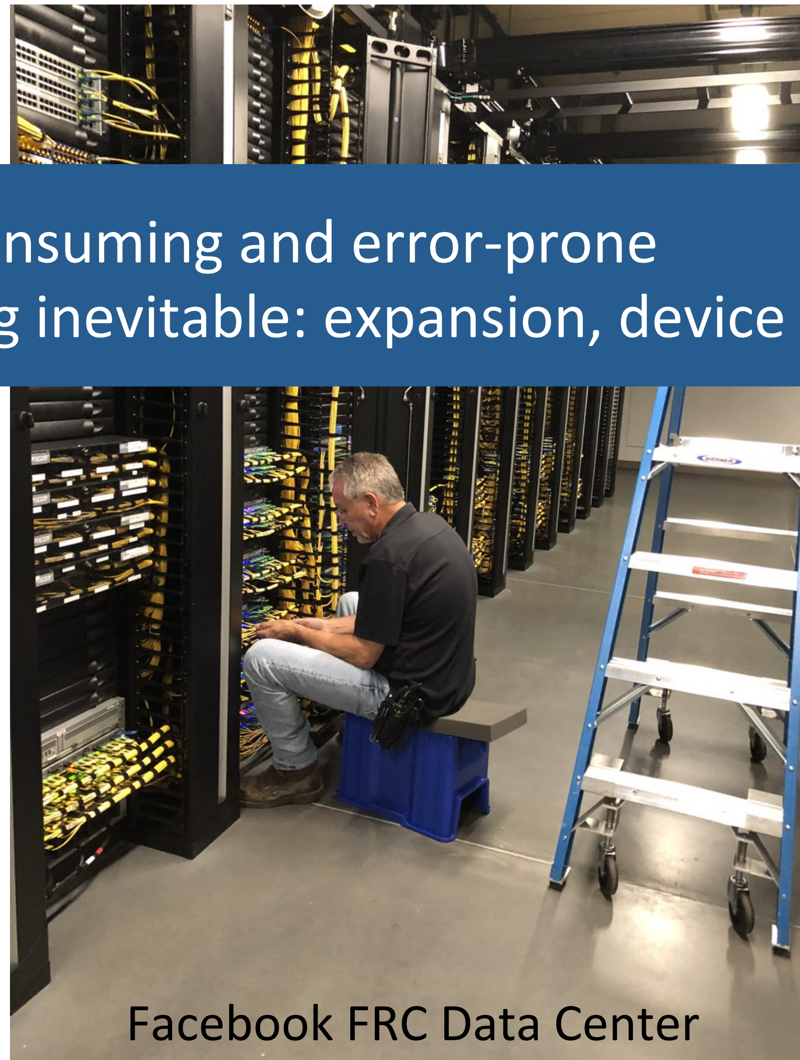
Network operation is hard

- Wiring


Facebook FRC Data Center

Network operation is hard

- Wiring



- Time-consuming and error-prone
- Rewiring inevitable: expansion, device upgrade

Facebook FRC Data Center

Network Operation  Topology Change

Network Operation  Topology Change

Operational Problems	Topology Change
Failure	Bypass or fix failures
Service Provisioning	Change into the right topology
Maintenance	Partition the graph
Wiring	Automatic wiring with software

Network Operation  Topology Change

- If fast enough, the change should be hidden from upper layers of the network stack

Operational Problems	Topology Change
Failure	Bypass or fix failures
Service Provisioning	Change into the right topology
Maintenance	Partition the graph
Wiring	Automatic wiring with software

Physical-Layer Programmability

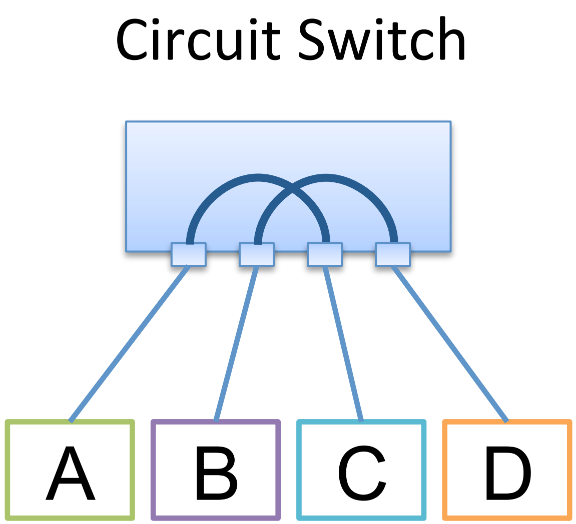
Physical-Layer Programmability

- The network topology is configurable

Physical-Layer Programmability

- The network topology is configurable

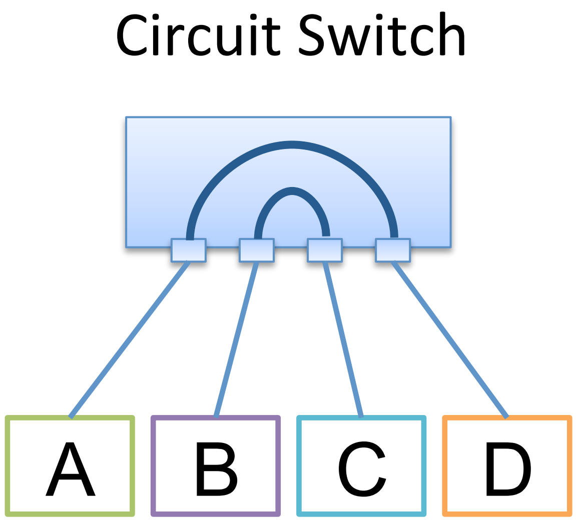
- Circuit switching
  - *optical or wireless*
  - *reconfigure internal connections*



Physical-Layer Programmability

- The network topology is configurable

- Circuit switching
  - *optical or wireless*
  - *reconfigure internal connections*

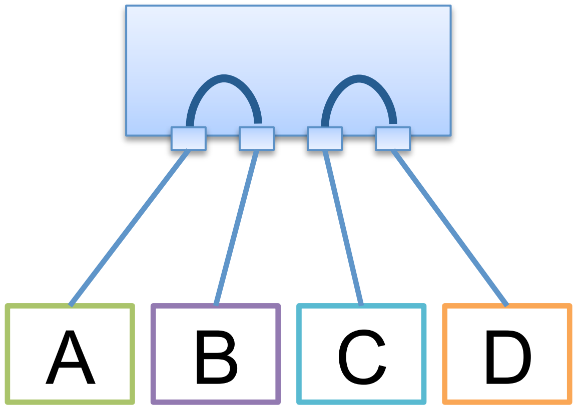


Physical-Layer Programmability

- The network topology is configurable

- Circuit switching
  - *optical or wireless*
  - *reconfigure internal connections*

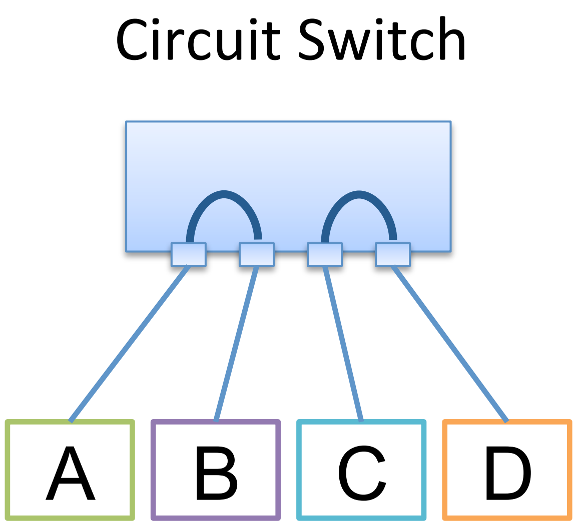
Circuit Switch



Physical-Layer Programmability

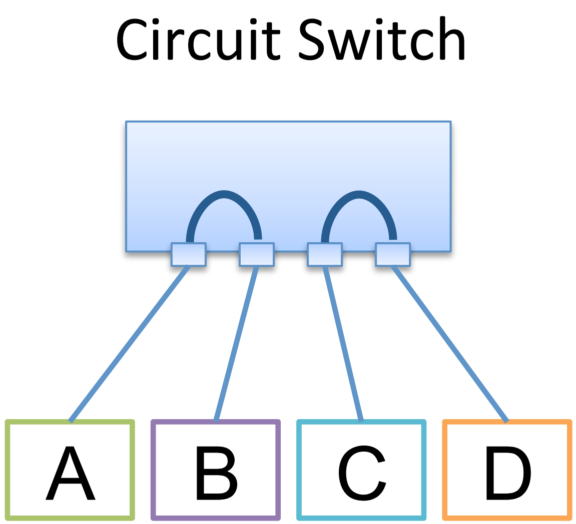
- The network topology is configurable

- Circuit switching
  - *optical or wireless*
  - *reconfigure internal connections*
- Fast topology change
  - *ms or us*



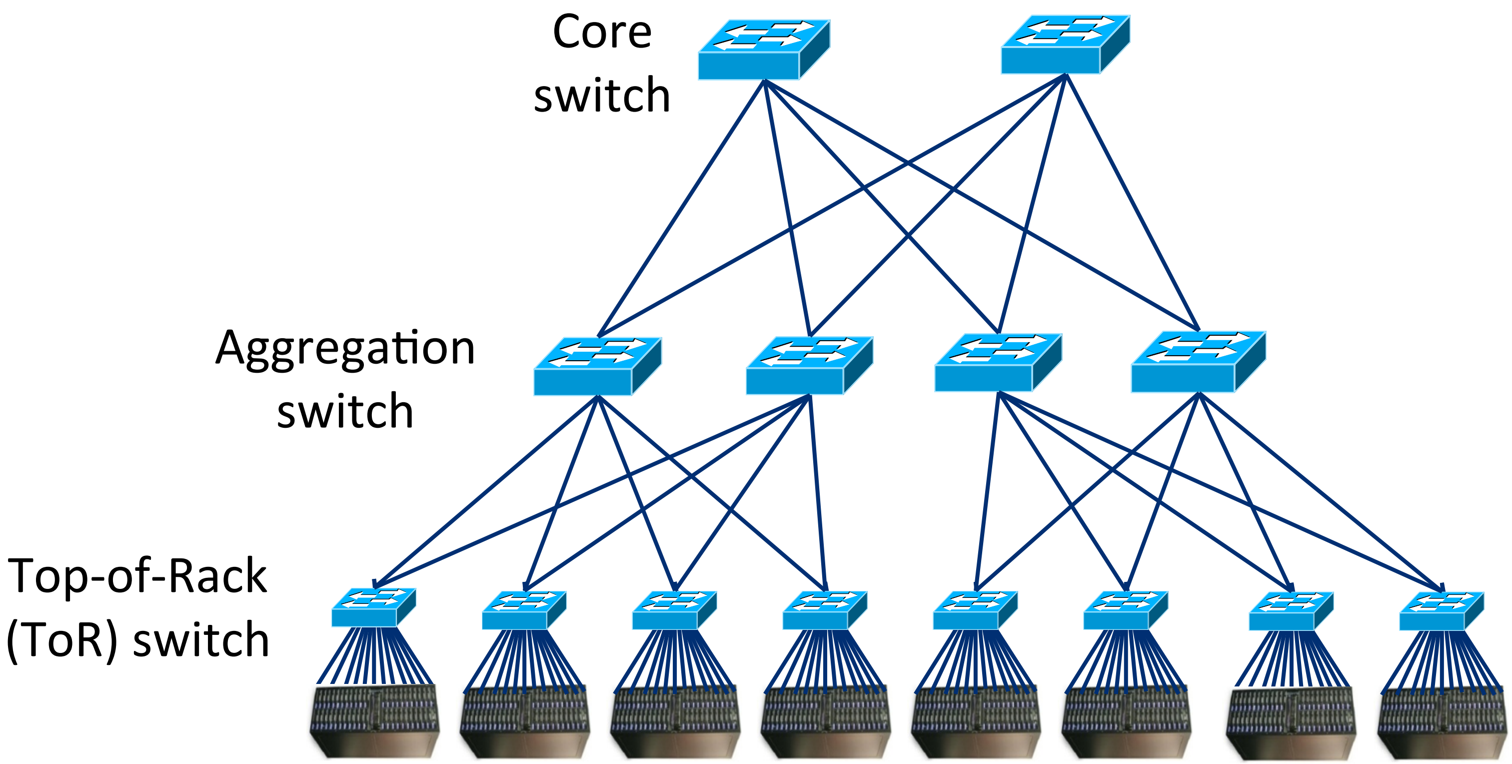
Physical-Layer Programmability

- The network topology is configurable
- Circuit switching
  - *optical or wireless*
  - *reconfigure internal connections*
- Fast topology change
  - *ms or us*
- Controlled by software



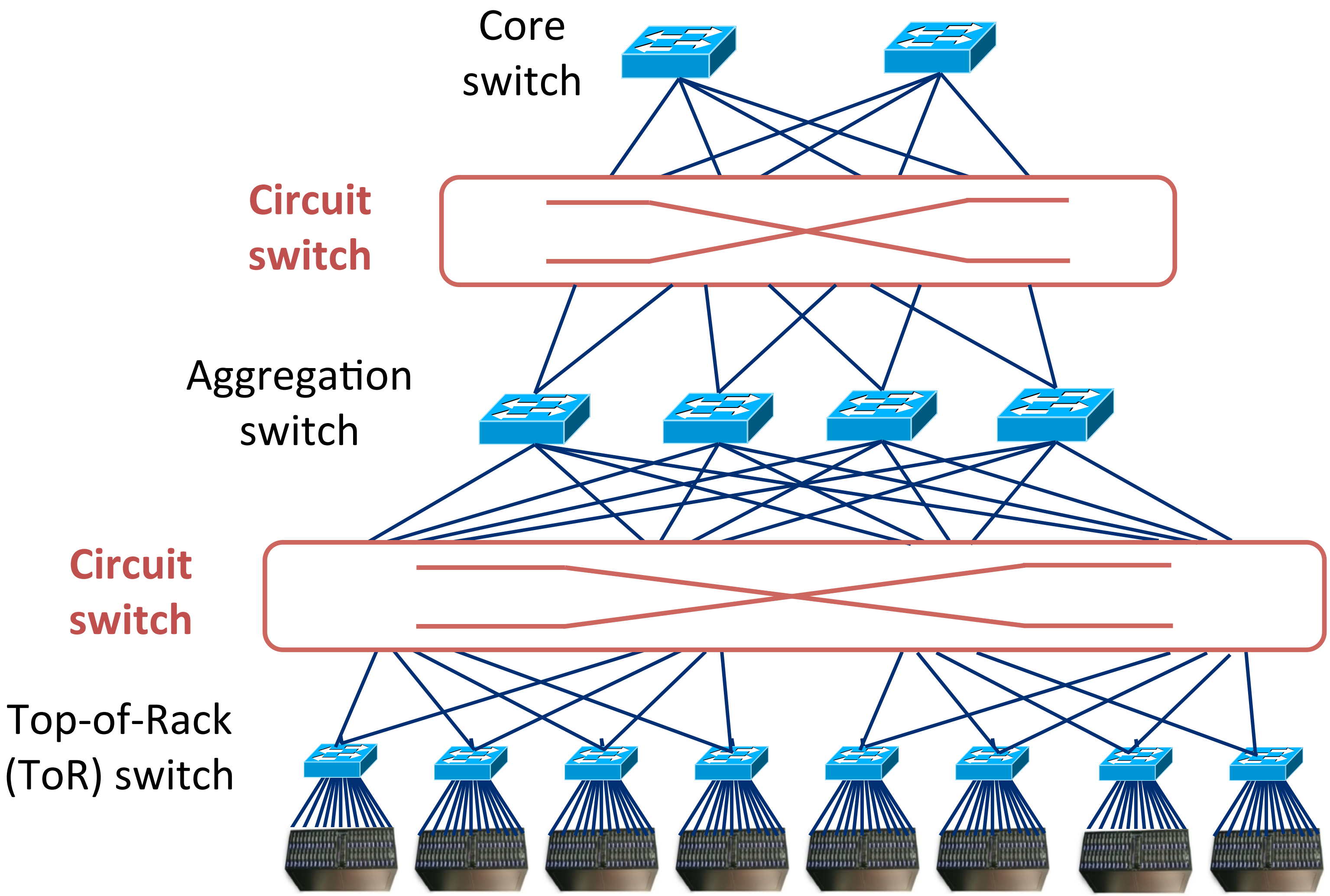
High-Level Idea: New Network Model

High-Level Idea: New Network Model

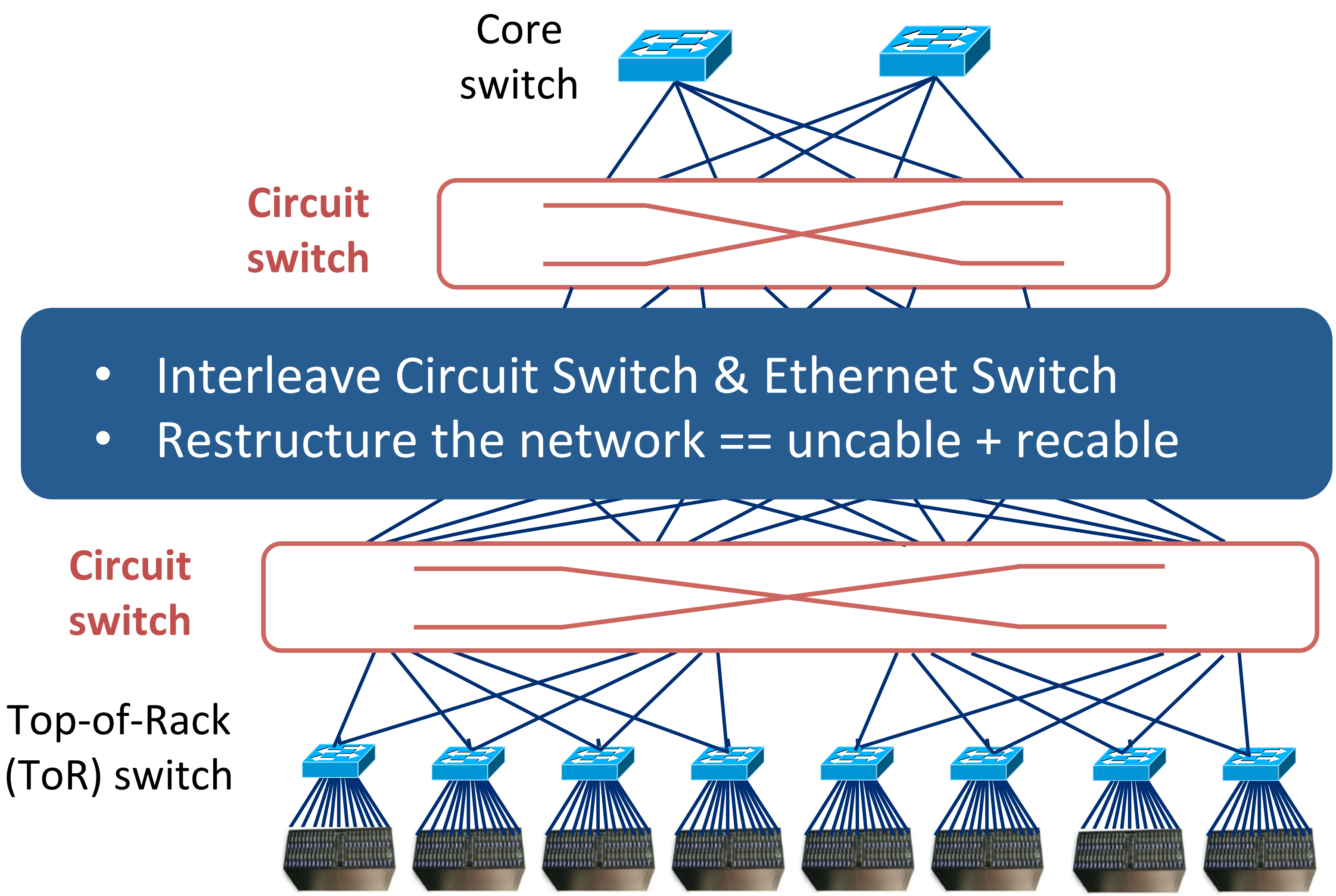


Today's Data Center

High-Level Idea: New Network Model

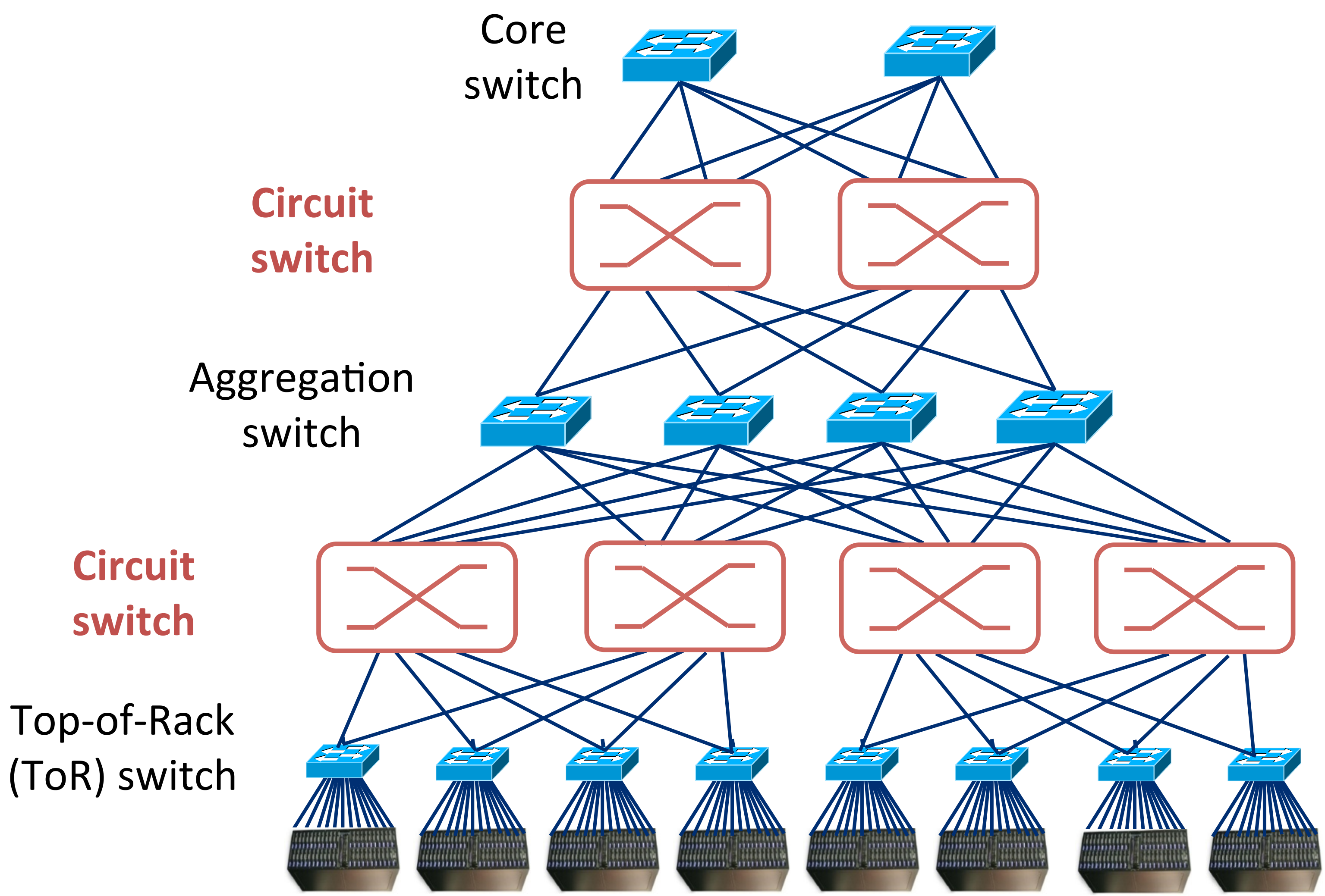


High-Level Idea: New Network Model

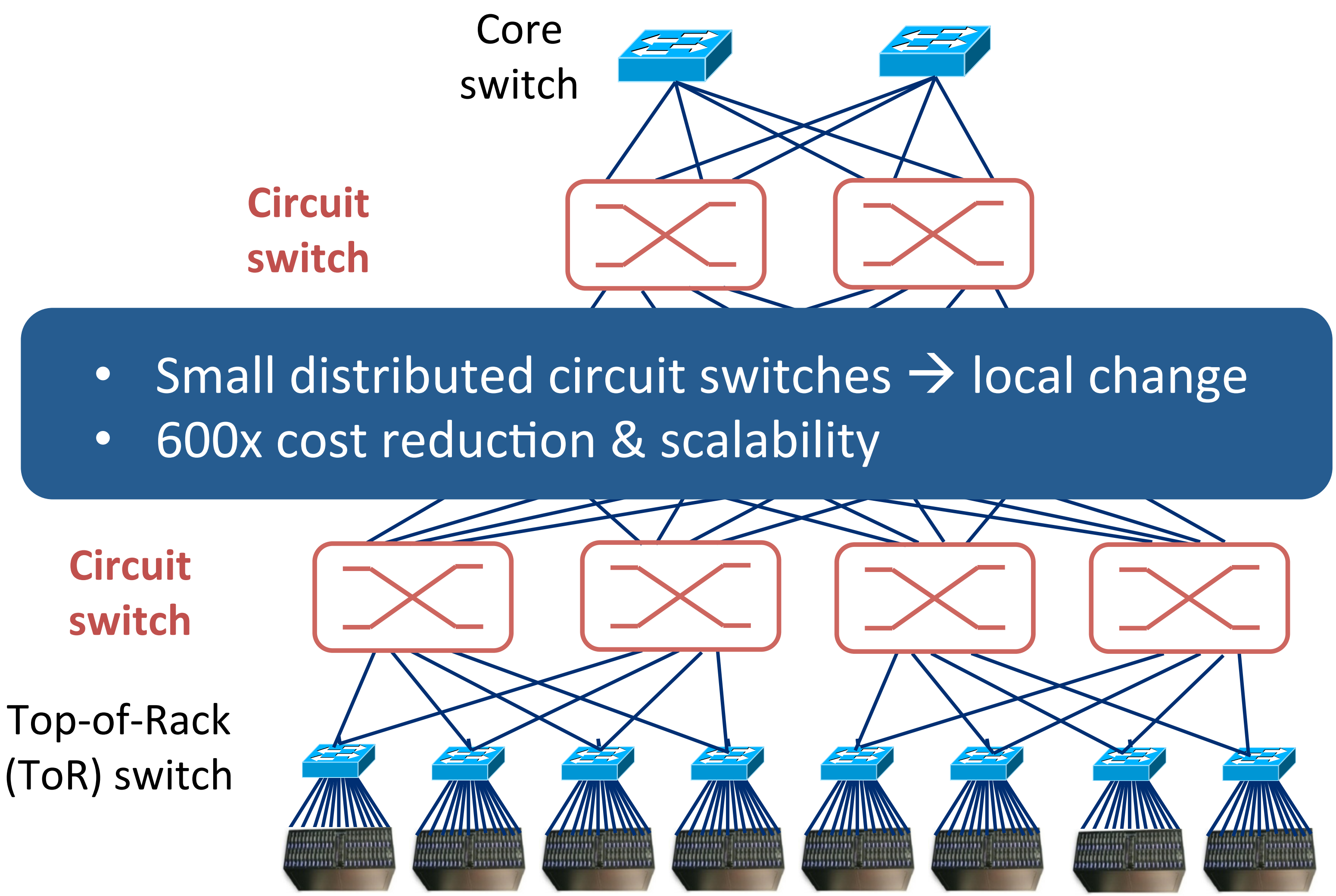


- Interleave Circuit Switch & Ethernet Switch
- Restructure the network == uncable + recable

High-Level Idea: New Network Model



High-Level Idea: New Network Model



- Small distributed circuit switches → local change
- 600x cost reduction & scalability

Outline

ShareBackup
[HotNets'17, SIGCOMM'18]

Failure Recovery

Flat-tree
[HotNets'16, SIGCOMM'17]

Service Provisioning

OmniSwitch
[HotCloud'15]

Wiring & Maintenance

Lighthouse
(In submission)

Physical-Layer Programmability in WAN

Outline

ShareBackup
[HotNets'17, SIGCOMM'18]

Failure Recovery

Flat-tree
[HotNets'16, SIGCOMM'17]

Service Provisioning

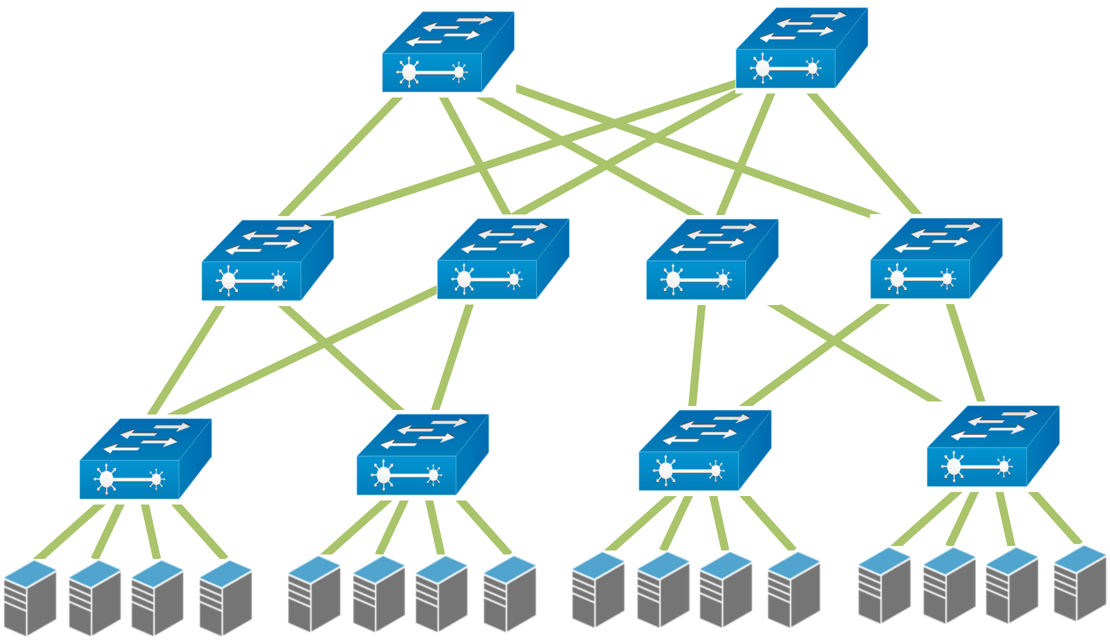
OmniSwitch
[HotCloud'15]

Wiring & Maintenance

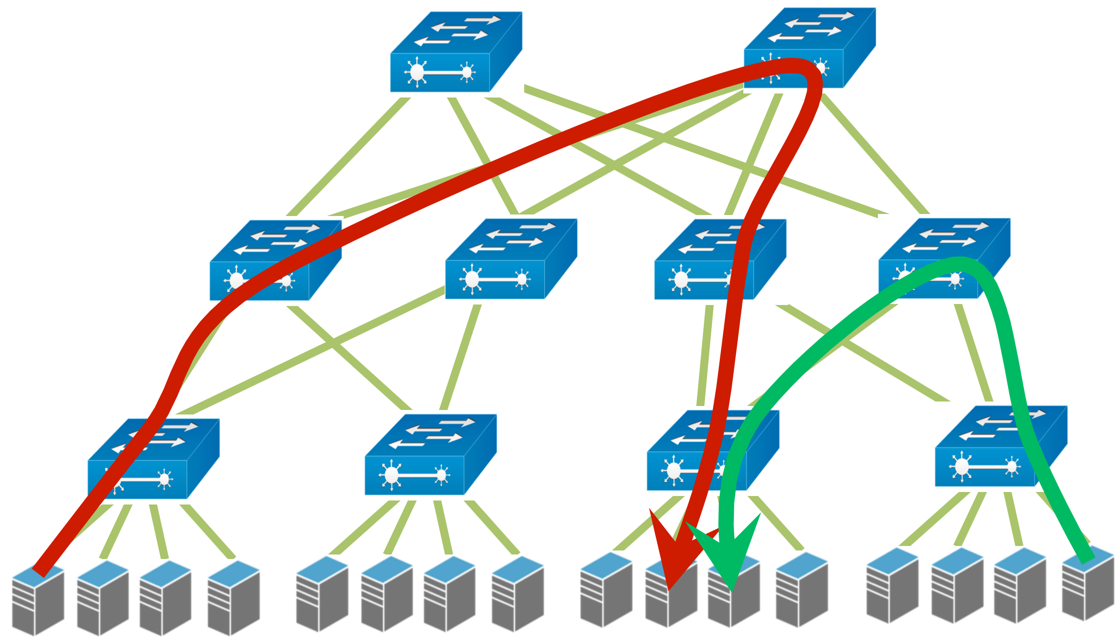
Lighthouse
(In submission)

Physical-Layer Programmability in WAN

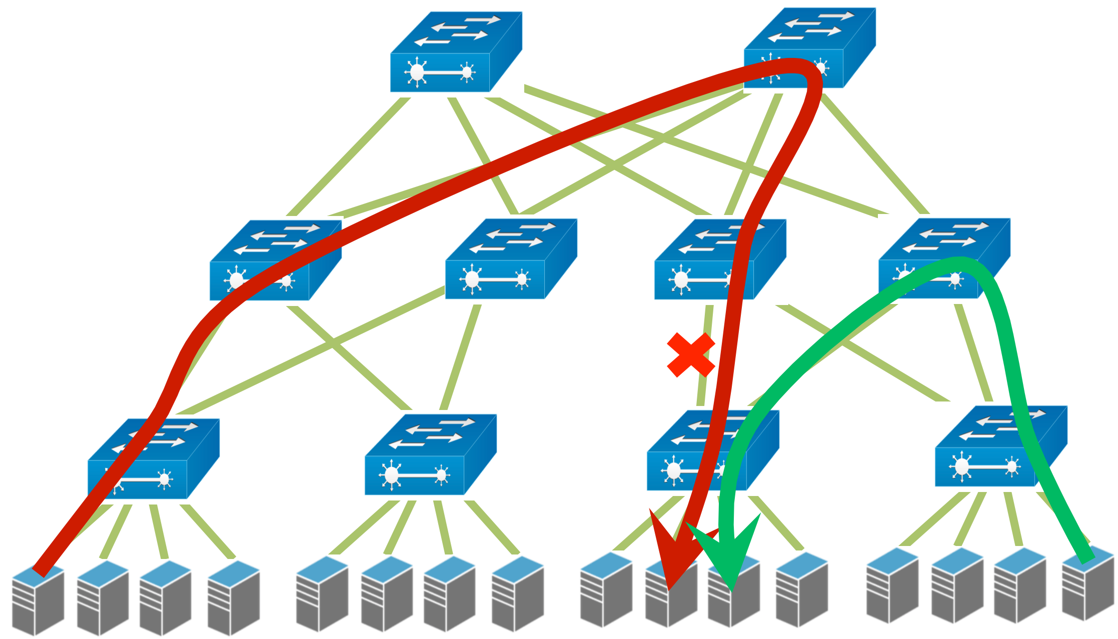
Default Failure Recovery: Rerouting



Default Failure Recovery: Rerouting

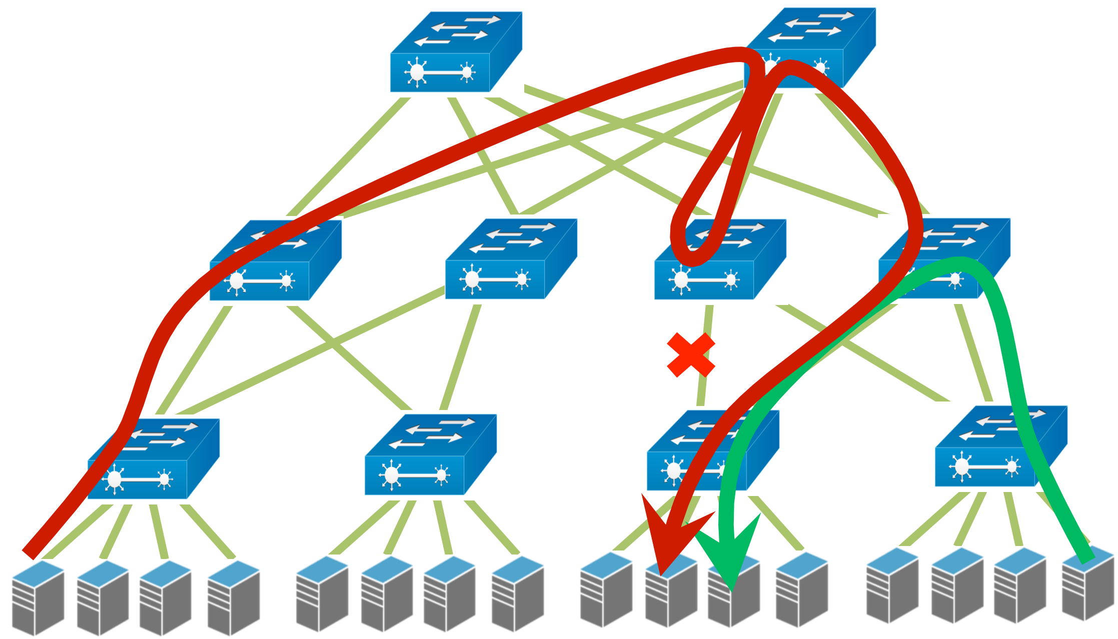


Default Failure Recovery: Rerouting



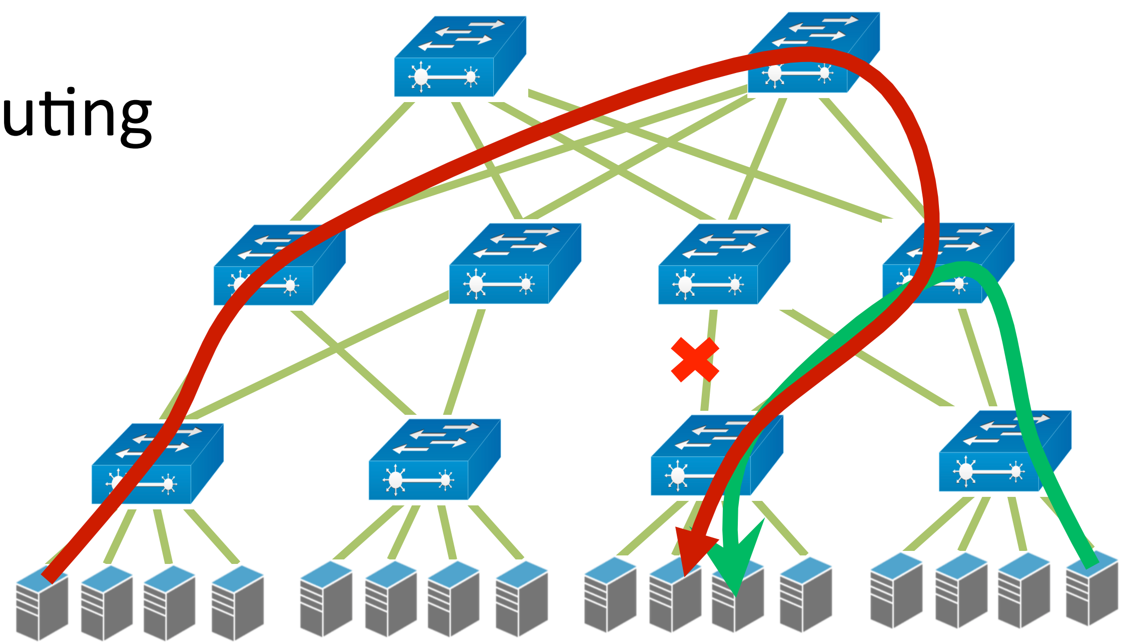
Default Failure Recovery: Rerouting

- Fast local rerouting
  - *Inflated path length*



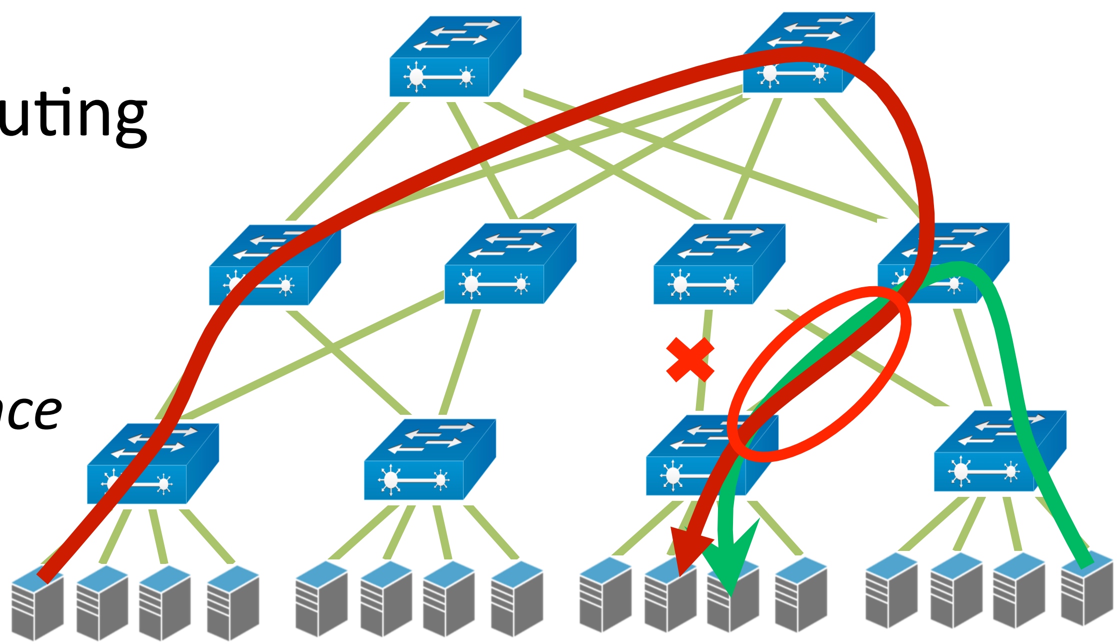
Default Failure Recovery: Rerouting

- Fast local rerouting
  - *Inflated path length*
- Global optimal rerouting
  - *High latency*



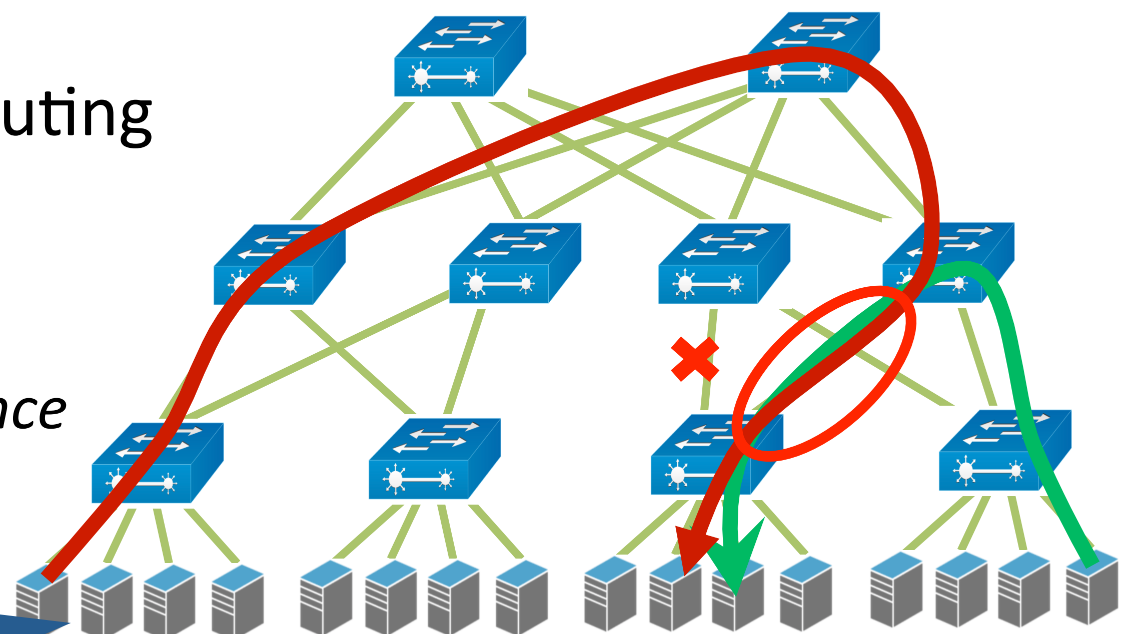
Default Failure Recovery: Rerouting

- Fast local rerouting
  - *Inflated path length*
- Global optimal rerouting
  - *High latency*
- Impact other flows
  - *Degraded performance*



Default Failure Recovery: Rerouting

- Fast local rerouting
  - *- Inflated path length*
- Global optimal rerouting
  - *- High latency*
- Impact other flows
  - *- Degraded performance*



Restore bandwidth immediately!

Shareable Backup

No backup — 1:1 backup

Shareable Backup

No backup **Different backup ratios** 1:1 backup

Shareable Backup

No backup **Different backup ratios** 1:1 backup

Circuit Switches → a pool of backup switches

Shareable Backup

1. Failures are rare
2. Failed switch replaced in ms

No backup

1:1 backup

Circuit Switches → a pool of backup switches

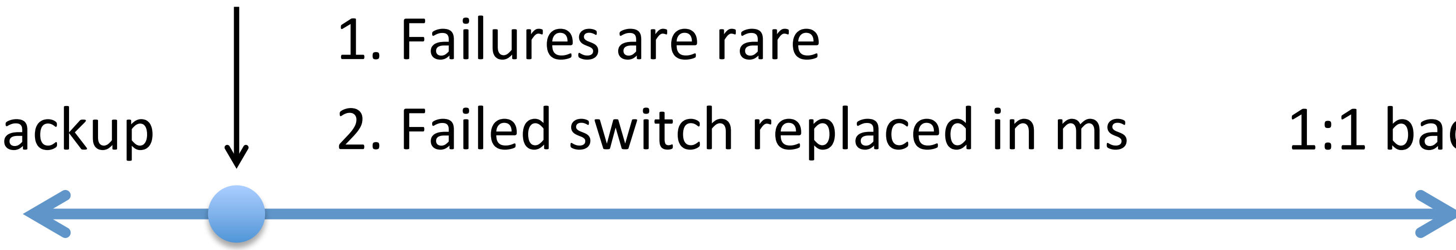
Gill et al., Understanding Network Failures in Data Centers: Measurement, Analysis, and Implications, SIGCOMM 2011

Shareable Backup

1. Failures are rare
2. Failed switch replaced in ms

No backup

1:1 backup

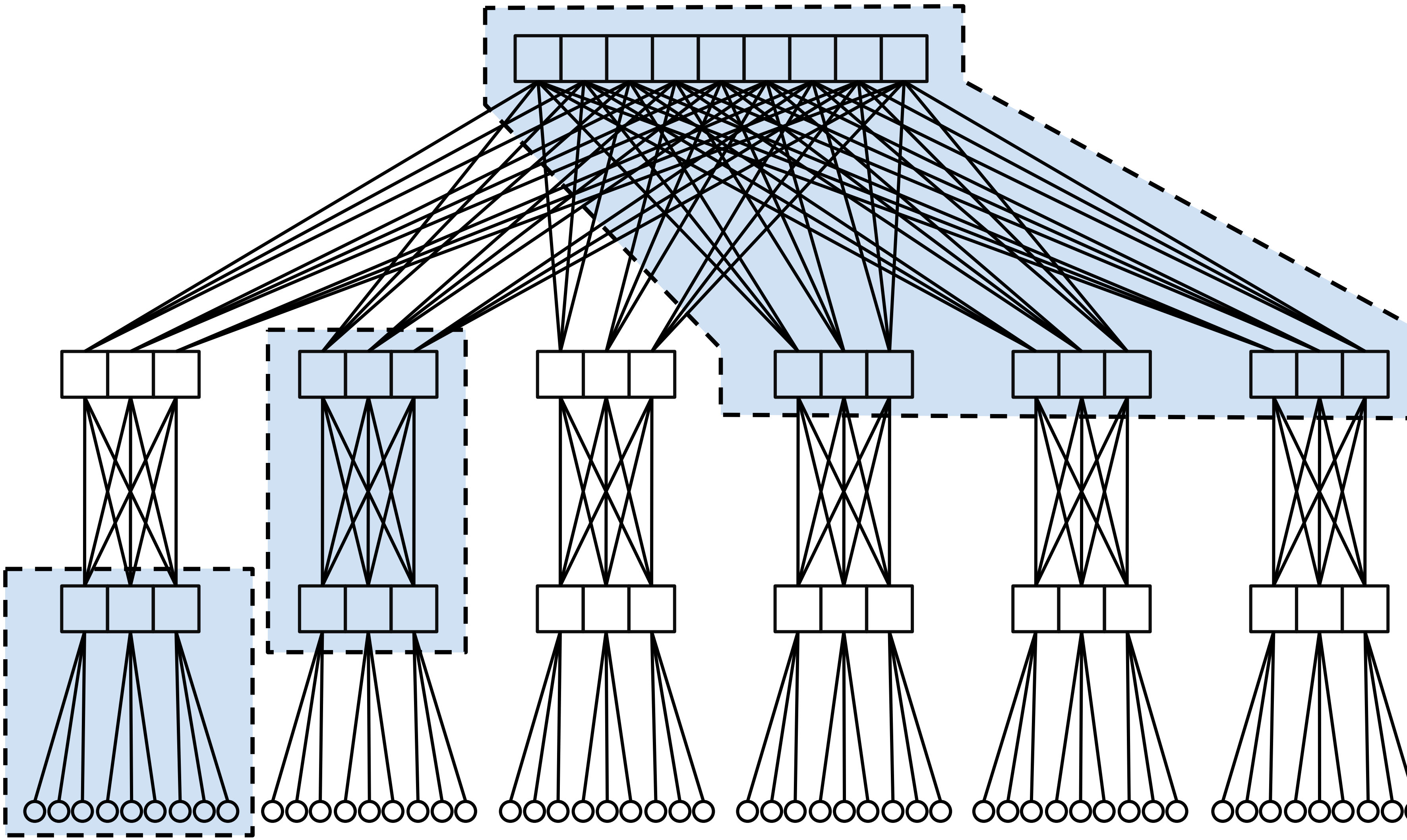


Circuit Switches → a pool of backup switches

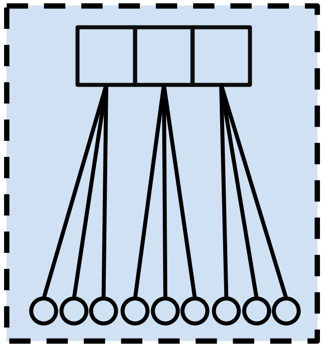
ShareBackup: Data Center with Shareable Backup

Gill et al., Understanding Network Failures in Data Centers: Measurement, Analysis, and Implications, SIGCOMM 2011

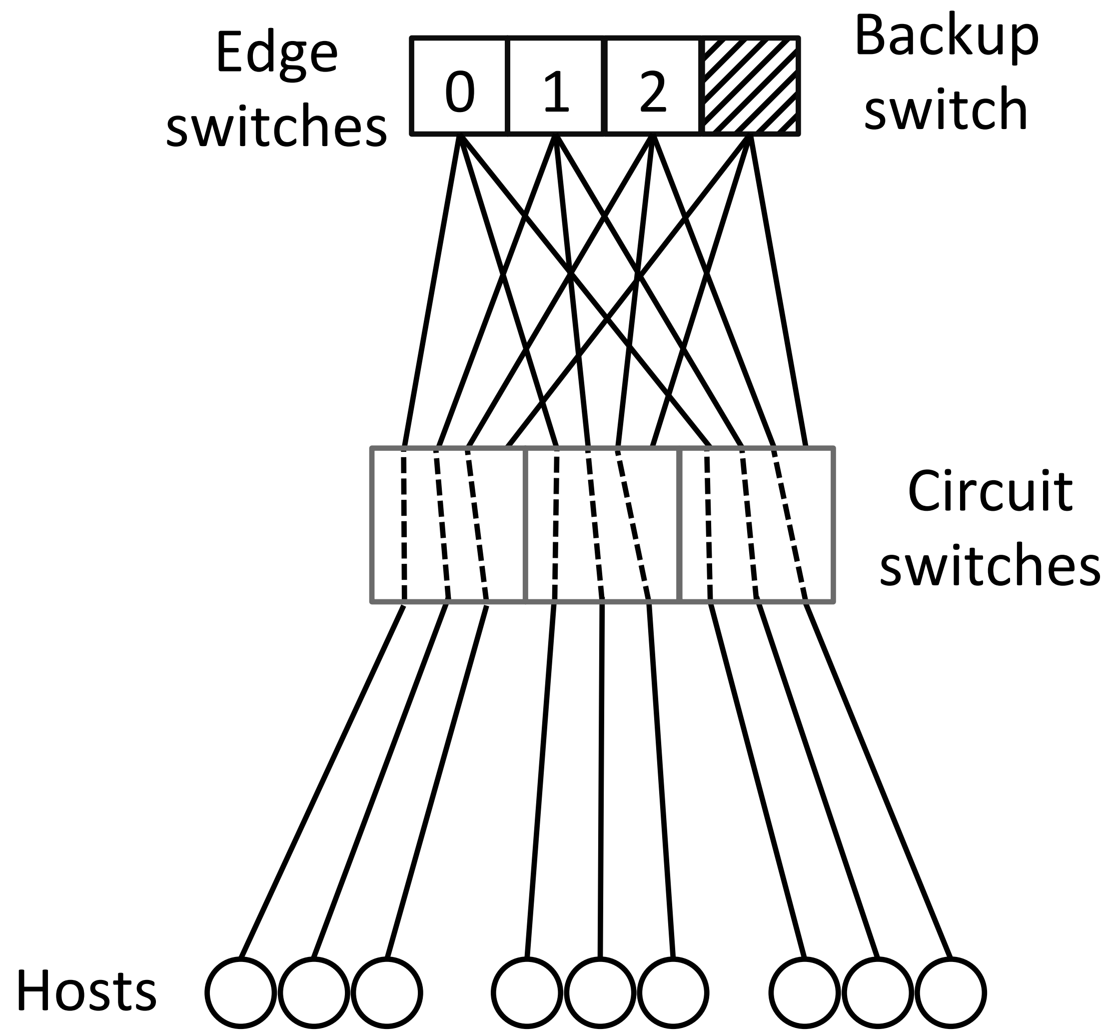
Architecture Design

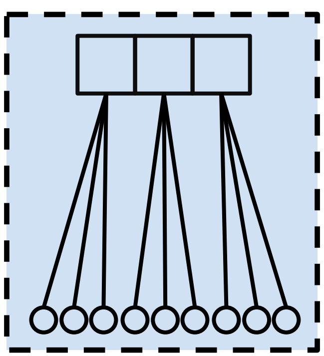


Architecture Design

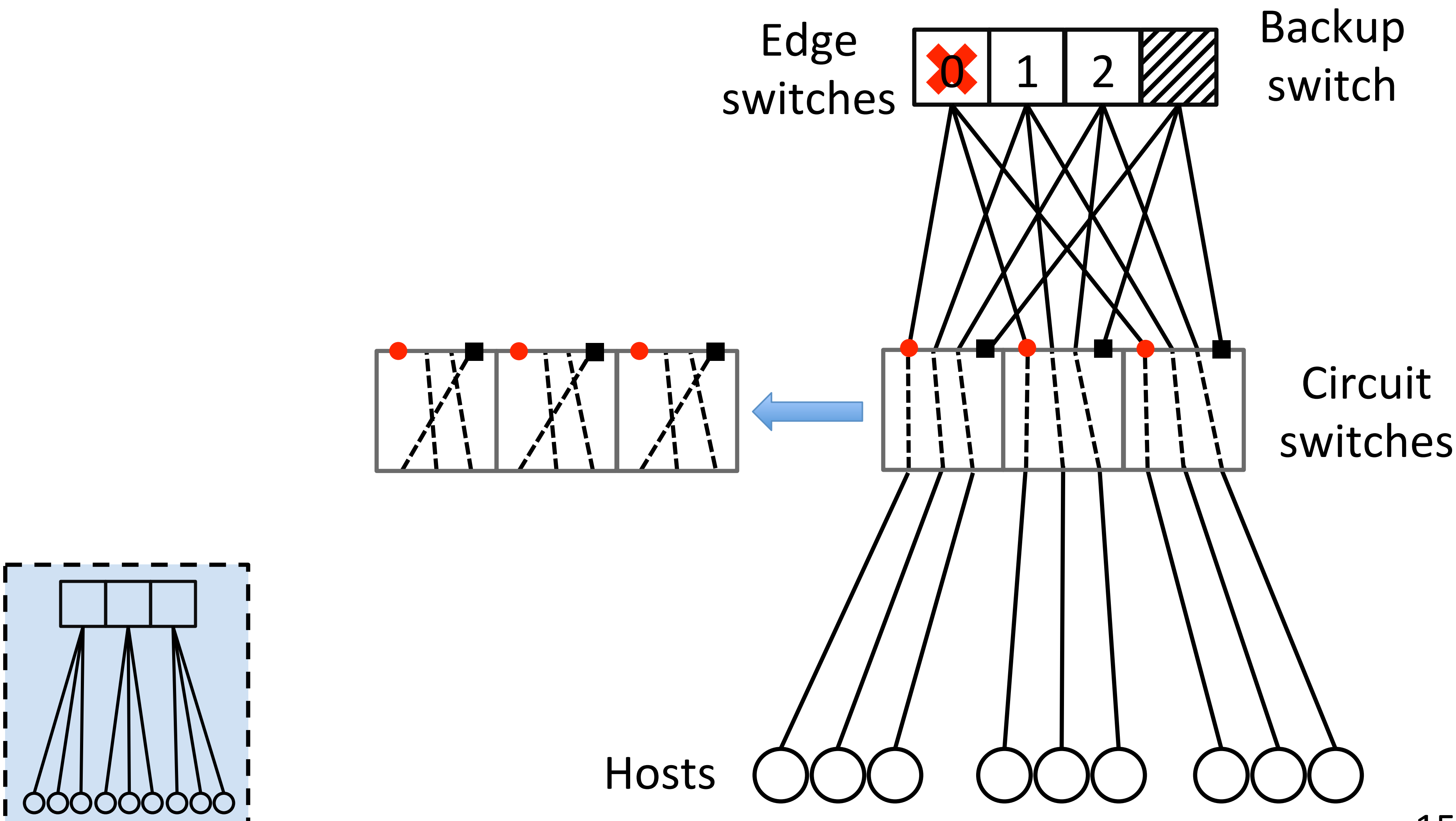


Architecture Design

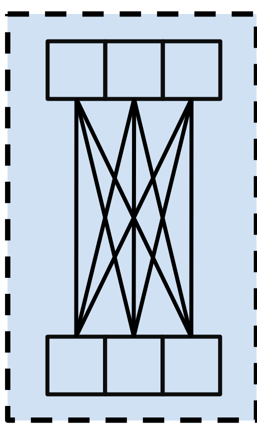


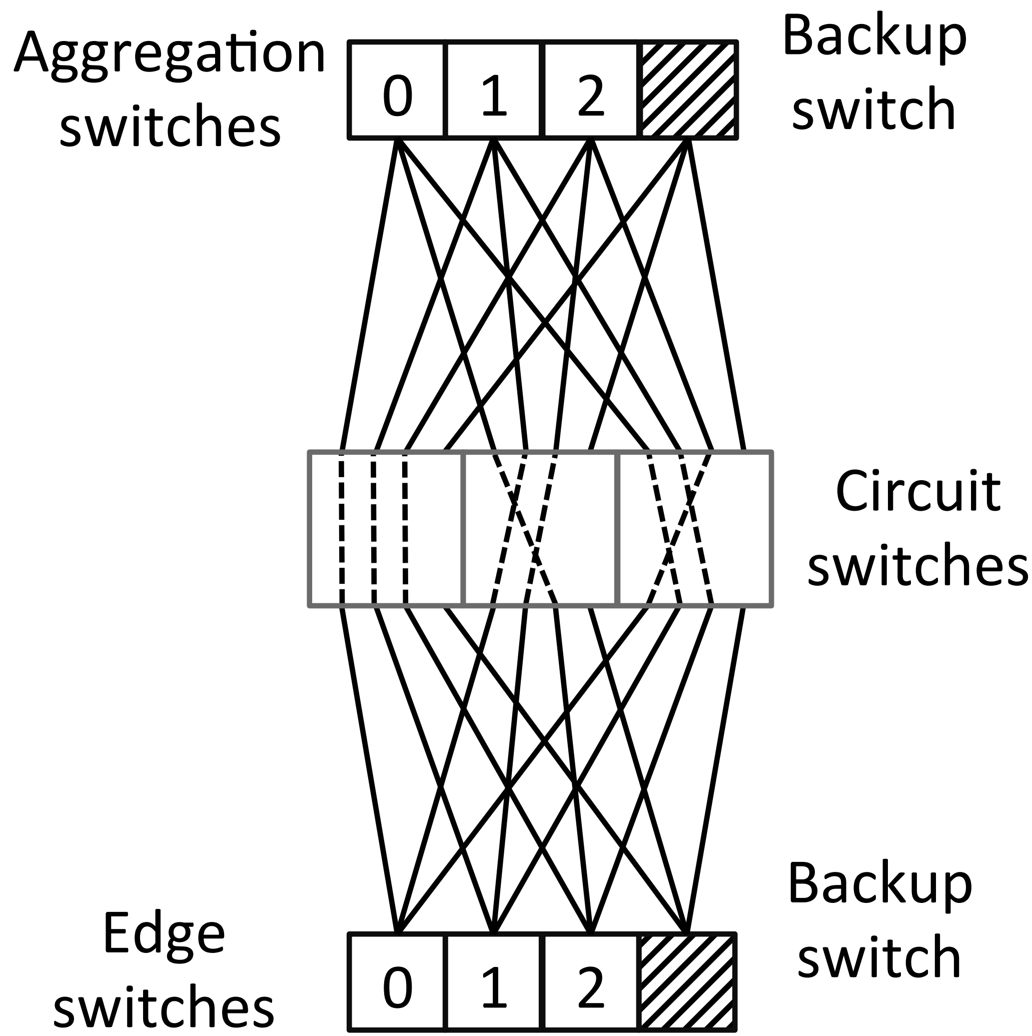


Architecture Design

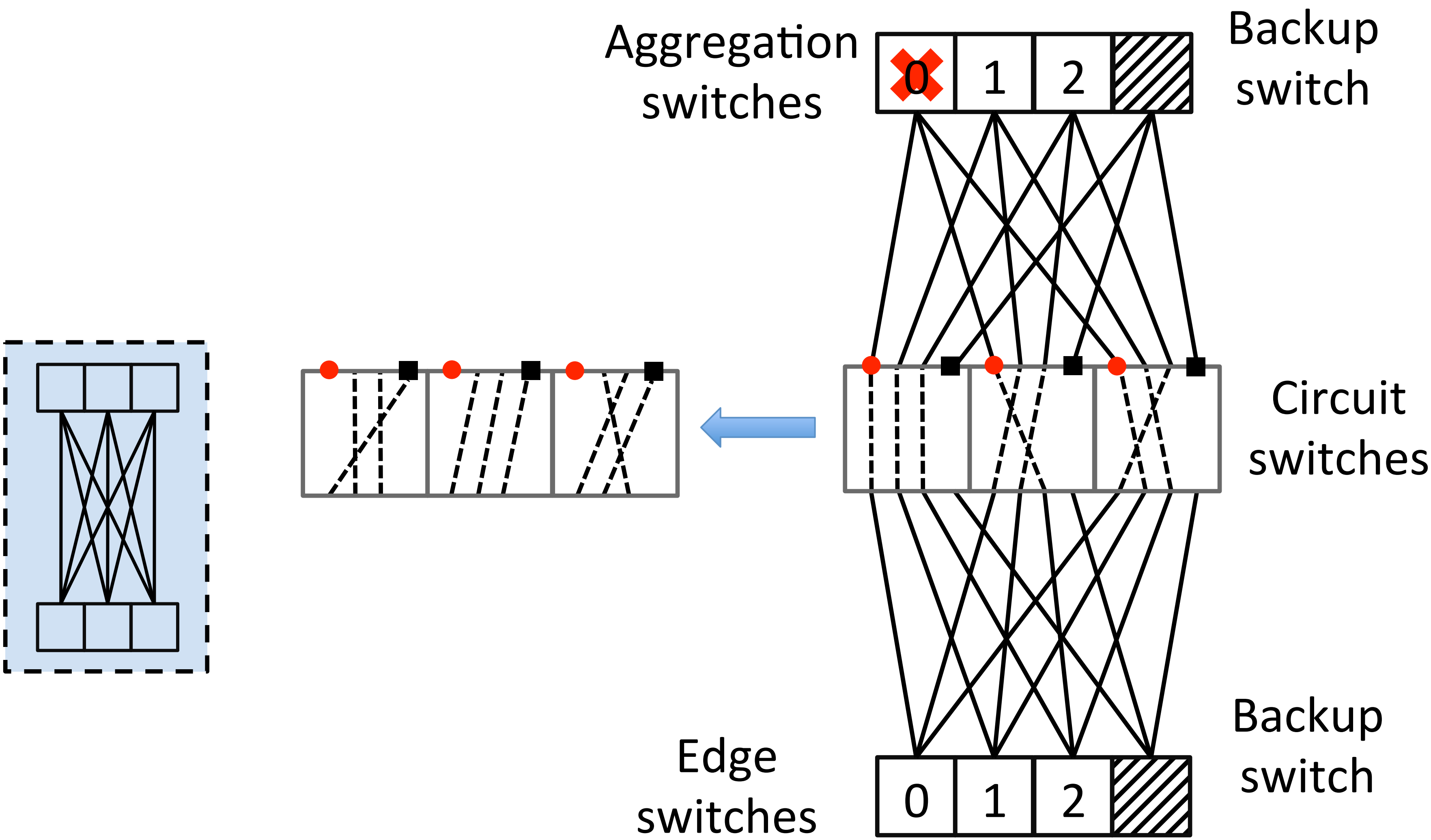


Architecture Design

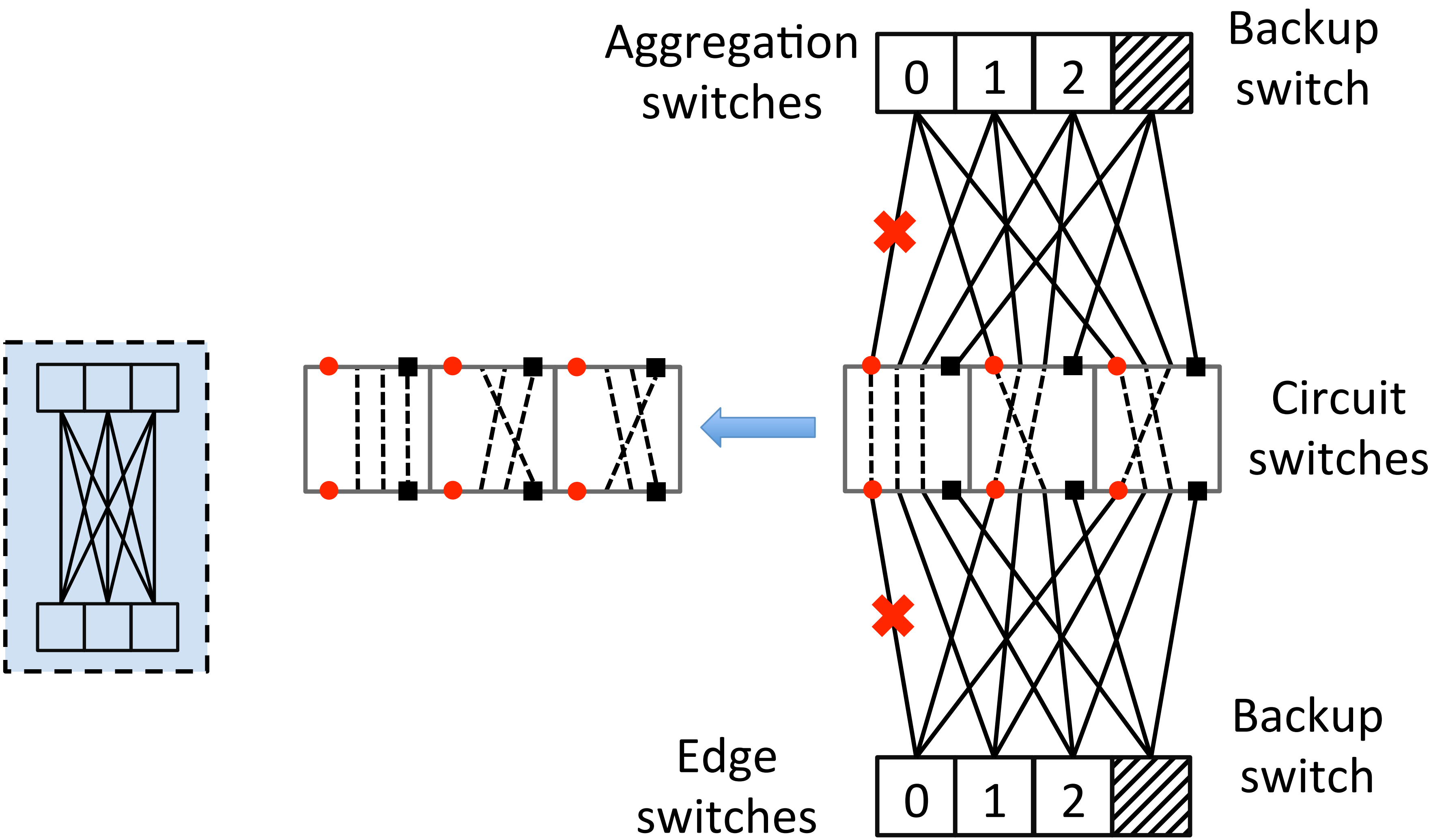




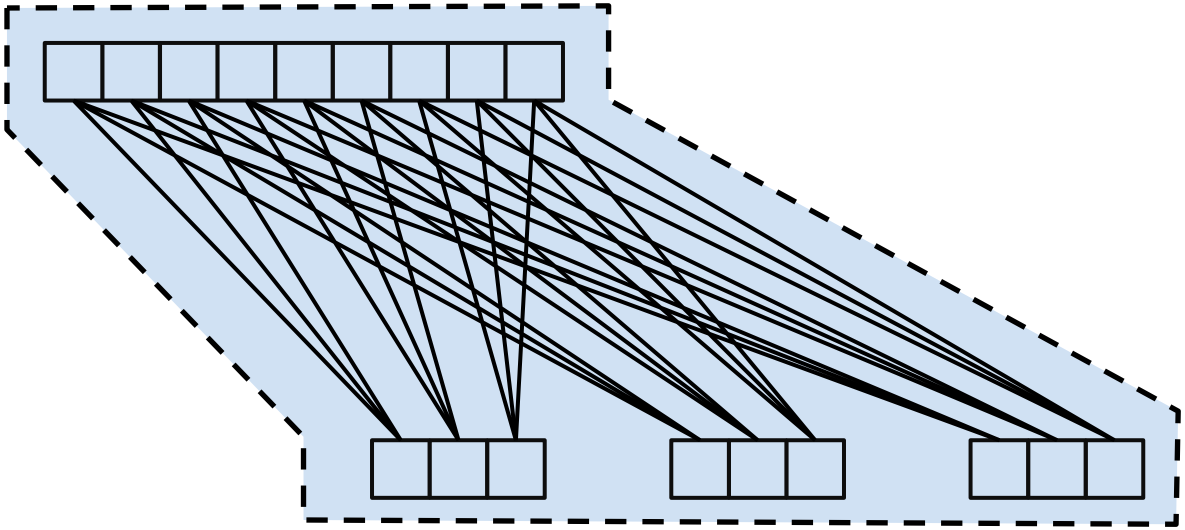
Architecture Design



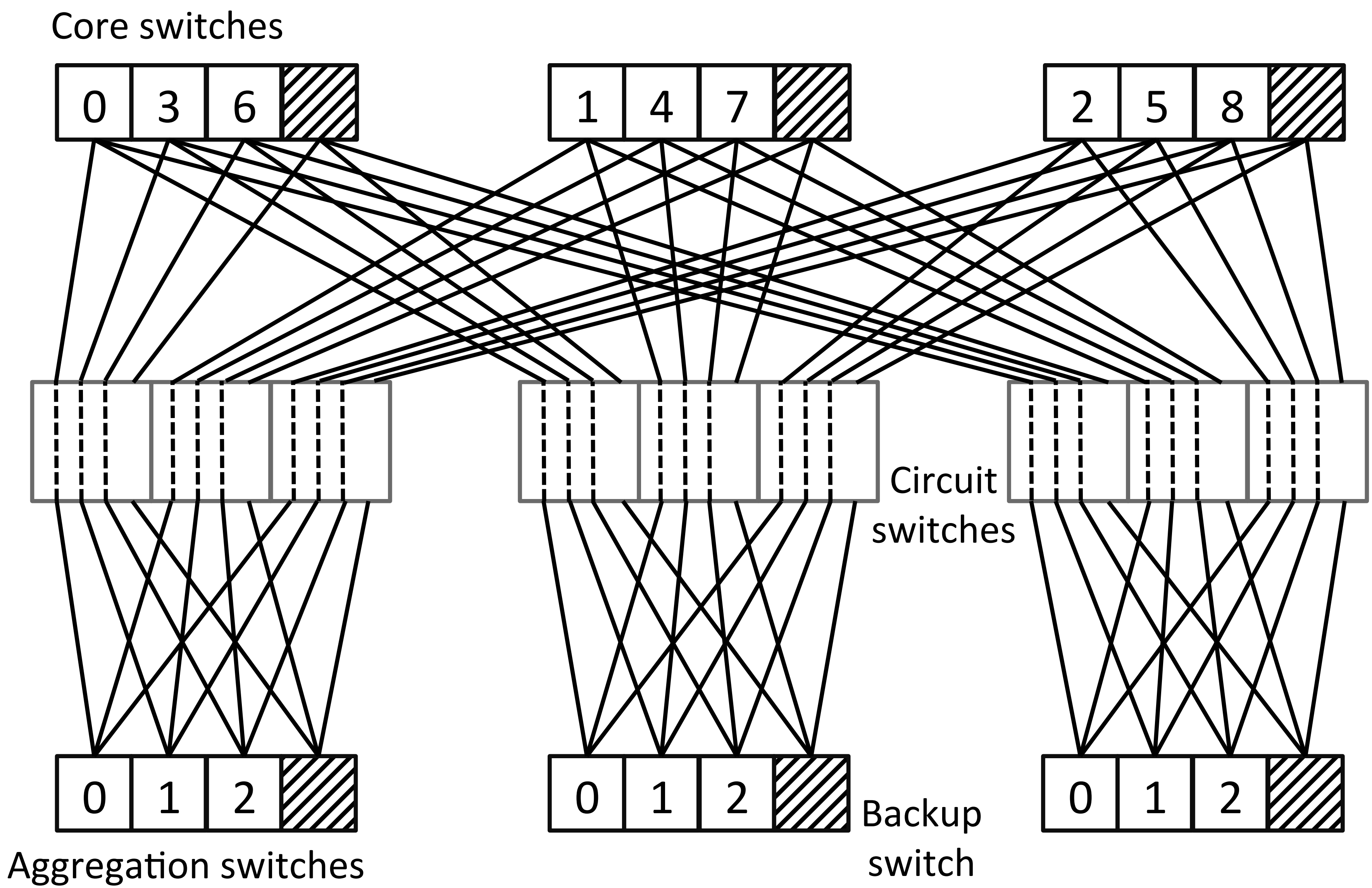
Architecture Design

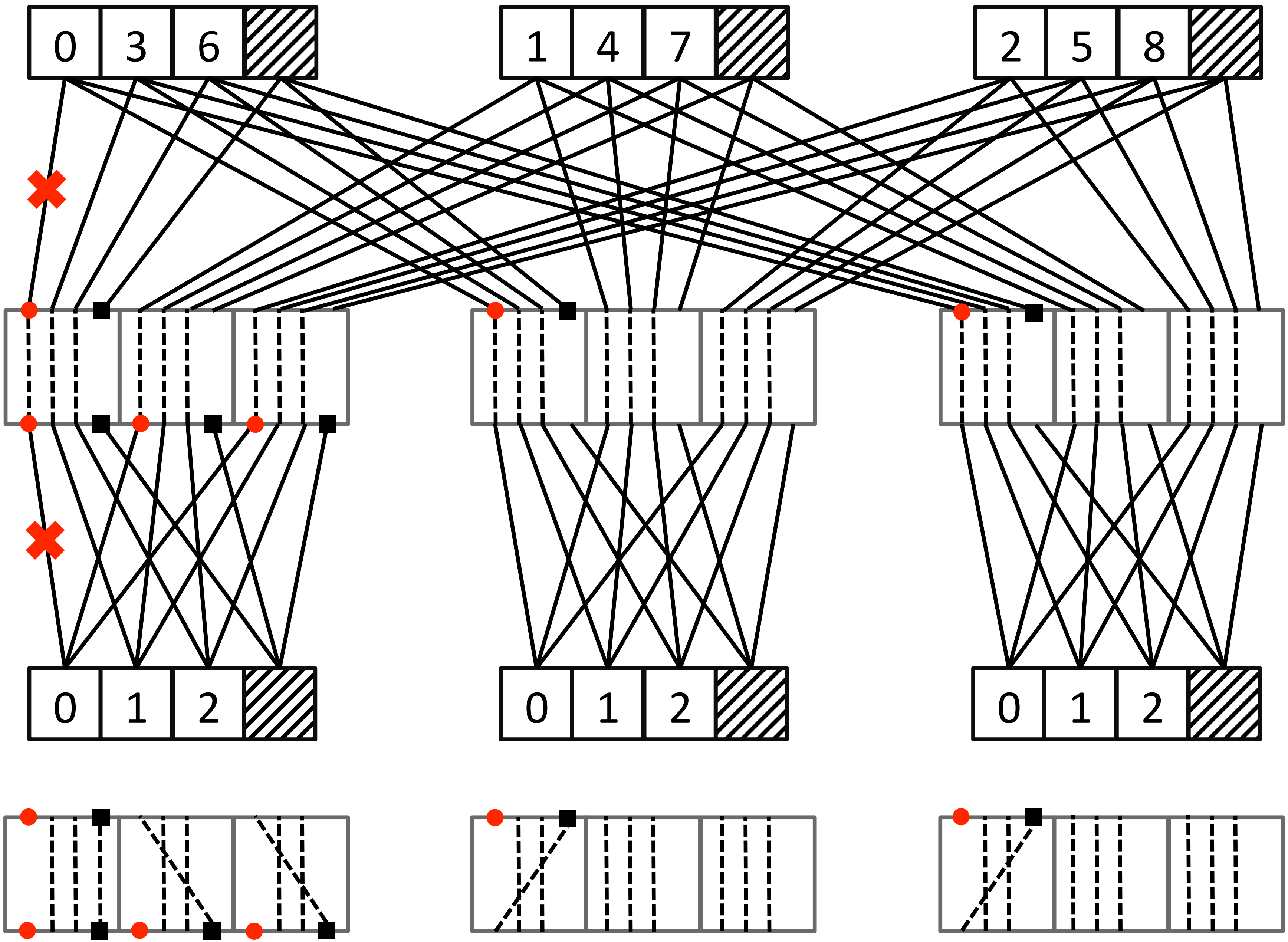


Architecture Design



Architecture Design




0
3
6
1
4
7
2
5
8
0
1
2
0
1
2
0
1
2

Challenge: Fast Failure Recovery

Challenge: Fast Failure Recovery

- Distributed controllers
  - *Configure circuit switches quickly*

Challenge: Fast Failure Recovery

- Distributed controllers
  - *Configure circuit switches quickly*
- Live impersonation
  - *Backup switch as hot standby*

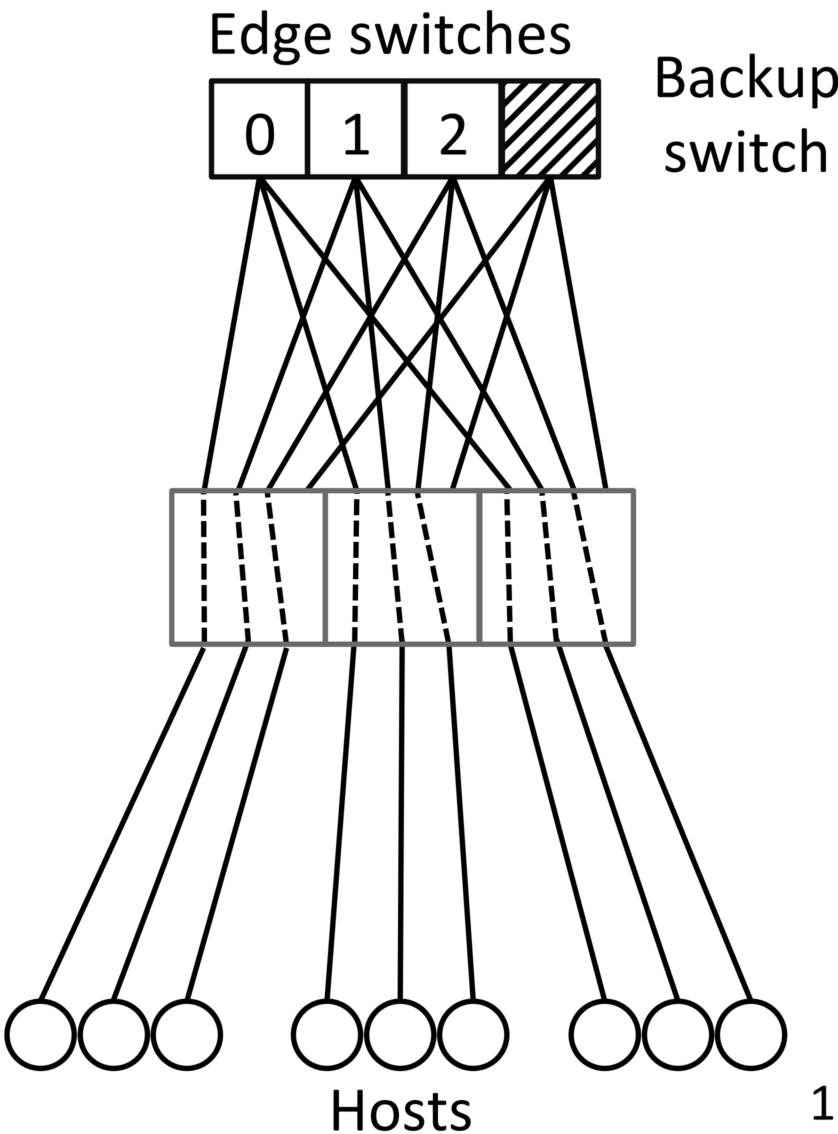
Challenge: Fast Failure Recovery

- Distributed controllers
  - *Configure circuit switches quickly*
- Live impersonation
  - *Backup switch as hot standby*
- Routing table in place
  - *Save the time of setting forwarding rules*

Live Impersonation

Routing Table of Every Edge Switch

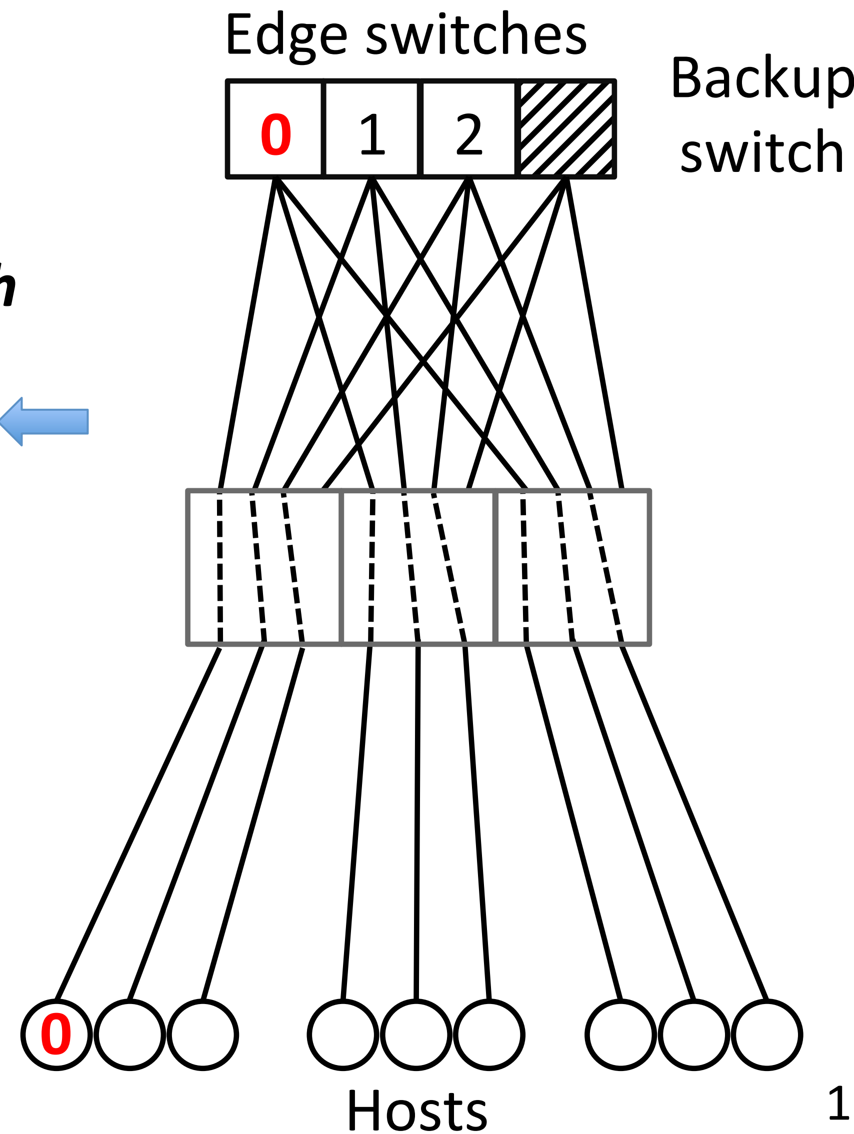
Routing Table 0	**VLAN 0**
Routing Table 1	**VLAN 1**
Routing Table 2	**VLAN 2**



Live Impersonation

Routing Table of Every Edge Switch

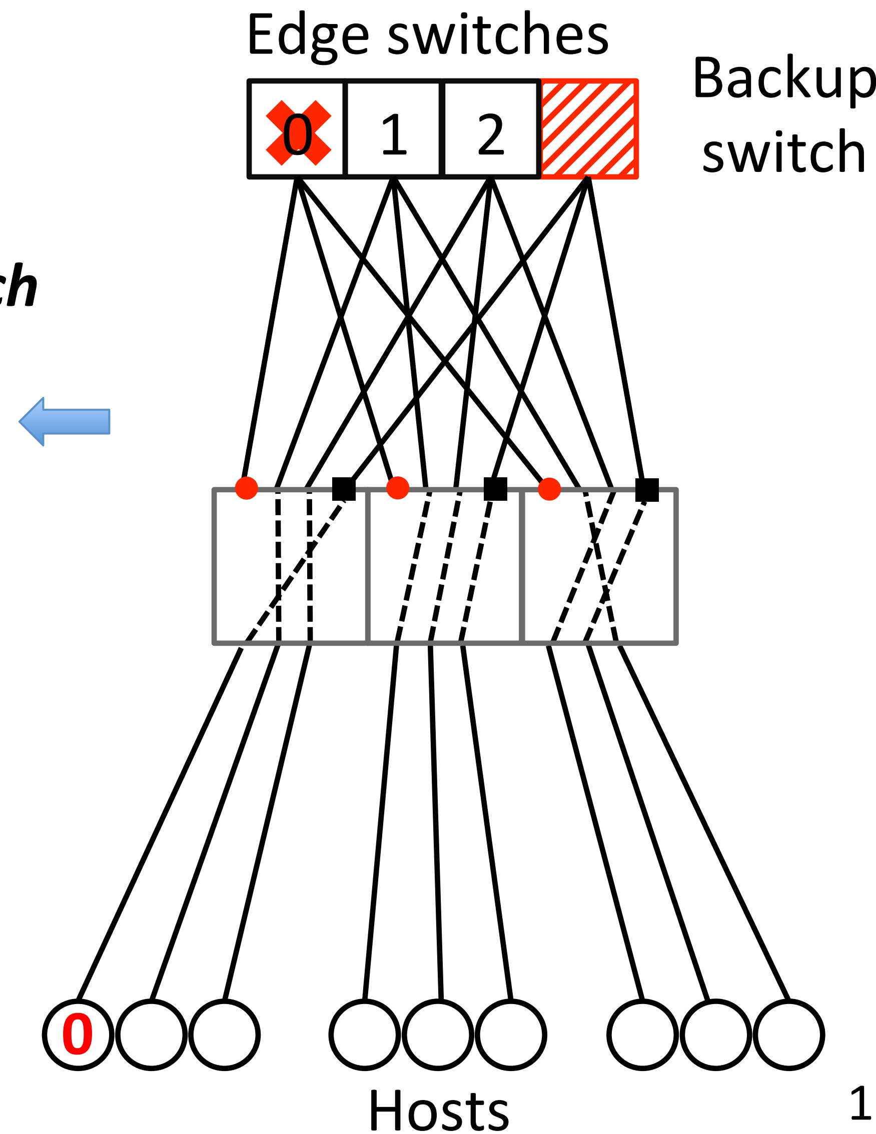
Routing Table 0	**VLAN 0**
Routing Table 1	**VLAN 1**
Routing Table 2	**VLAN 2**



Live Impersonation

Routing Table of Every Edge Switch

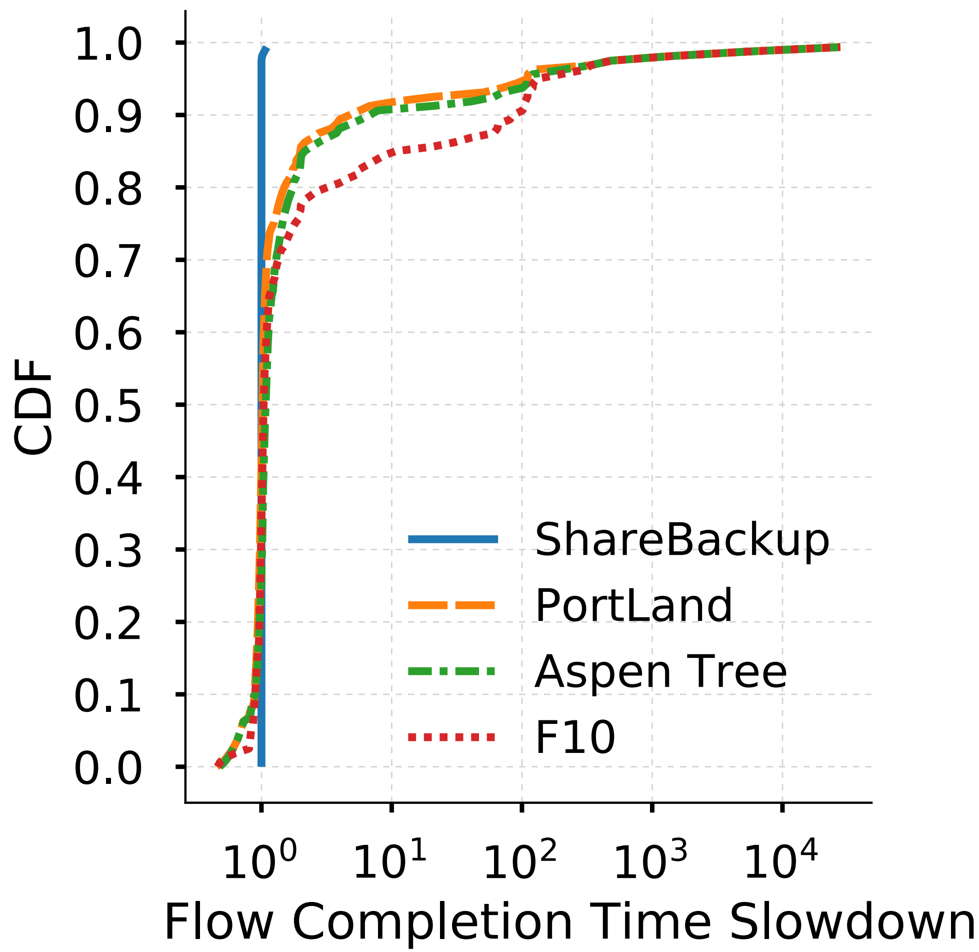
Routing Table 0	VLAN 0
Routing Table 1	VLAN 1
Routing Table 2	VLAN 2



Simulation

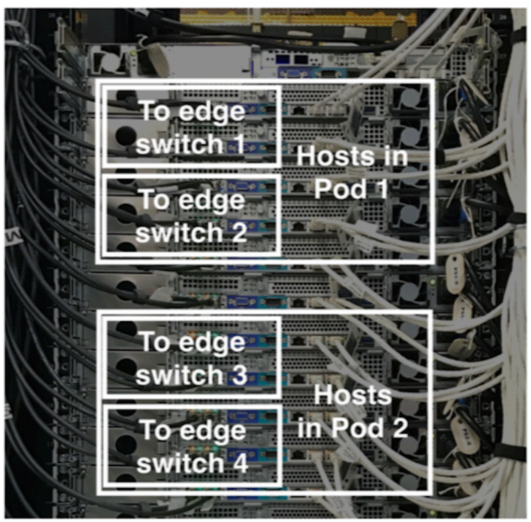
- Facebook prod traffic trace
- Microsoft prod failure distribution

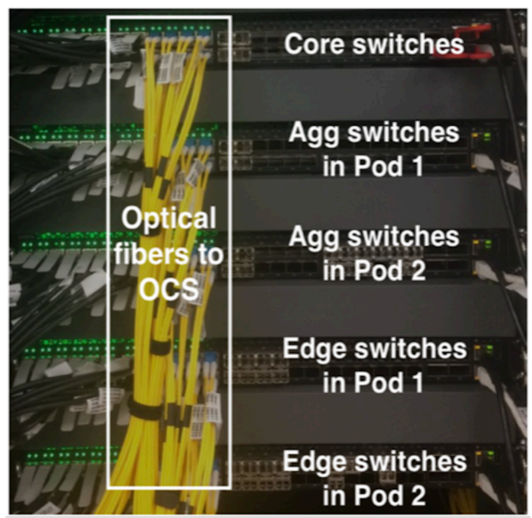
- Near-zero slow down during failures

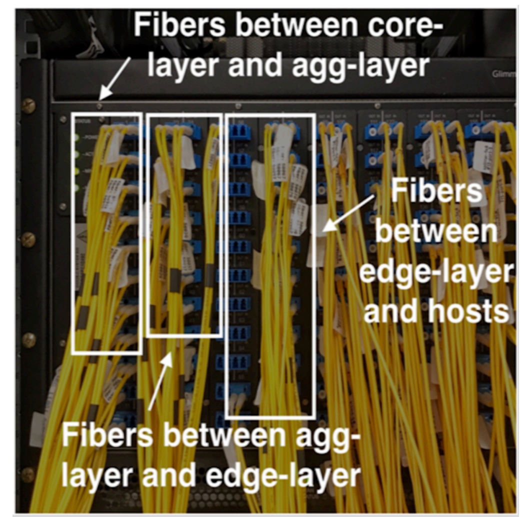


Testbed

- 24 hosts, 12 regular switches, 6 backup switches
- Hadoop & Spark applications

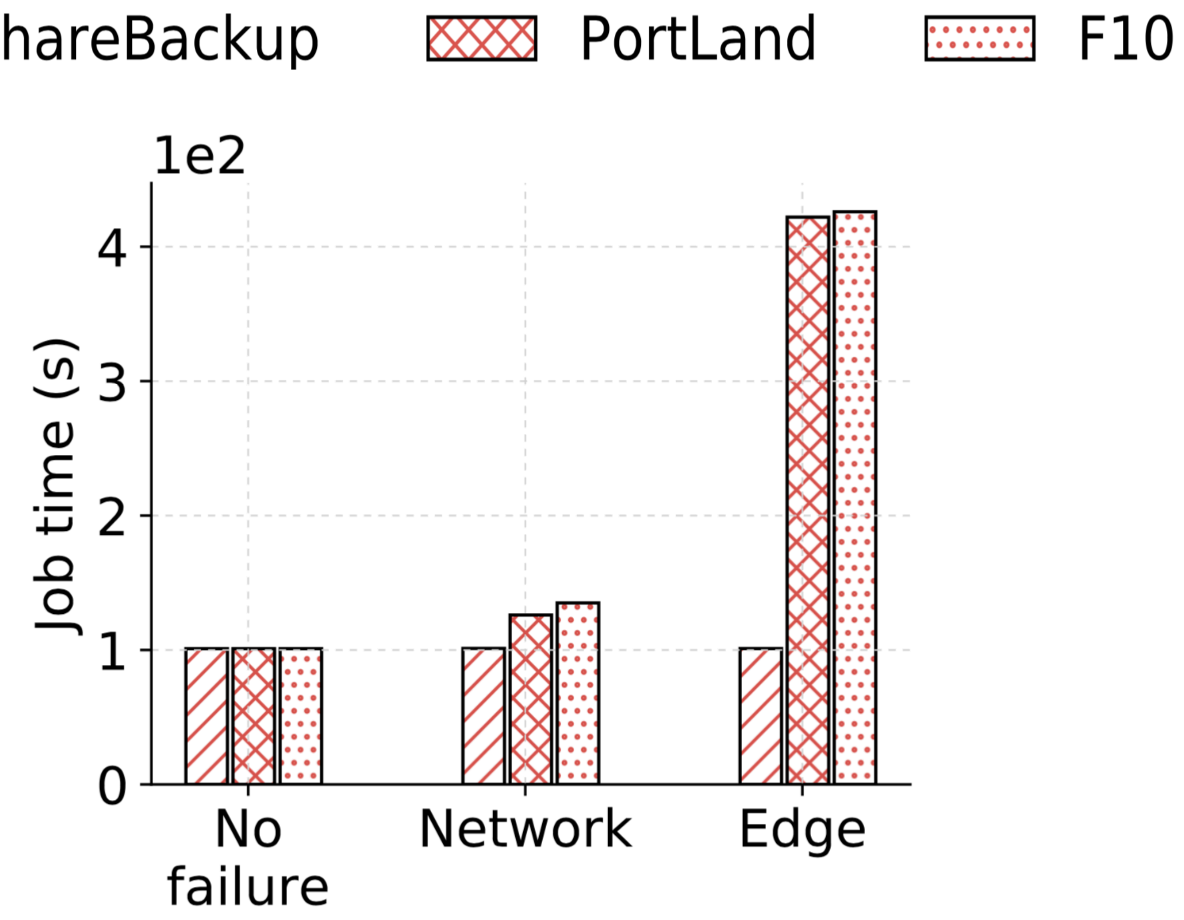






Testbed

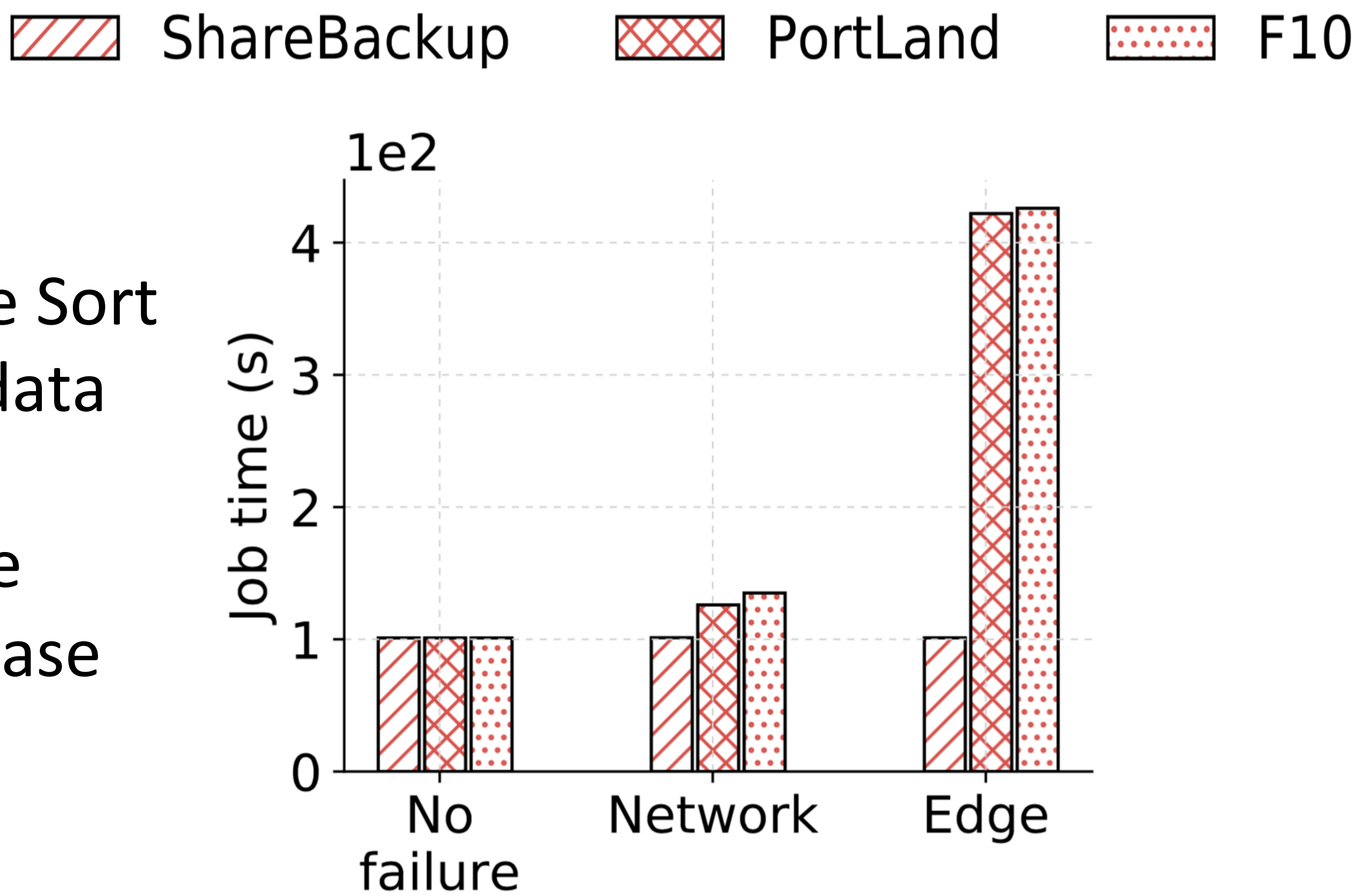
- MapReduce Sort w/ 100GB data



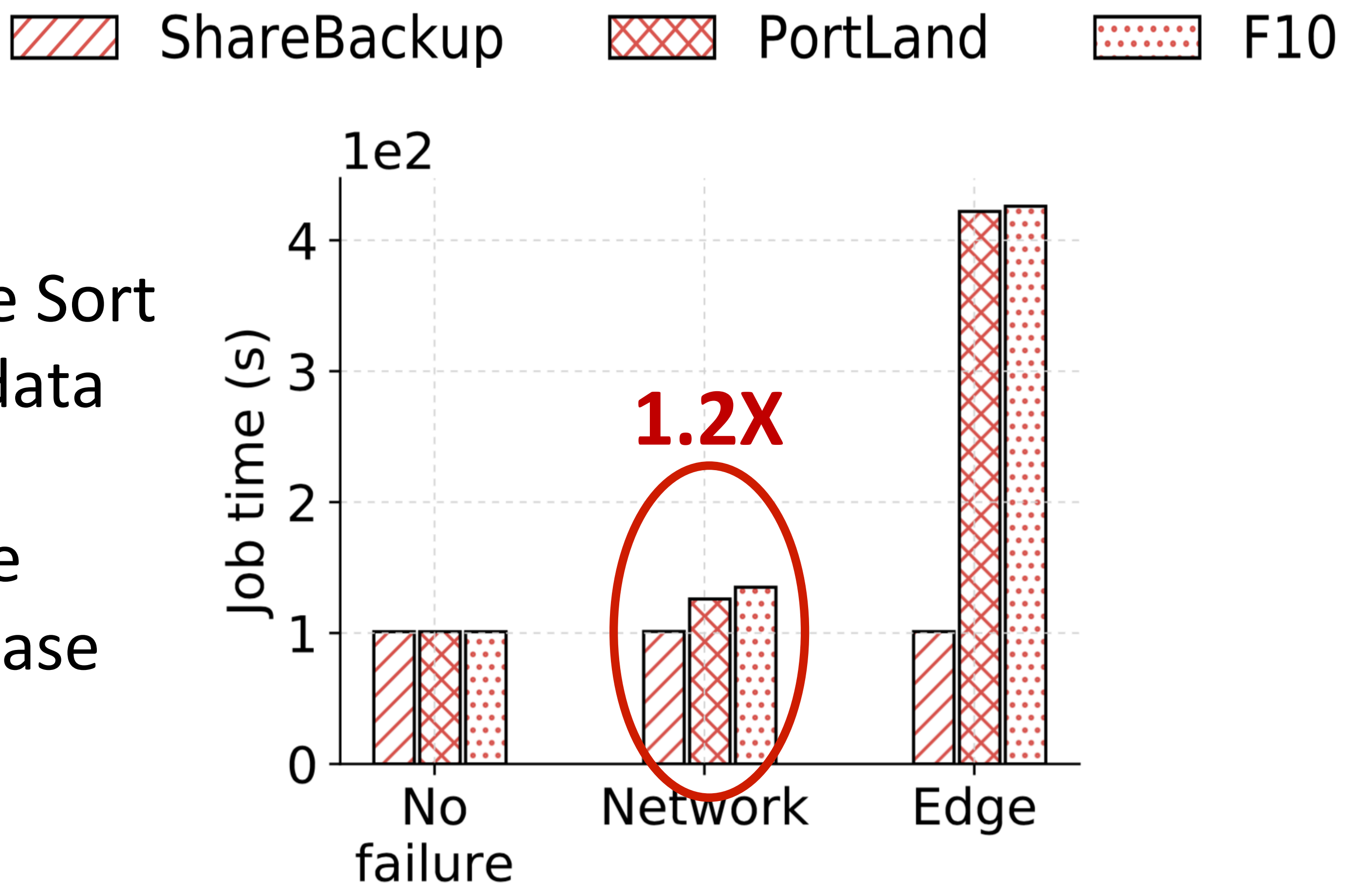
Testbed

- MapReduce Sort w/ 100GB data

- Same as the no-failure case



Testbed

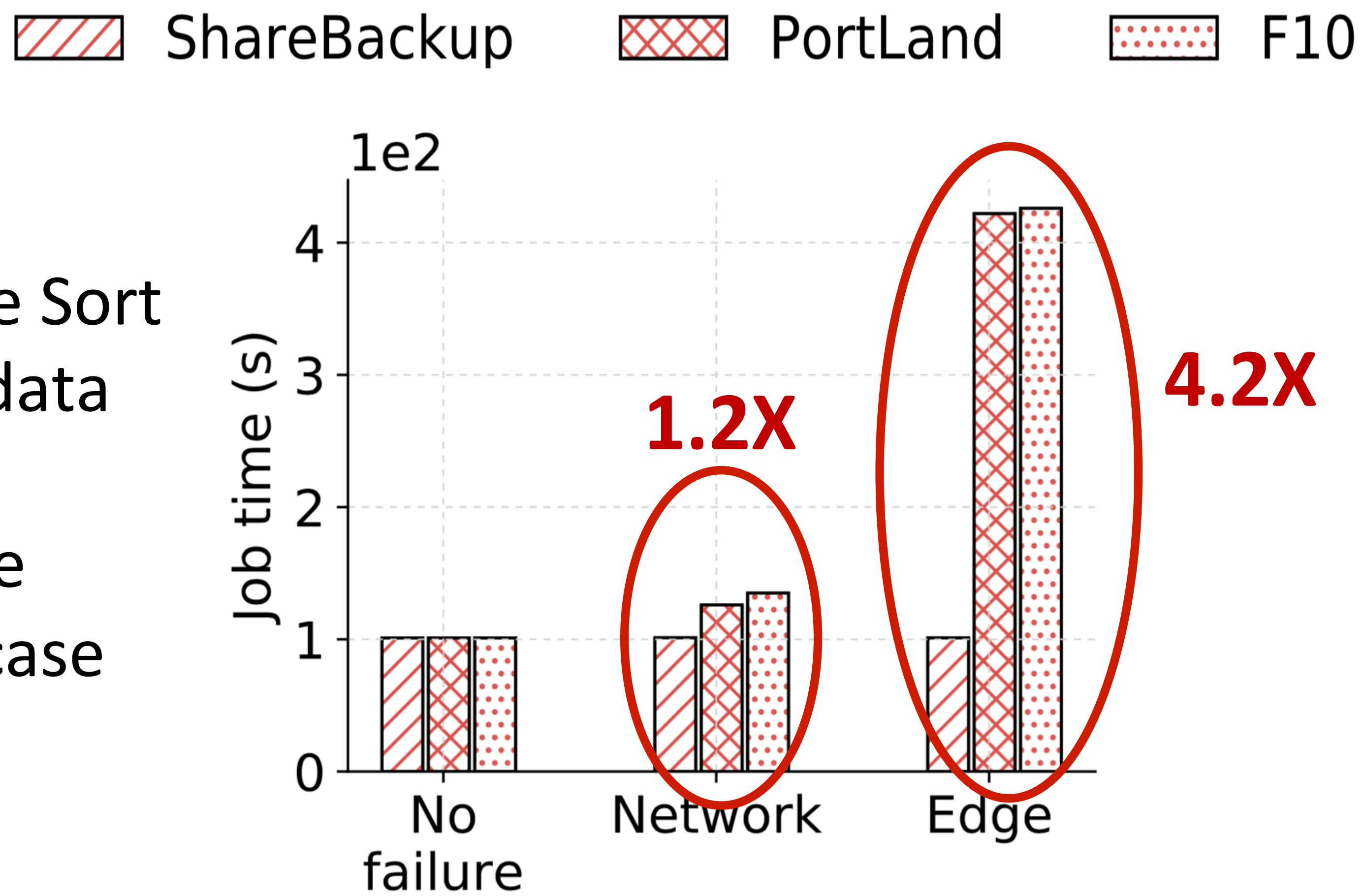


- MapReduce Sort w/ 100GB data

- Same as the no-failure case

Testbed

- MapReduce Sort w/ 100GB data
- Same as the no-failure case



Outline

ShareBackup
[HotNets'17, SIGCOMM'18]

Failure Recovery

Flat-tree
[HotNets'16, SIGCOMM'17]

Service Provisioning

OmniSwitch
[HotCloud'15]

Wiring & Maintenance

Lighthouse
(In submission)

Physical-Layer Programmability in WAN

Outline

ShareBackup
[HotNets'17, SIGCOMM'18]

Failure Recovery

Flat-tree
[HotNets'16, SIGCOMM'17]

Service Provisioning

OmniSwitch
[HotCloud'15]

Wiring & Maintenance

Lighthouse
(In submission)

Physical-Layer Programmability in WAN

Different Topologies Needed

- Public cloud
  - *- VM clusters have different traffic characteristics*
  - *- Cloud providers should meet SLA*

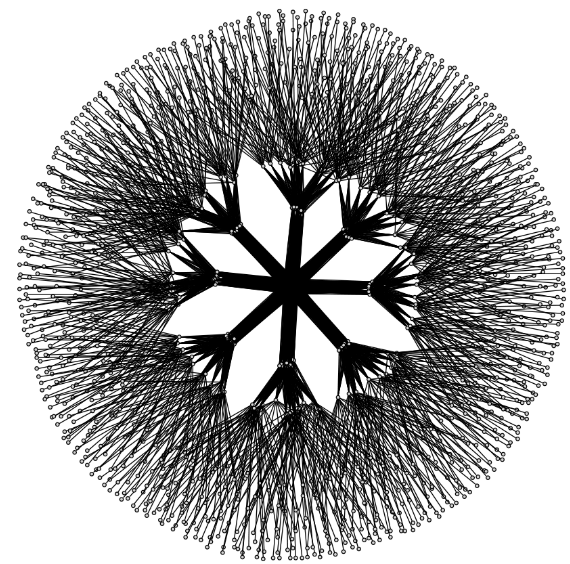
Different Topologies Needed

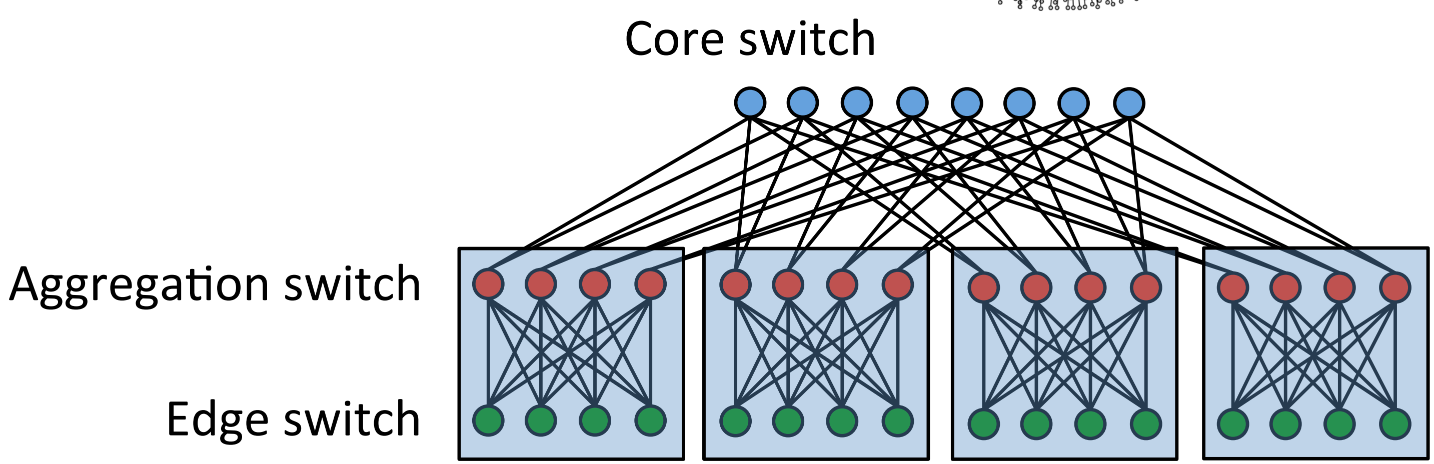
- Public cloud
  - *VM clusters have different traffic characteristics*
  - *Cloud providers should meet SLA*
- Private cloud
  - *Sub-systems of the service create different clustering features*
  - *Content providers should ensure service availability*

Different Topologies Needed

- Public cloud
  - *VM clusters have different traffic characteristics*
  - *Cloud providers should meet SLA*
- Private cloud
  - *Sub-systems of the service create different clustering features*
  - *Content providers should ensure service availability*
- Network vulnerable during service provisioning
  - *Utilization increases*

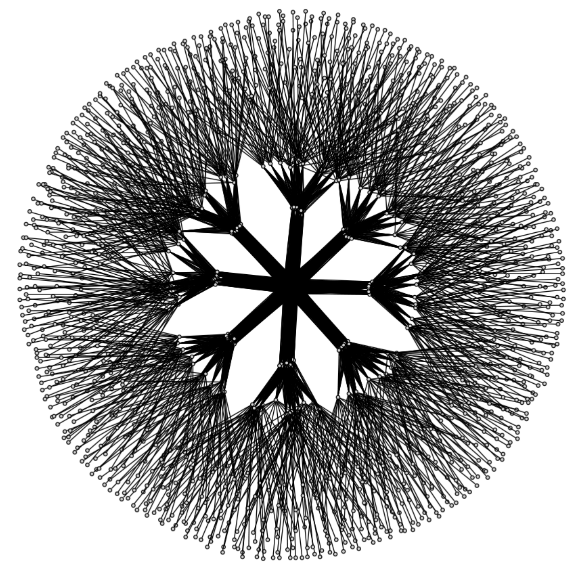
Clos Topology

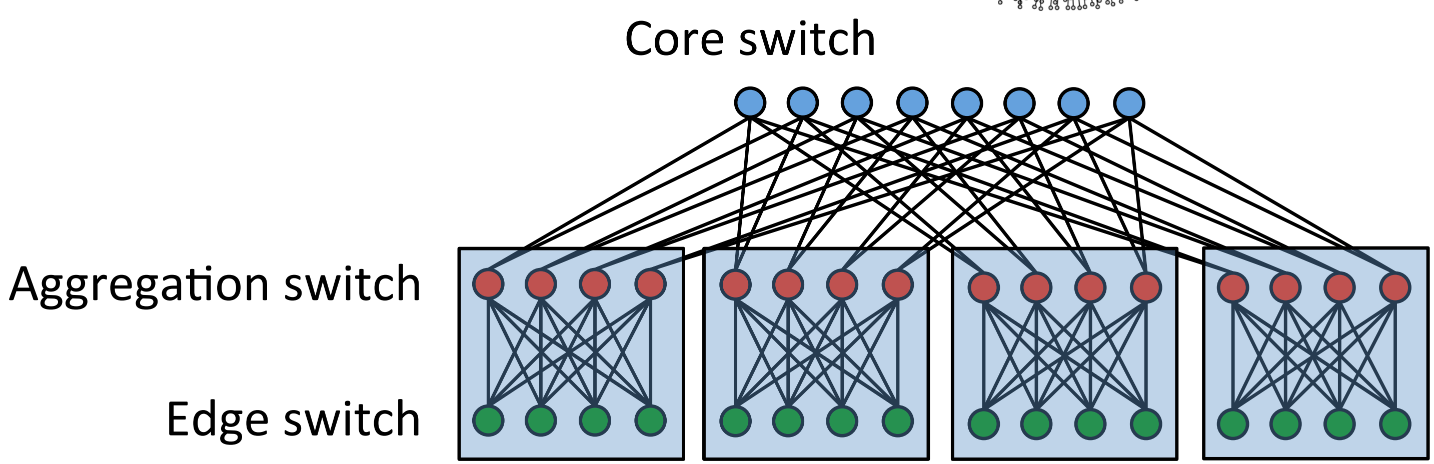




Clos Topology

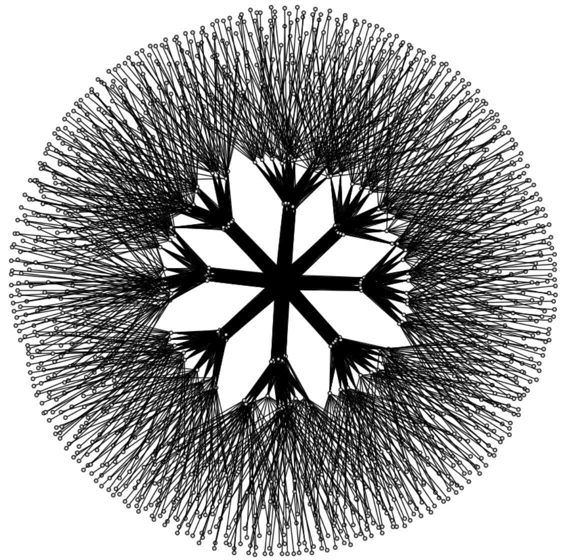
- Implementation friendly
  - *Central wiring*
  - *Flexible scale and oversubscription*
  - *Pod modular design*

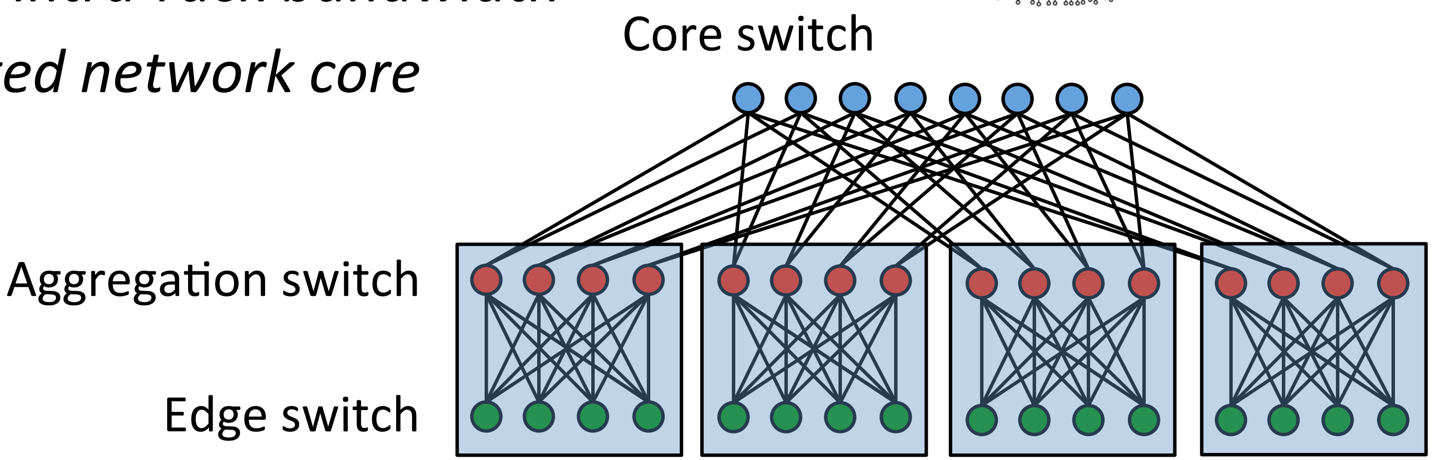




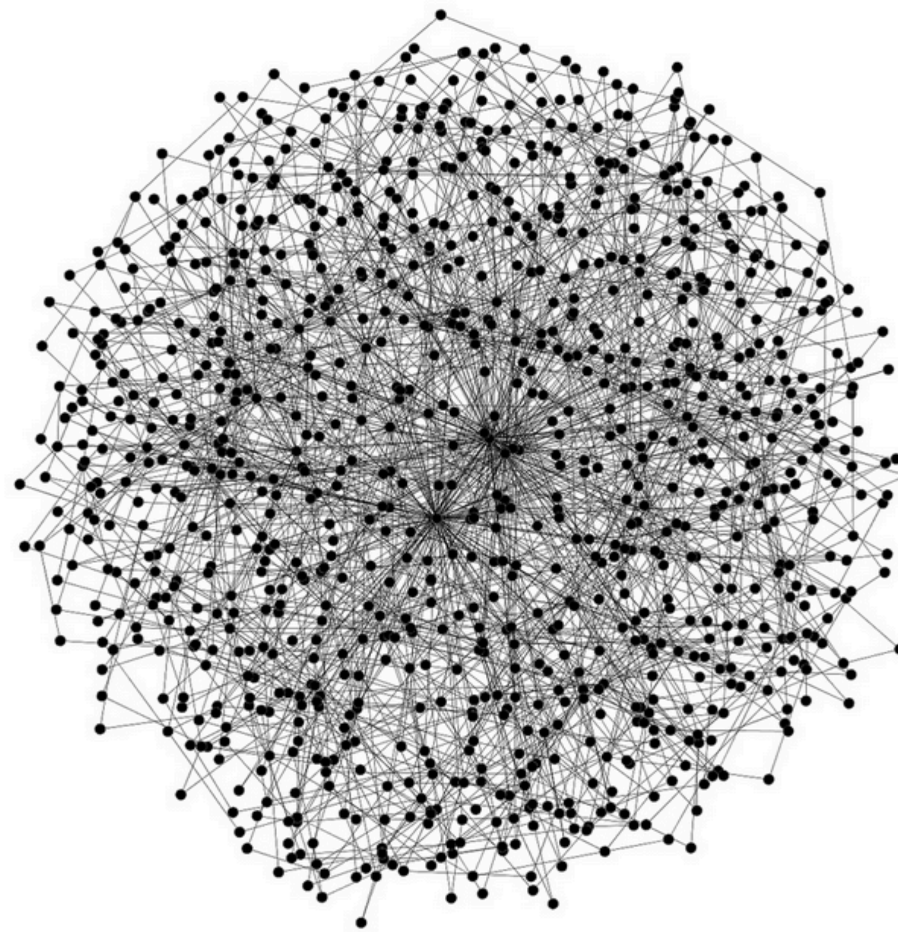
Clos Topology

- Implementation friendly
  - *Central wiring*
  - *Flexible scale and oversubscription*
  - *Pod modular design*
- Good rack-level performance
  - *Affluent intra-rack bandwidth*
  - *Congested network core*





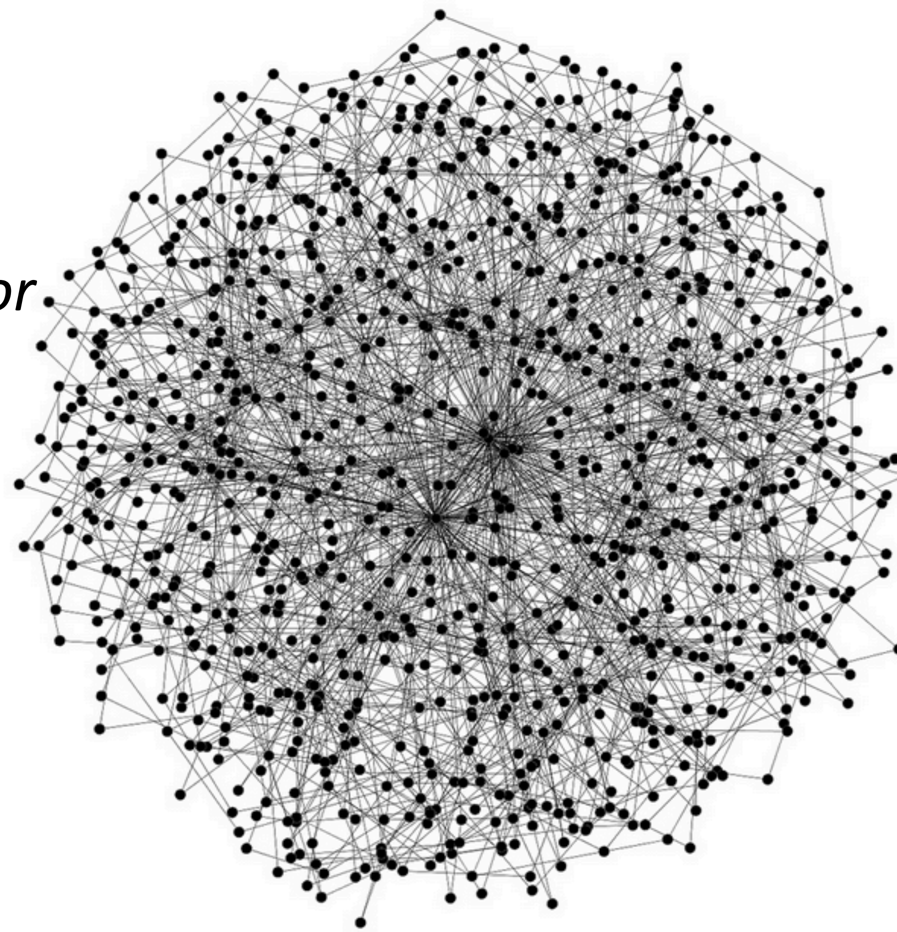
Random Graph



[*Jellyfish NSDI'12*]

Random Graph

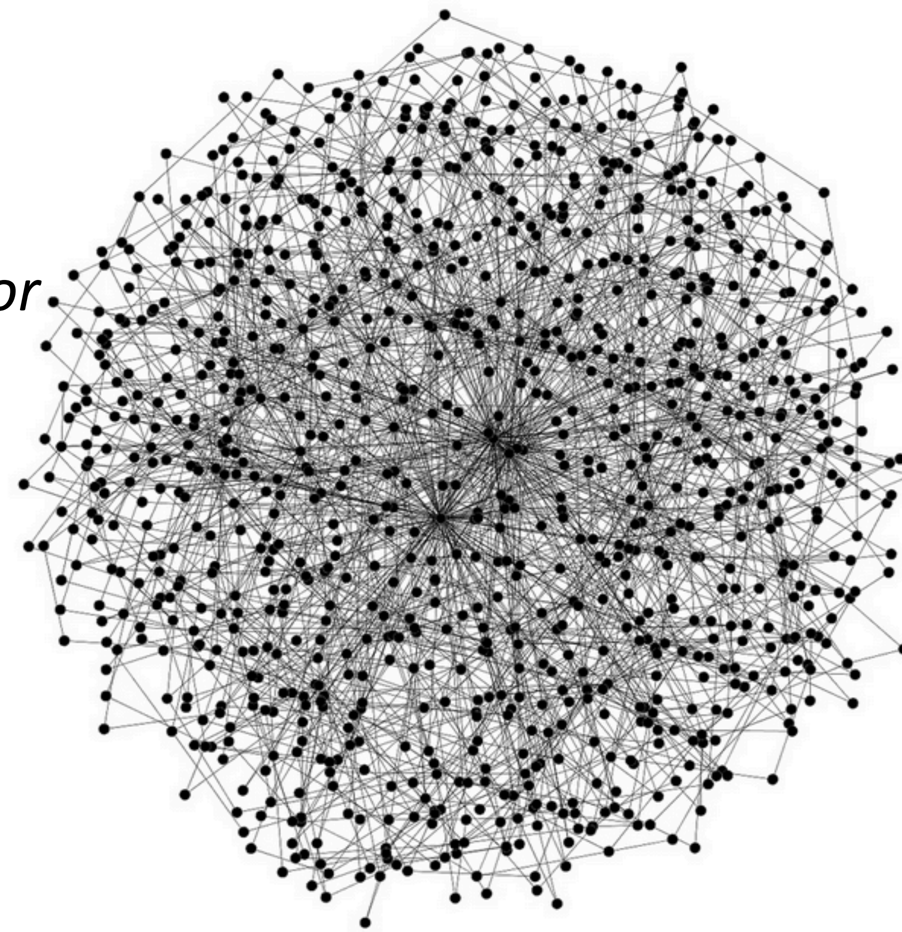
- Good connectivity
  - *Low average path length*
  - *Rich bandwidth*
  - *Near optimal throughput for uniform traffic*



[*Jellyfish NSDI'12*]

Random Graph

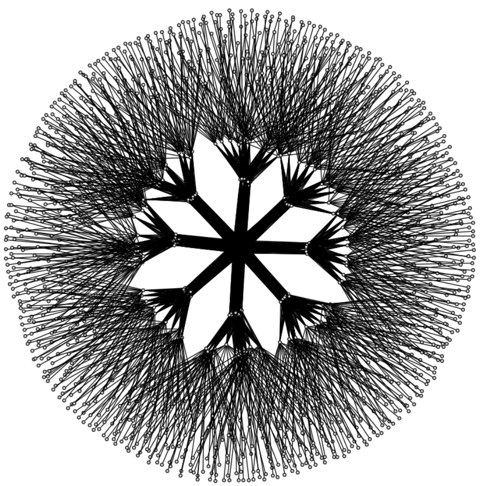
- Good connectivity
  - *Low average path length*
  - *Rich bandwidth*
  - *Near optimal throughput for uniform traffic*
- Hard to implement
  - *Neighbor-to-neighbor wiring complicated*



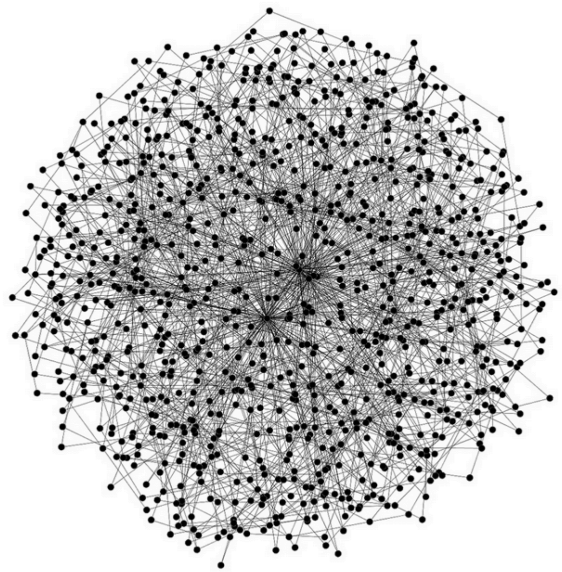
[*Jellyfish NSDI'12*]

Flat-tree

Tree Network



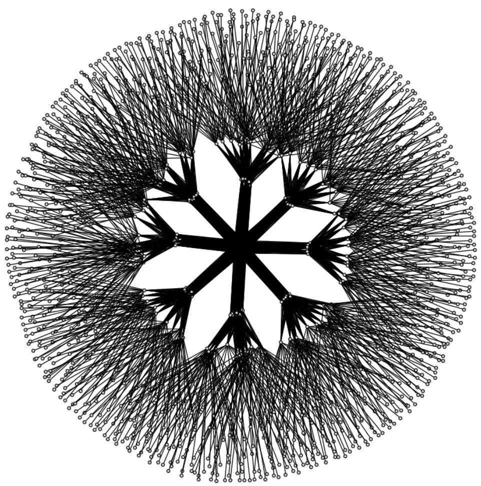
vs.



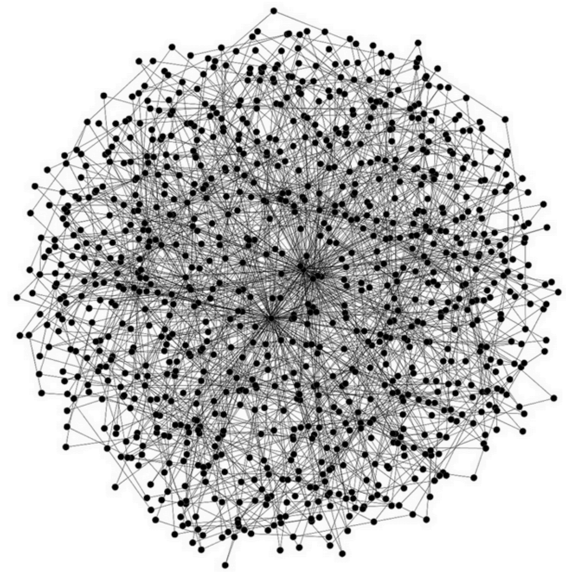
Flat Network

Flat-tree

Tree Network



vs.



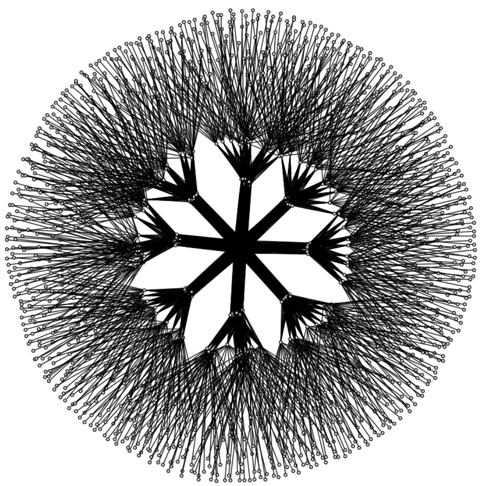
Flat Network

Easy implementation

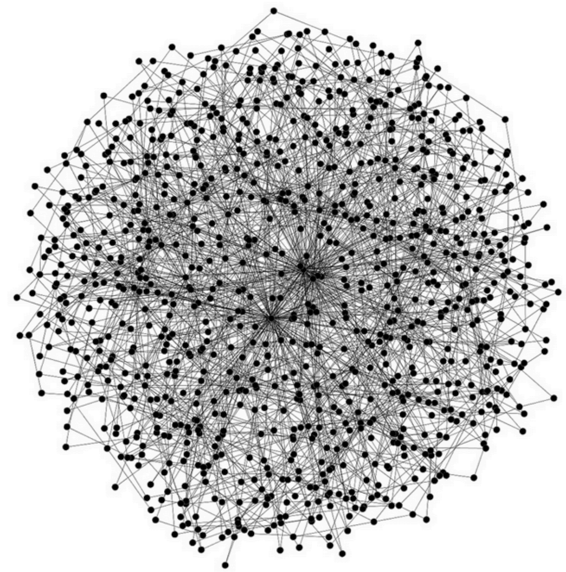
Good connectivity

Flat-tree

Tree Network



vs.



Flat Network

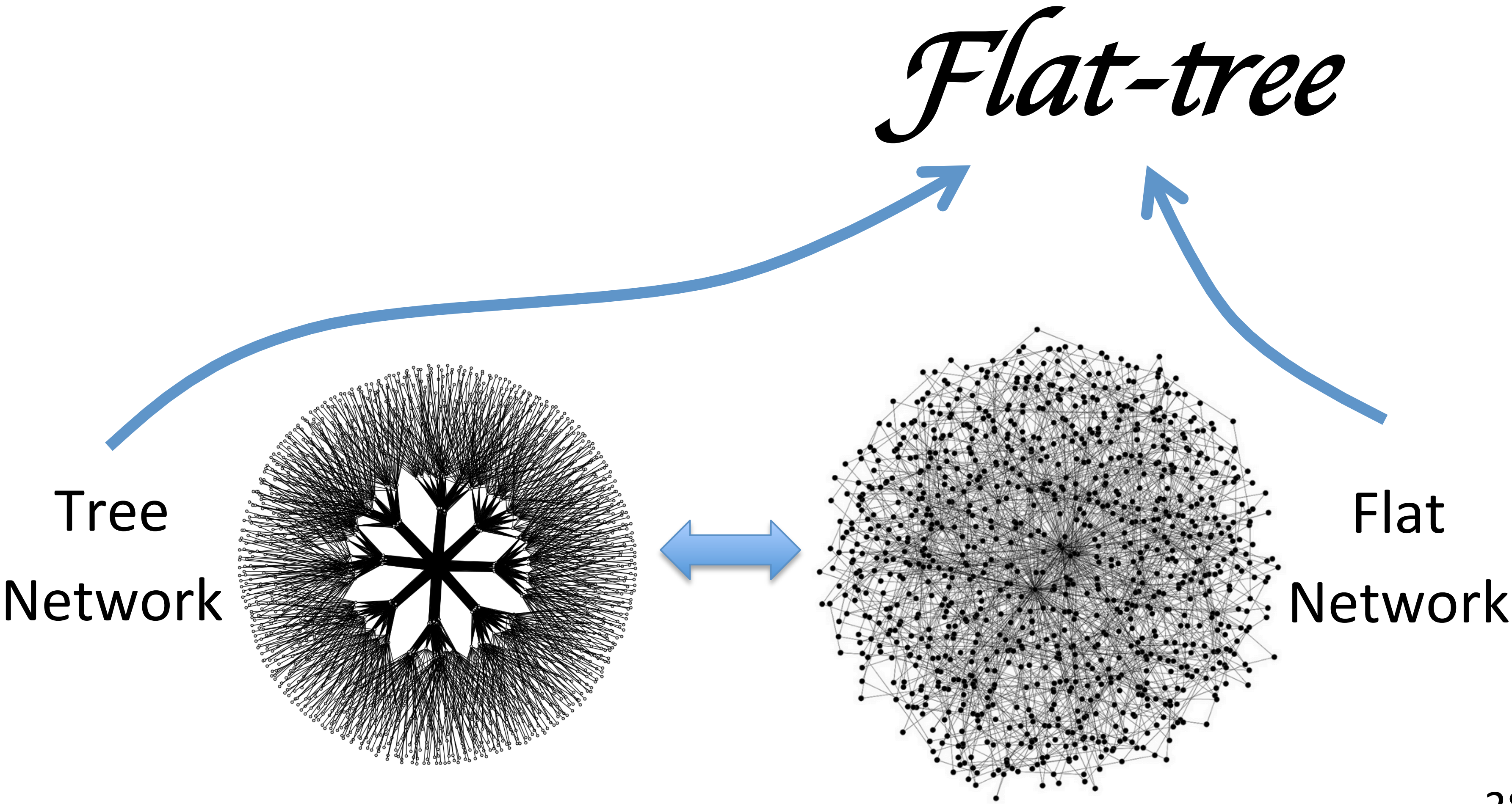
Easy implementation

Clustered traffic

Good connectivity

Uniform traffic

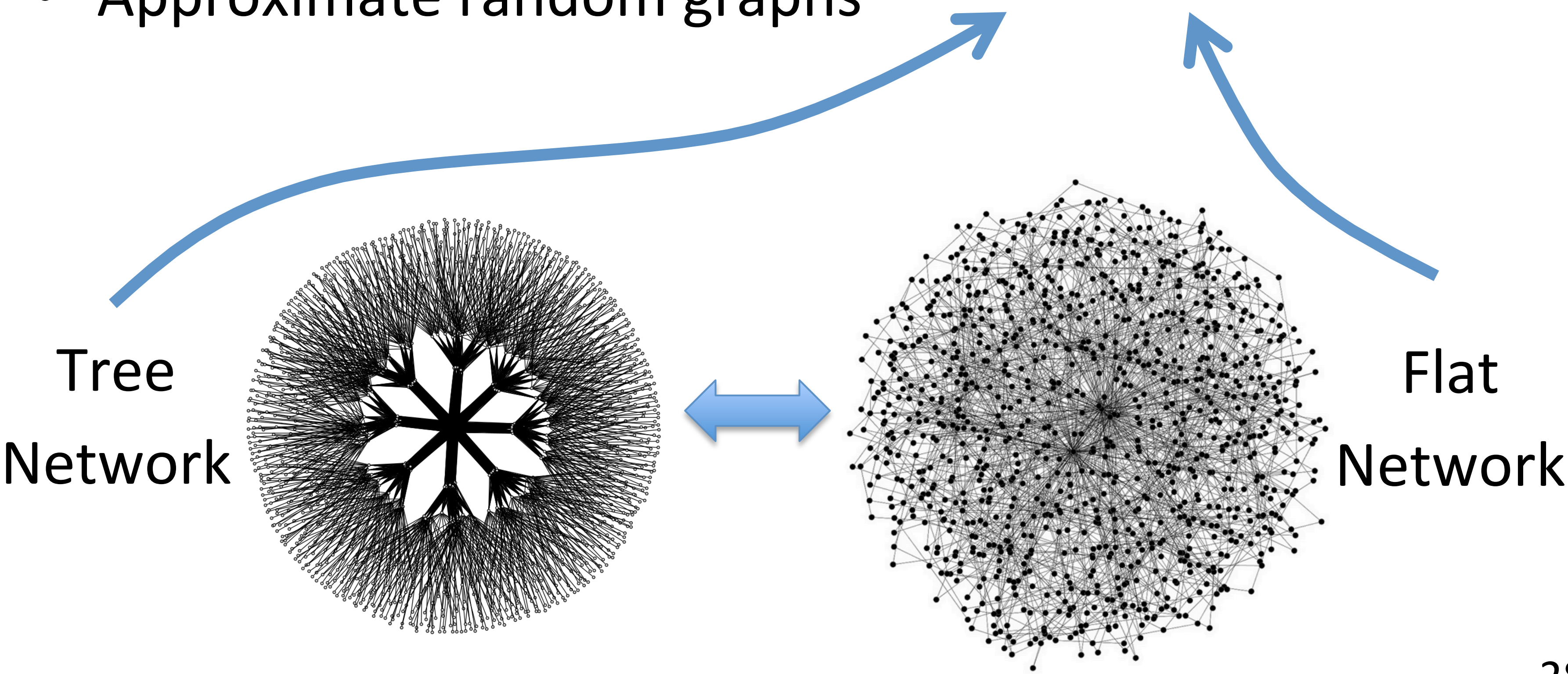
Flat-tree



Flat-tree

- Start from Clos
- Flatten tree structure
- Approximate random graphs

Flat-tree



Flatten the Tree

- How to flatten the tree structure?

Flatten the Tree

- How to flatten the tree structure?

Difference	Clos	Random graph	Solution
Server distribution	Edge switches	All switches	Relocate servers
Wiring	Central	Neighbor-to-neighbor	Diversify connections

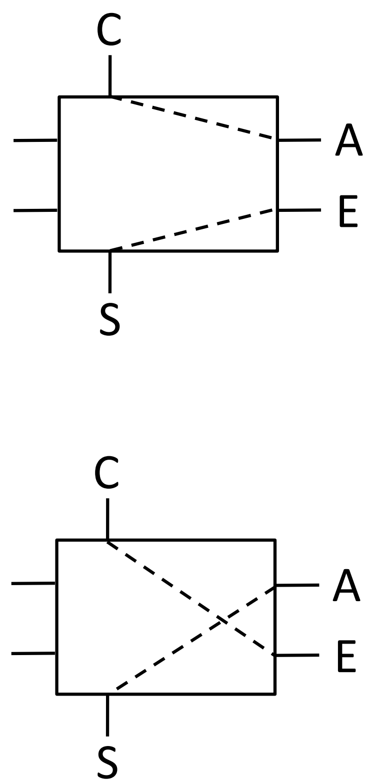
Circuit Switch Configurations

C: core switch

A: aggregation switch

E: edge switch

S: server



6-port Circuit Switch

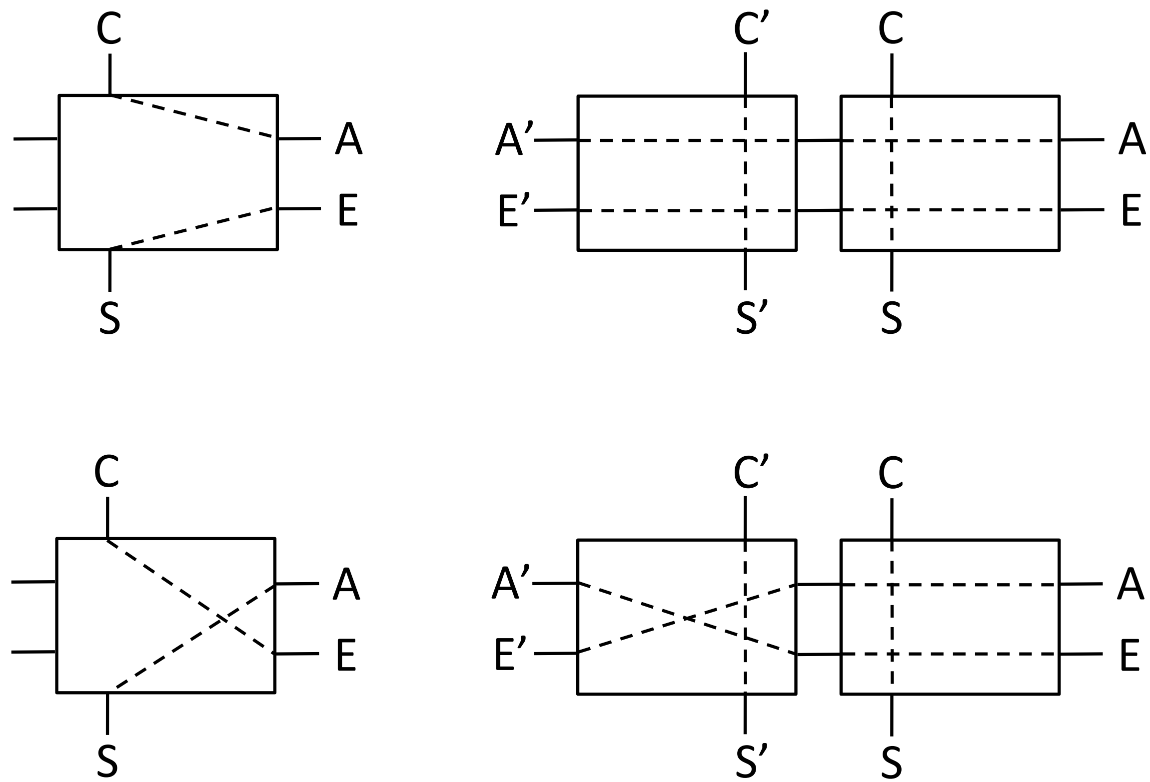
Circuit Switch Configurations

C: core switch

A: aggregation switch

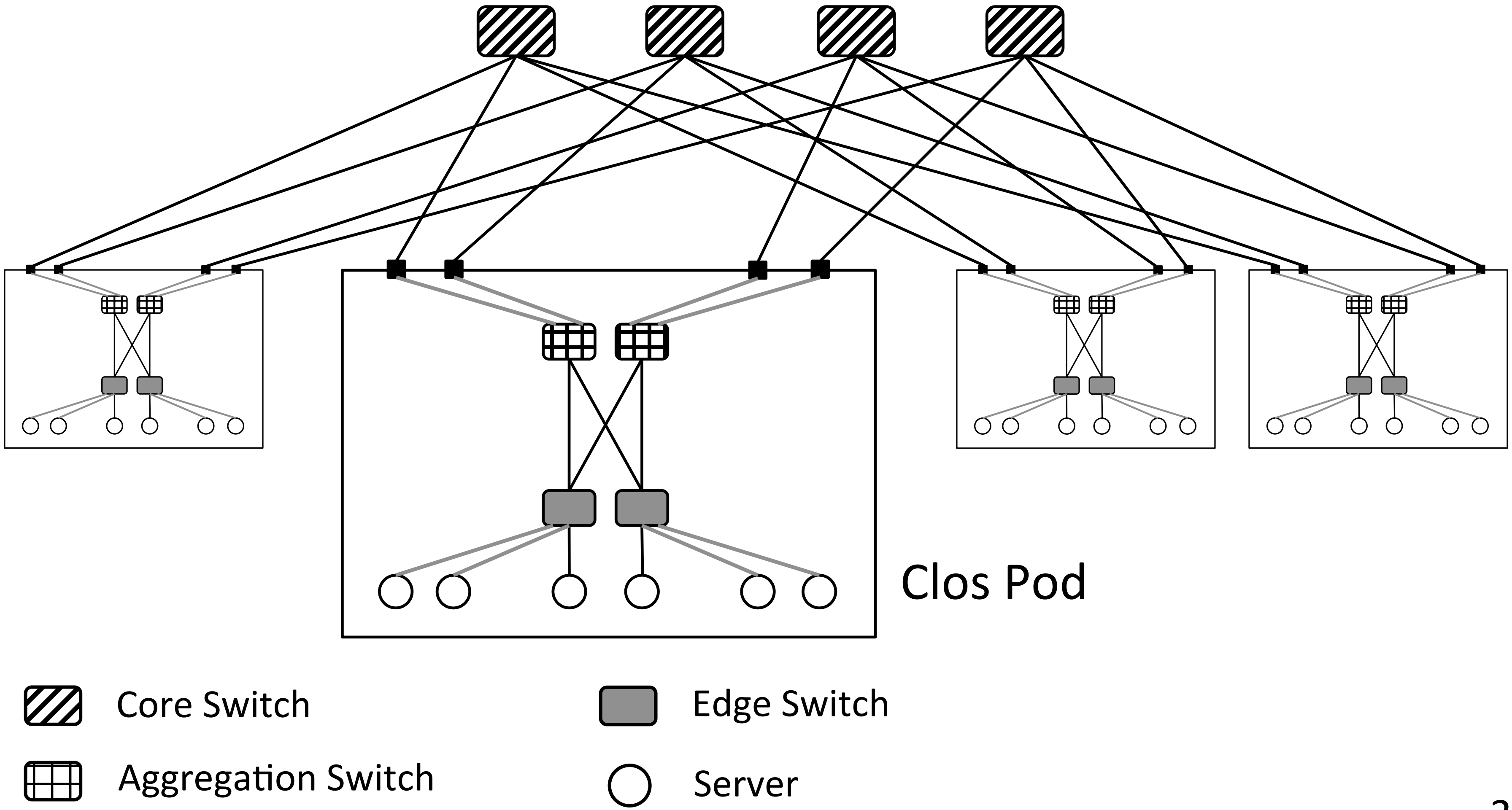
E: edge switch

S: server

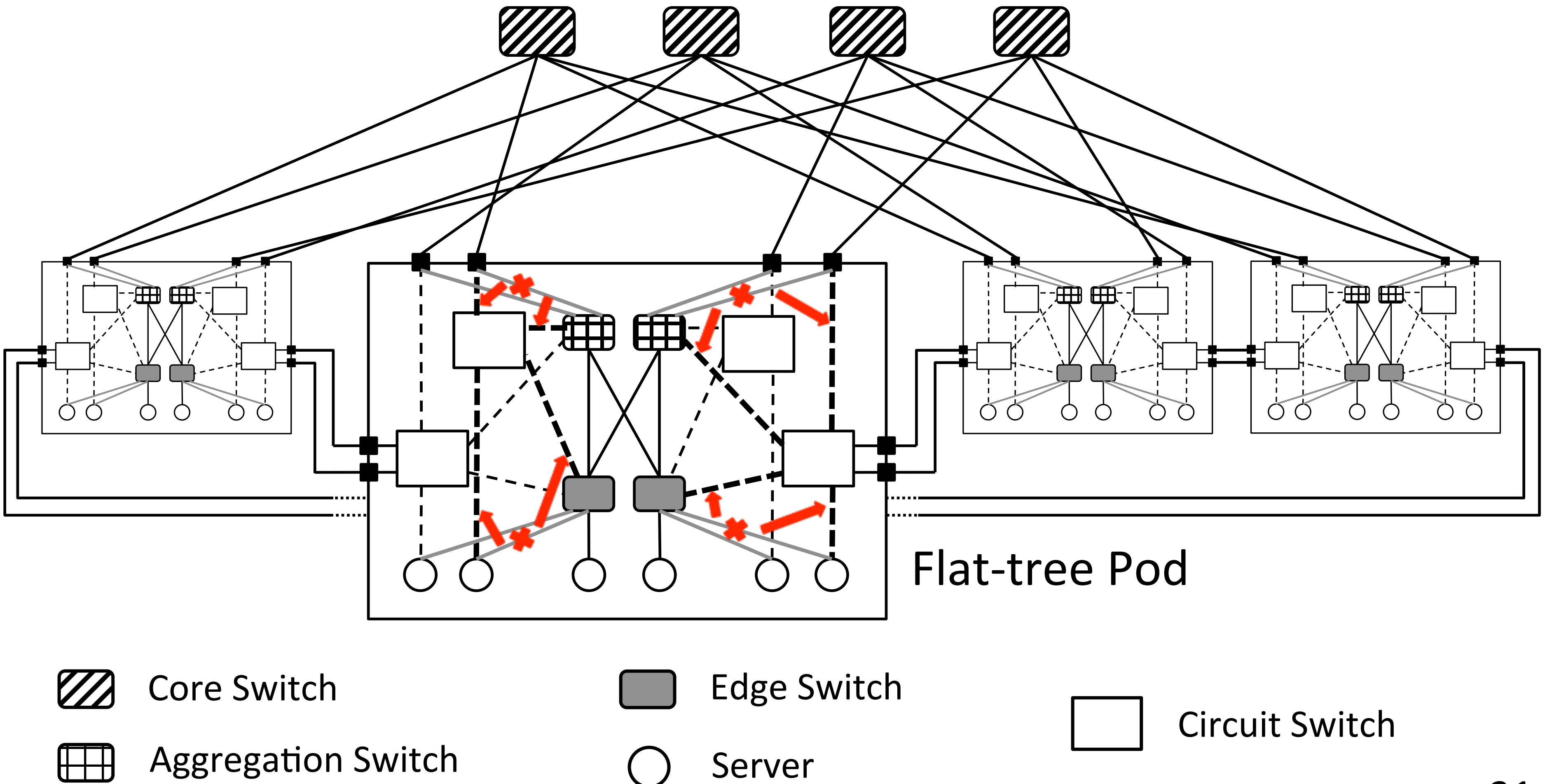


6-port Circuit Switch

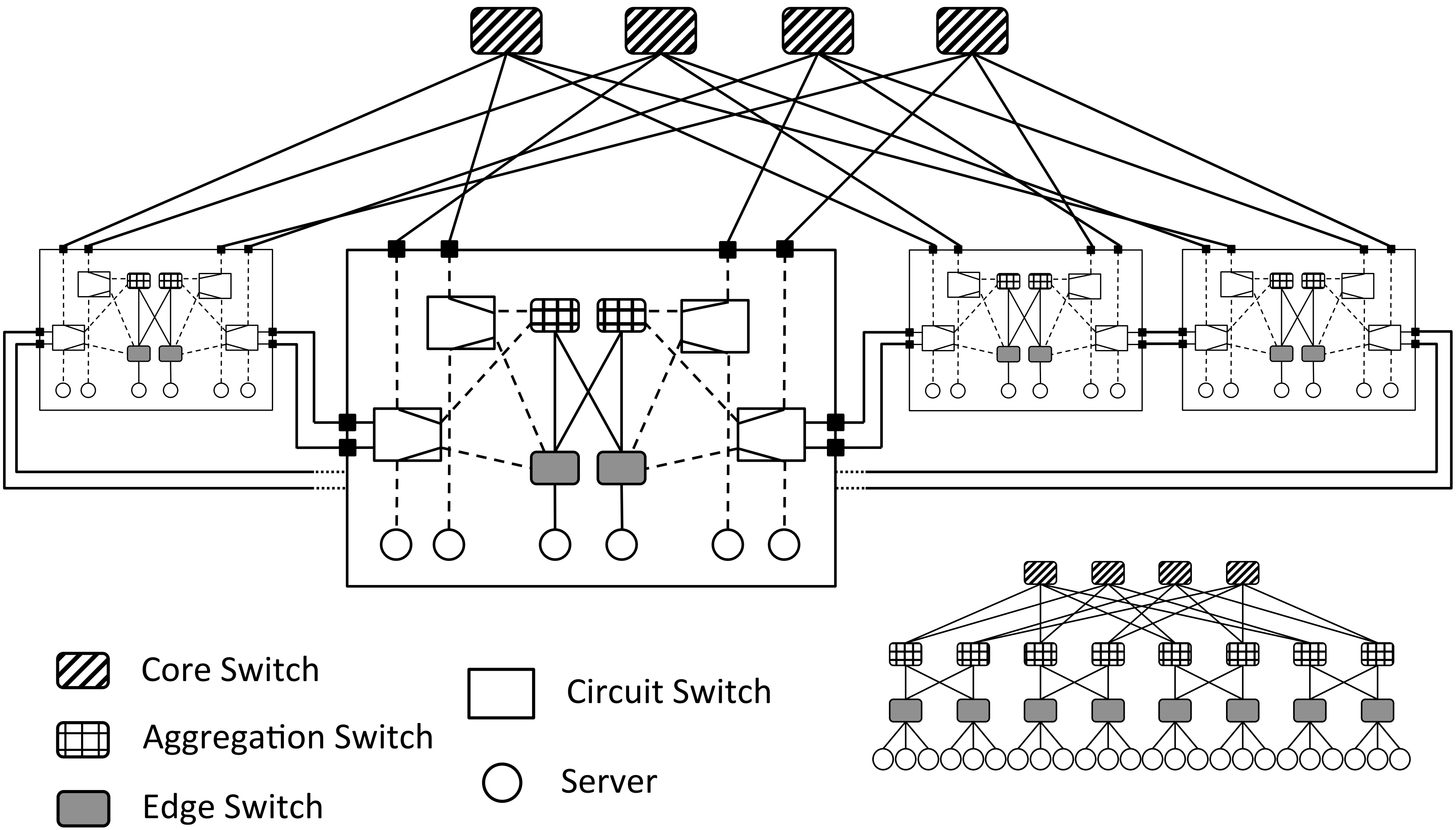
Flat-tree Example



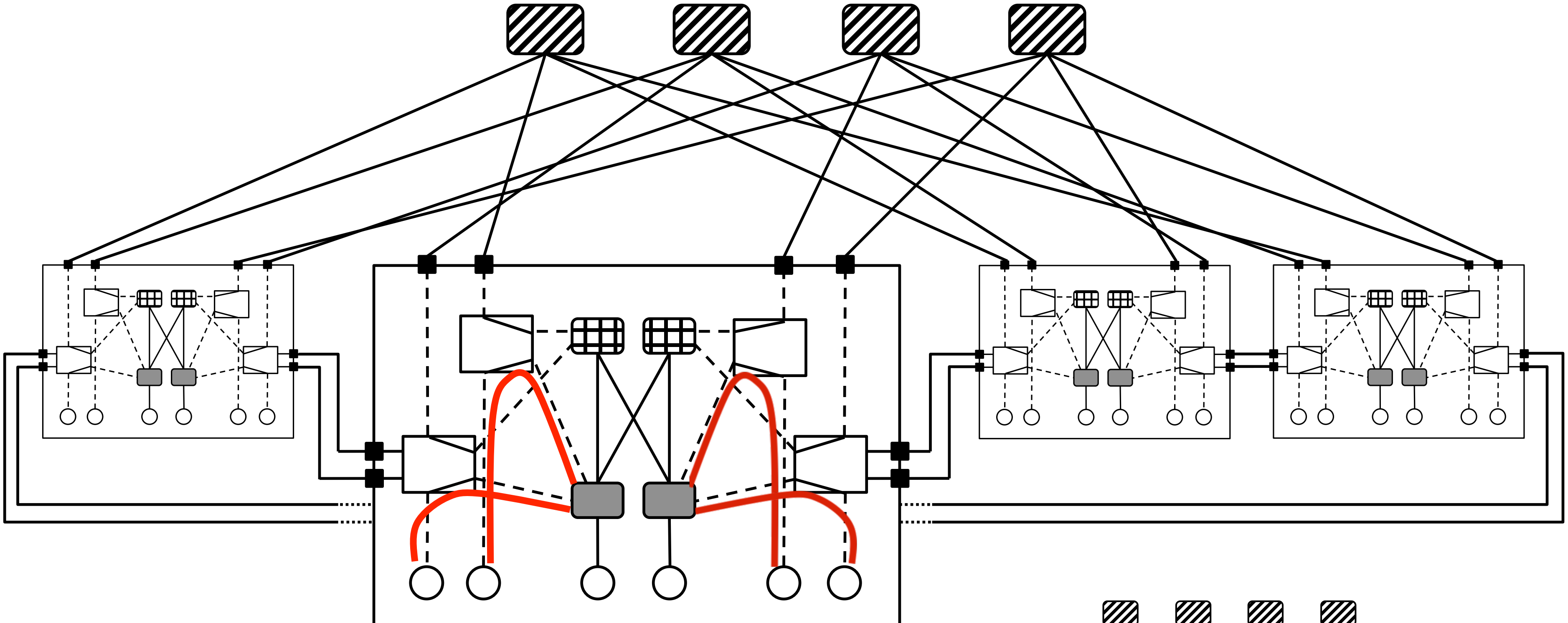
Flat-tree Example

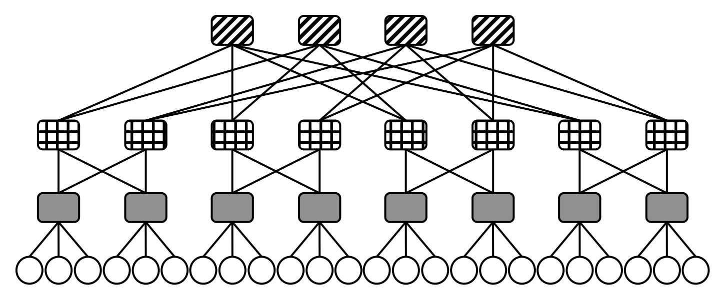


Clos Network



Clos Network





Core Switch

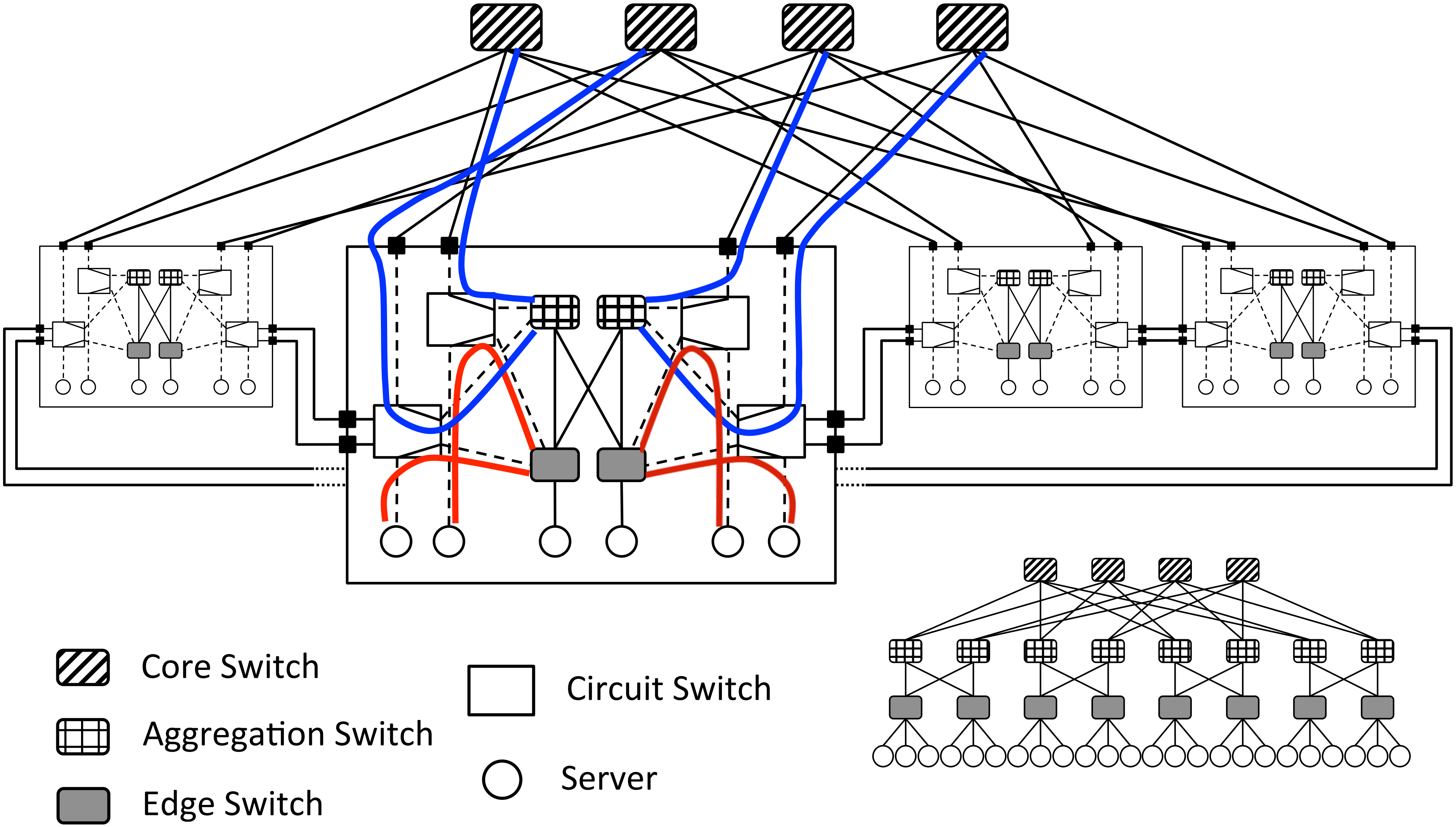
Aggregation Switch

Edge Switch

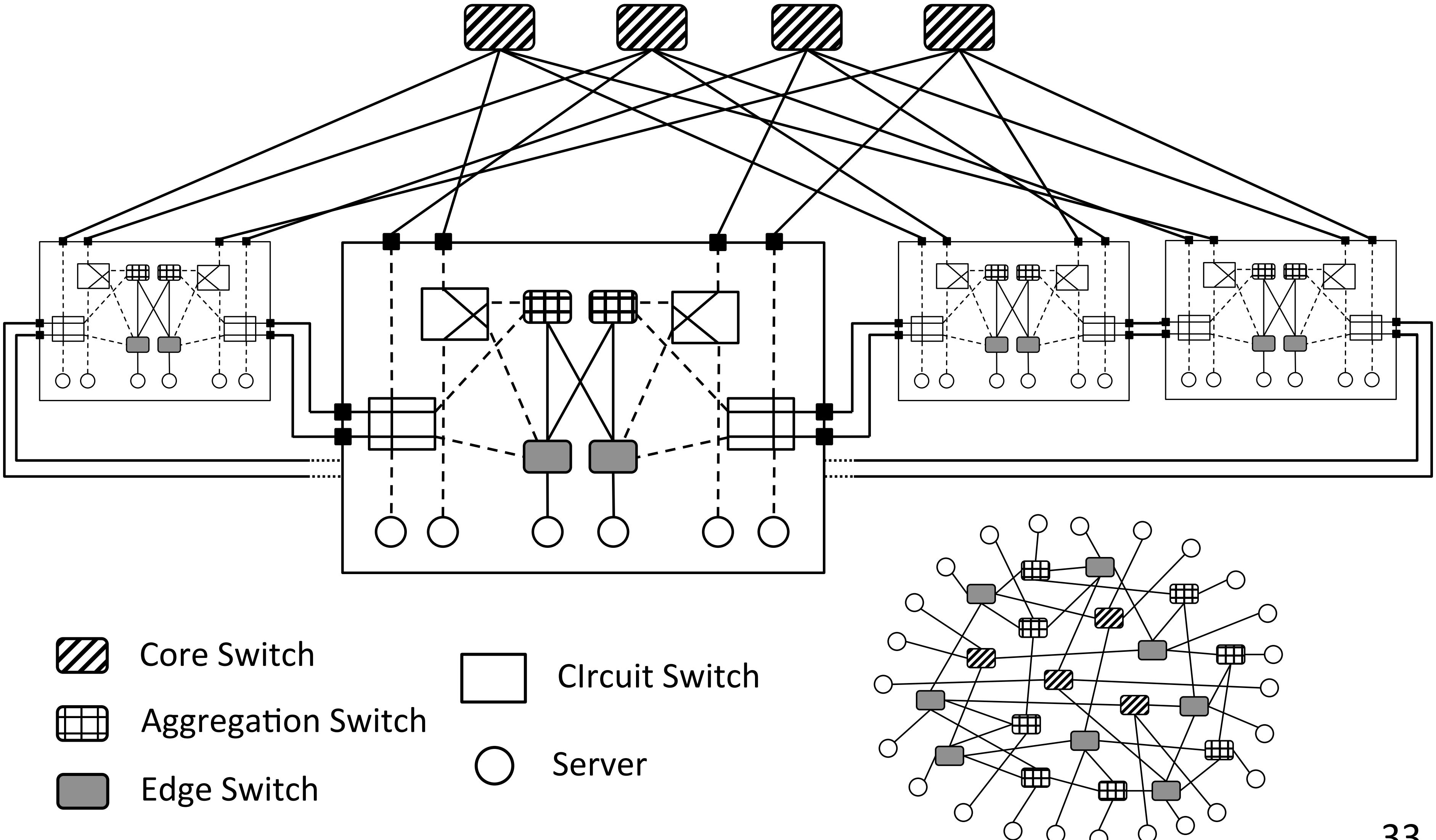
Circuit Switch

Server

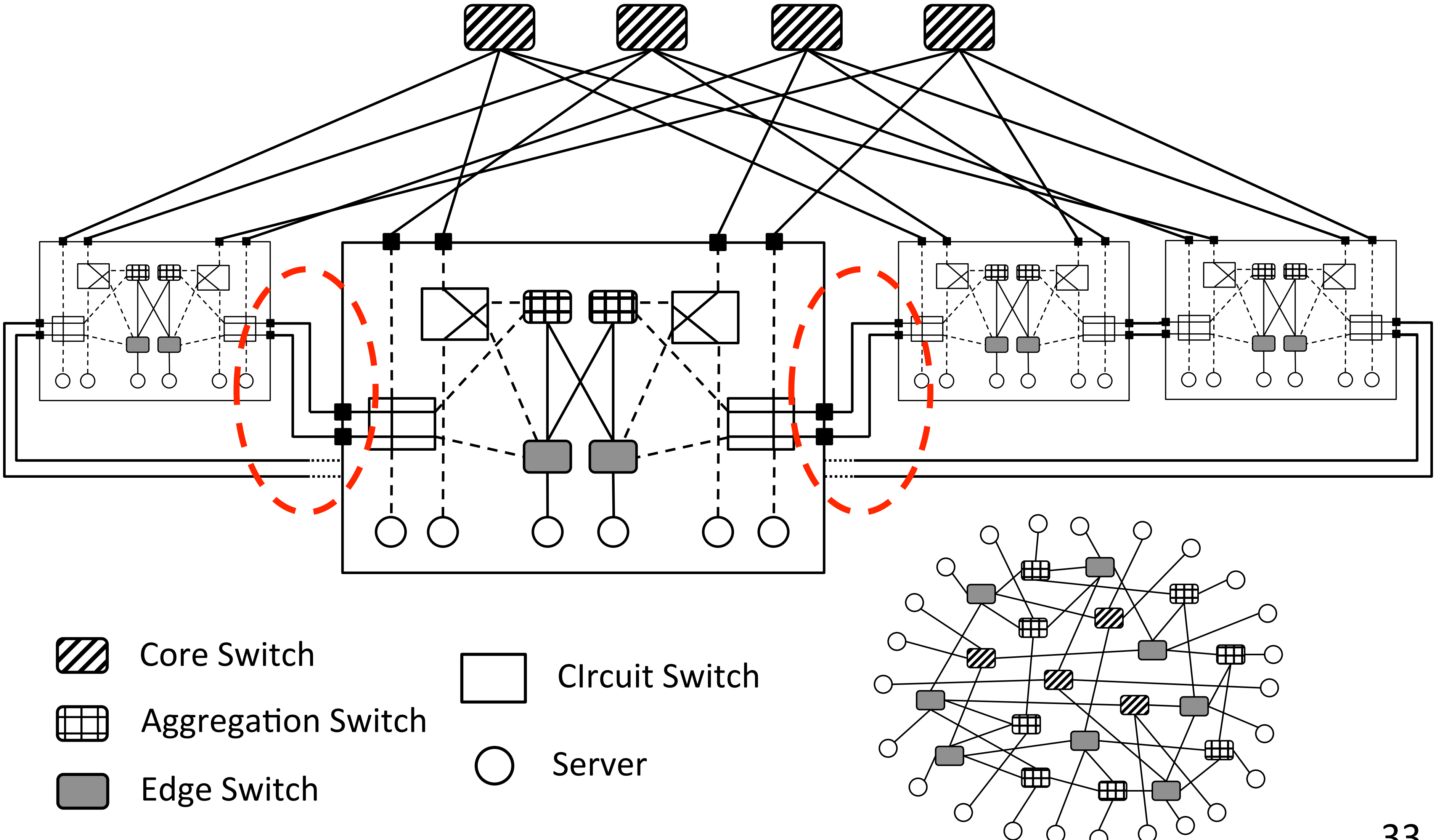
Clos Network



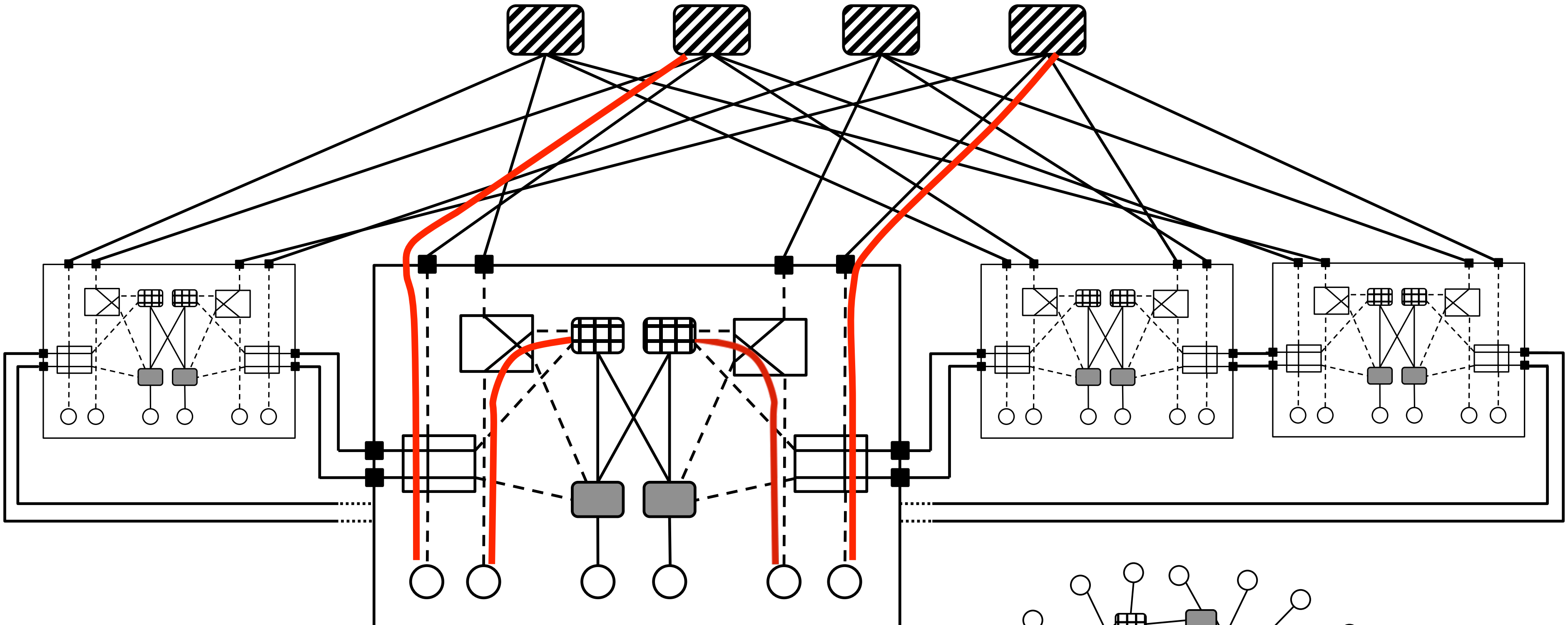
Approximate Random Graph



Approximate Random Graph



Approximate Random Graph



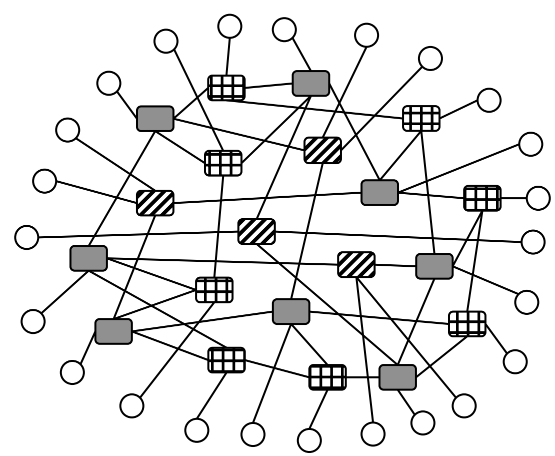
Core Switch

Aggregation Switch

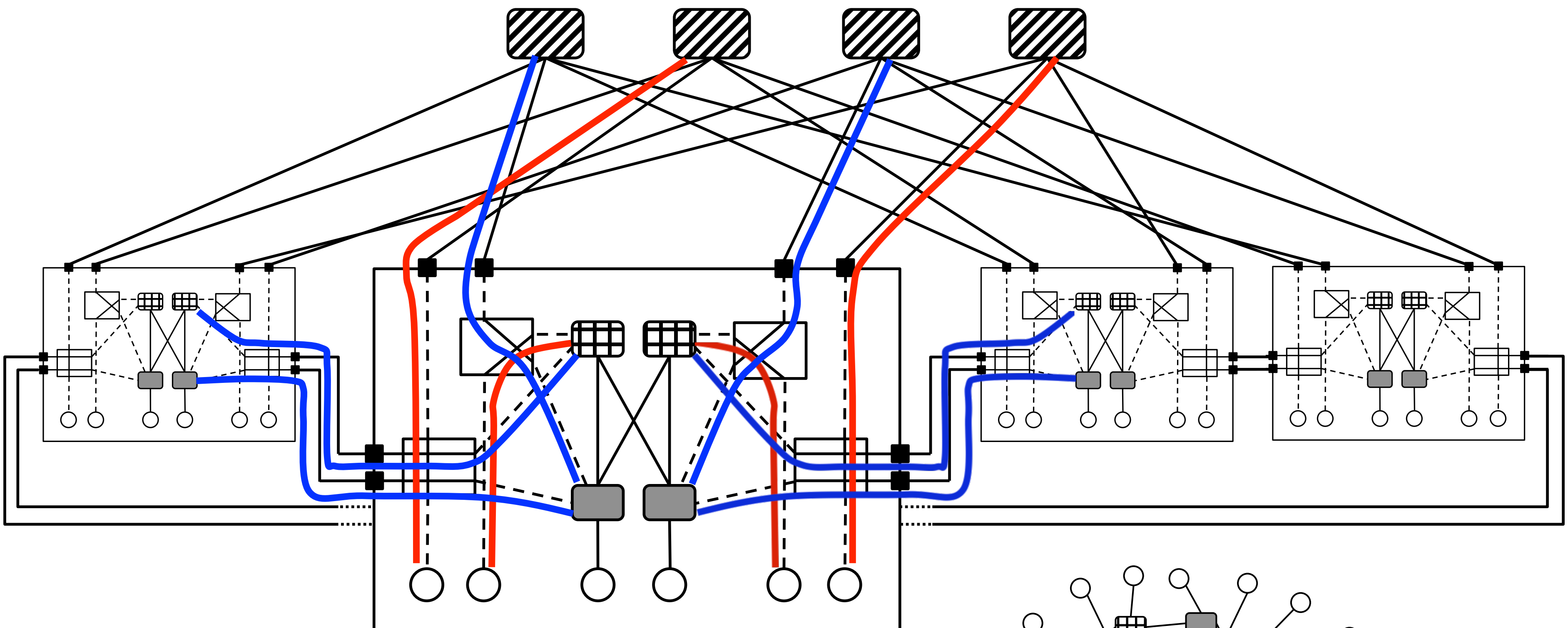
Edge Switch

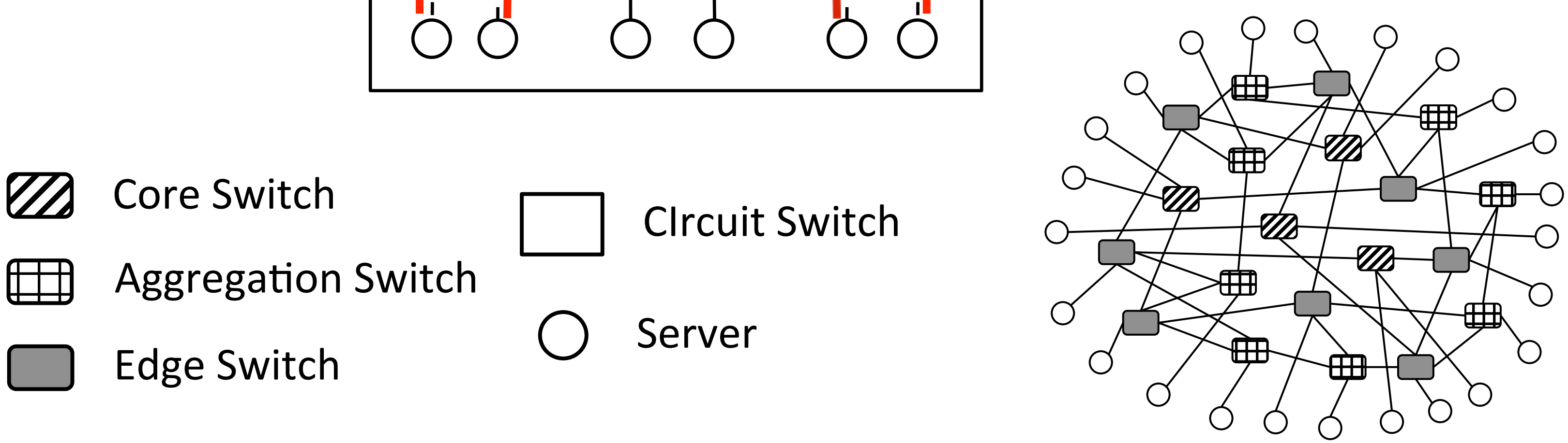
CIrcuit Switch

Server

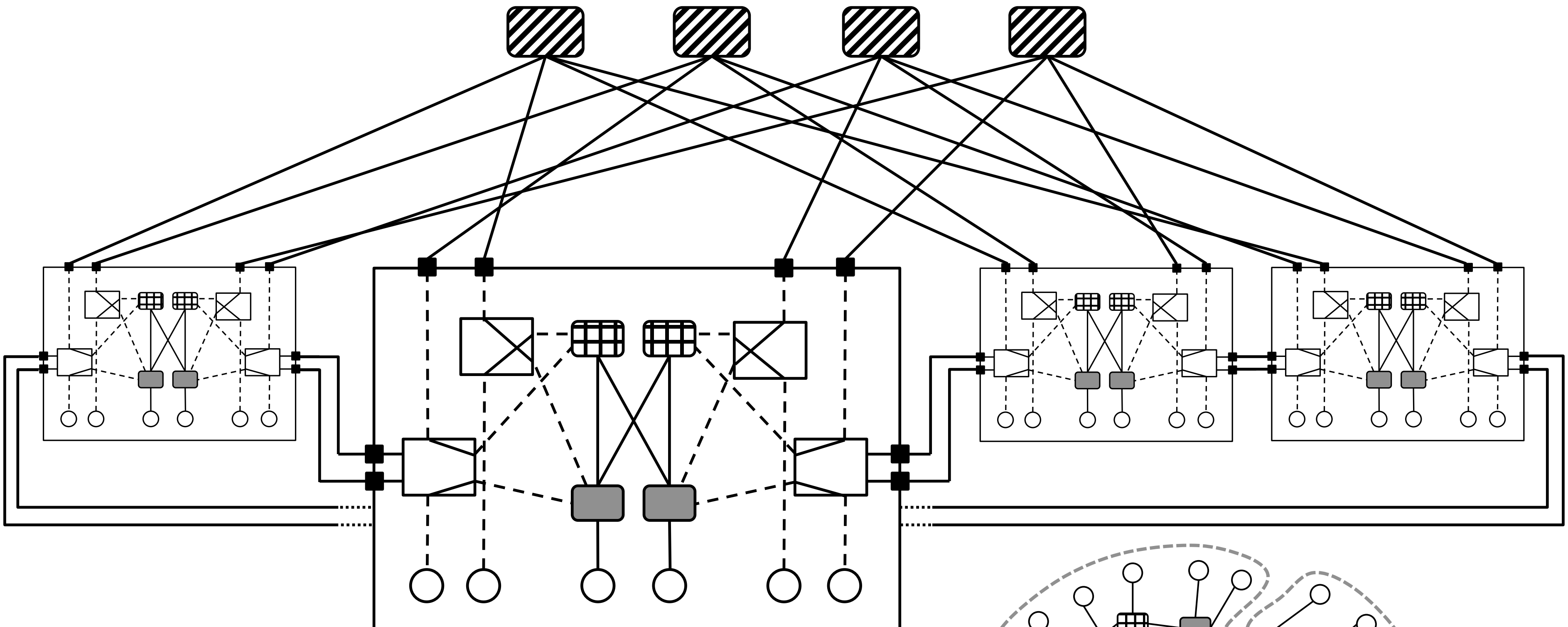


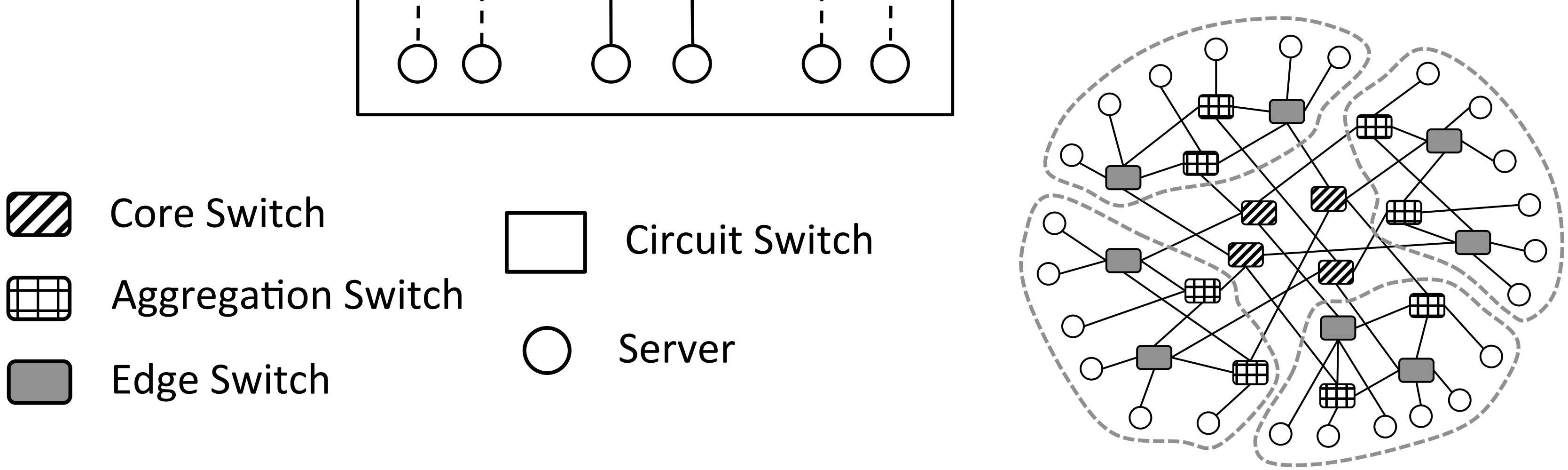
Approximate Random Graph



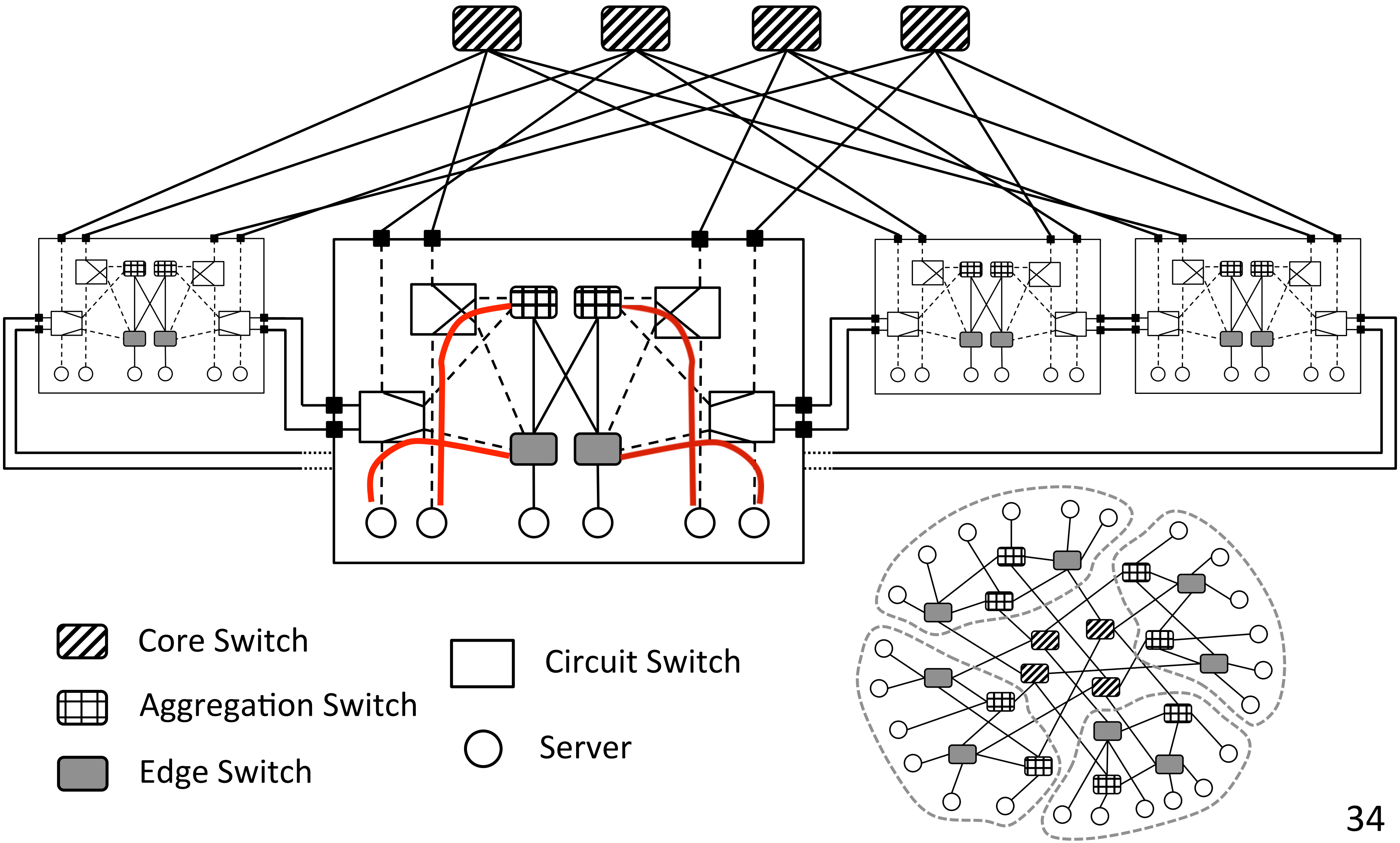


Approximate Local Random Graph

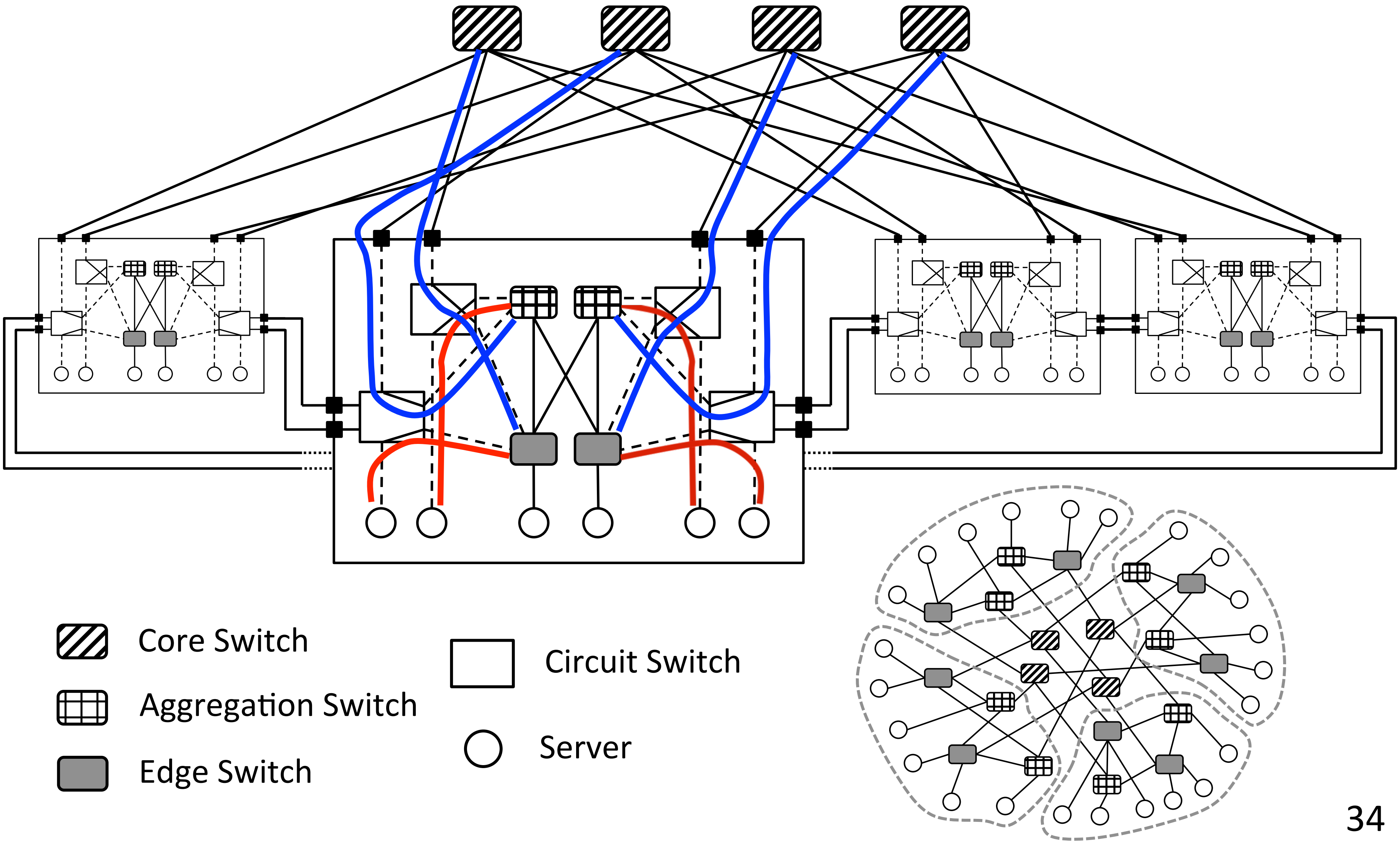




Approximate Local Random Graph



Approximate Local Random Graph



Routing Challenges

- Server mobility
  - *Server relocated to different switches*
  - *Assign IP addresses to support prefix matching*

Routing Challenges

- Server mobility
  - *Server relocated to different switches*
  - *Assign IP addresses to support prefix matching*
- Different routing schemes per topology
  - *Clos: ECMP, two-level routing, SDN*
  - *Random graph: k-shortest-path routing + MPTCP*

Routing Challenges

- Server mobility
  - *Server relocated to different switches*
  - *Assign IP addresses to support prefix matching*
- Different routing schemes per topology
  - *Clos: ECMP, two-level routing, SDN*
  - *Random graph: k-shortest-path routing + MPTCP*
- k-shortest-path routing
  - *k paths for every server pairs*
  - *Enormous number of states → exceed switch capacity*
  - *No solution from random graph networks*

Routing Challenges

- Server mobility
  - Customized addressing scheme
  - Different sets of IP addresses per topology
- Different routing schemes per topology

  - Clos: ECMP, two-level routing, SDN

  - Random graph: k-shortest-path routing + MPTCP
- k-shortest-path routing

  - k paths for every server pairs

  - Enormous number of states → exceed switch capacity

  - No solution from random graph networks

Routing Challenges

- Server mobility
  - Customized addressing scheme
  - Different sets of IP addresses per topology
- Different routing schemes per topology
  - k-shortest-path routing for all topologies
  - ECMP/two-level routing/SDN encoded as k paths
- k-shortest-path routing

  - k paths for every server pairs

  - Enormous number of states → exceed switch capacity

  - No solution from random graph networks

Routing Challenges

- Server mobility
  - Customized addressing scheme
  - Different sets of IP addresses per topology
- Different routing schemes per topology
  - k-shortest-path routing for all topologies
  - ECMP/two-level routing/SDN encoded as k paths
- k-shortest-path routing
  - Addressing: server-level → switch-level k paths
  - Source routing: further reduce network states

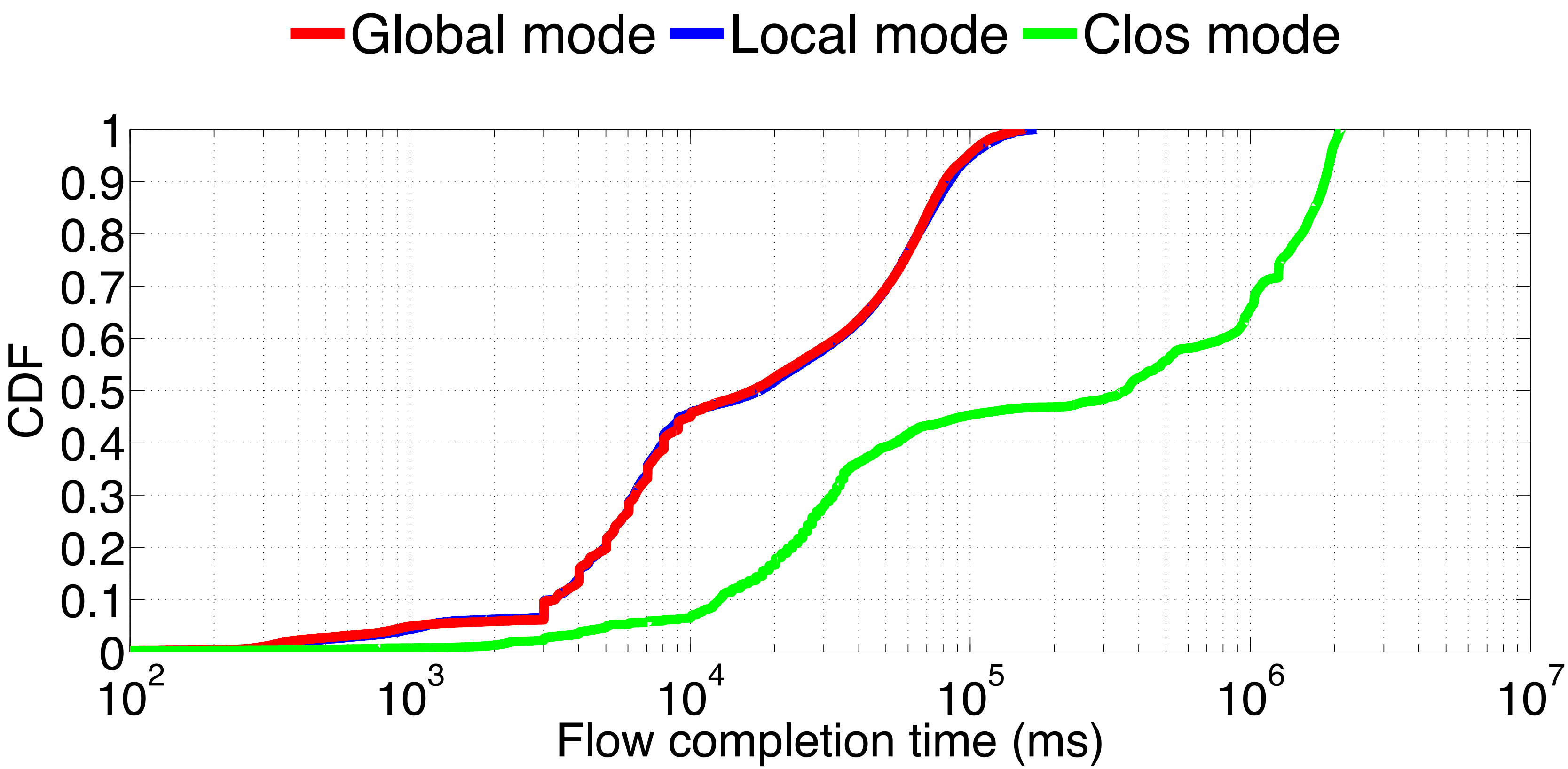
- No solution from random graph networks

Transmission Performance

- Packet-level simulation
- Traffic traces from 4 Facebook data centers
  - *Hadoop-1: no locality*
  - *Hadoop-2 : rack-level locality*
  - *Web: Pod-level locality*
  - *Cache: Pod-level locality*

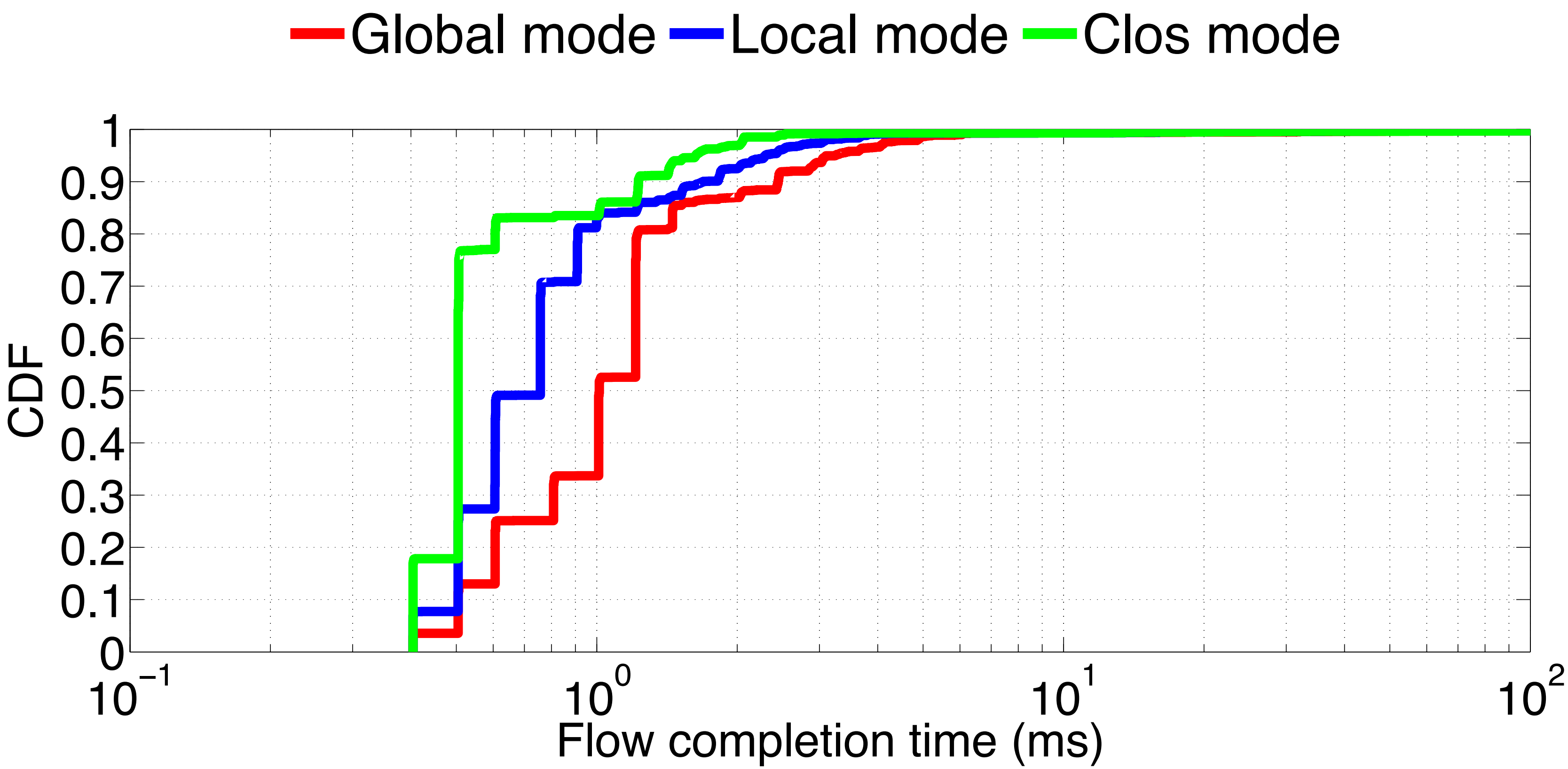
Network-wide Traffic

- Hadoop-1: no locality



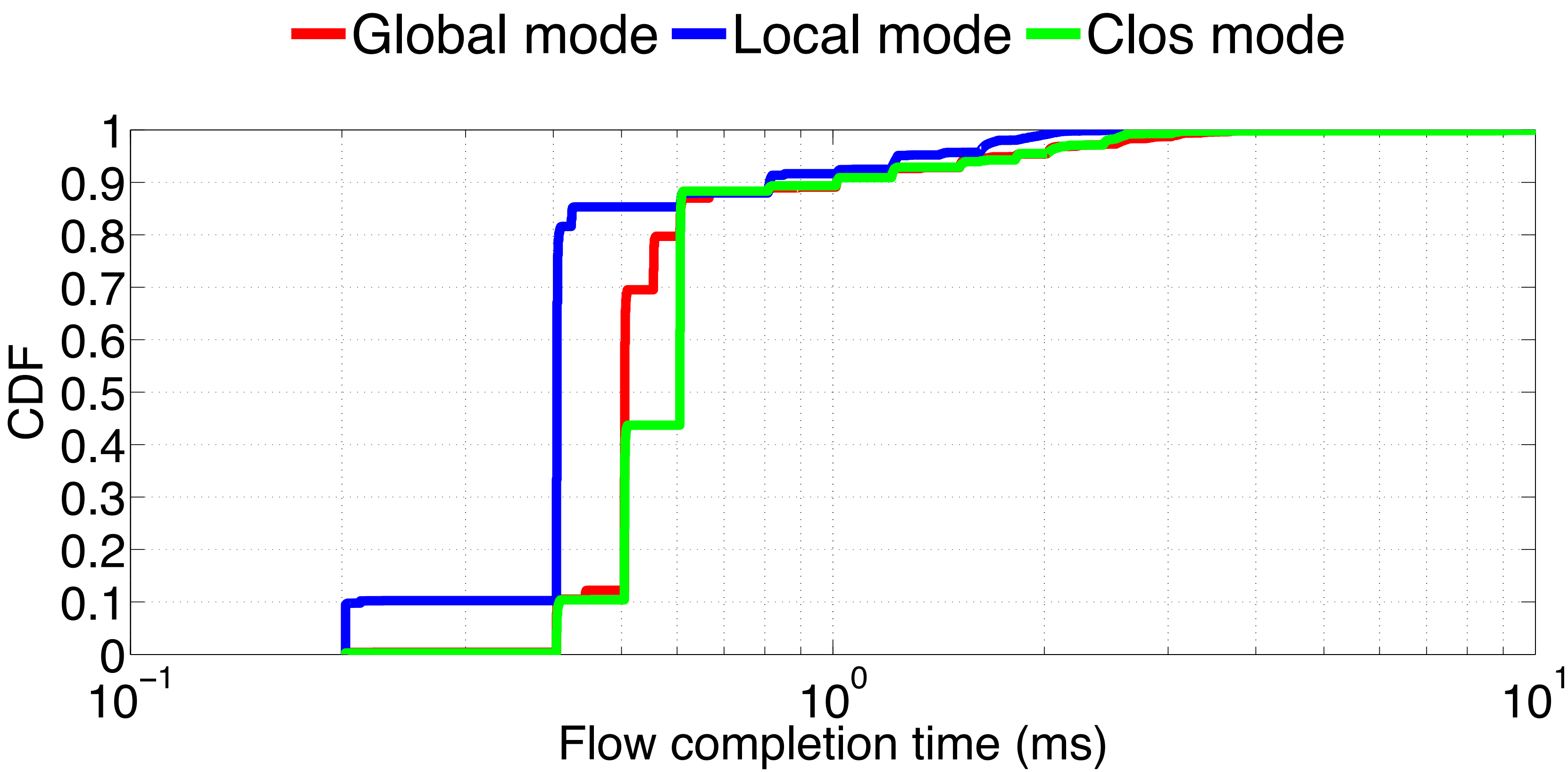
Rack-level Locality

- Hadoop-2: rack-level locality



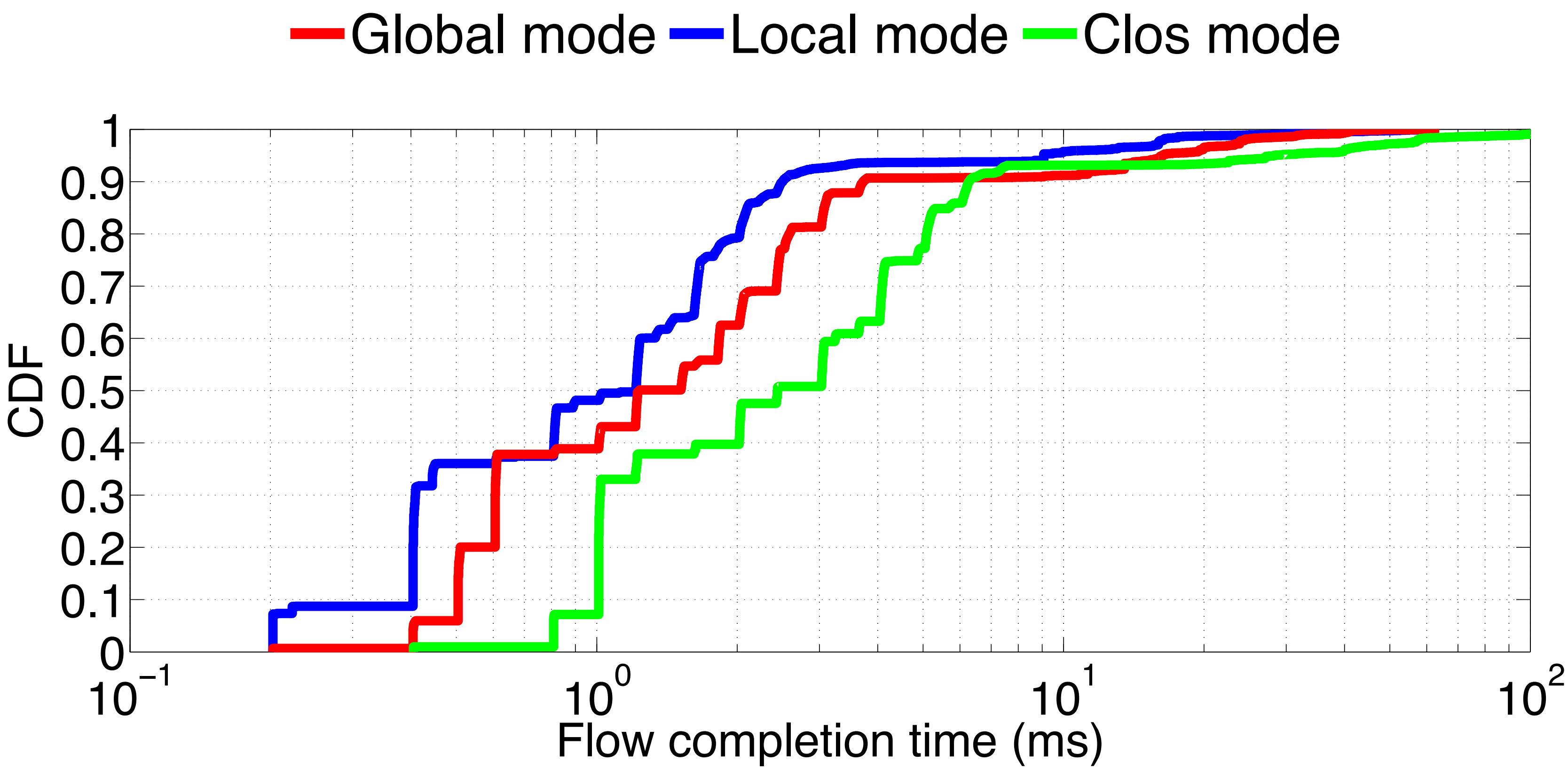
Pod-level Locality

- Web: Pod-level locality



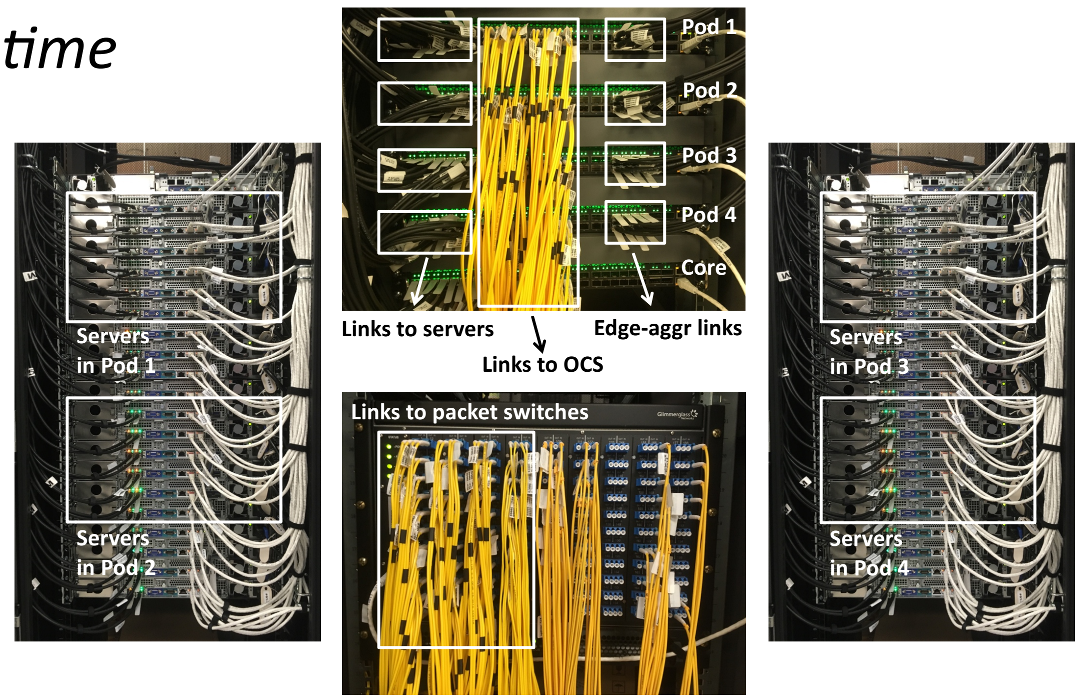
Pod-level Locality

- Cache: Pod-level locality



Testbed

- Implementation of motivating example
  - *Hadoop & Spark*
  - *27.6% more bandwidth*
  - *10% less data read time*



Outline

ShareBackup
[HotNets'17, SIGCOMM'18]

Failure Recovery

Flat-tree
[HotNets'16, SIGCOMM'17]

Service Provisioning

OmniSwitch
[HotCloud'15]

Wiring & Maintenance

Lighthouse
(In submission)

Physical-Layer Programmability in WAN

Outline

ShareBackup
[HotNets'17, SIGCOMM'18]

Failure Recovery

Flat-tree
[HotNets'16, SIGCOMM'17]

Service Provisioning

OmniSwitch
[HotCloud'15]

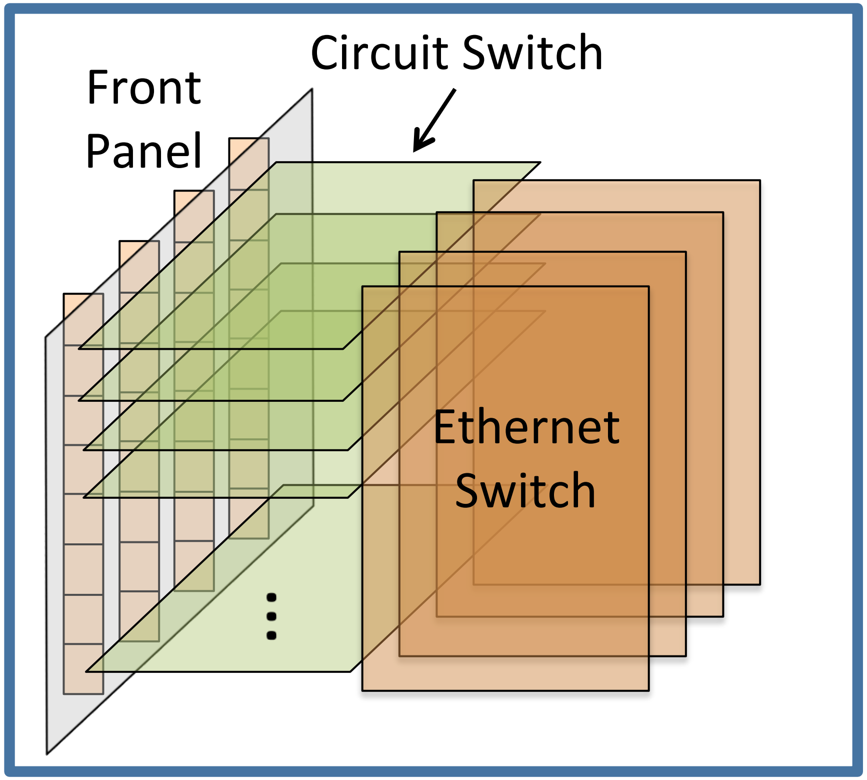
Wiring & Maintenance

Lighthouse
(In submission)

Physical-Layer Programmability in WAN

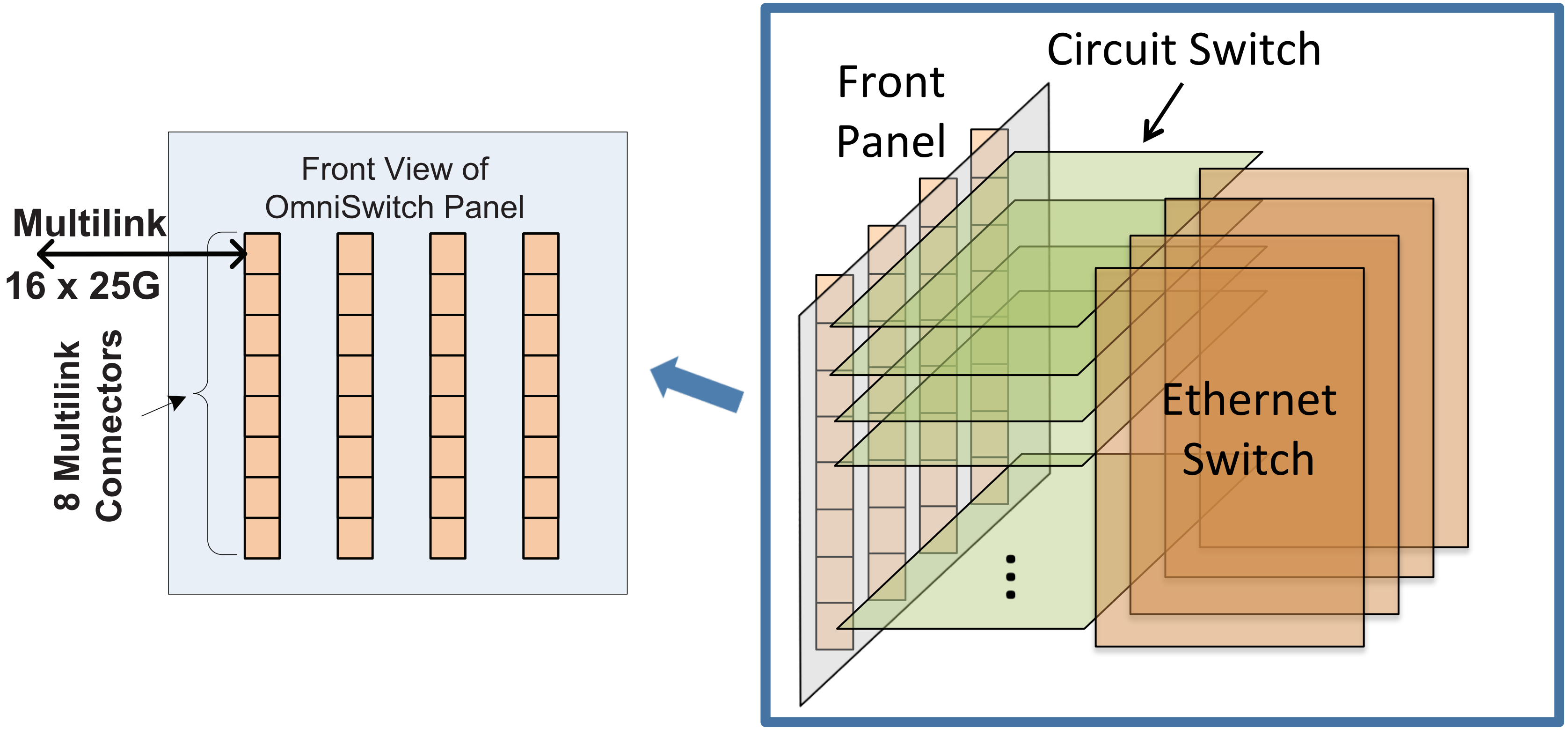
OmniSwitch

- Universal Building Block



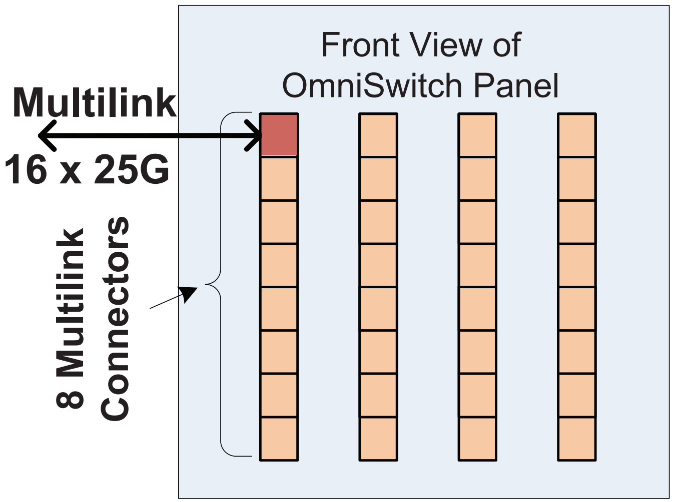
OmniSwitch

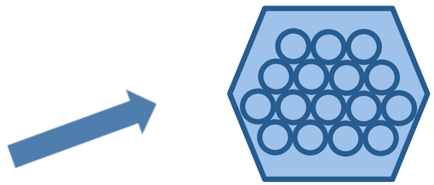
- Universal Building Block



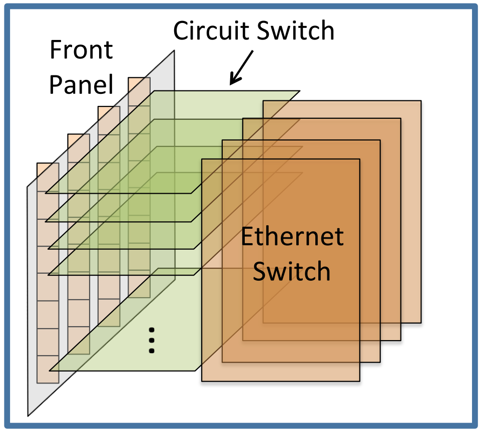
OmniSwitch

- Universal Building Block

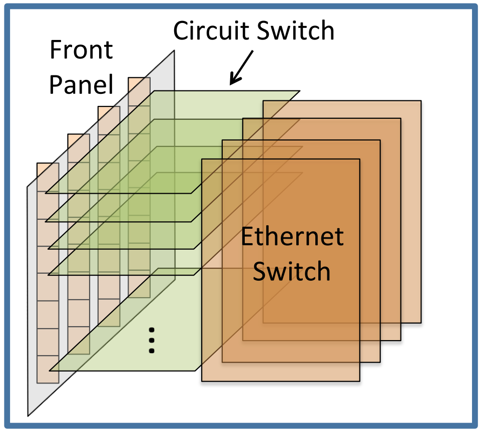




Automatic Wiring

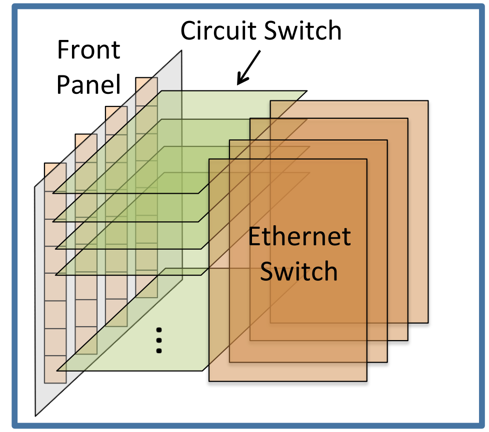


Automatic Wiring

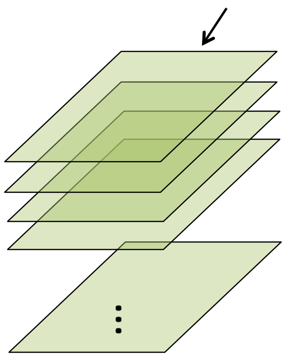


4 Ethernet Switches (128 ports each)

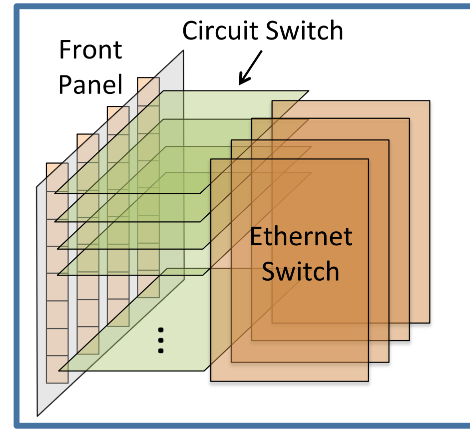
Automatic Wiring



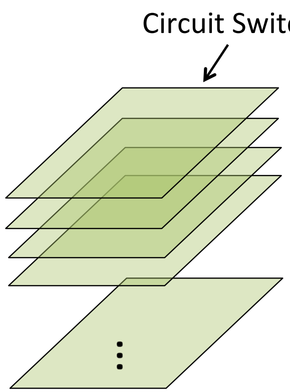
4 Ethernet Switches (128 ports each)

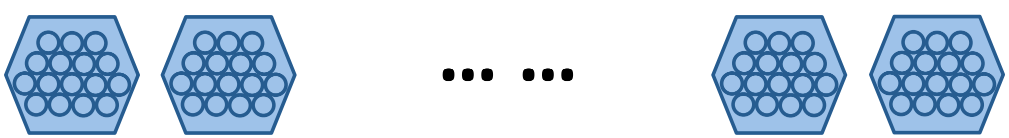


Automatic Wiring



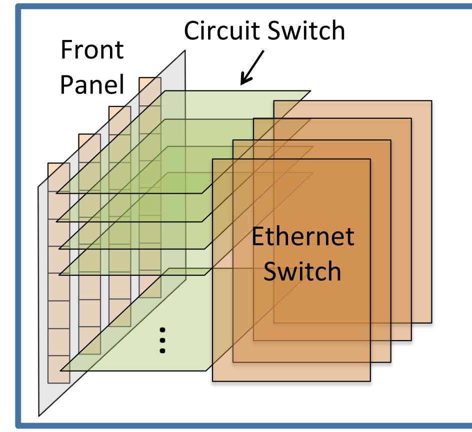
4 Ethernet Switches (128 ports each)



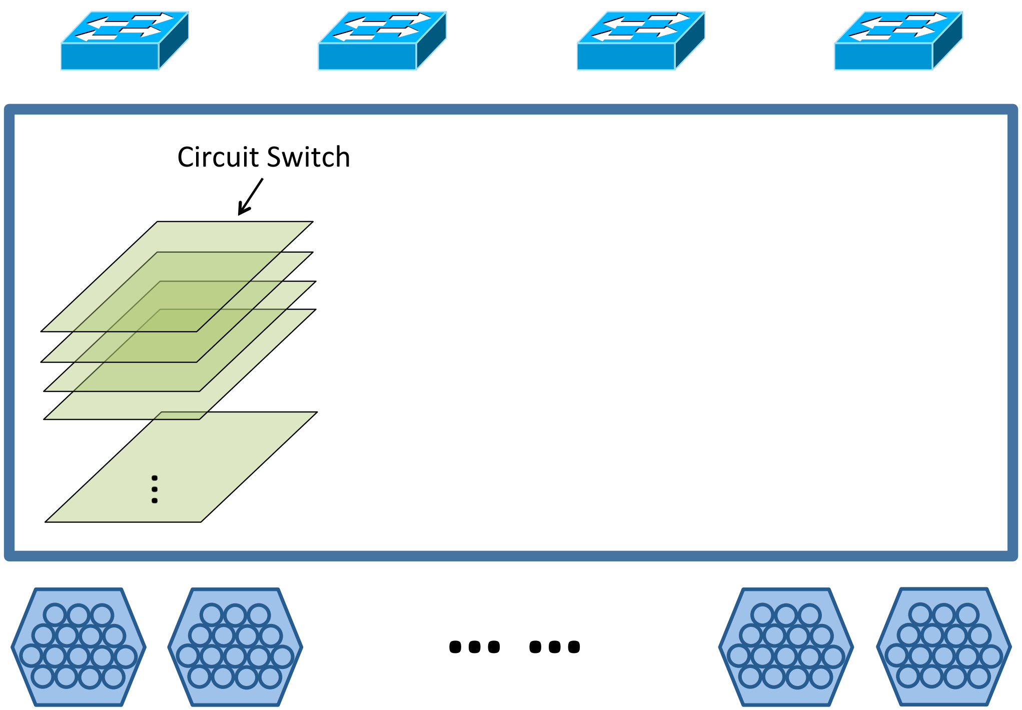


32 Multilink Connectors
(16 individual links each)

Automatic Wiring



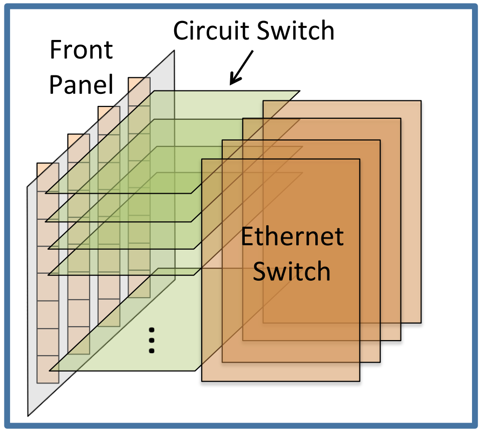
4 Ethernet Switches (128 ports each)

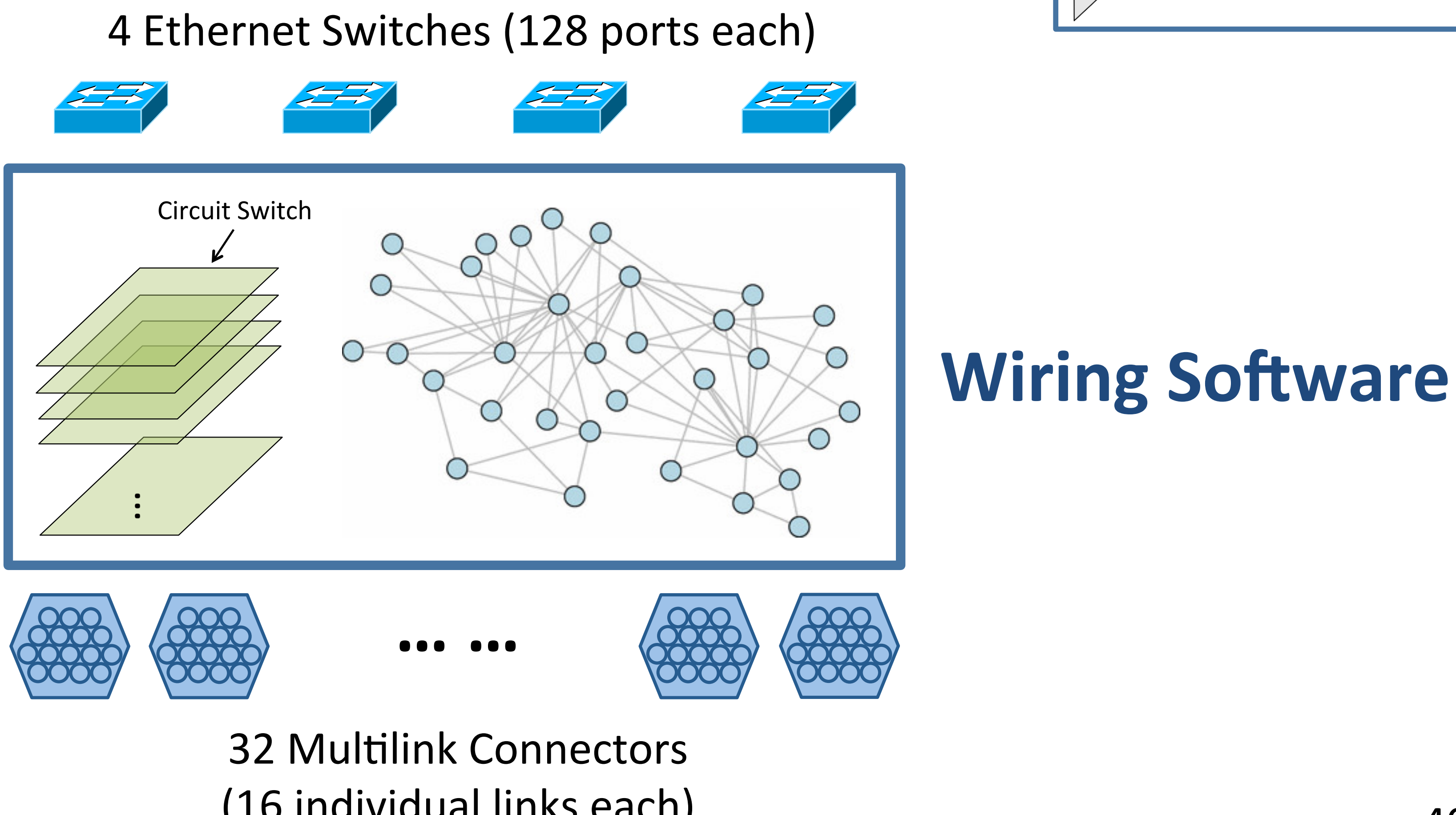


Wiring Software

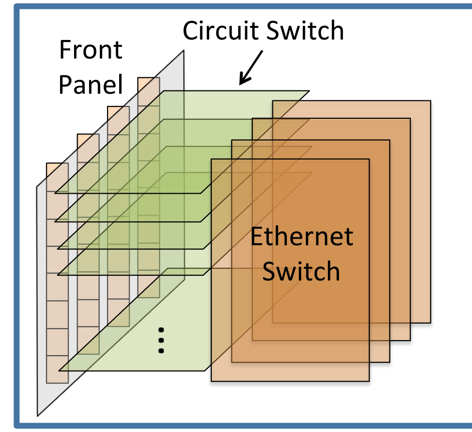
32 Multilink Connectors
(16 individual links each)

Automatic Wiring

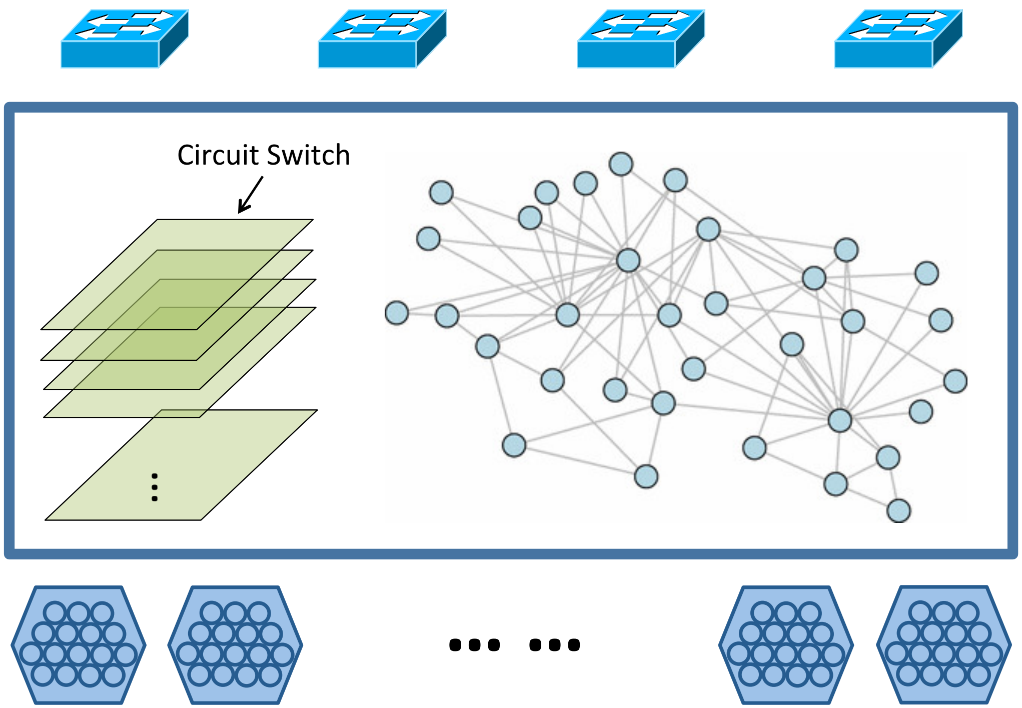




Easy Maintenance



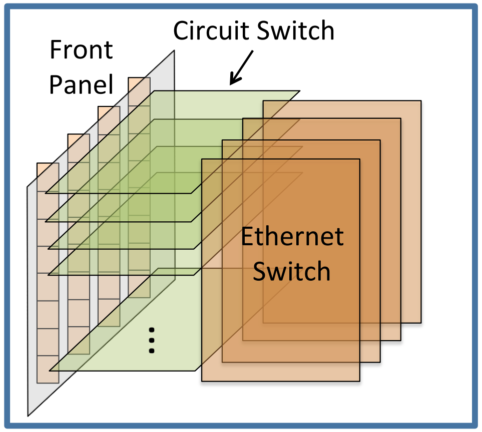
4 Ethernet Switches (128 ports each)

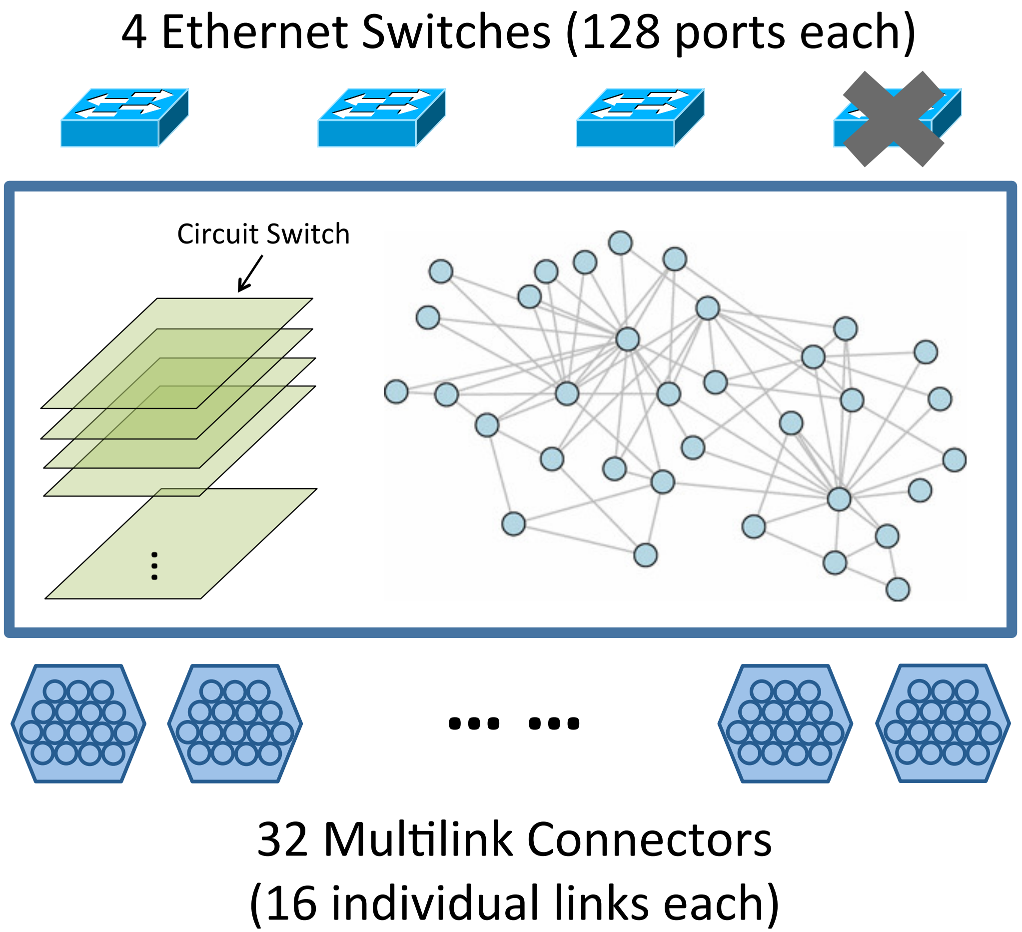


Wiring Software

32 Multilink Connectors
(16 individual links each)

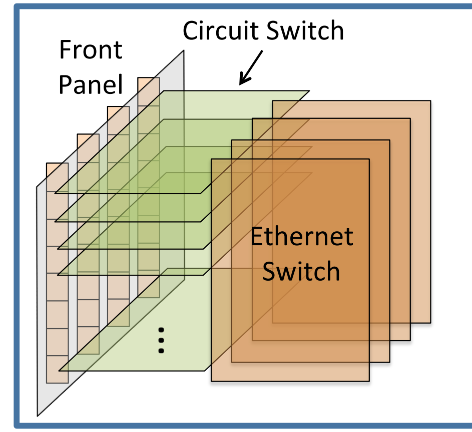
Easy Maintenance



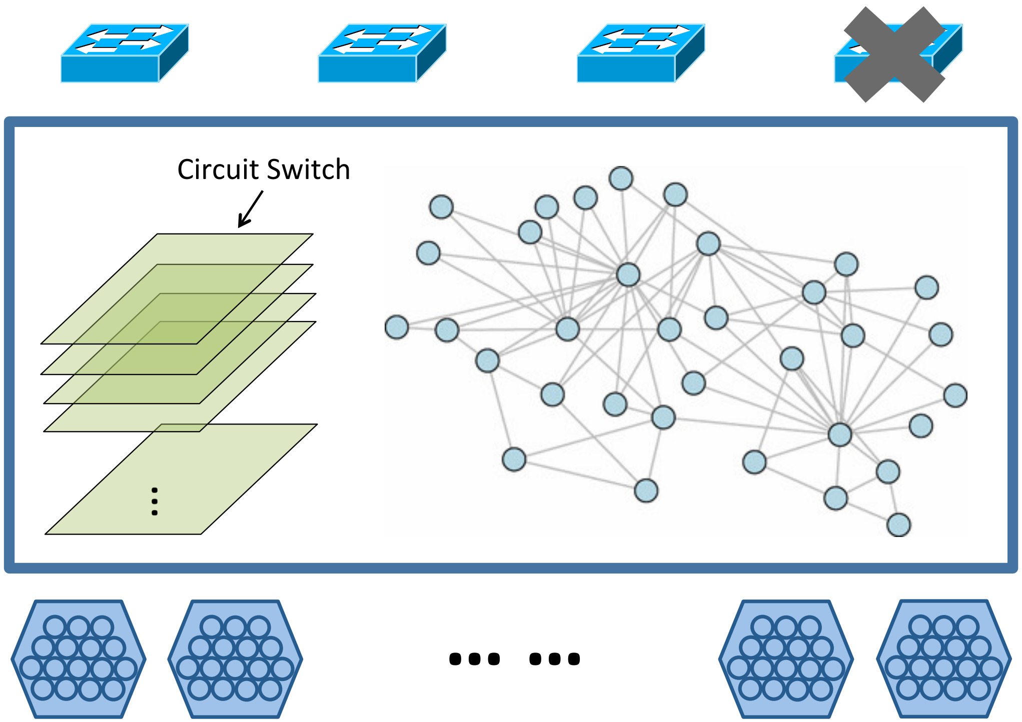


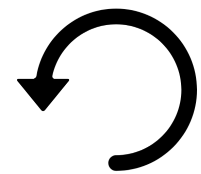
Wiring Software

Easy Maintenance



4 Ethernet Switches (128 ports each)

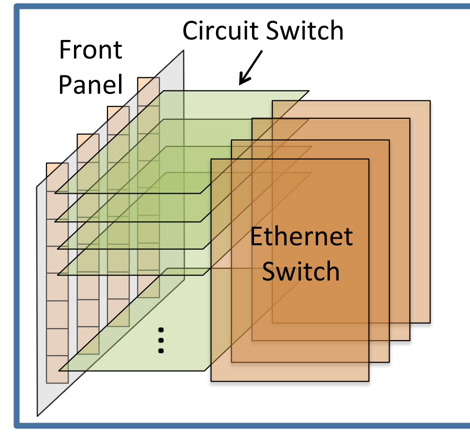




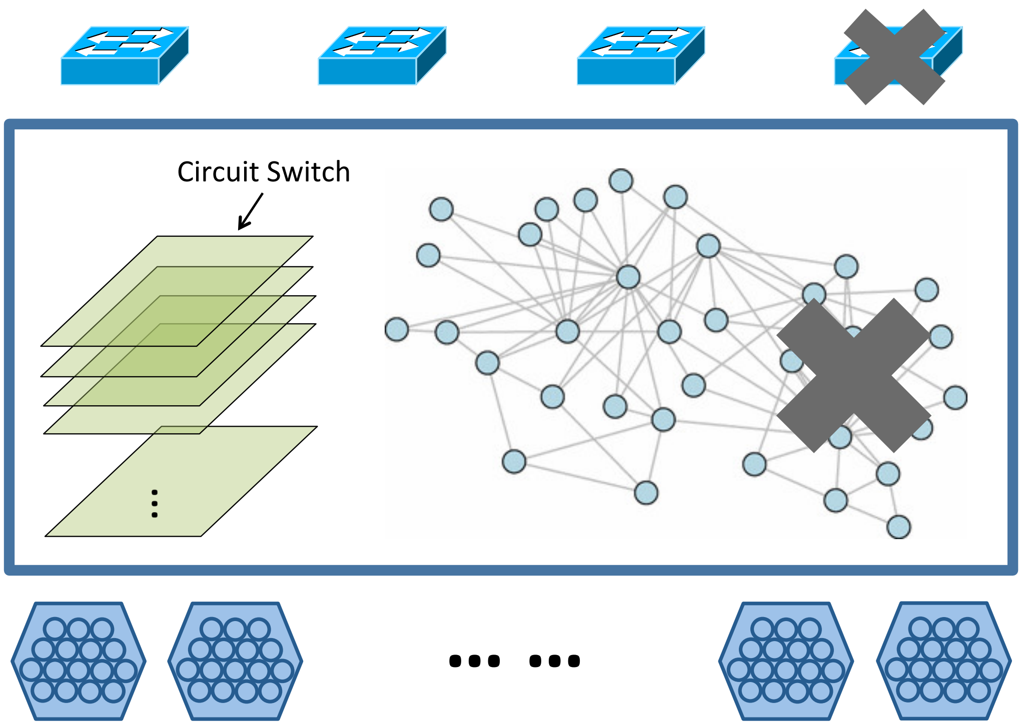
Wiring Software

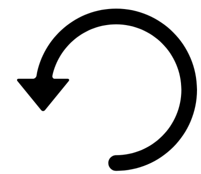
32 Multilink Connectors
(16 individual links each)

Easy Maintenance



4 Ethernet Switches (128 ports each)





Wiring Software

32 Multilink Connectors
(16 individual links each)

Outline

ShareBackup
[HotNets'17, SIGCOMM'18]

Failure Recovery

Flat-tree
[HotNets'16, SIGCOMM'17]

Service Provisioning

OmniSwitch
[HotCloud'15]

Wiring & Maintenance

Lighthouse
(In submission)

Physical-Layer Programmability in WAN

Outline

ShareBackup
[HotNets'17, SIGCOMM'18]

Failure Recovery

Flat-tree
[HotNets'16, SIGCOMM'17]

Service Provisioning

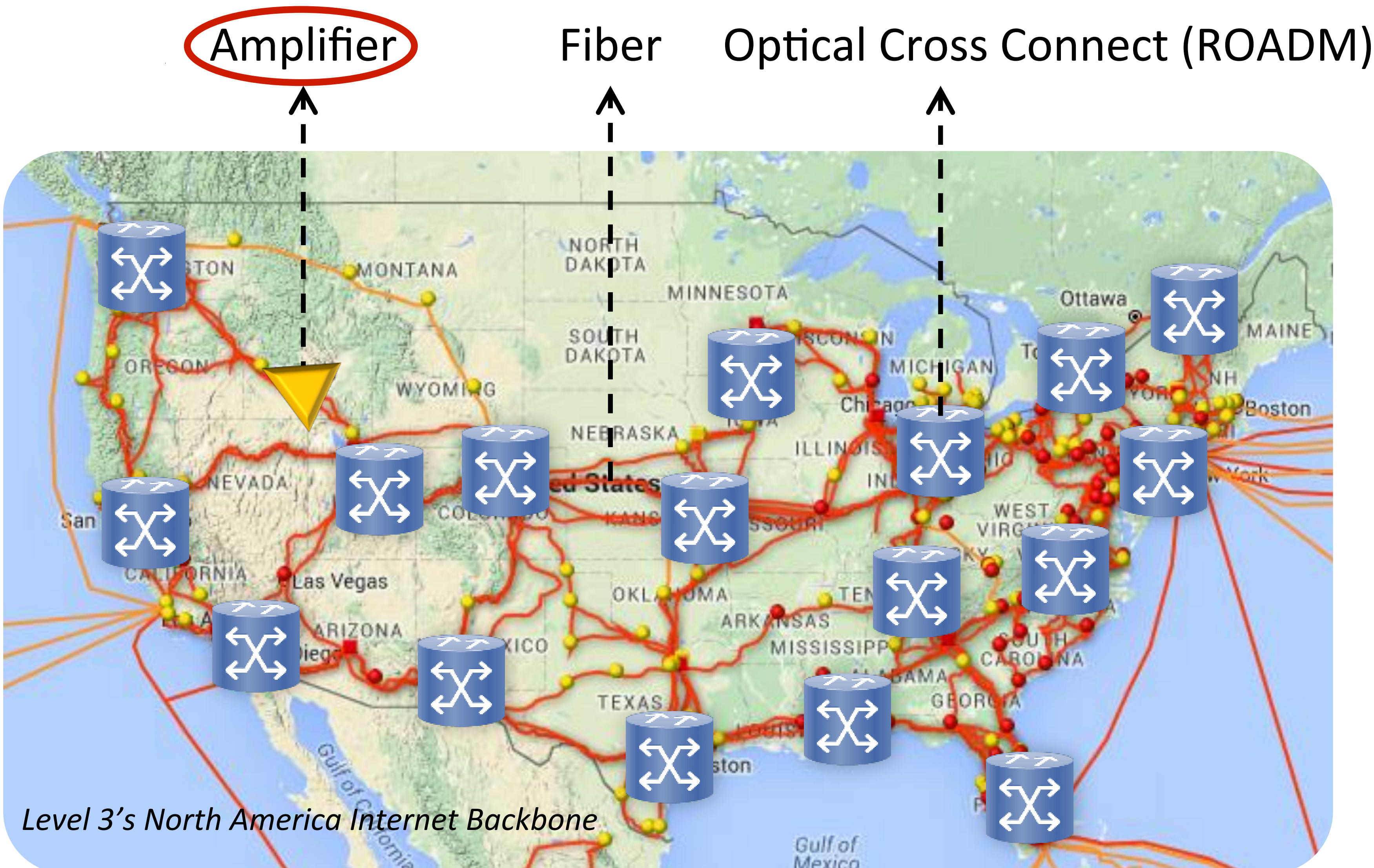
OmniSwitch
[HotCloud'15]

Wiring & Maintenance

Lighthouse
(In submission)

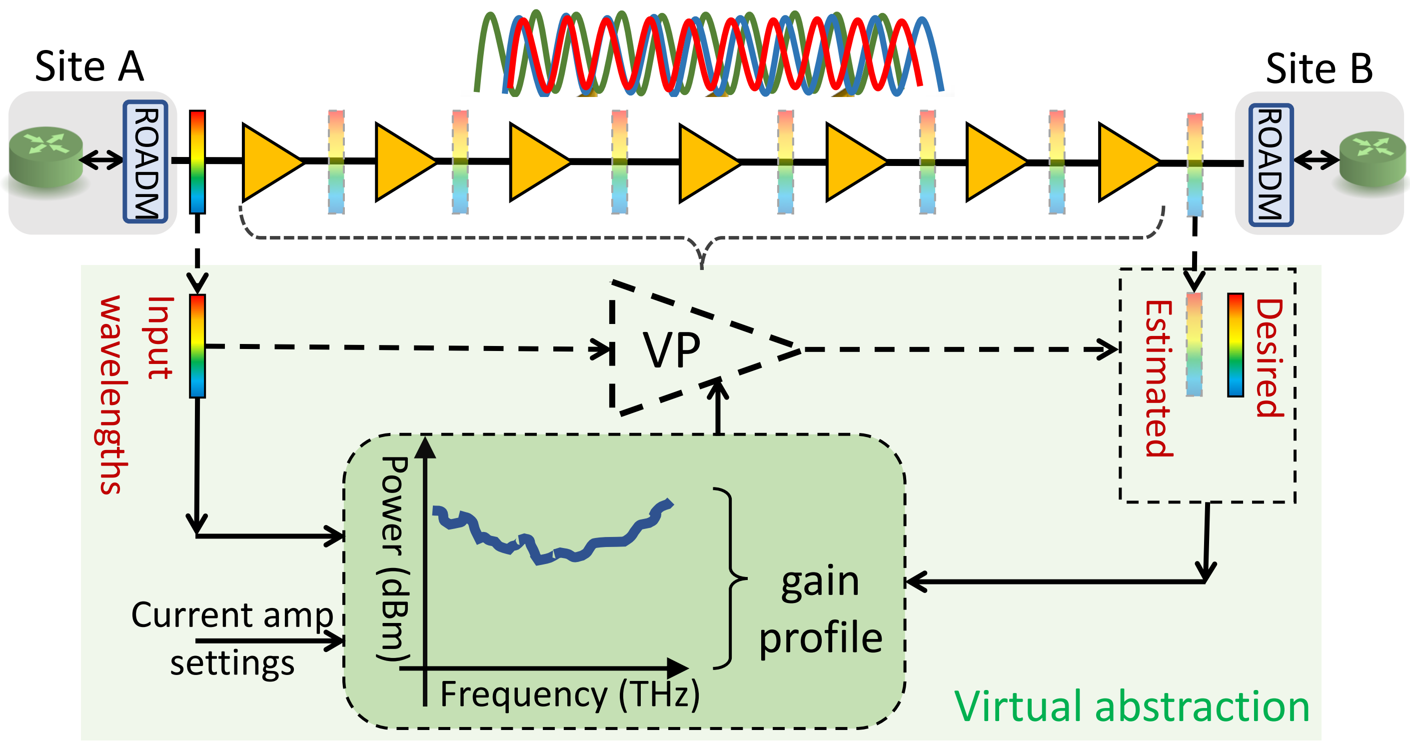
Physical-Layer Programmability in WAN

Wide Area Network (WAN)



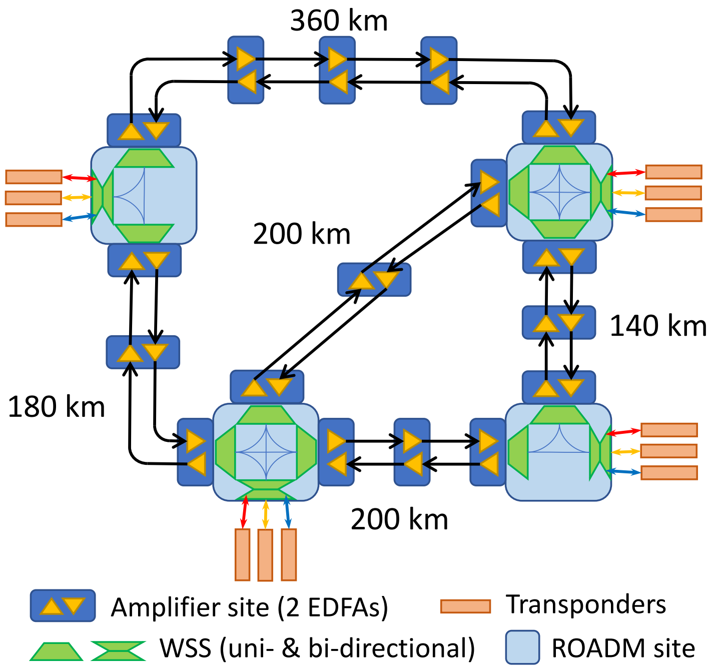
Fast Wavelength Shifting

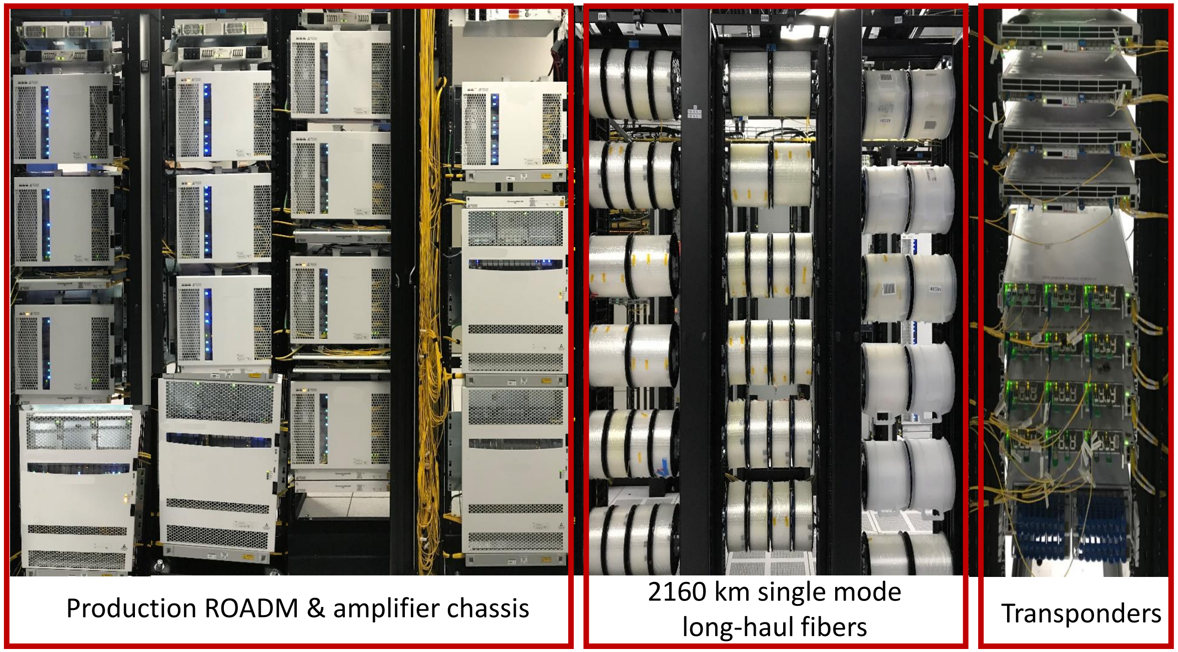
- Wavelength shifting is slow: ~10min
- Model power profile with Virtual Amplifier



Testbed Demo

- Wavelength shifting in 8 seconds





20+ engineers, 1 month time

Summary

- Physical-layer programmability for network operation
- Four example architectures

	ShareBackup	Flat-tree	OmniSwitch	Lighthouse
Design purpose	Failure recovery	Service provisioning	Wiring & maintenance	Programmability in WAN
Intuition	Shareable backup	Topology conversion	Universal building block	Wavelength shifting
Key ideas	Failure group	1. Server mobility 2. Link diversification	Wiring software	Model power profile

Social Impact

Social Impact

- Connect the world
  - *1st week of social distancing: 15% increase of FB utilization*
  - *Reliability and availability at highest priority*
  - *More oncall efforts to guard our infrastructures*

Social Impact

- Connect the world
  - *1st week of social distancing: 15% increase of FB utilization*
  - *Reliability and availability at highest priority*
  - *More oncall efforts to guard our infrastructures*

- Simplify & automate network management

Social Impact

- Connect the world
  - *1st week of social distancing: 15% increase of FB utilization*
  - *Reliability and availability at highest priority*
  - *More oncall efforts to guard our infrastructures*
  - Simplify & automate network management
- Provide high-quality service
  - *Netflix and Amazon ceased HD content streaming*
  - *Zhihu (Chinese Quora) down for overload*

Social Impact

- Connect the world
  - *1st week of social distancing: 15% increase of FB utilization*
  - *Reliability and availability at highest priority*
  - *More oncall efforts to guard our infrastructures*

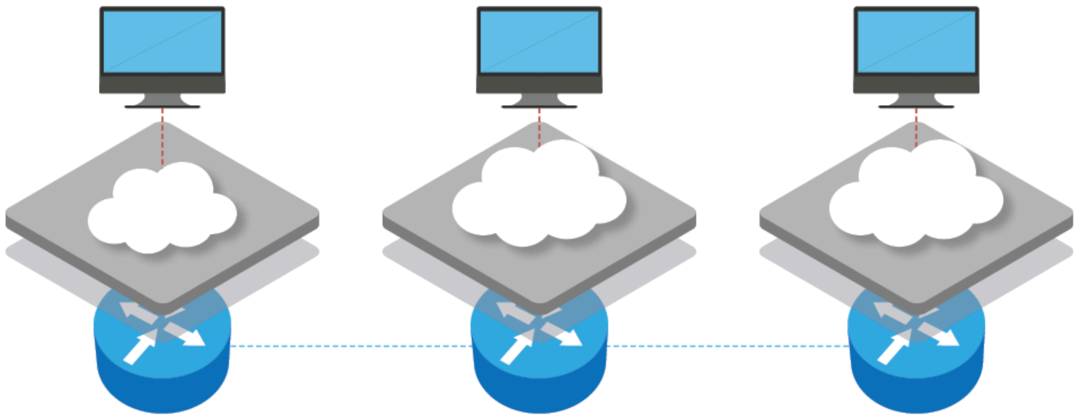
- Simplify & automate network management

- Provide high-quality service
  - *Netflix and Amazon ceased HD content streaming*
  - *Zhihu (Chinese Quora) down for overload*

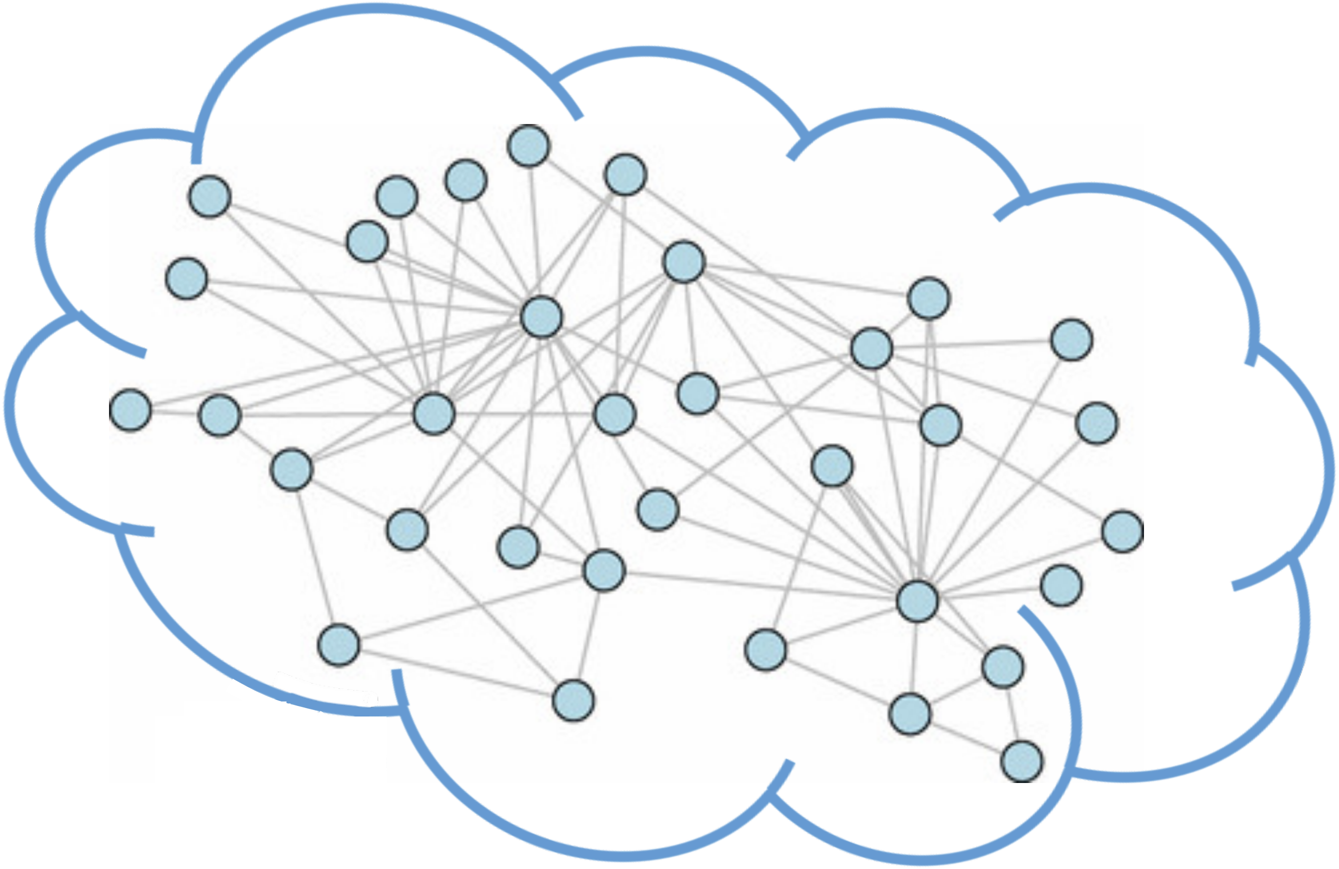
- Make elastic capacity of hardware possible

Bright Future: Truly Flexible Cloud

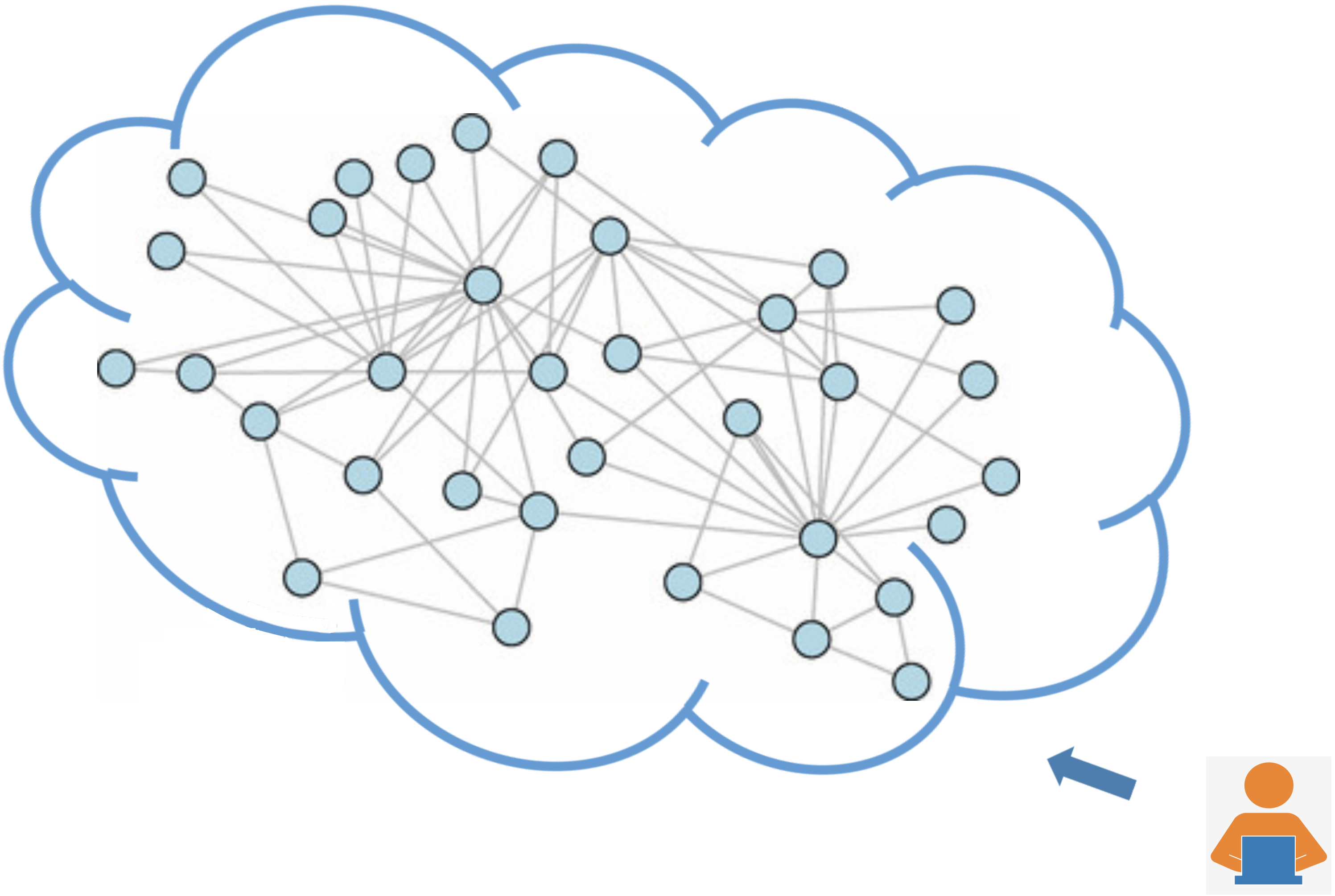
Virtualization



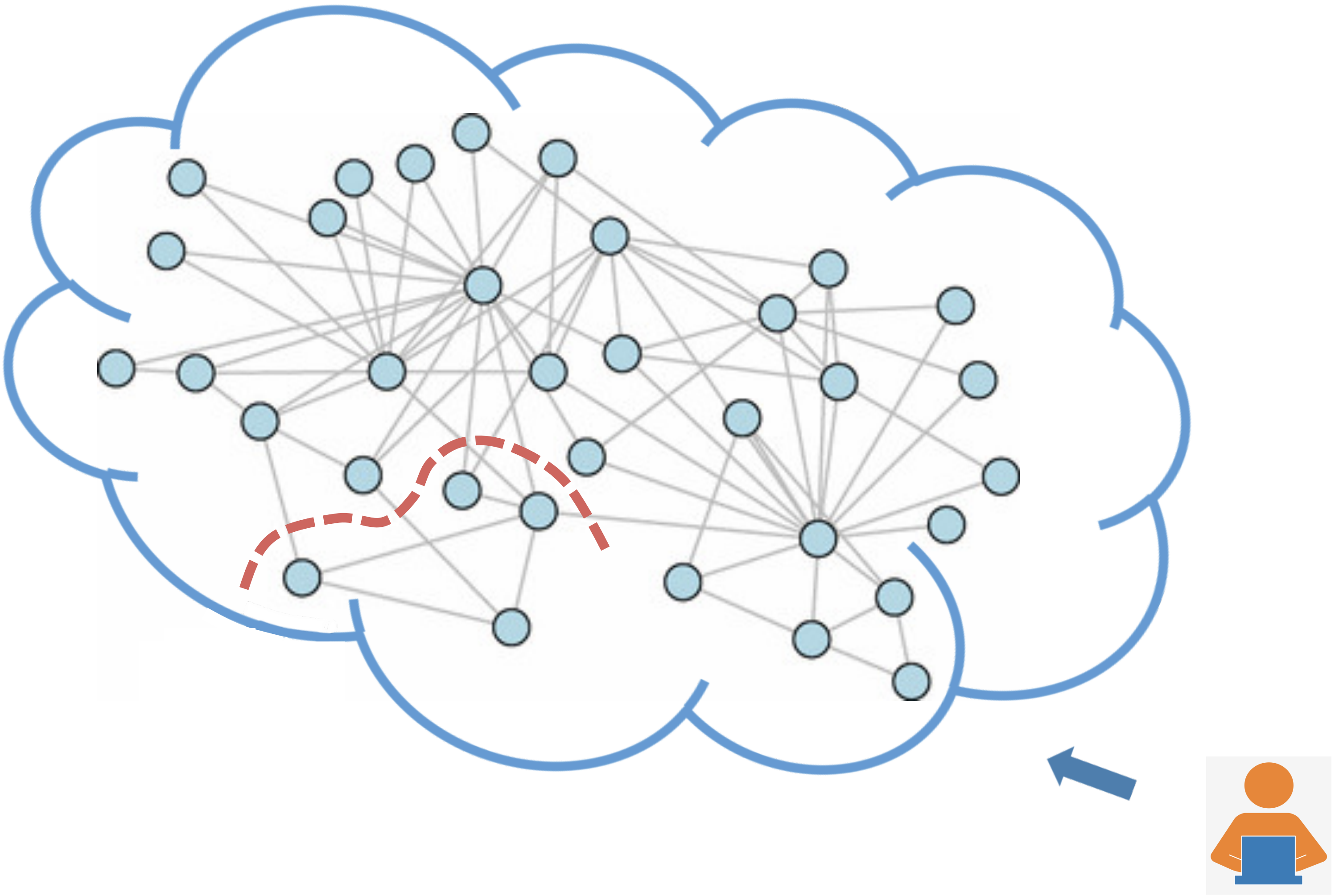
Bright Future: Truly Flexible Cloud



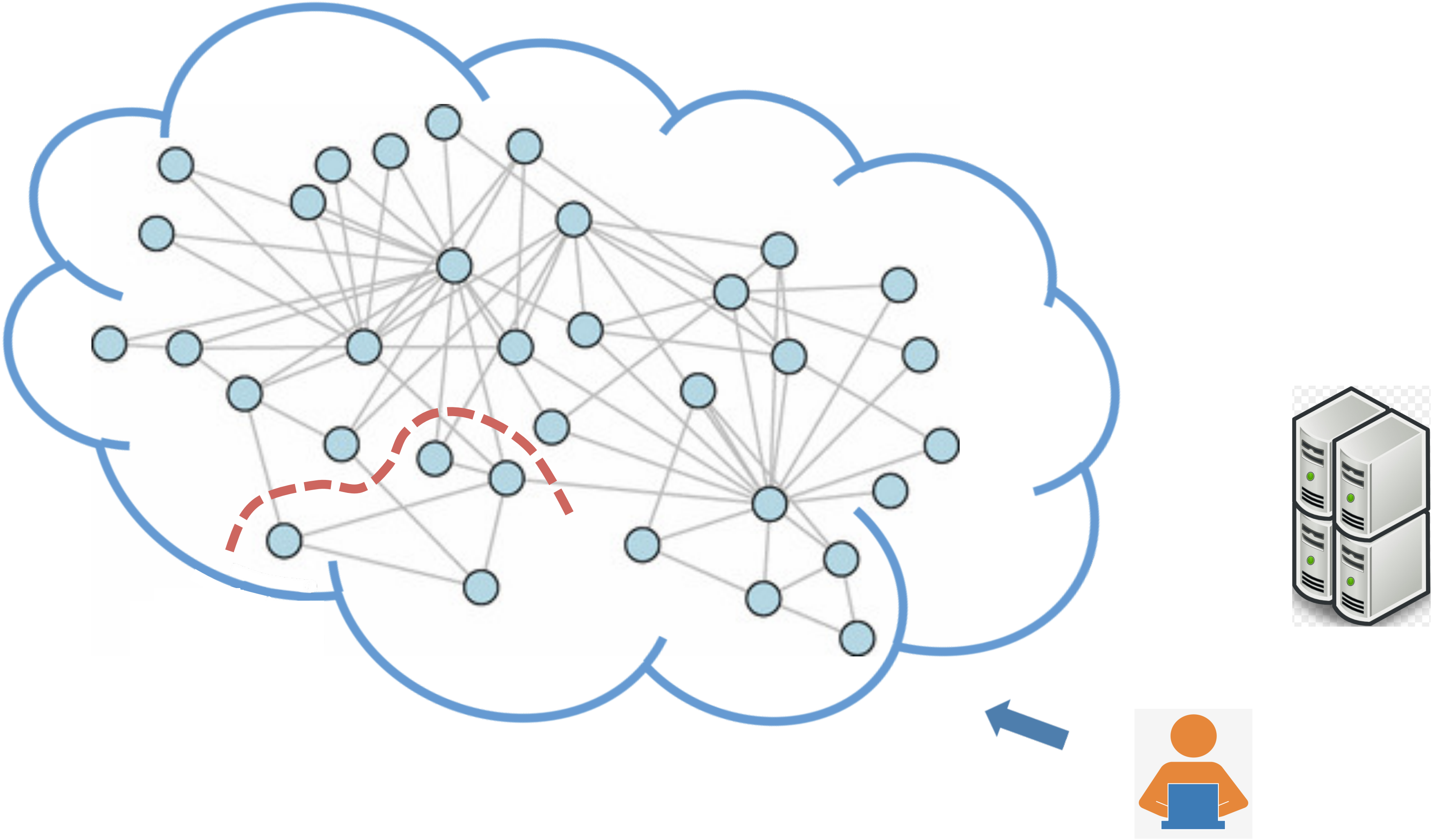
Bright Future: Truly Flexible Cloud



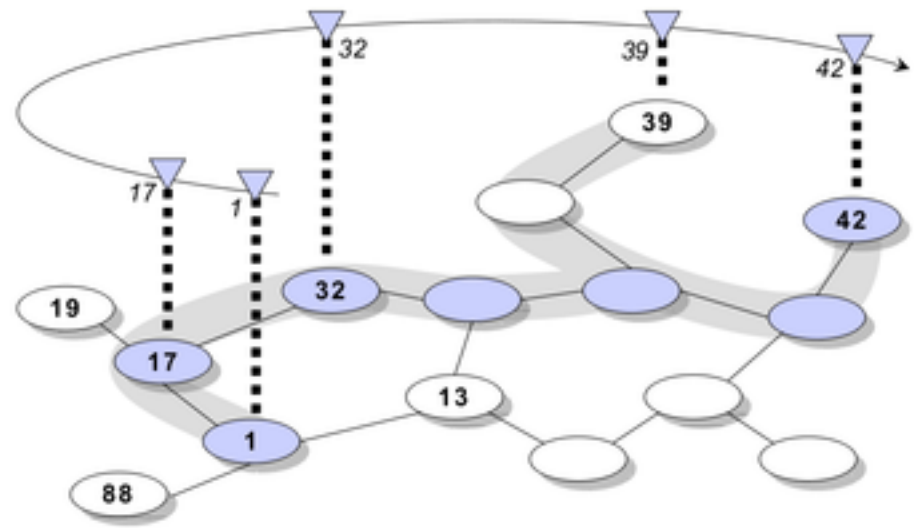
Bright Future: Truly Flexible Cloud



Bright Future: Truly Flexible Cloud

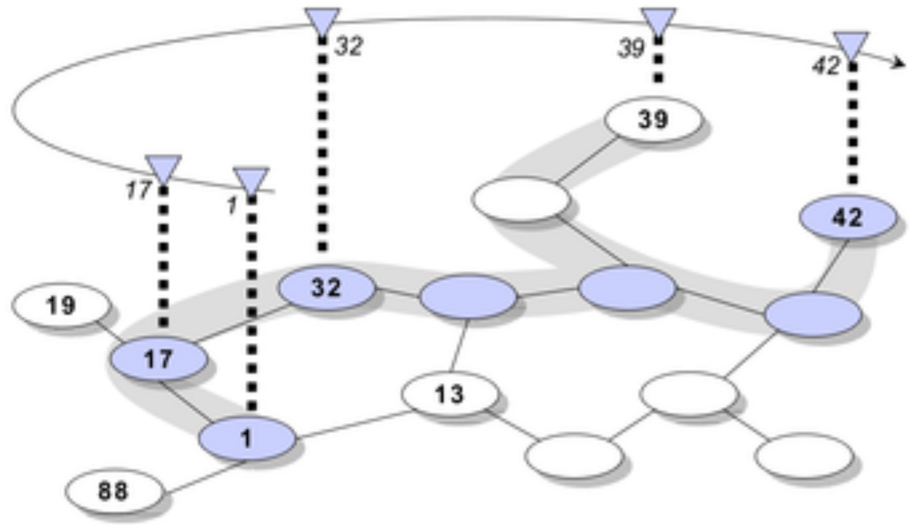


Direction 1: Network Verification



Direction 1: Network Verification

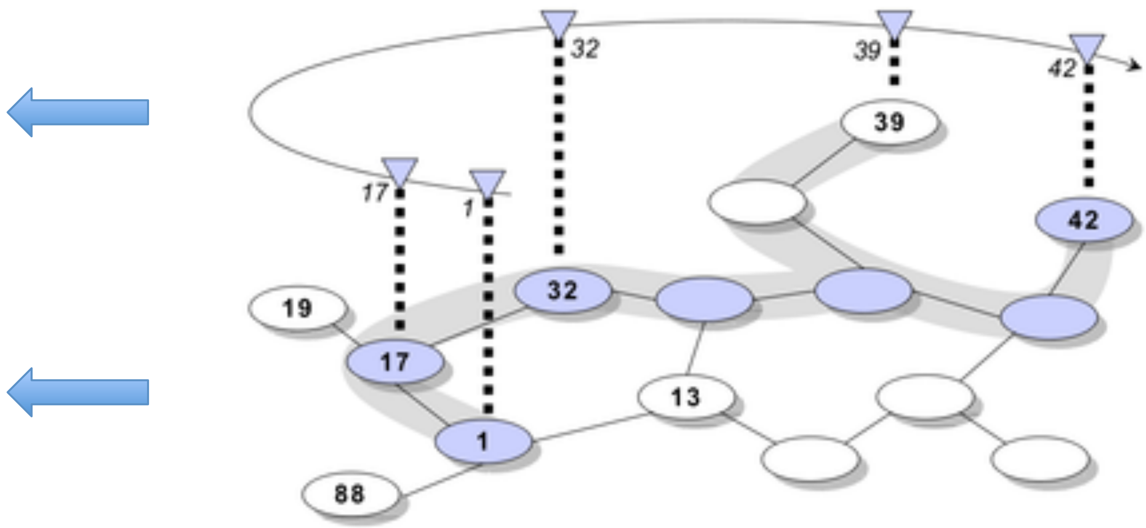
Validate connectivity



Direction 1: Network Verification

Validate routing

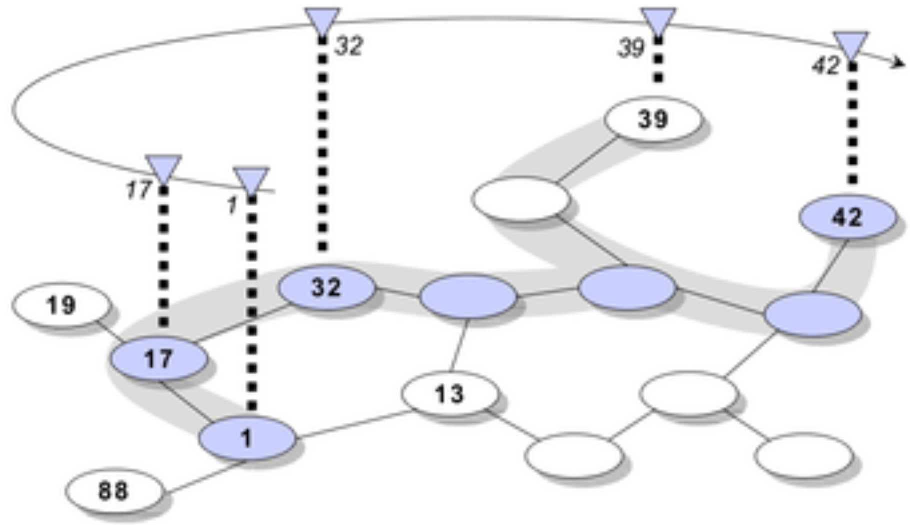
Validate connectivity



Direction 1: Network Verification

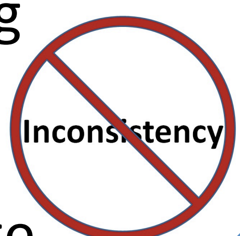
Validate routing

Validate connectivity

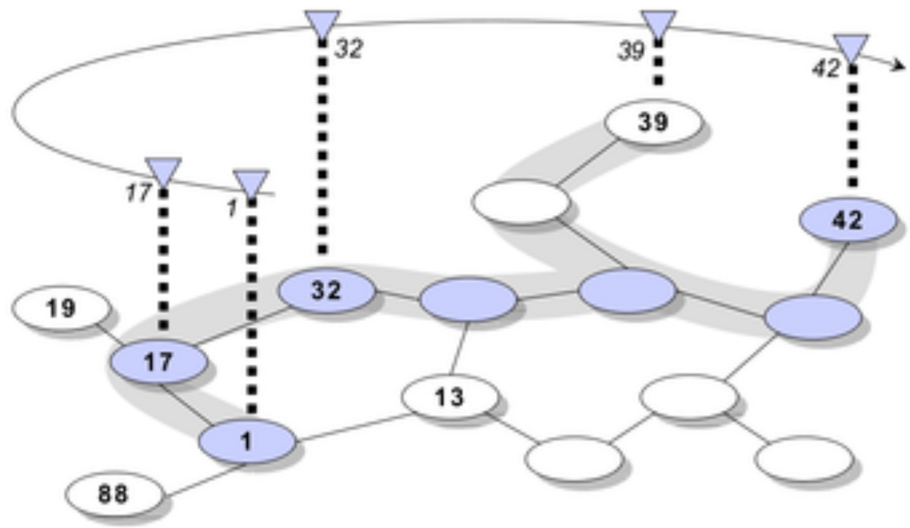


Direction 1: Network Verification

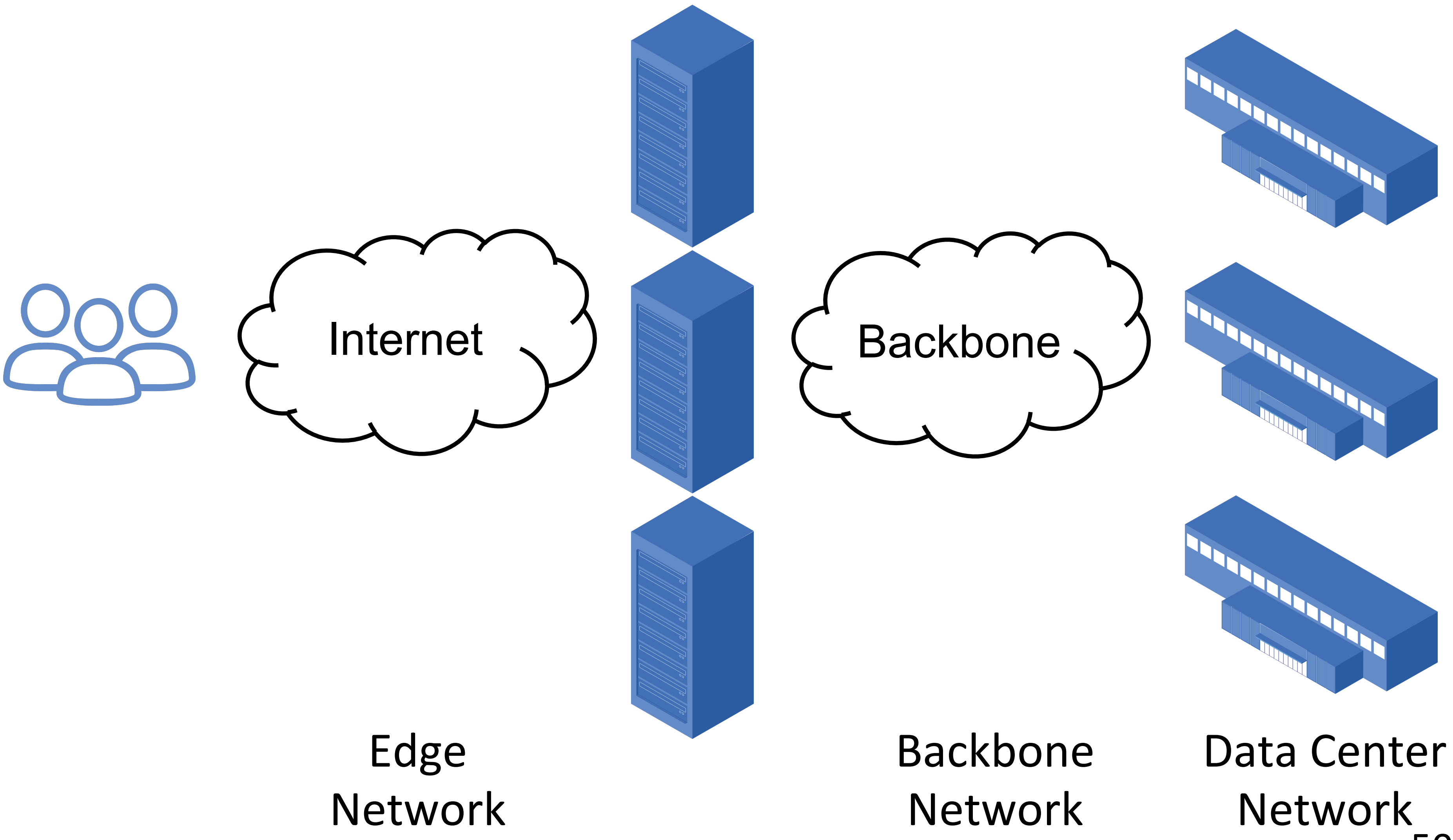
Validate routing



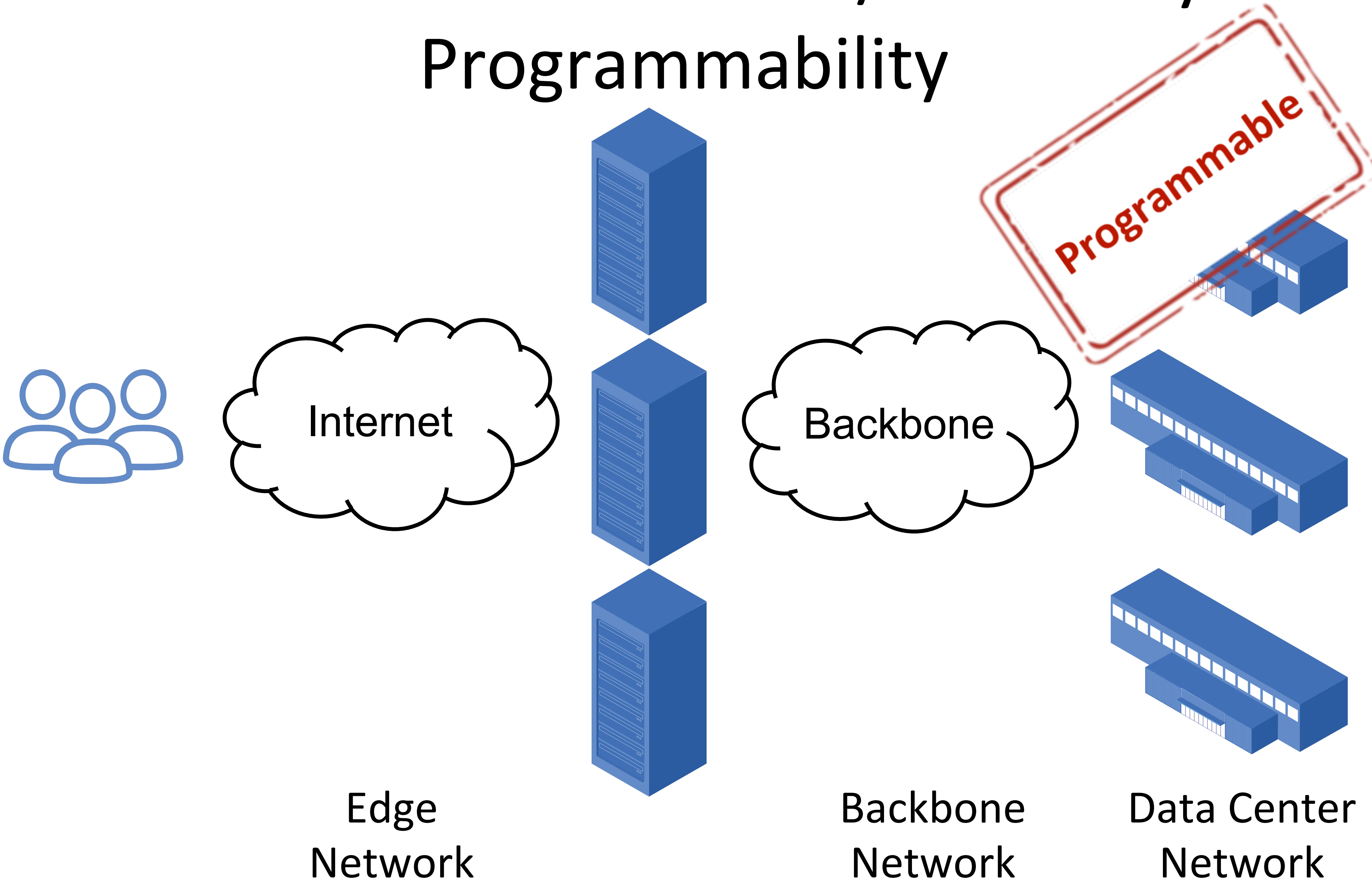
Validate connectivity



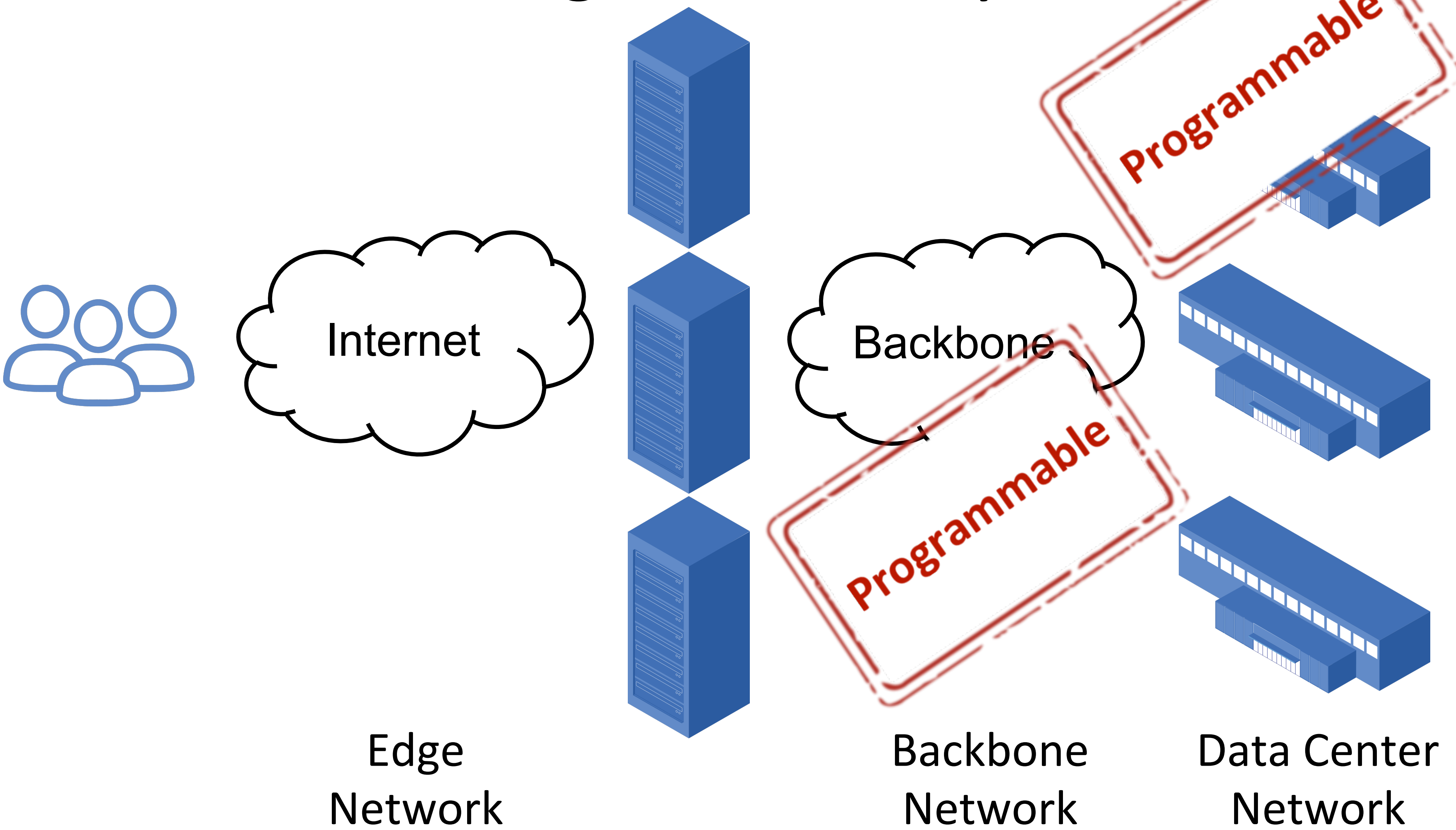
Direction 2: End-to-End / Cross-Layer Programmability



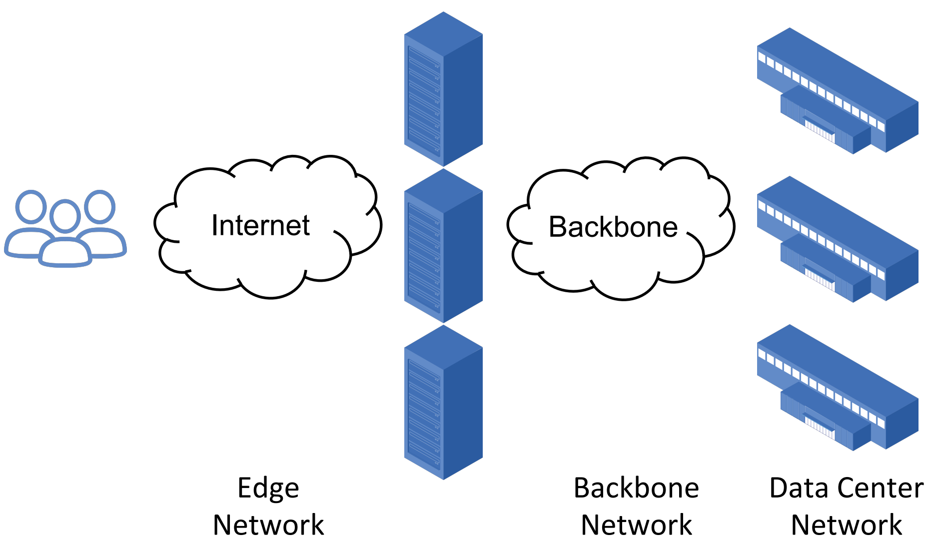
Direction 2: End-to-End / Cross-Layer Programmability



Direction 2: End-to-End / Cross-Layer Programmability

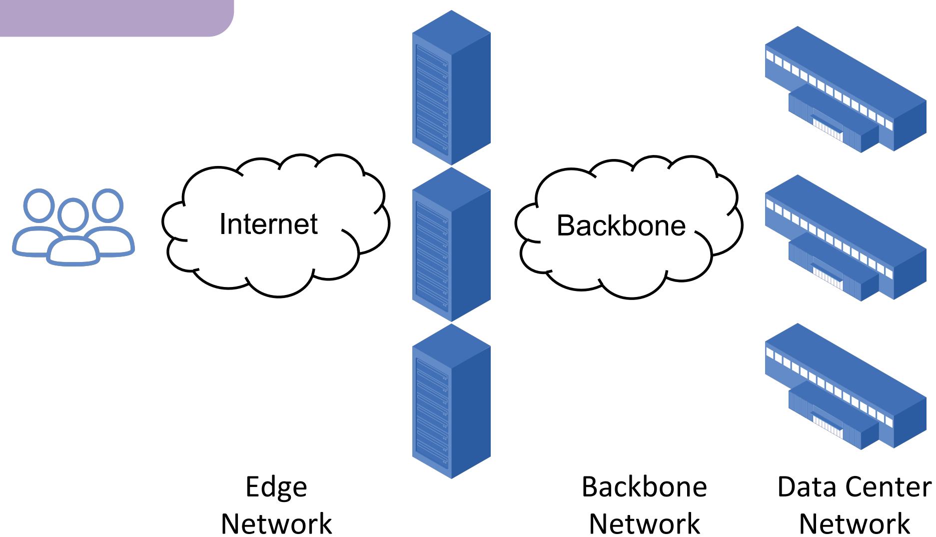


Direction 2: End-to-End / Cross-Layer Programmability



Direction 2: End-to-End / Cross-Layer Programmability

Smart NIC



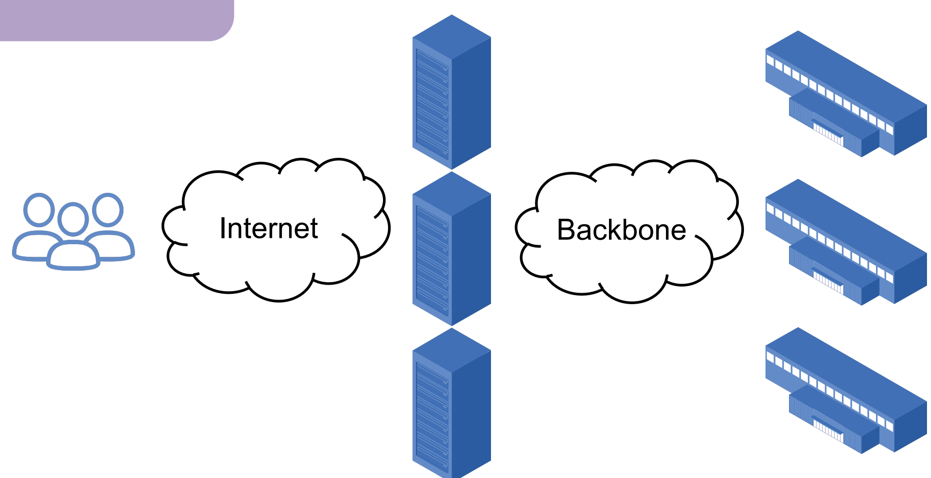
Direction 2: End-to-End / Cross-Layer Programmability

SDN

P4

Physical-Layer Programmability

Smart NIC



Edge Network

Backbone Network

Data Center Network

Direction 3: Joint Optimization of Traffic and Network Topology

Fit traffic to network topology

Direction 3: Joint Optimization of Traffic and Network Topology

Fit traffic to network topology

Fit network topology to traffic

Direction 3: Joint Optimization of Traffic and Network Topology

Fit traffic to network topology

Fit network topology to traffic

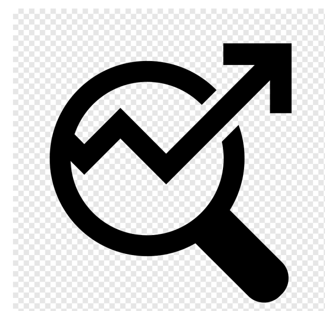
Formulation

Direction 3: Joint Optimization of Traffic and Network Topology

Fit traffic to network topology

Fit network topology to traffic

Formulation



Optimization