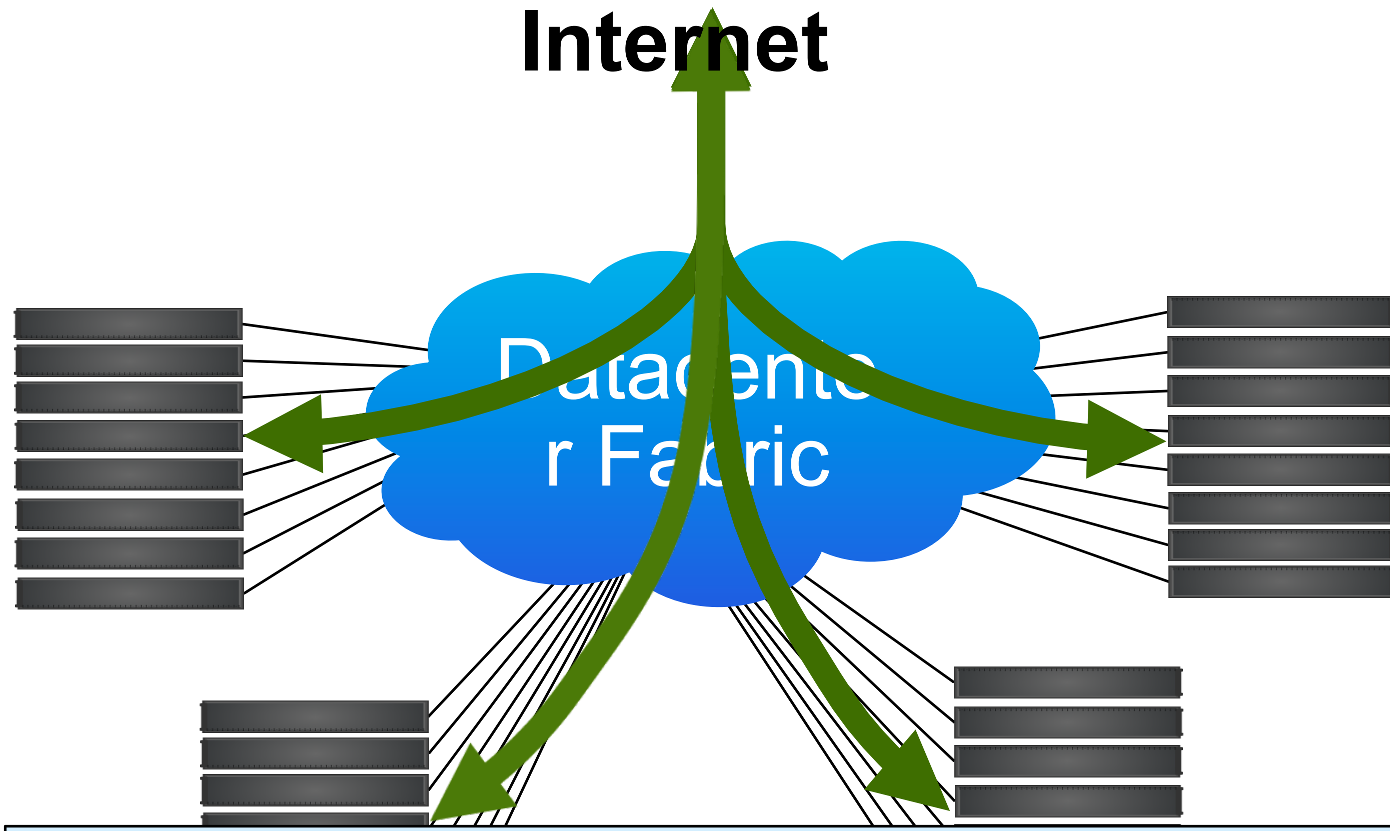
SPAIN:

High BW Data-Center Ethernet with Unmodified Switches

Jayaram Mudigonda, HP Labs
Praveen Yalagandula, HP Labs
Mohammad Al-Fares, UCSD
Jeff Mogul, HP Labs

Traditional Datacenter



Internet-facing applications:
E-Mail, Web Servers, etc.

DC Trends

HPC Applications

Information Explosion

hadoop Map Reduce

hadoop HDFS

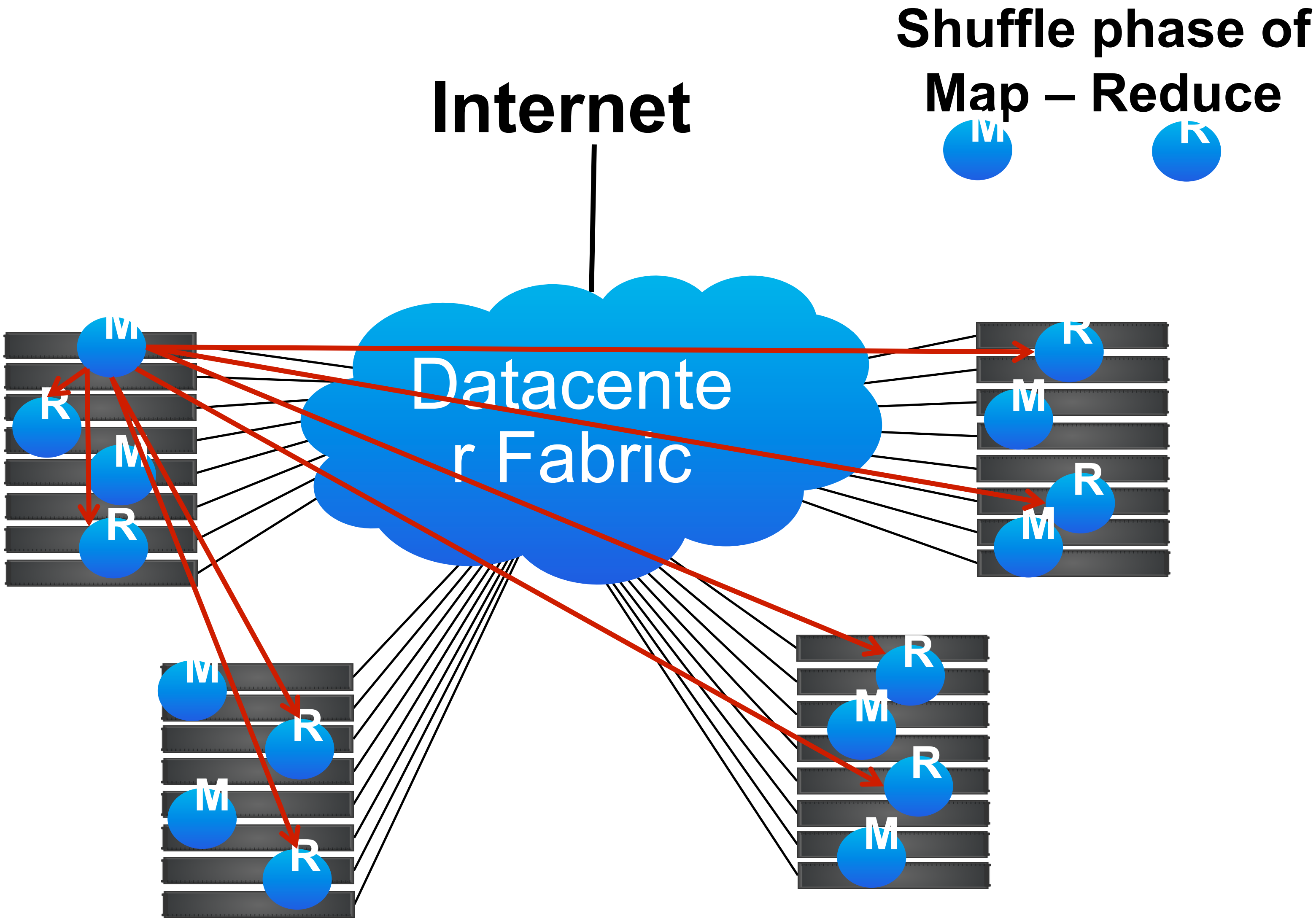
HBASE

Application Consolidation

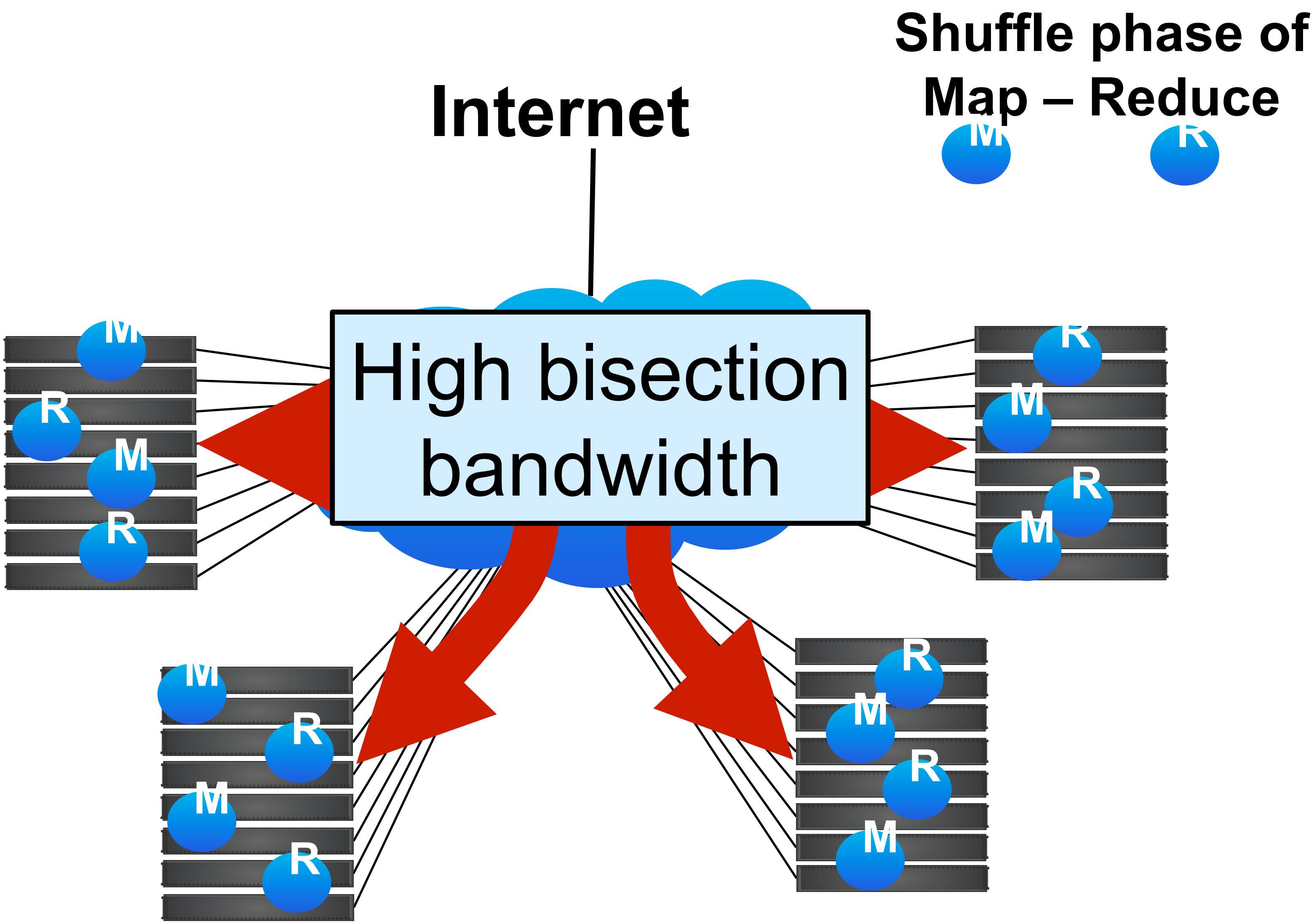
vmware™ Xen™

Virtualization

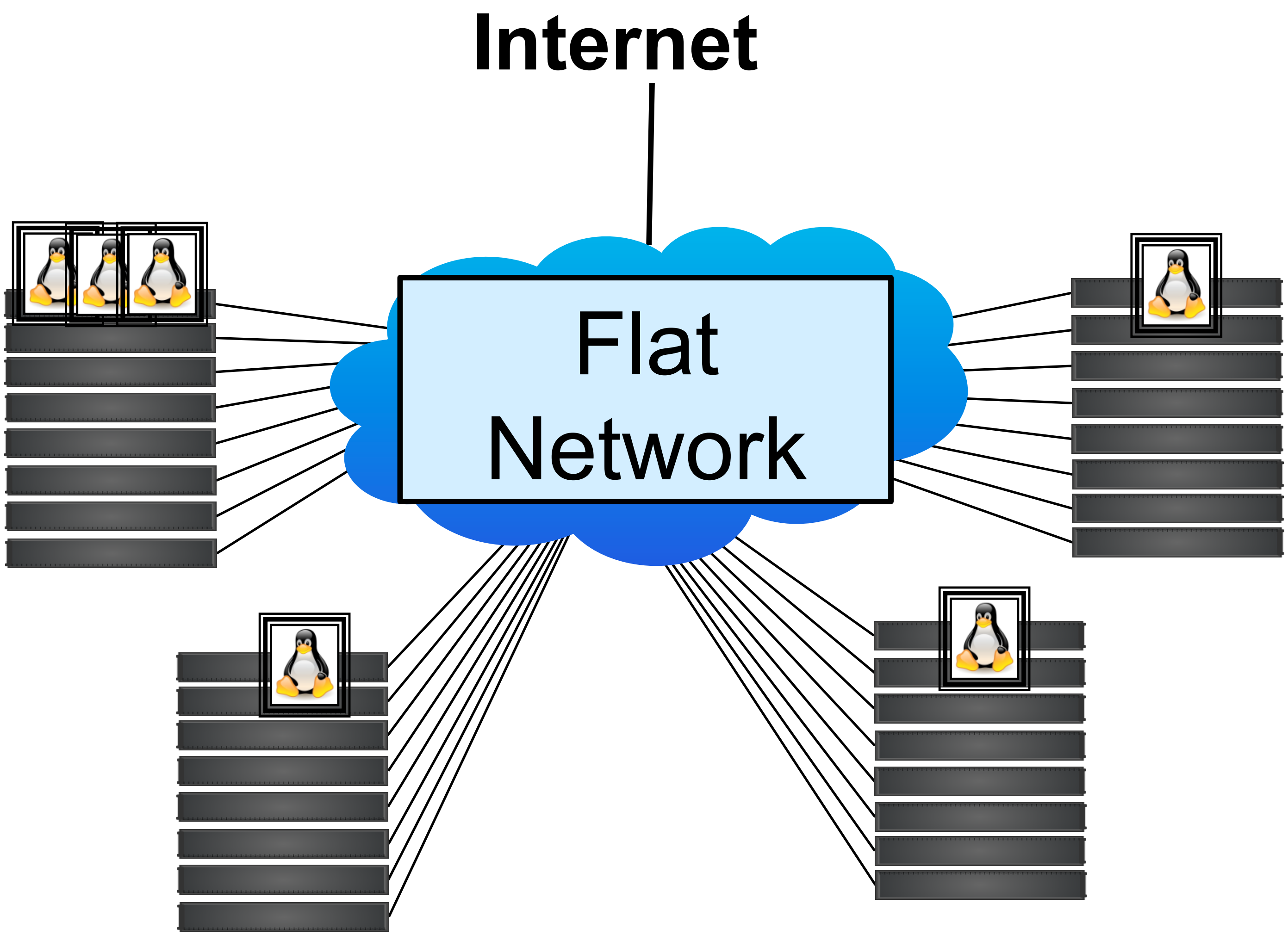
DC Trends


Shuffle phase of
Map – Reduce
M
R
Internet
Datacente
r Fabric
M
R
M
R
M
R
M
R
R
M
R
M
R
M
R
M

DC Trends


Shuffle phase of
Map – Reduce
Internet
High bisection bandwidth
M
R

DC Trends



DC Fabric Goals

High bisection BW
Flat network
Low-cost

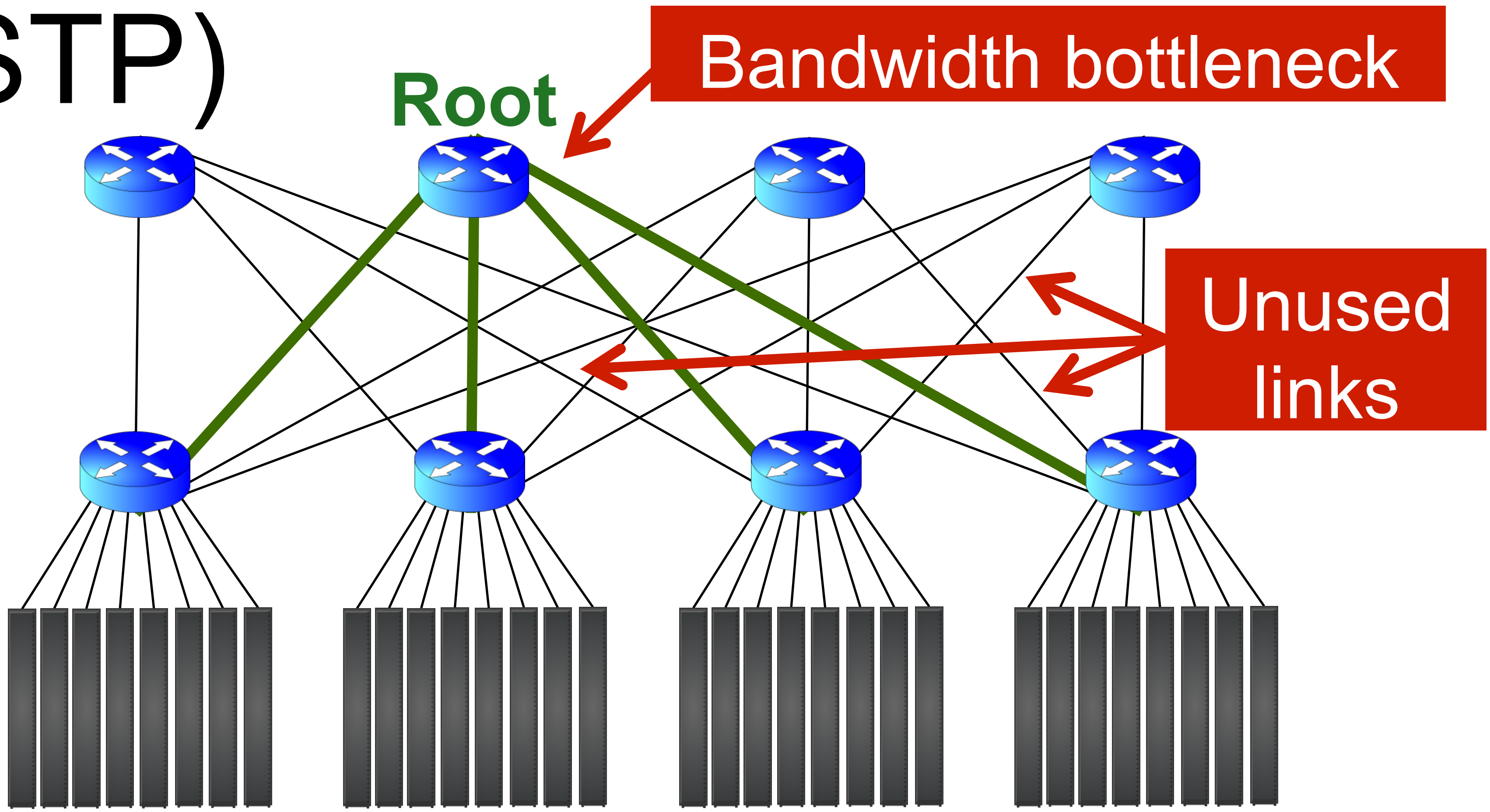
Ethernet: a good choice

- Commodity → Inexpensive
- Speeds:
  - 10G is here
  - 40G/100G soon
- Flat-addressing
- Self-configuring

But wait…

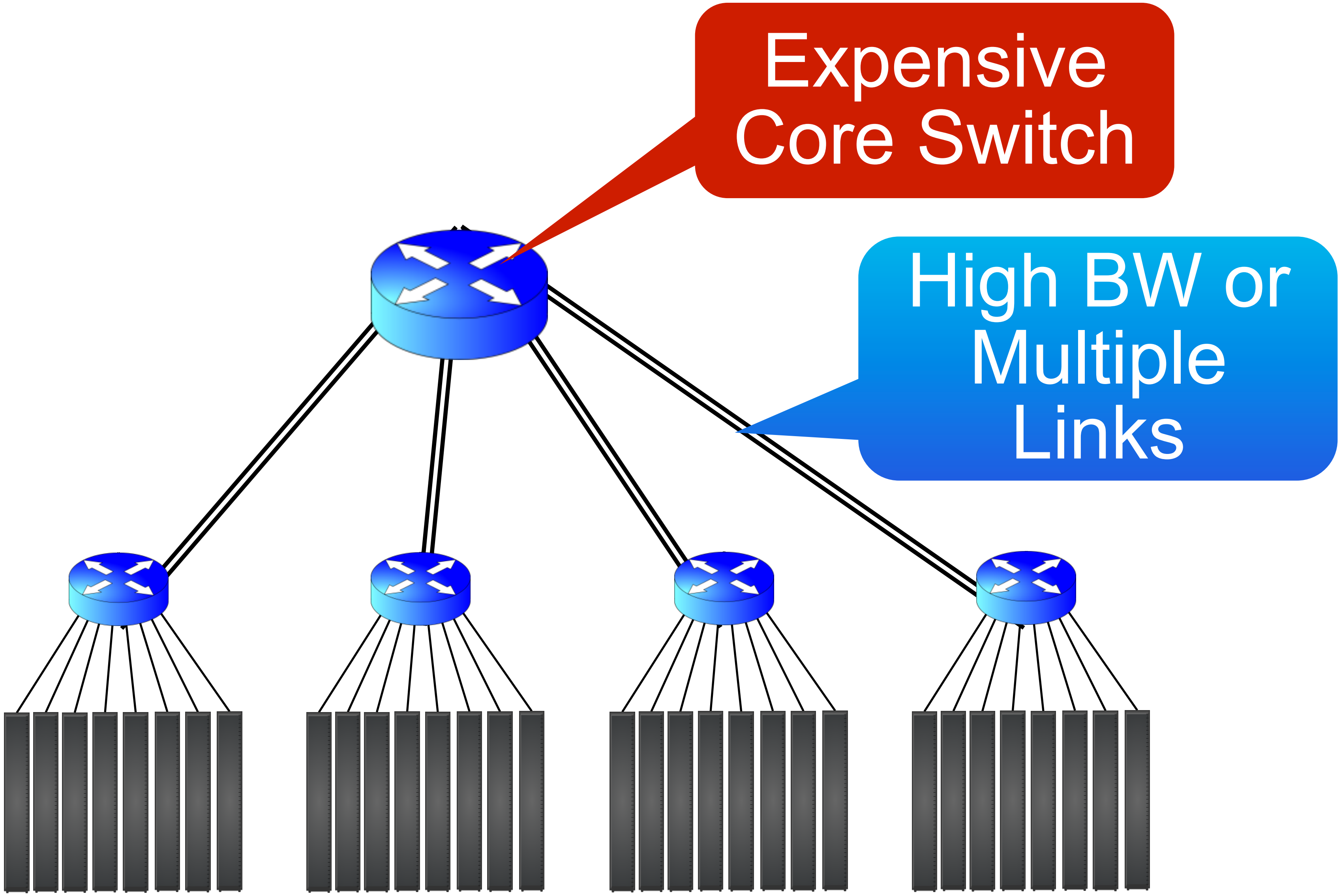
Spanning Tree Protocol (STP) makes Ethernet hard to scale!

Spanning Tree Protocol (STP)



Proposal 1: High-port core switch

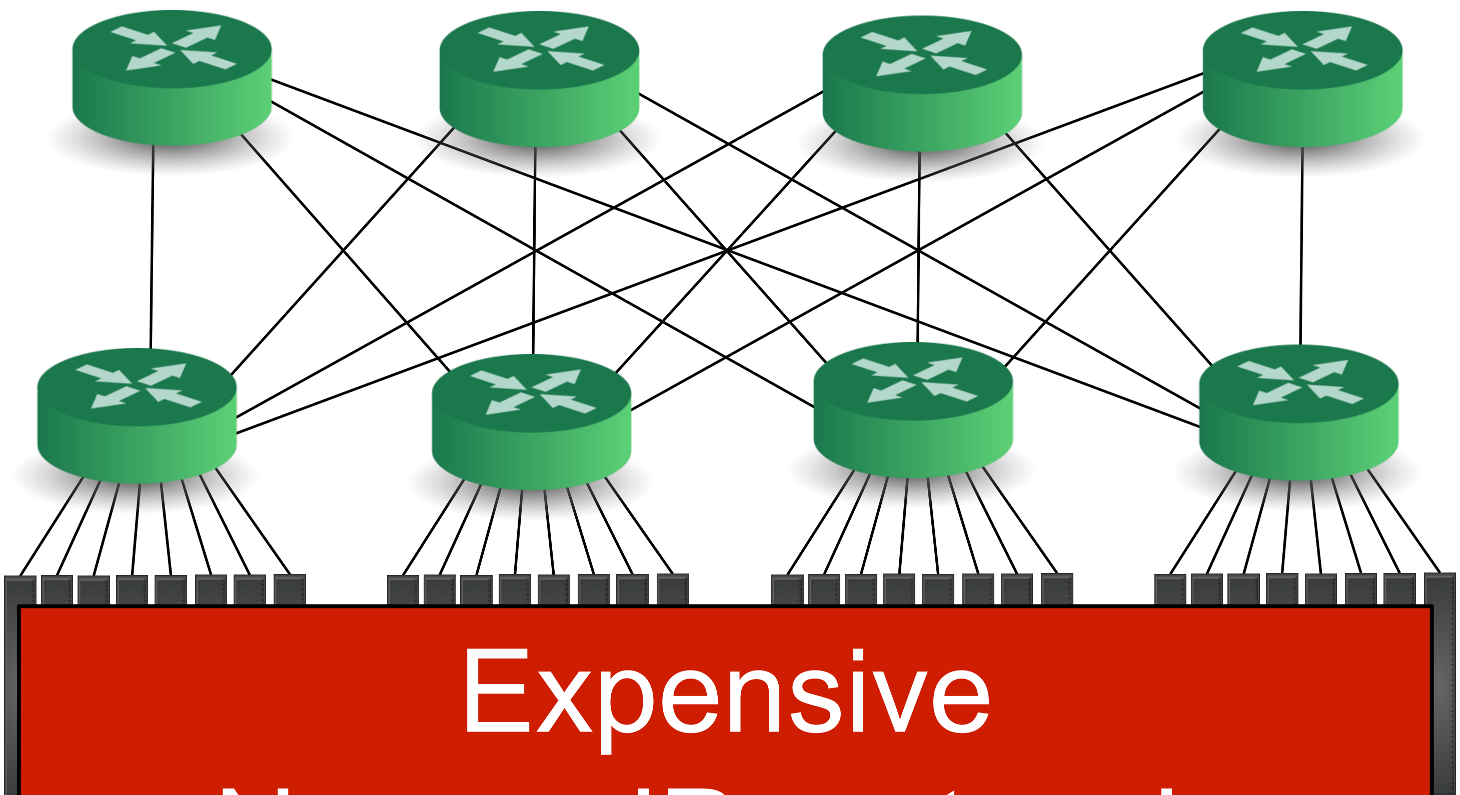
A common current approach


Expensive
Core Switch
High BW or
Multiple
Links

Proposal 2: L3

IP Subnetting
VL2 [SIGCOMM'09]

L3 routers



Expensive

No non-IP protocols

Proposal 3: Modify switches (HW/SW)

TRILL [IETF]

SEATTLE [SIGCOMM'08]

PortLand [SIGCOMM'09]

Not deployable today!

SPAIN

Unmodified L2 switches

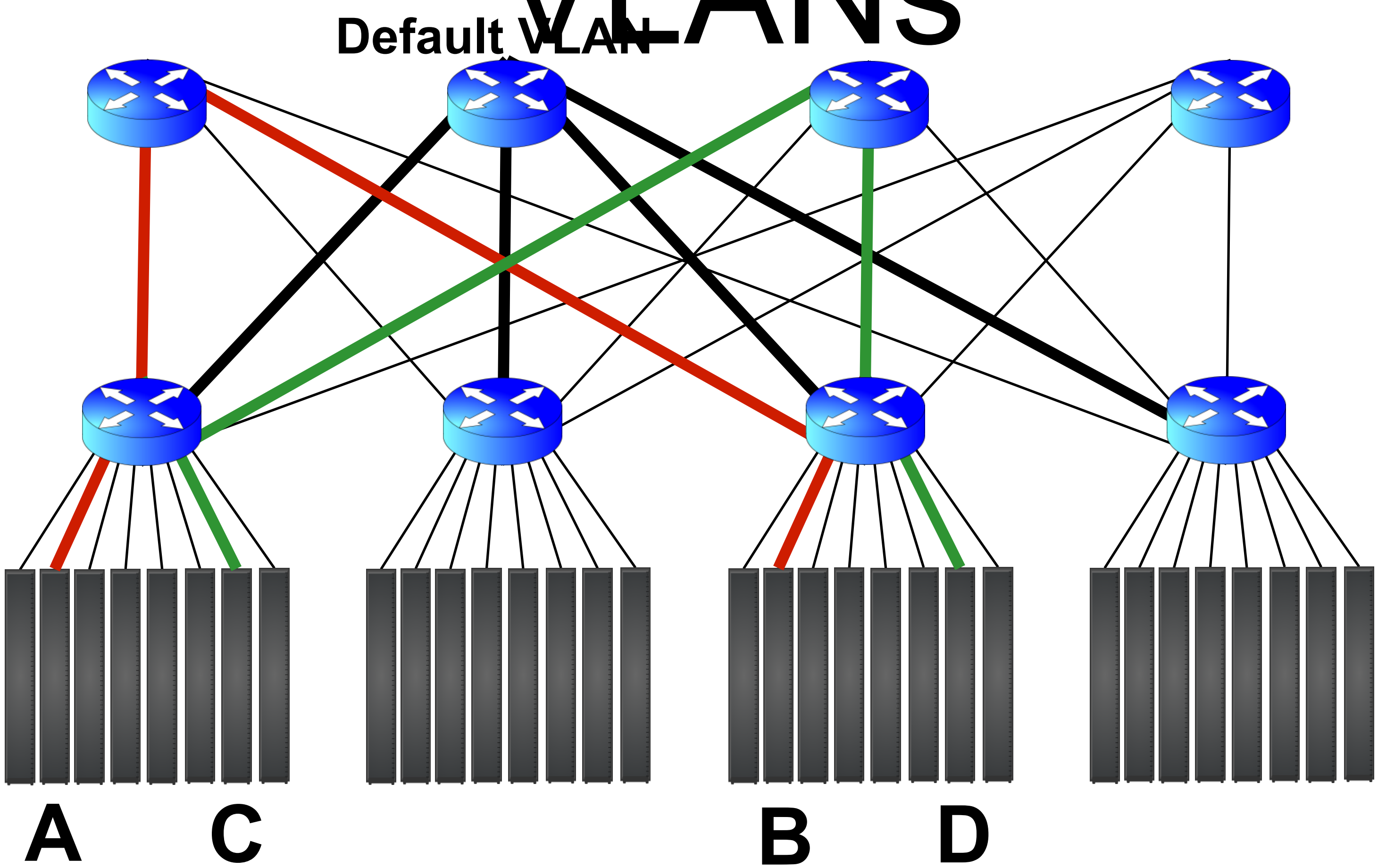
Multi-pathing

Arbitrary topologies

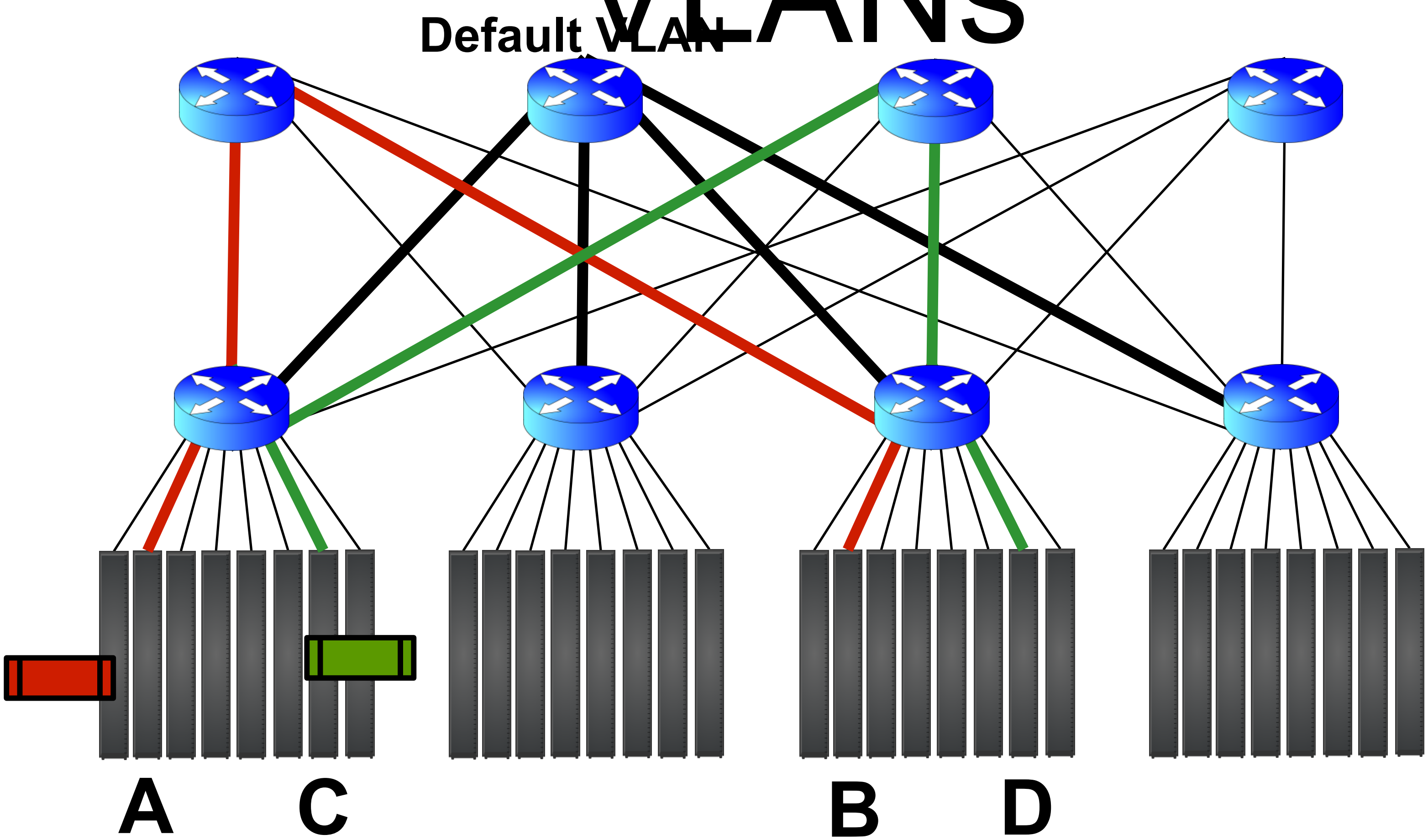
SPAIN Approach

Multi-pathing
via VLANs
+
End-host driver
to spread load

Multi-pathing via VLANs



Multi-pathing via VLANs



SPAIN

Unmodified L2 switches

Multi-pathing via VLANs

Arbitrary topologies

Minor End-host modifs

Low-cost

High-BW

DC Fabric

Today!

Outline

- Introduction
- SPAIN Components
  - Offline computation
  - End-host driver
- Evaluation
- Summary

Outline

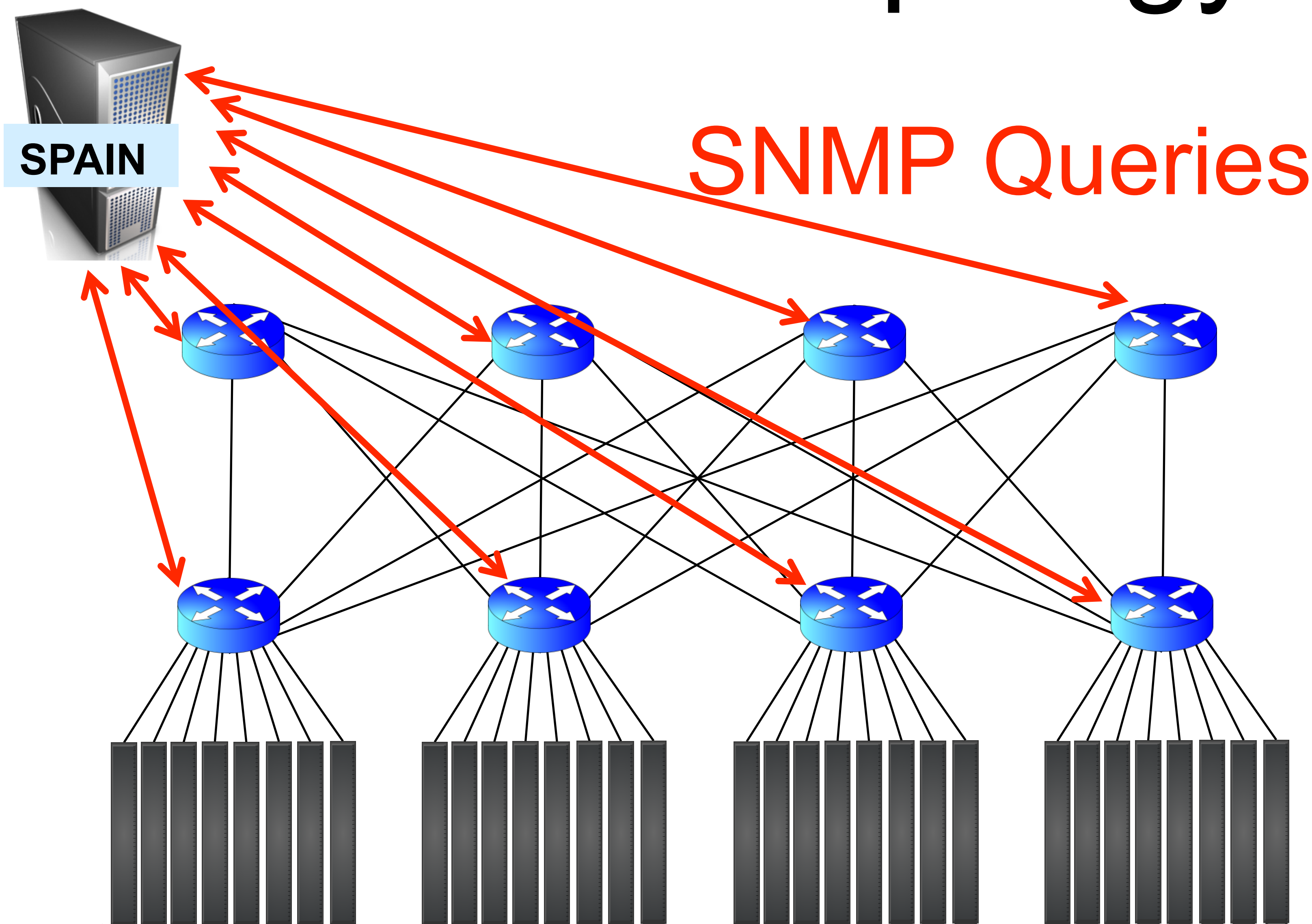
- Introduction
- SPAIN Components
  - **Offline computation**
  - End-host driver
- Evaluation
- Summary

Offline Computation

Steps:

1. Discover topology
2. Compute paths
3. Layout paths as VLANs

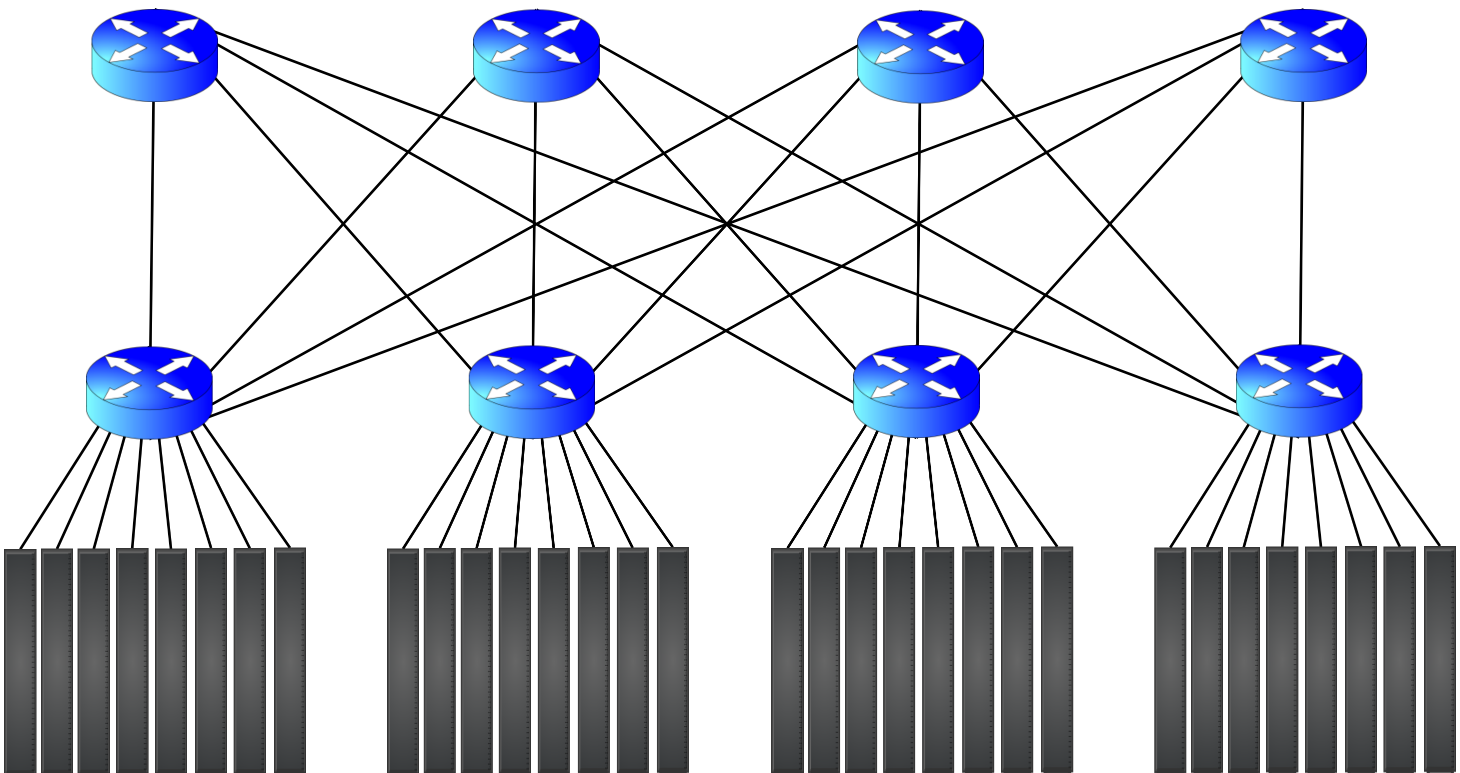
Discover topology



Compute paths

Goal: leverage redundancy; improve reliability

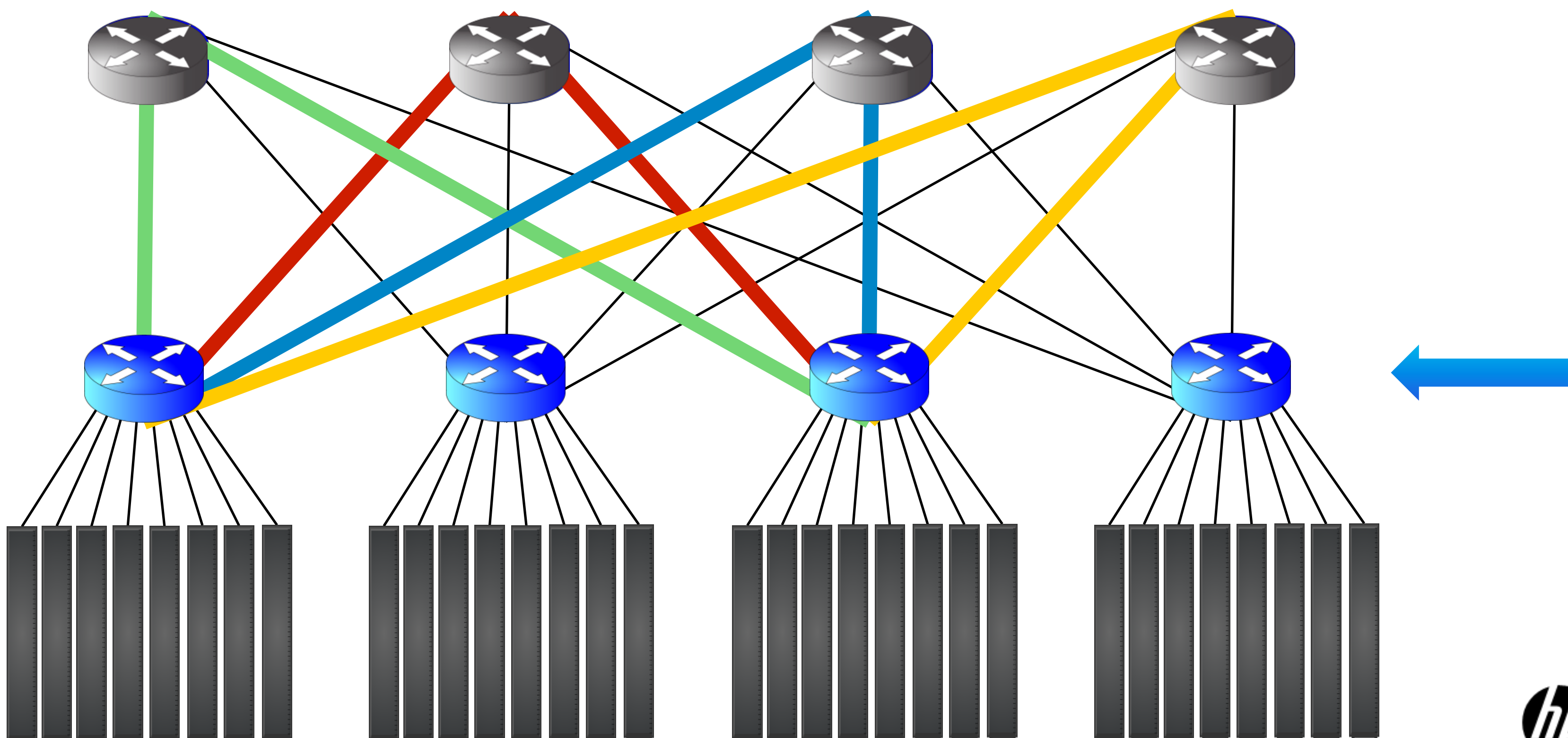
Challenges: large graphs;
more paths→more resources



Compute paths

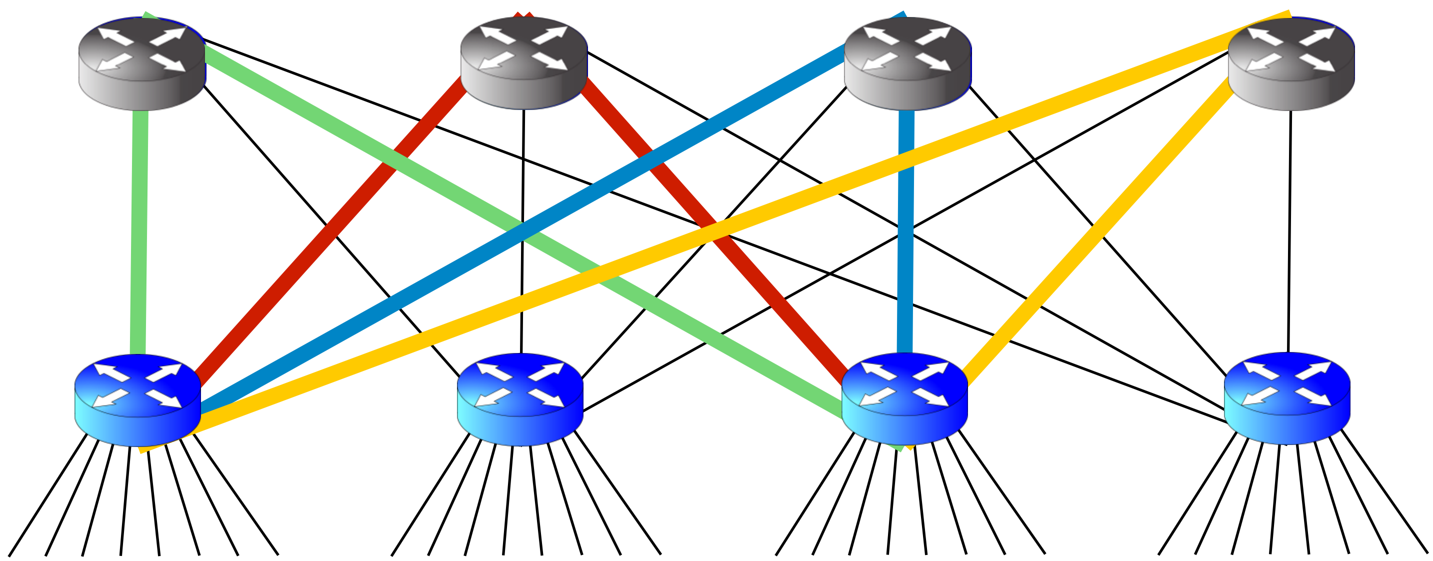
Only consider paths between edge-switches

Modified Dijkstra's; Prefer edge-disjoint paths



VLAN Layout

Simple scheme: Each Path as VLAN



But…

IEEE 802.1Q:

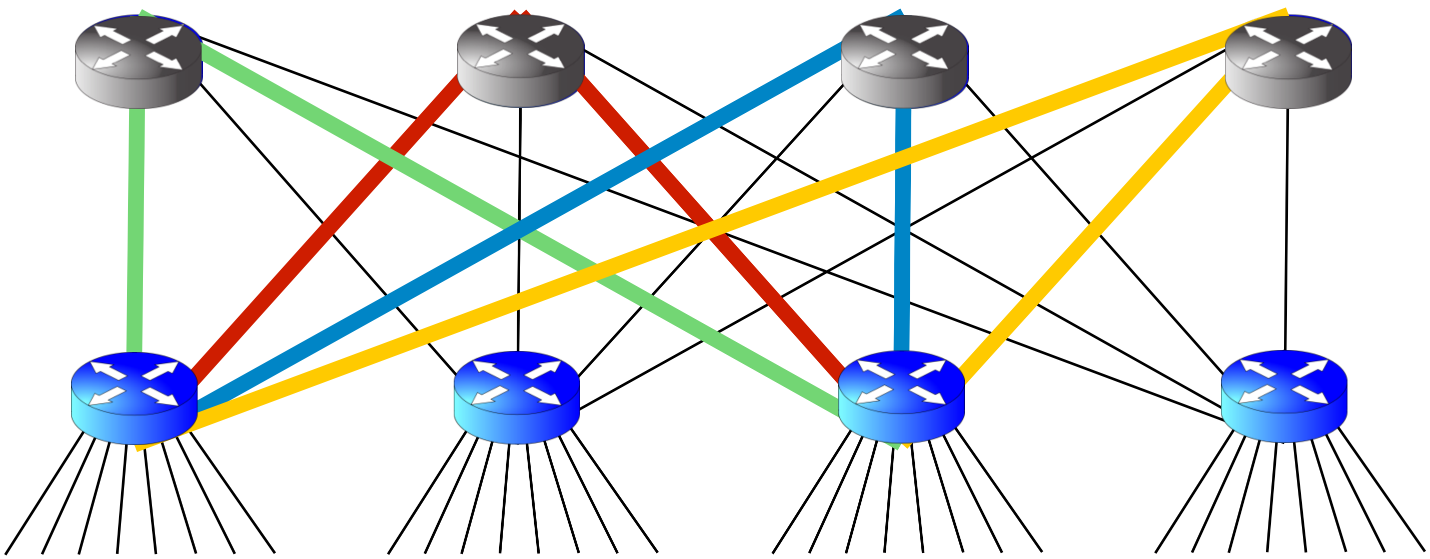
VLAN ID = 12 bits

➔ 4096 VLANs!

VLAN Layout

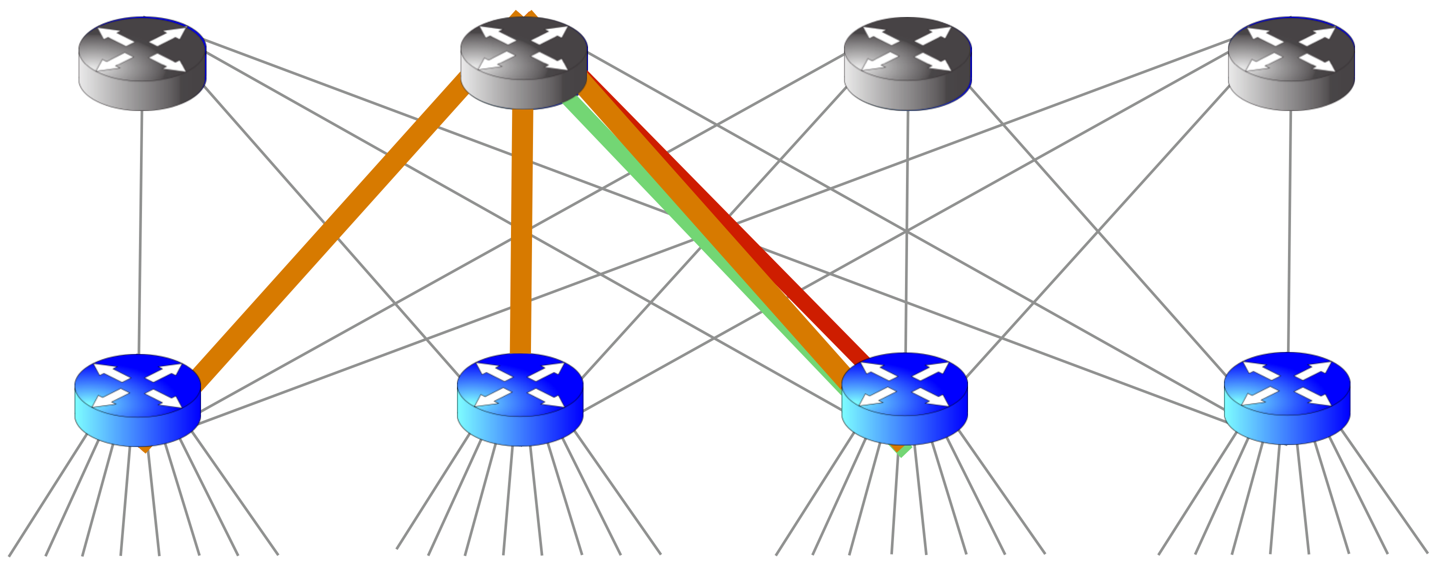
Simple scheme: Each Path as VLAN

Scales to only few switches



VLAN Layout

Our approach: 1 VLAN for a set of paths



Challenge:
Minimize VLANs

NP-Hard for
arbitrary topologies

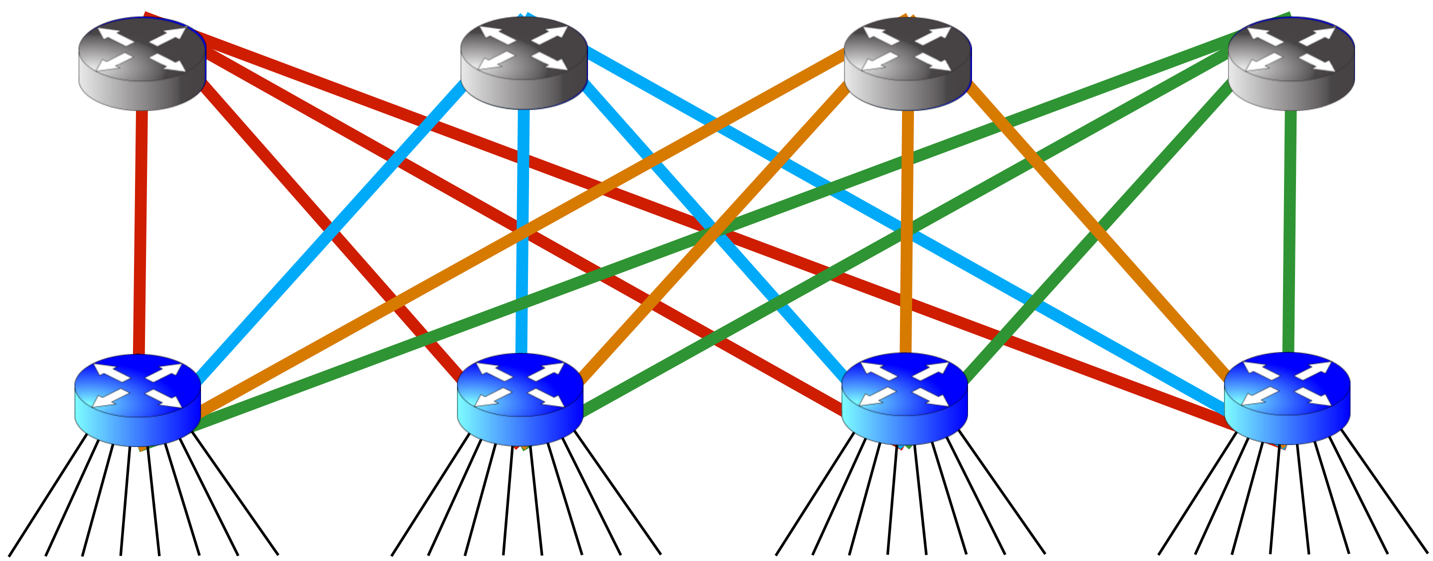
VLAN Layout

Heuristics:

1. Greedy path packing
2. Parallel graph-coloring

VLAN Layout

VLANs = 4



Outline

Introduction

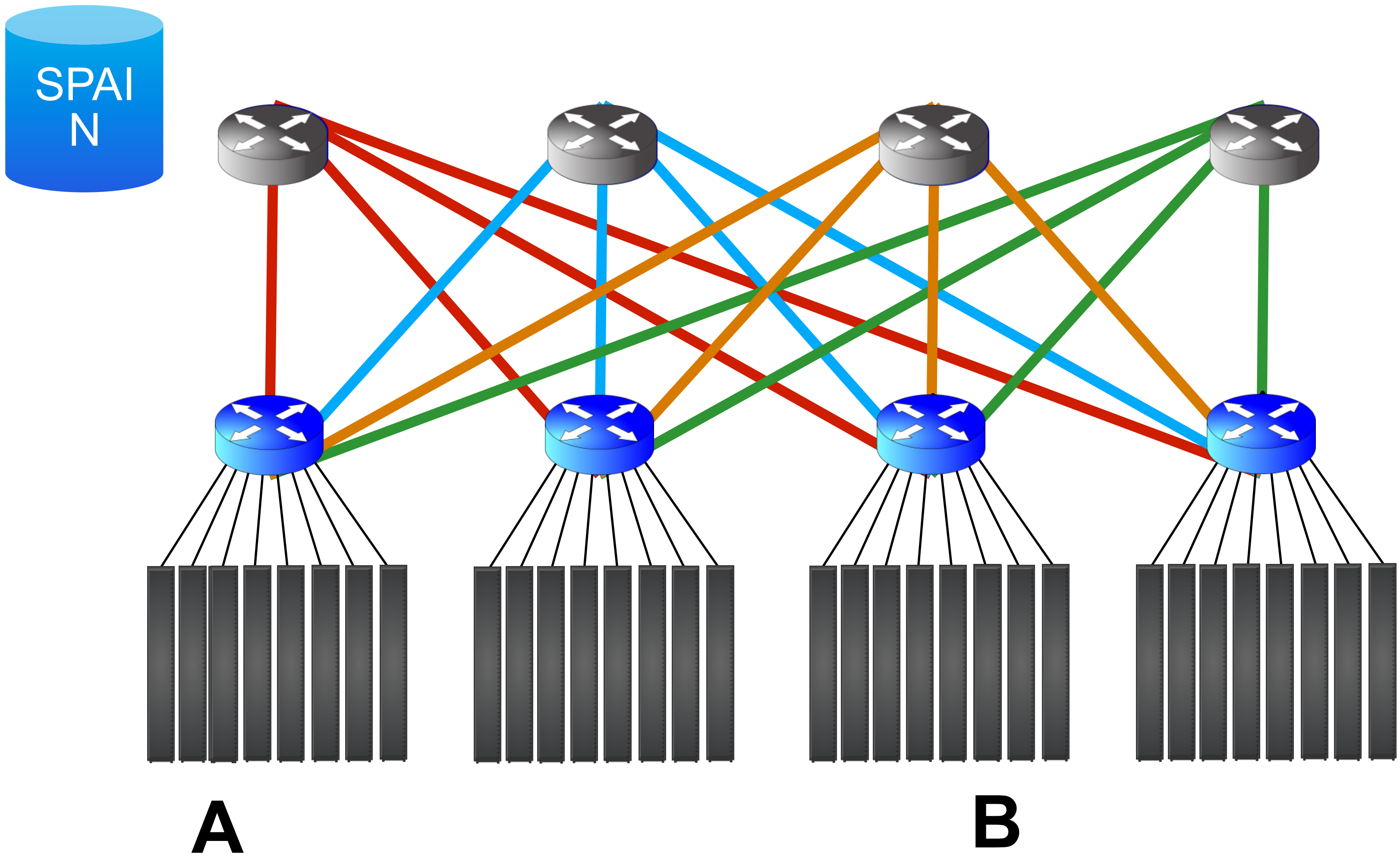
SPAIN Components

- Offline computation
- **End-host driver**

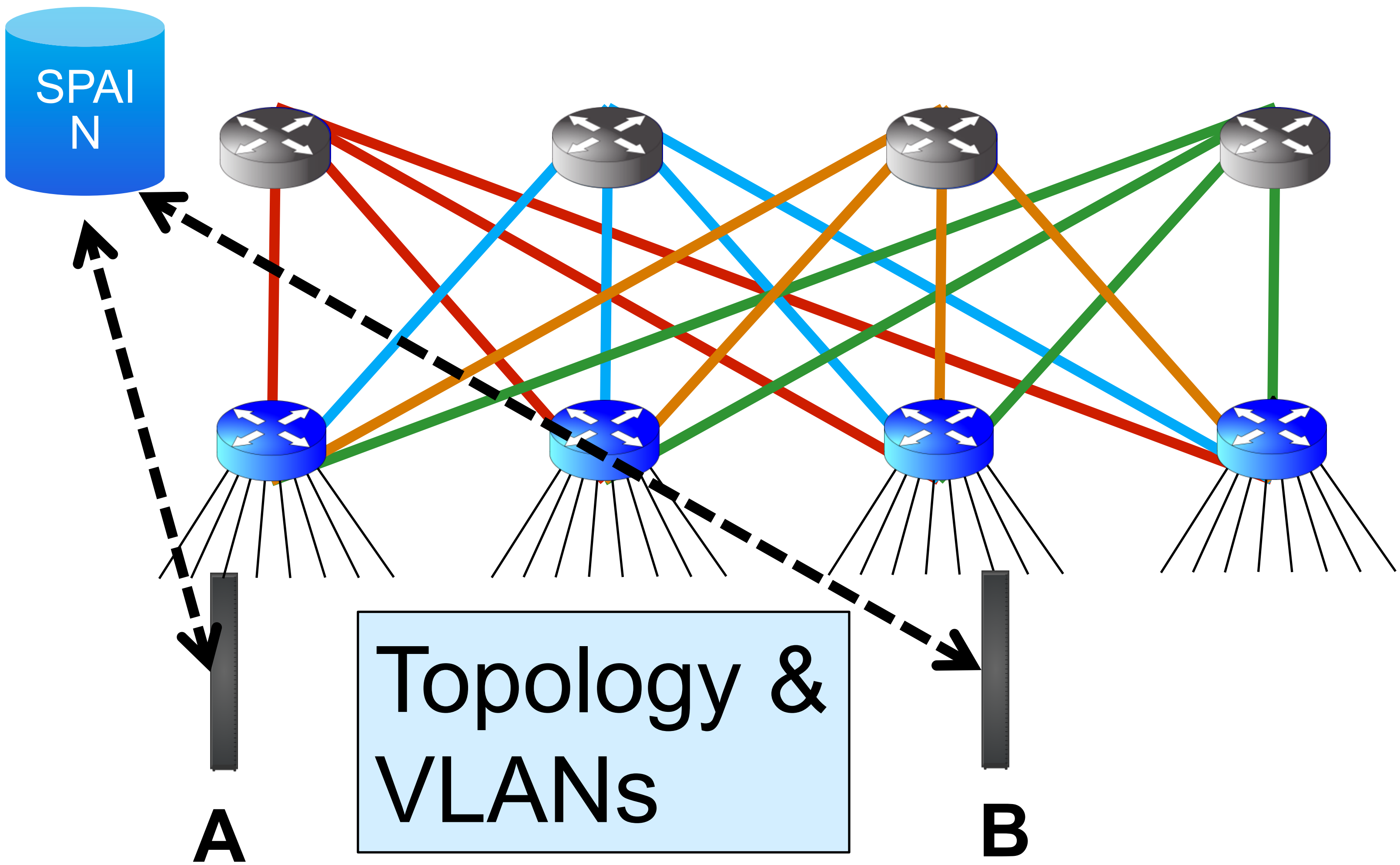
Evaluation

Summary

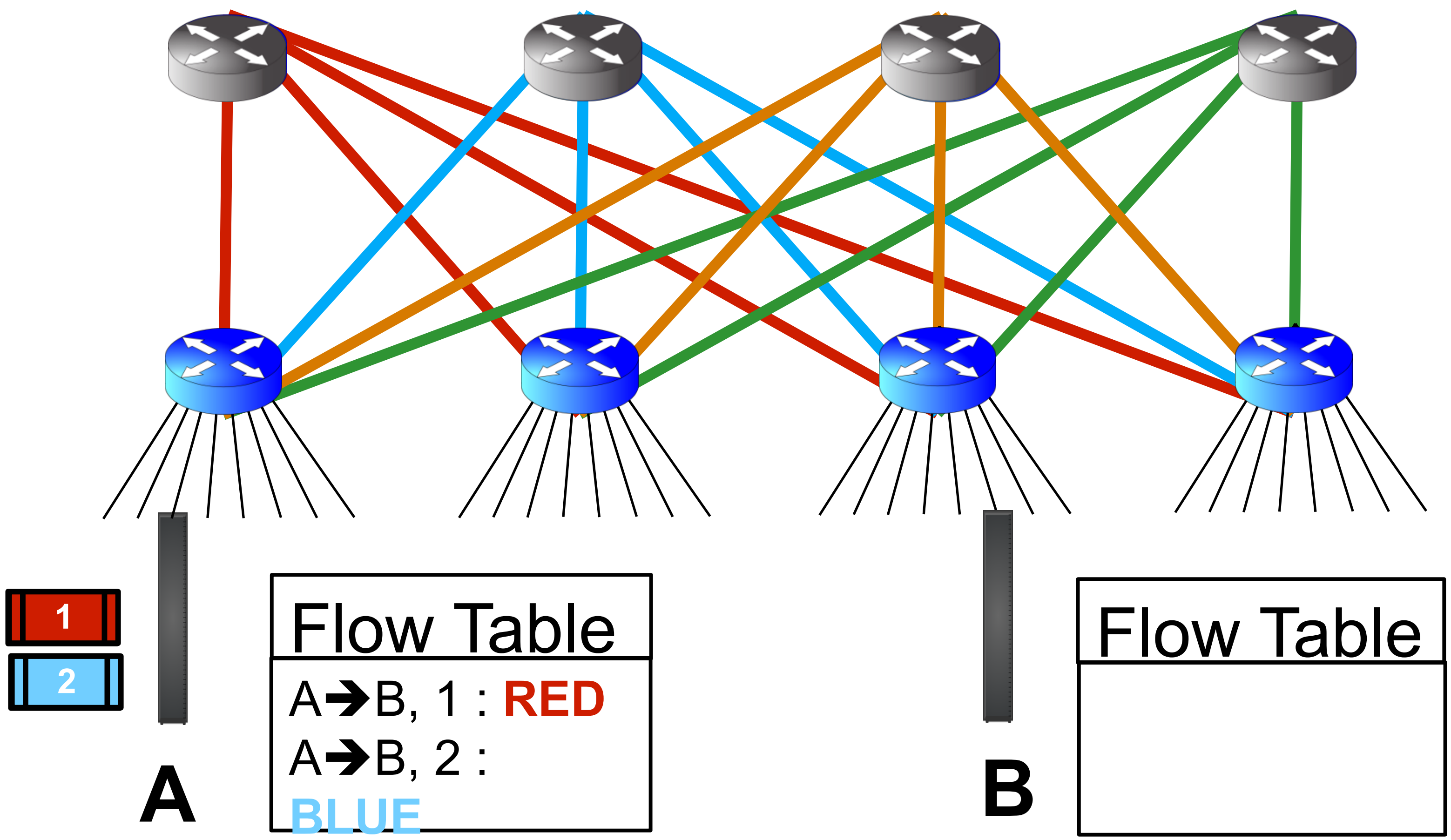
SPAIN End-host Driver



SPAIN End-host Driver



SPAIN End-host Driver



Challenges

Link & switch failures

Pathological flooding

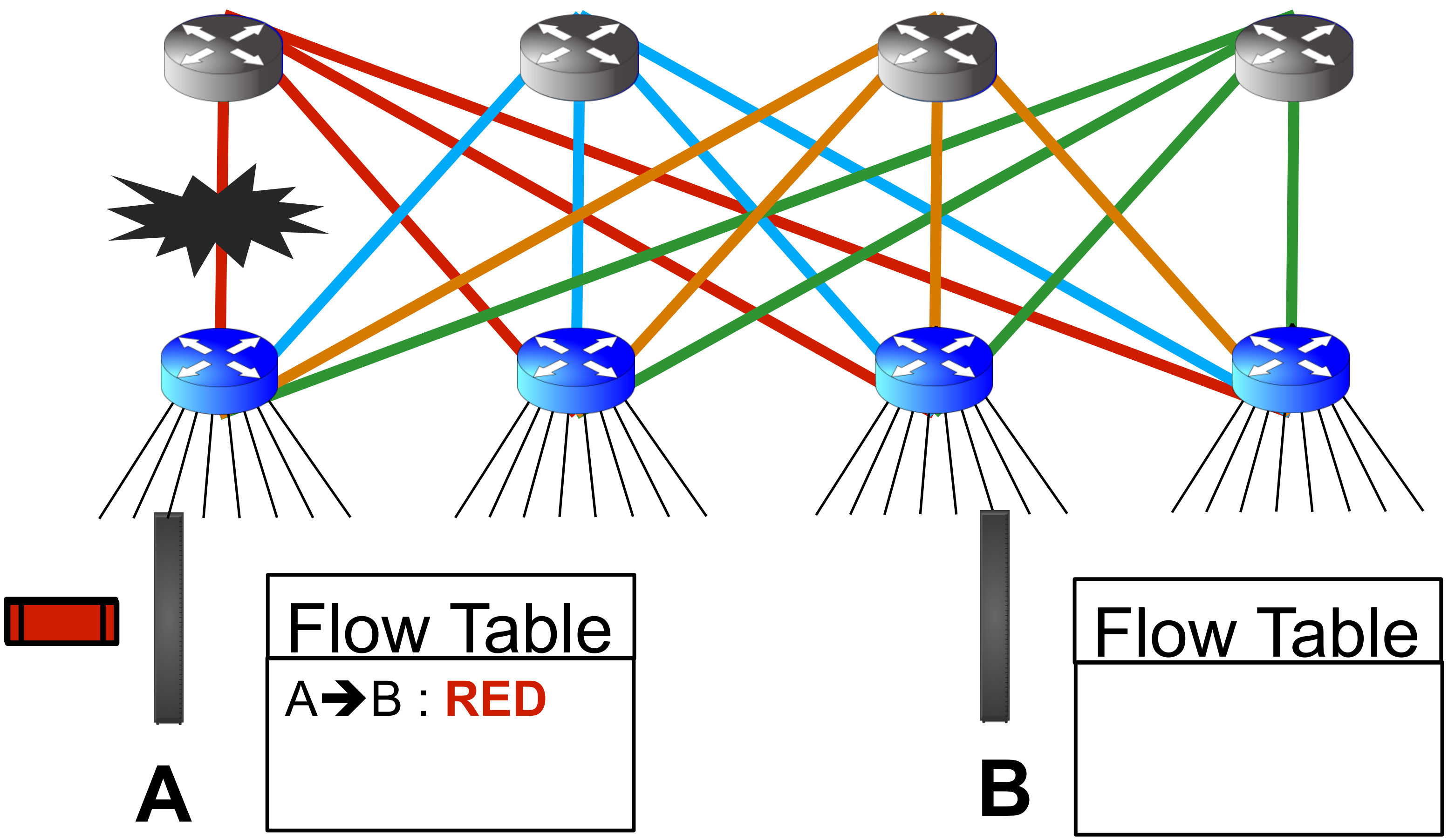
Interoperability

Host mobility

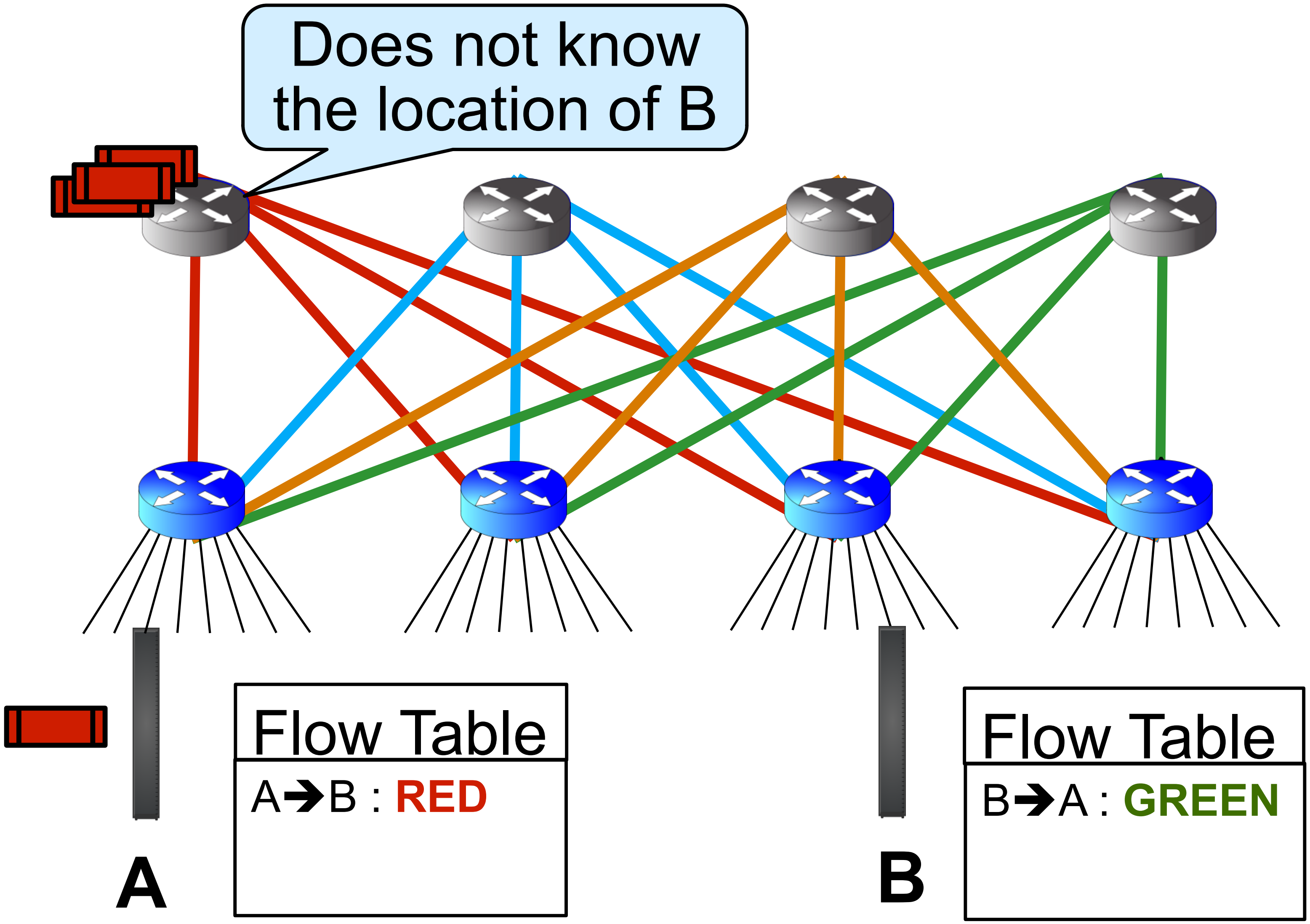
Load-balance

End-host state

Failures



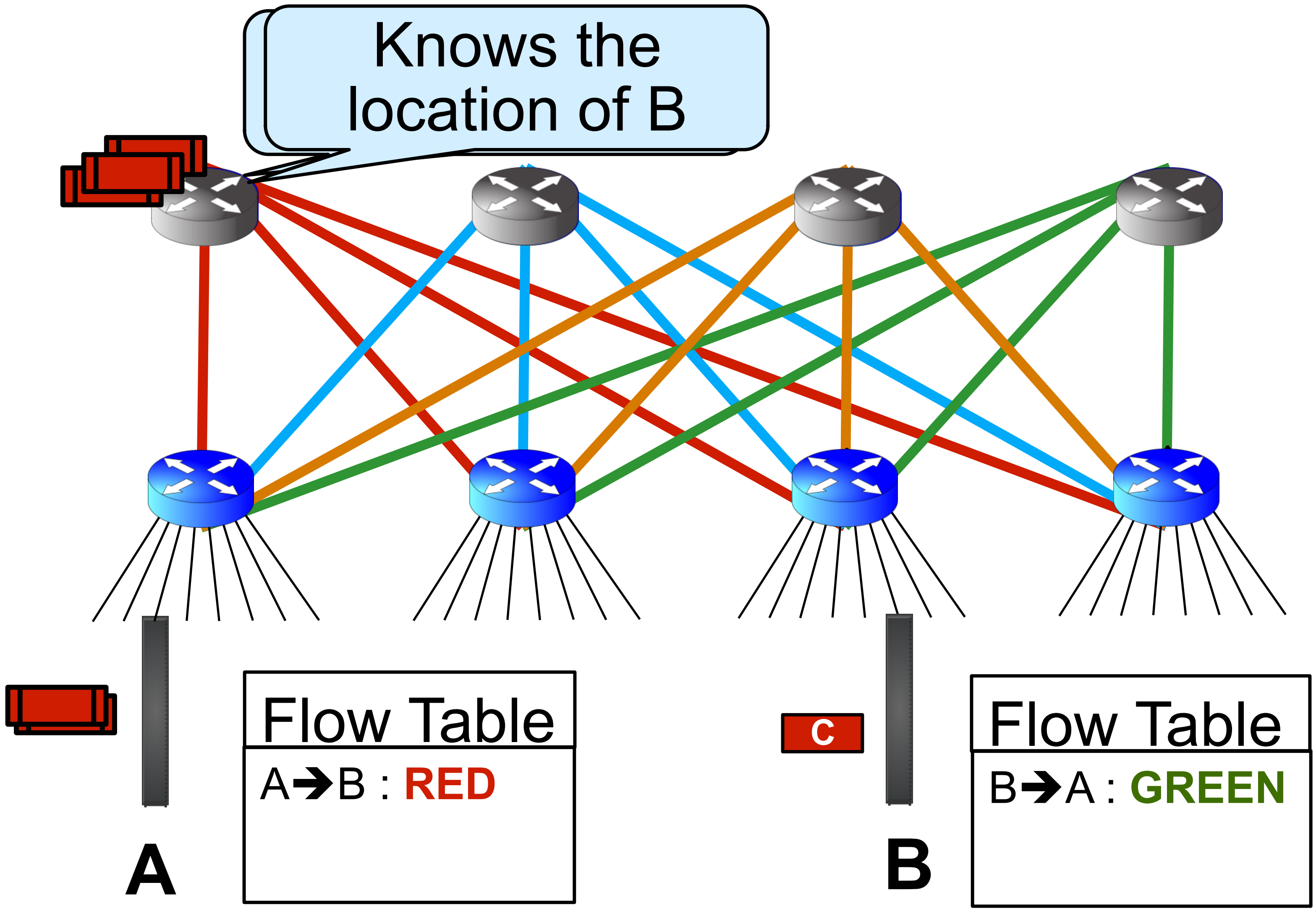
Pathological Flooding



Solution:

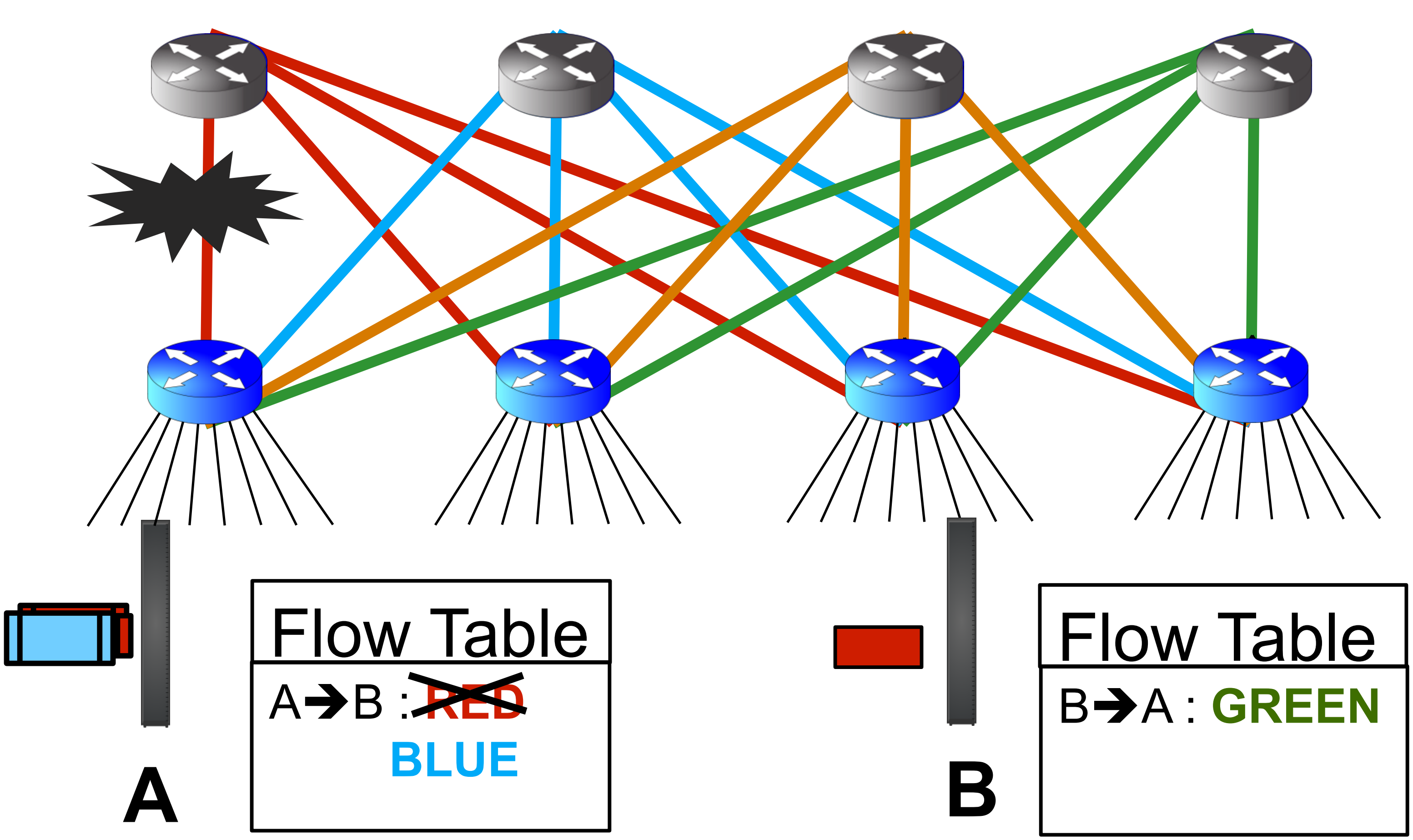
Chirping

Chirping



Knows the location of B
Flow Table
A→B : RED
C
Flow Table
B→A : GREEN
A
B

Chirping



Outline

- Introduction
- SPAIN Components
  - Offline computation
  - End-host driver
- **Evaluation**
- Summary

Evaluation

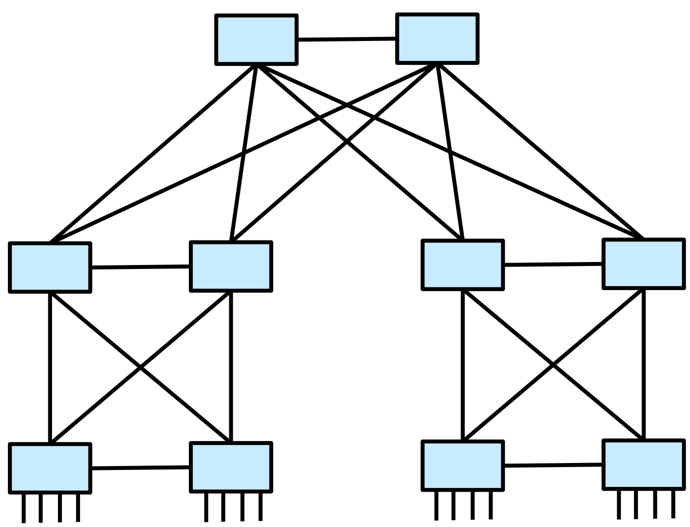
Simulations

Real testbed

Simulations

<u>Topologies</u>:

CiscoDC



Core switches

Aggregation modules
m = 2

Access switches per module
a = 2

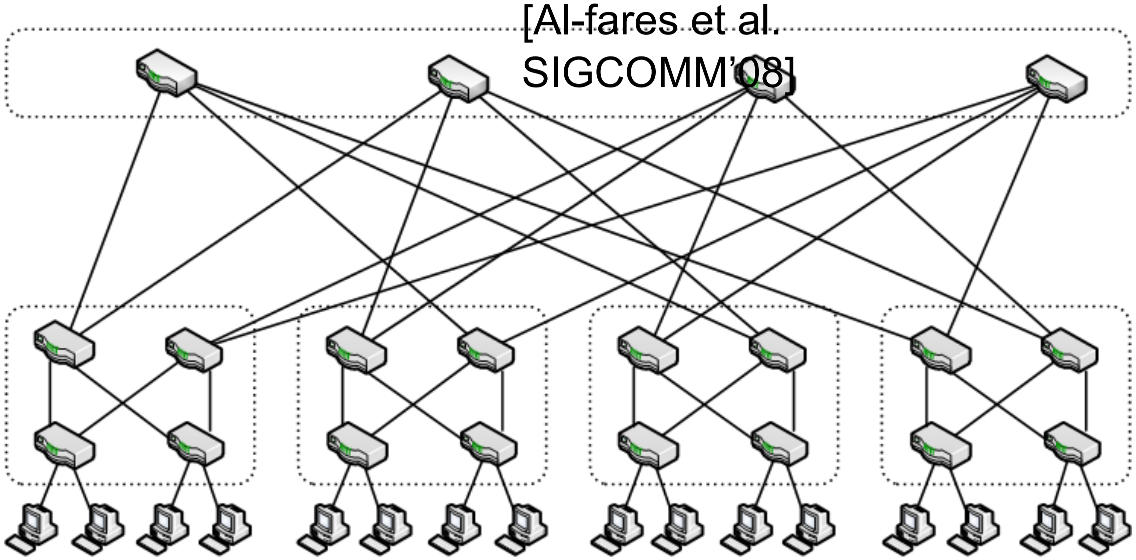
Simulations

Topologies:

CiscoDC

Fat-Tree

[Al-fares et al. SIGCOMM'08]



#ports/switch

p = 4

Simulations

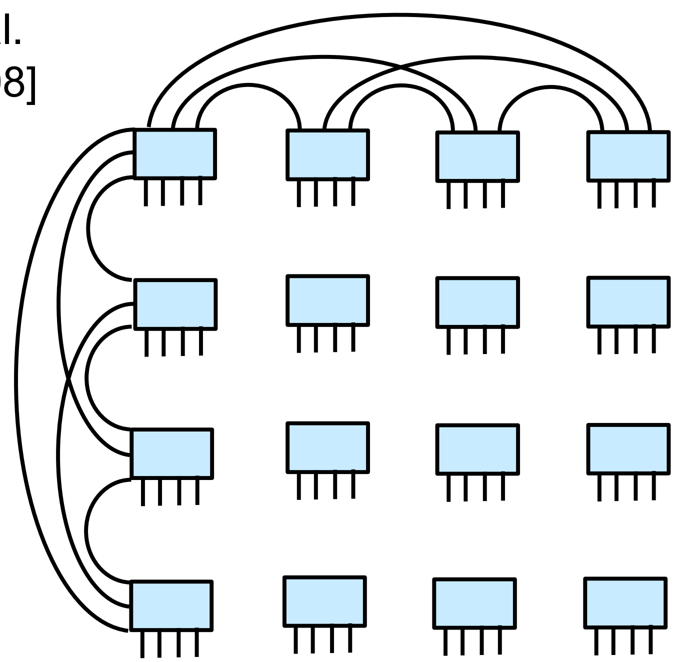
Topologies:

CiscoDC

Fat-Tree
[Al-fares et al. SIGCOMM'08]

HyperX
[Ahn et al. SC'09]

2D HyperX
k=4



Simulations

<u>Topologies</u>:

CiscoDC

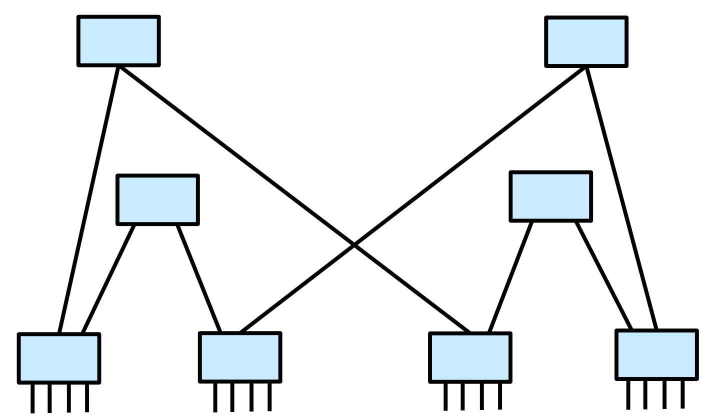
Fat-Tree
[Al-fares et al. SIGCOMM'08]

HyperX
[Ahn et al. SC'09]

B-Cube
[Guo et al. SIGCOMM'09]

#ports/switch (p) = 2

Levels (l) = 2



Simulations

Topologies:

CiscoDC

Fat-Tree
[Al-fares et al. SIGCOMM'08]

HyperX
[Ahn et al. SC'09]

B-Cube
[Guo et al. SIGCOMM'09]

Metrics:

#VLANs

Link-Coverage

Reliability

Throughput

Simulations

Topologies:

CiscoDC

Fat-Tree [Al-fares et al. SIGCOMM'08]

HyperX [Ahn et al. SC'09]

B-Cube [Guo et al. SIGCOMM'09]

Metrics:

#VLANs

Link-Coverage

Reliability

Throughput

Num. of VLANs

	#switches	#VLANs
CiscoDC (8,8)	146	38
Fat-Tree (48)	2880	576
HyperX (16)	256	971
B-Cube (48,2)	2048	2048

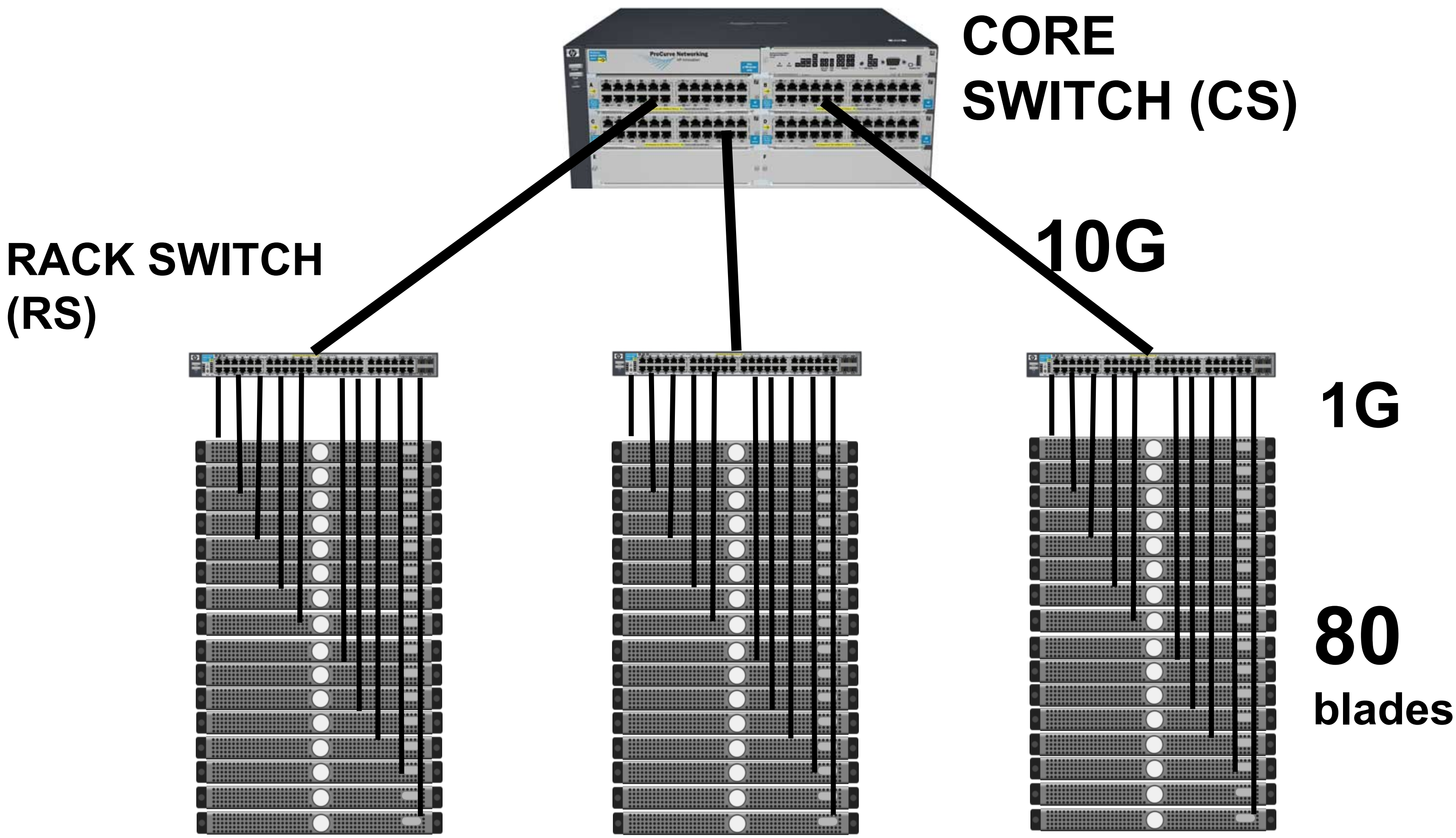
Throughput

CiscoDC	2x
Fat-Tree	24x
HyperX	10.5x
B-Cube	1.6x

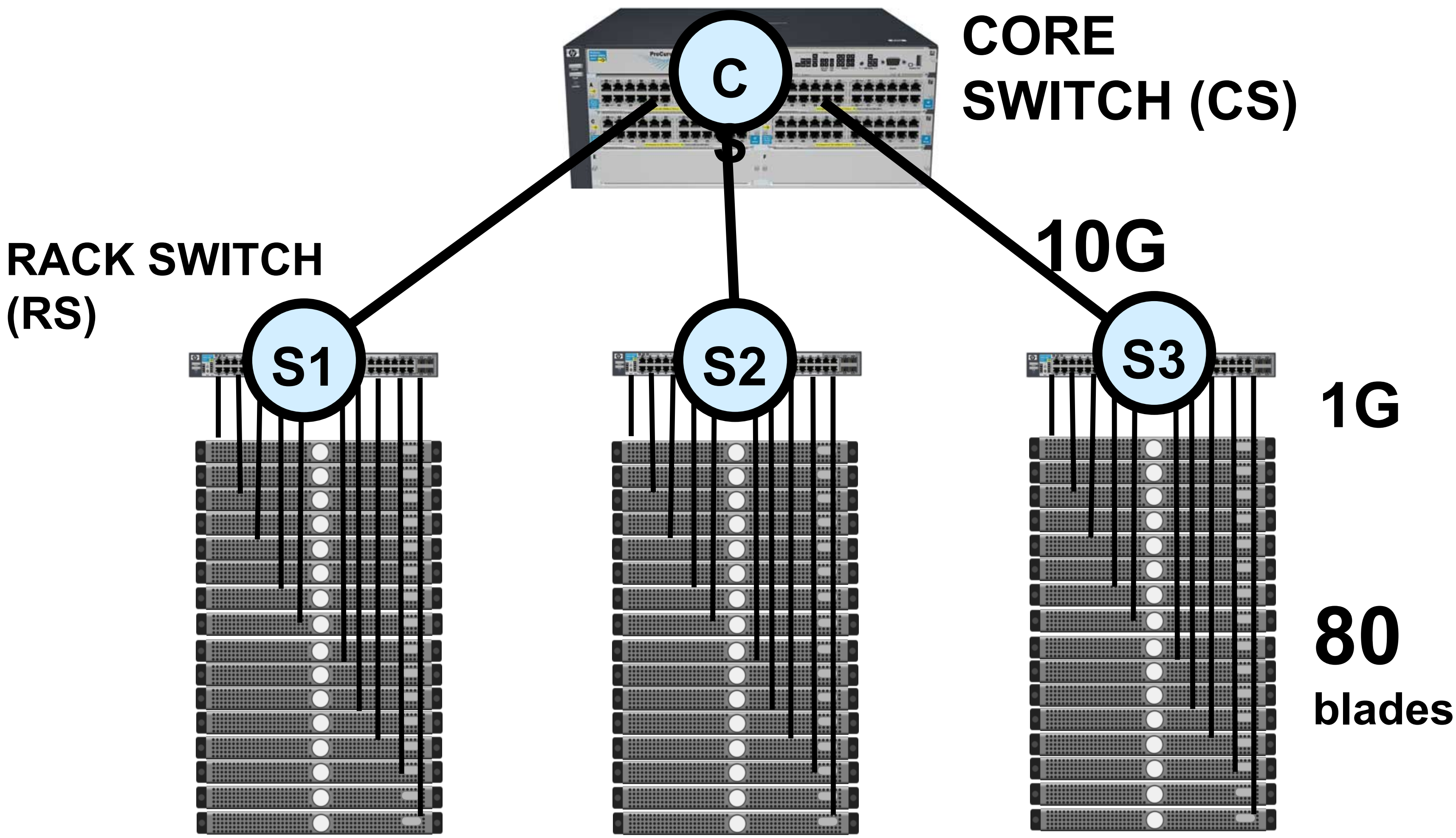
Improvement over STP

OpenCirrus Experiments

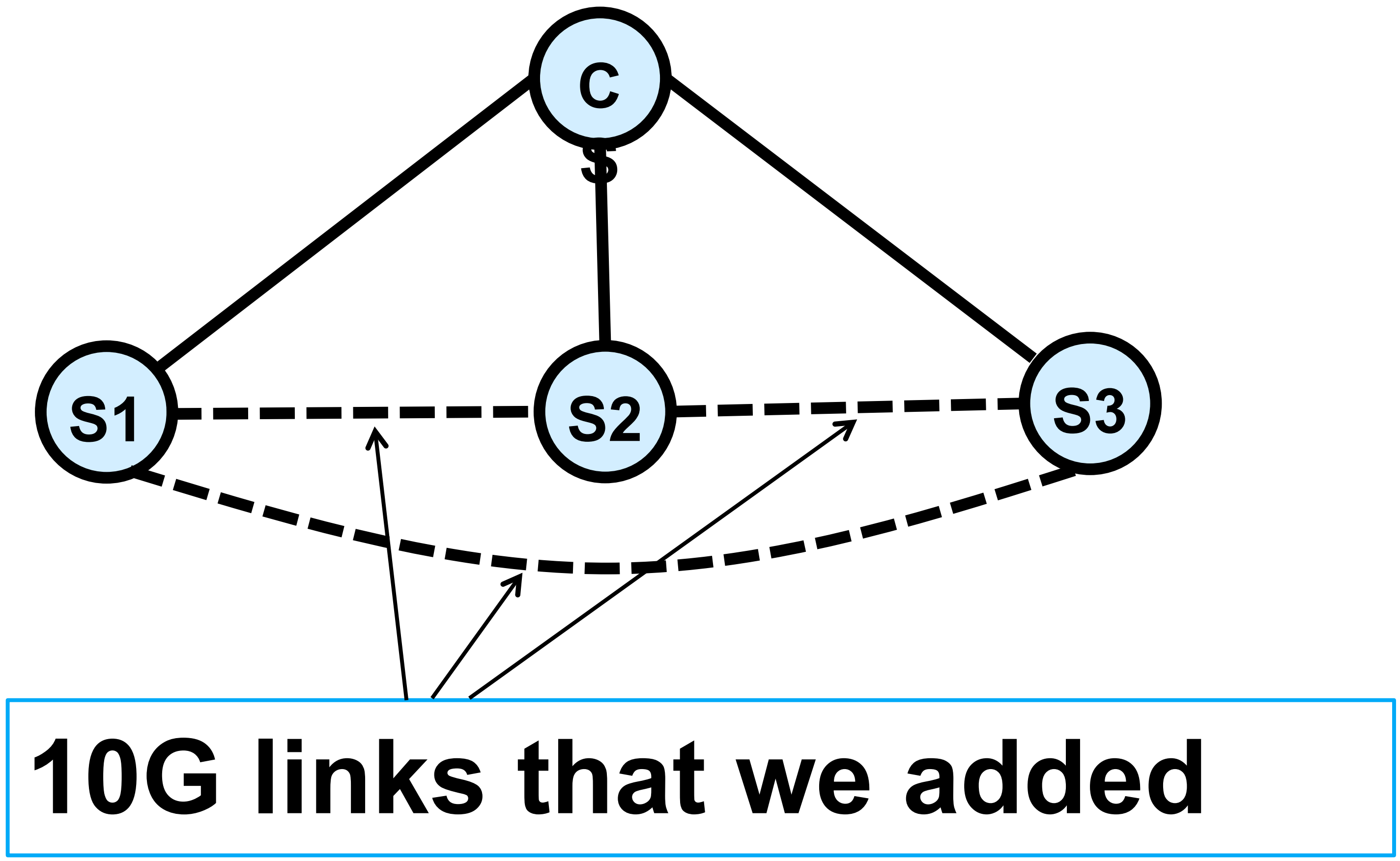
OpenCirrus Testbed



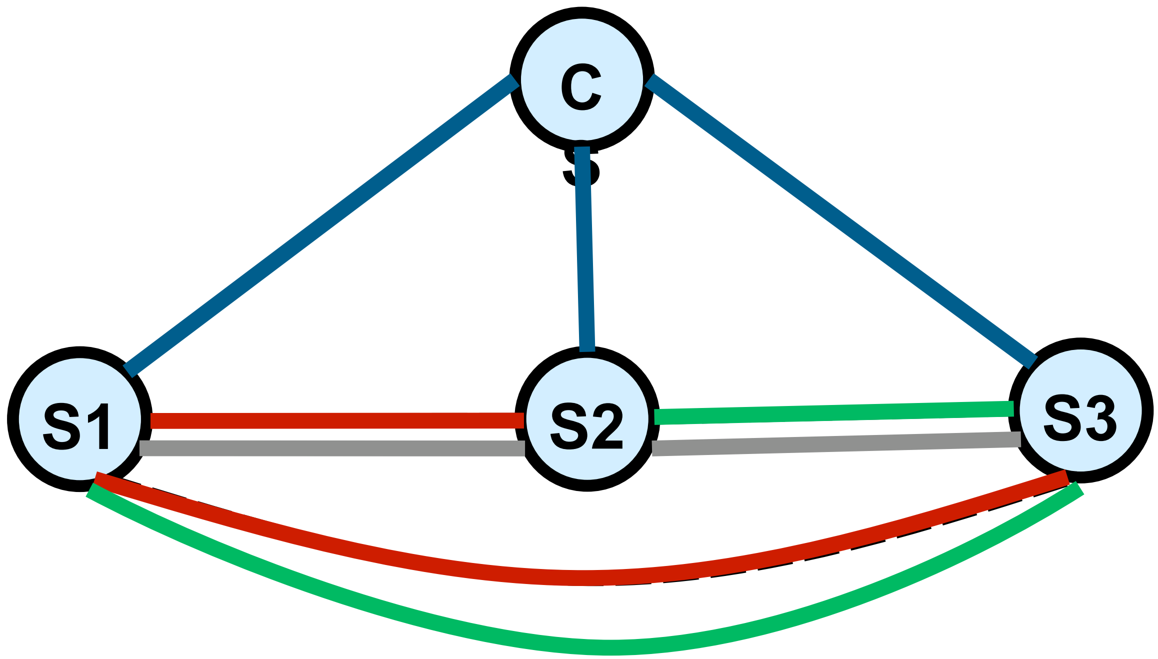
OpenCirrus Testbed



OpenCirrus Testbed



OpenCirrus Testbed



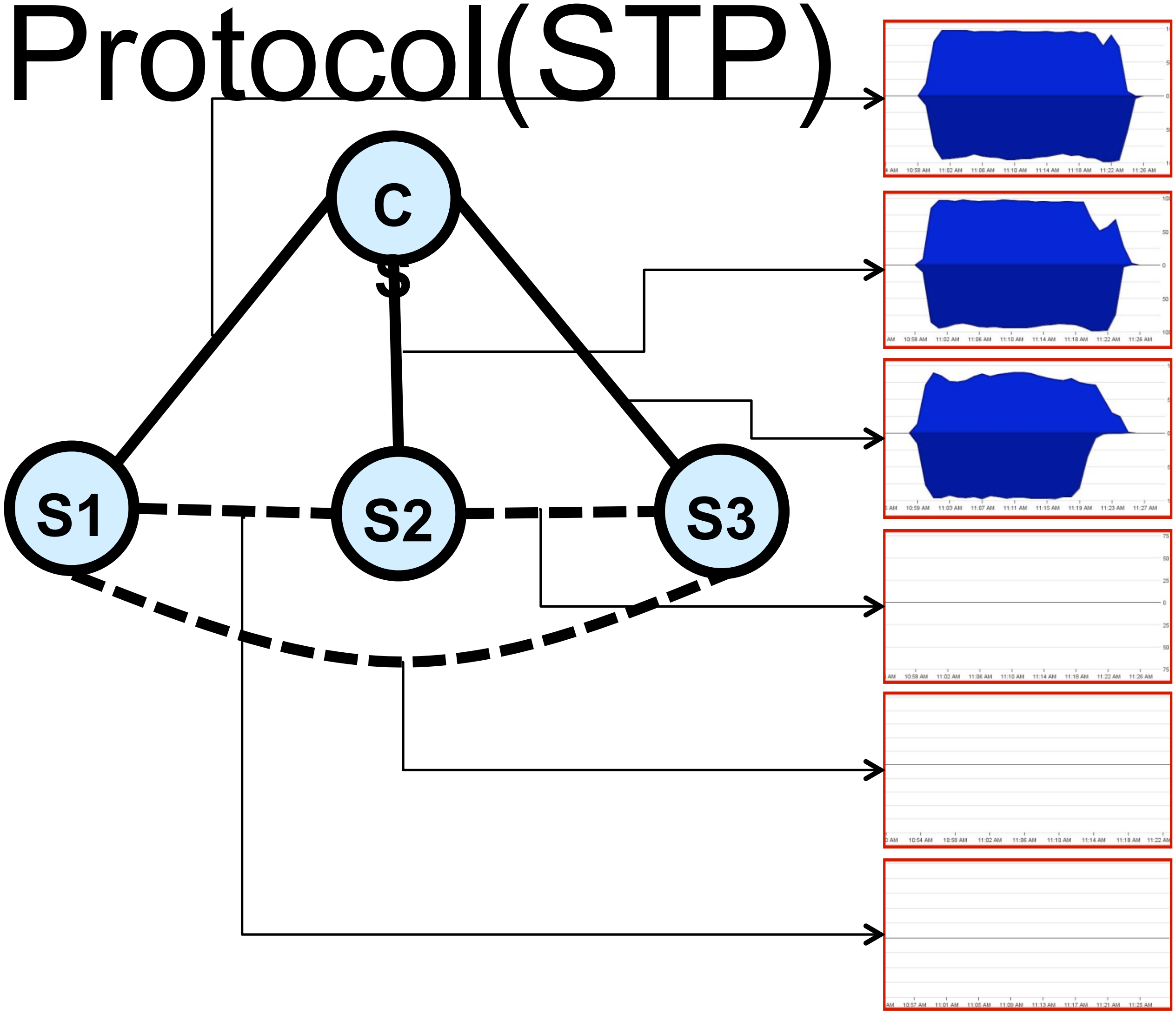
4 VLANs

Shuffle-like experiment

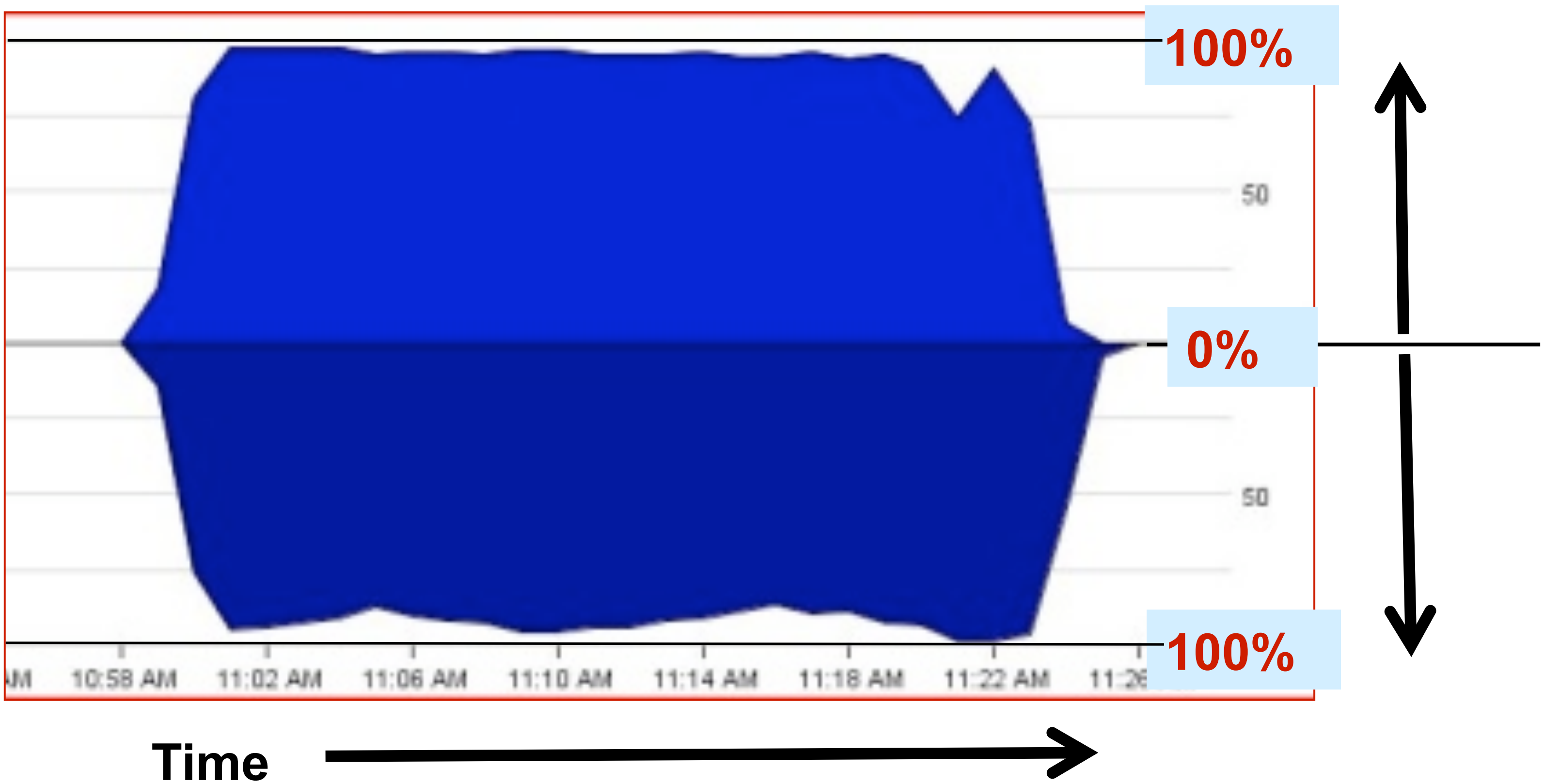
Every server to all other servers

500MB data transfer

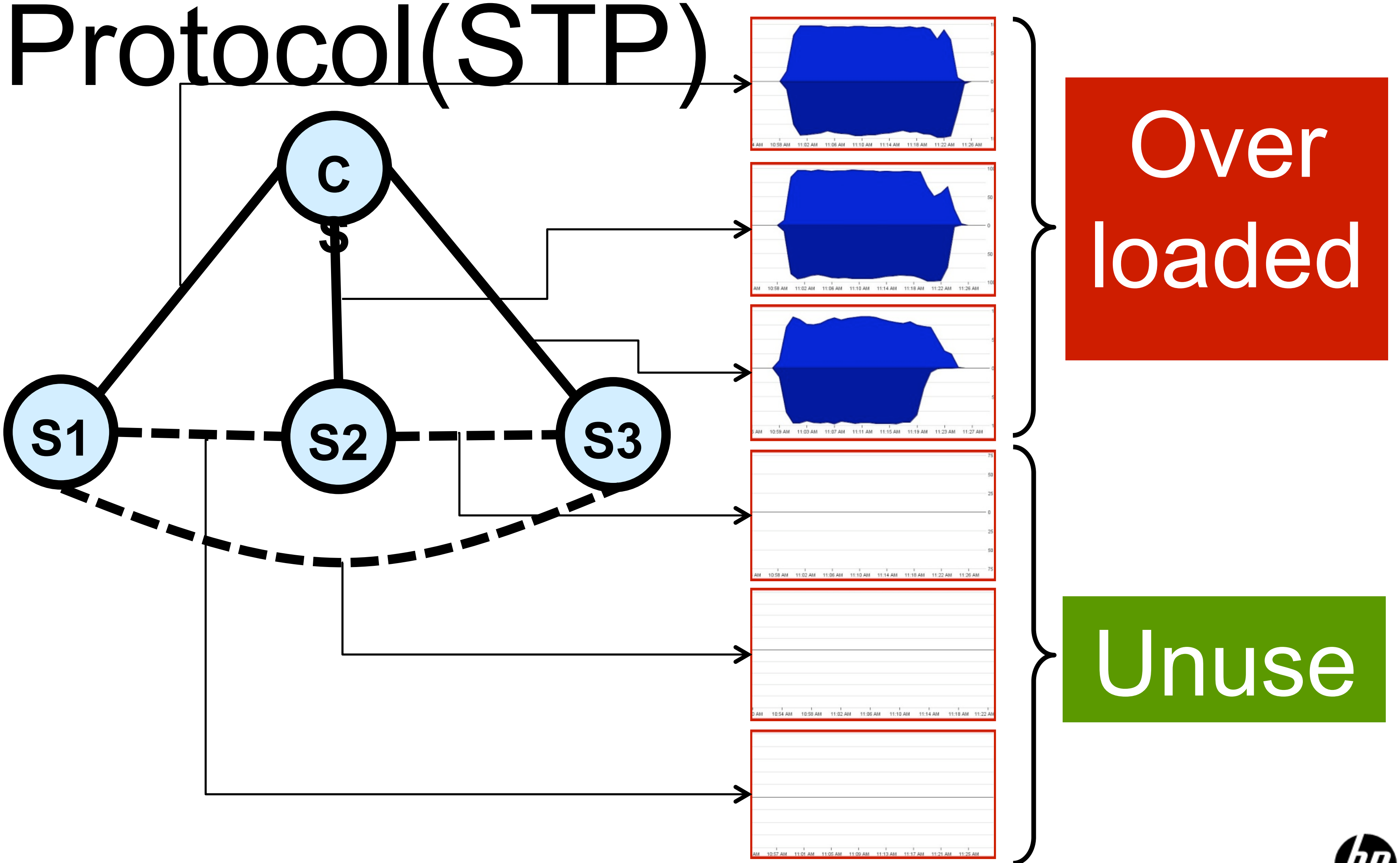
Spanning Tree Protocol(STP)



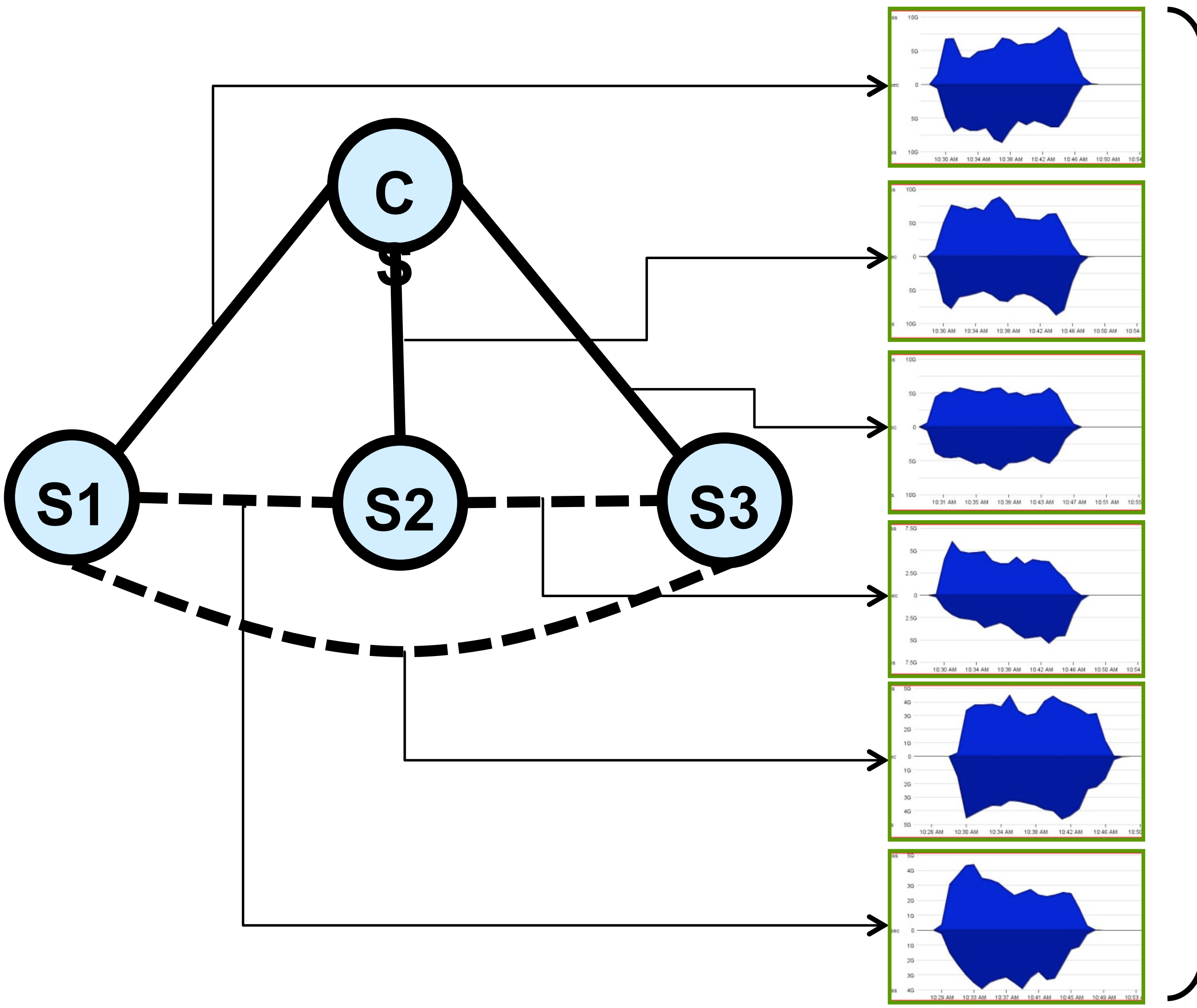
Link utilization in each direction



Spanning Tree Protocol(STP)

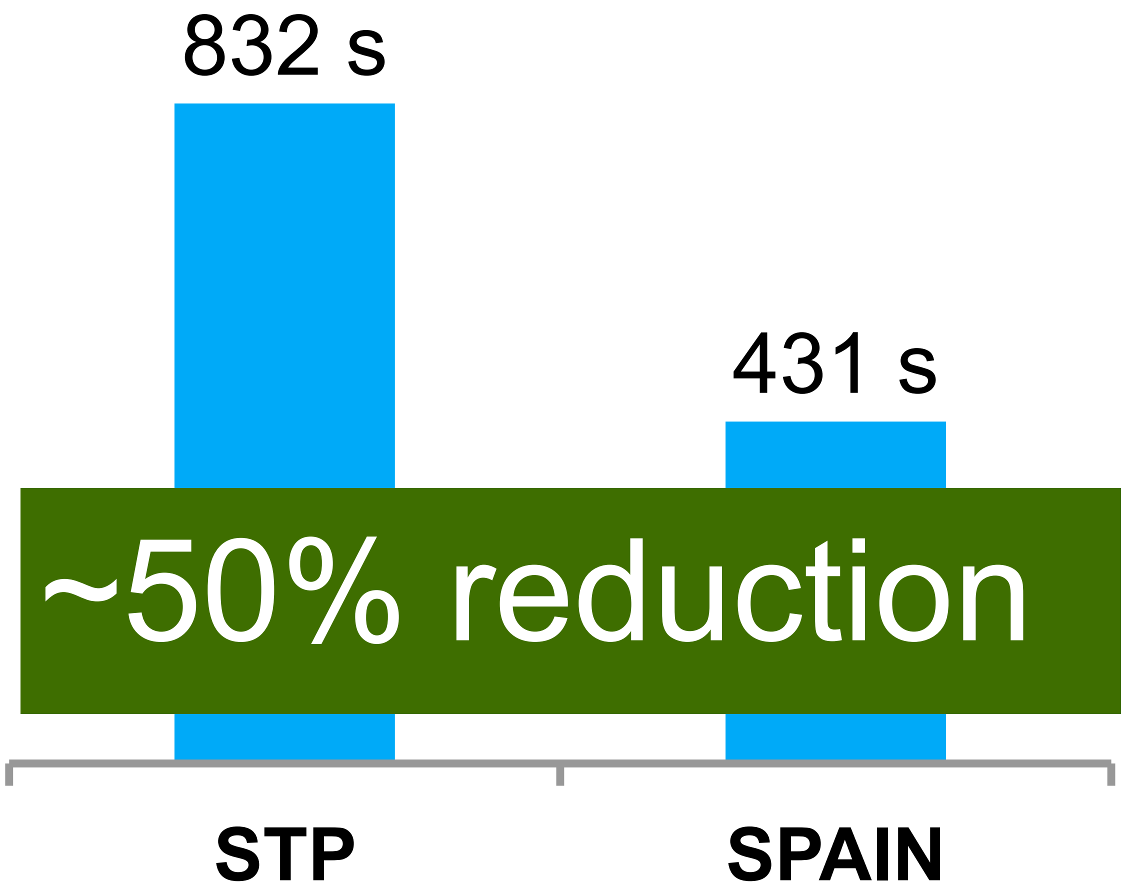


SPAIN

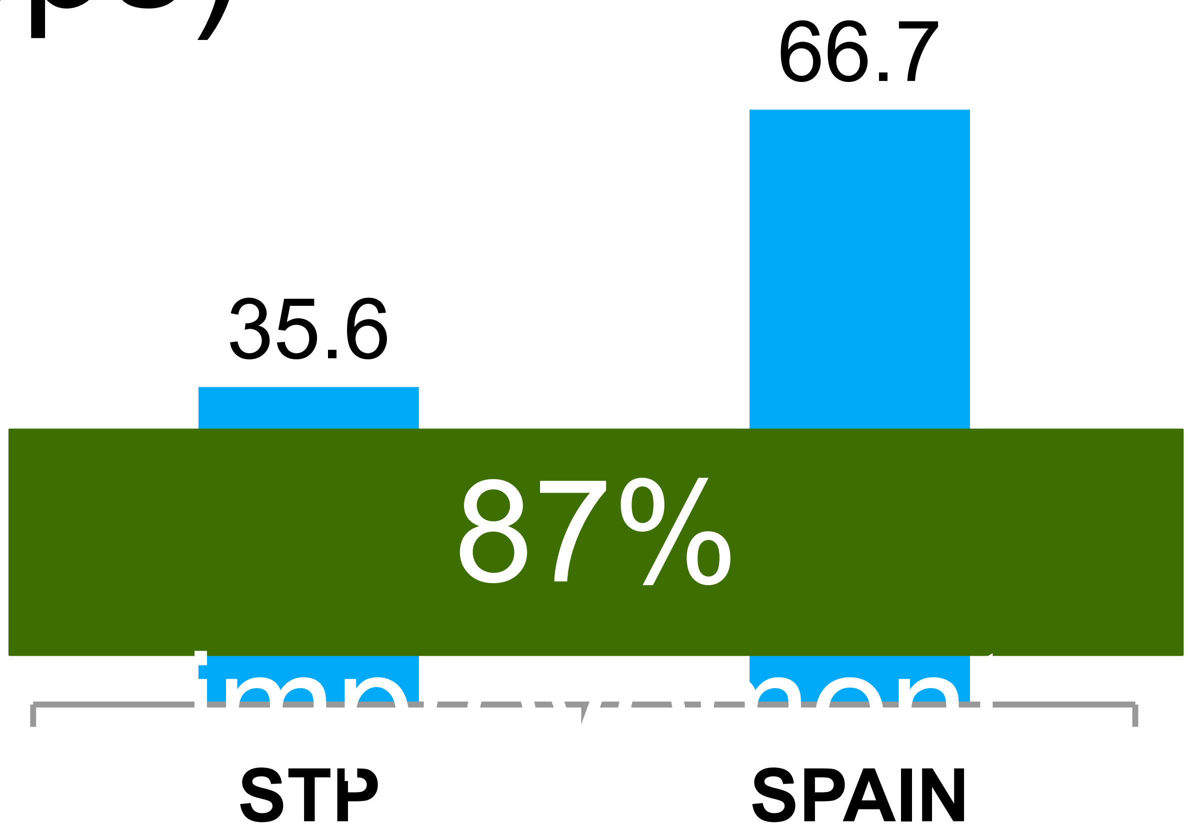


No bottle-necks

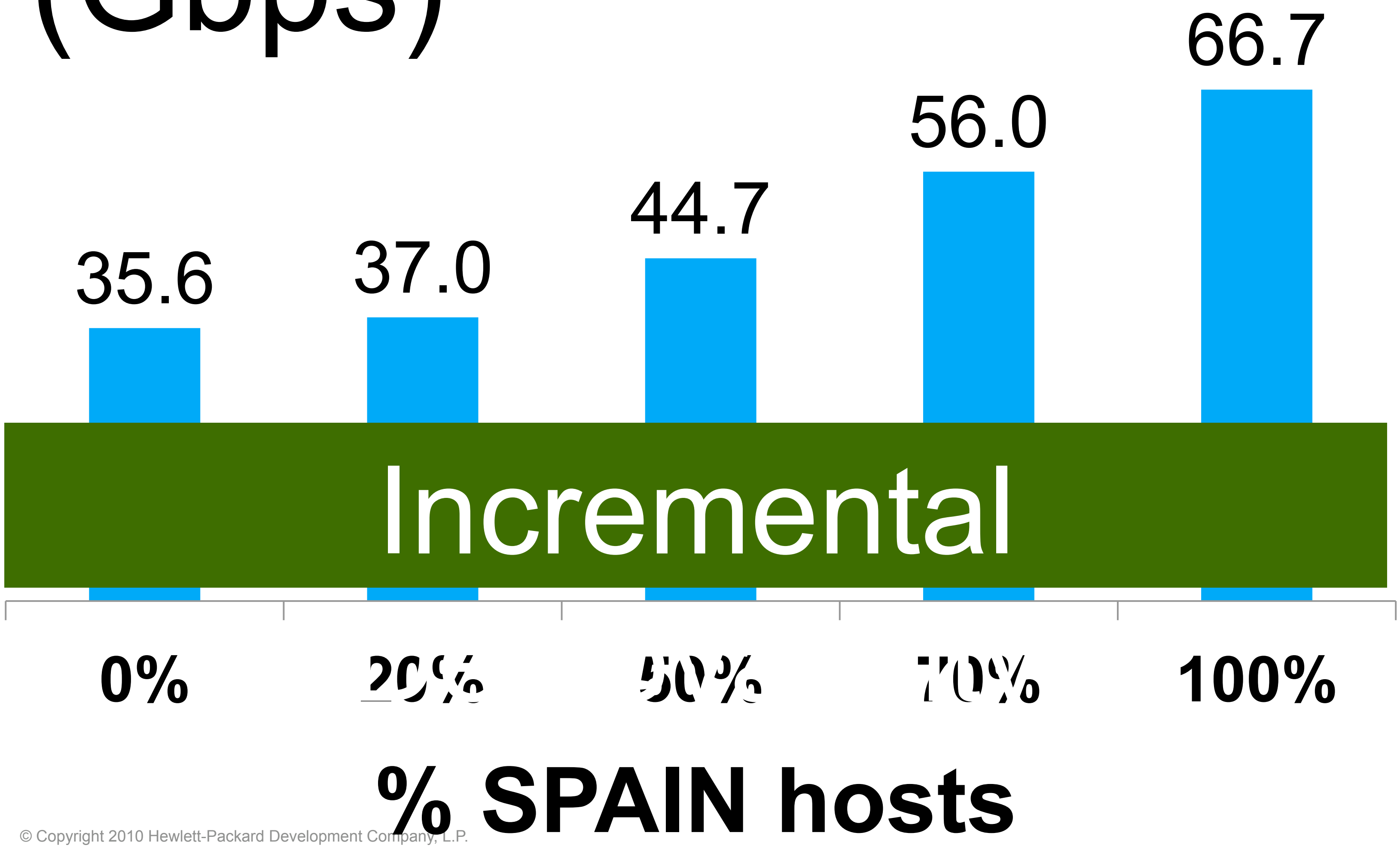
Completion times



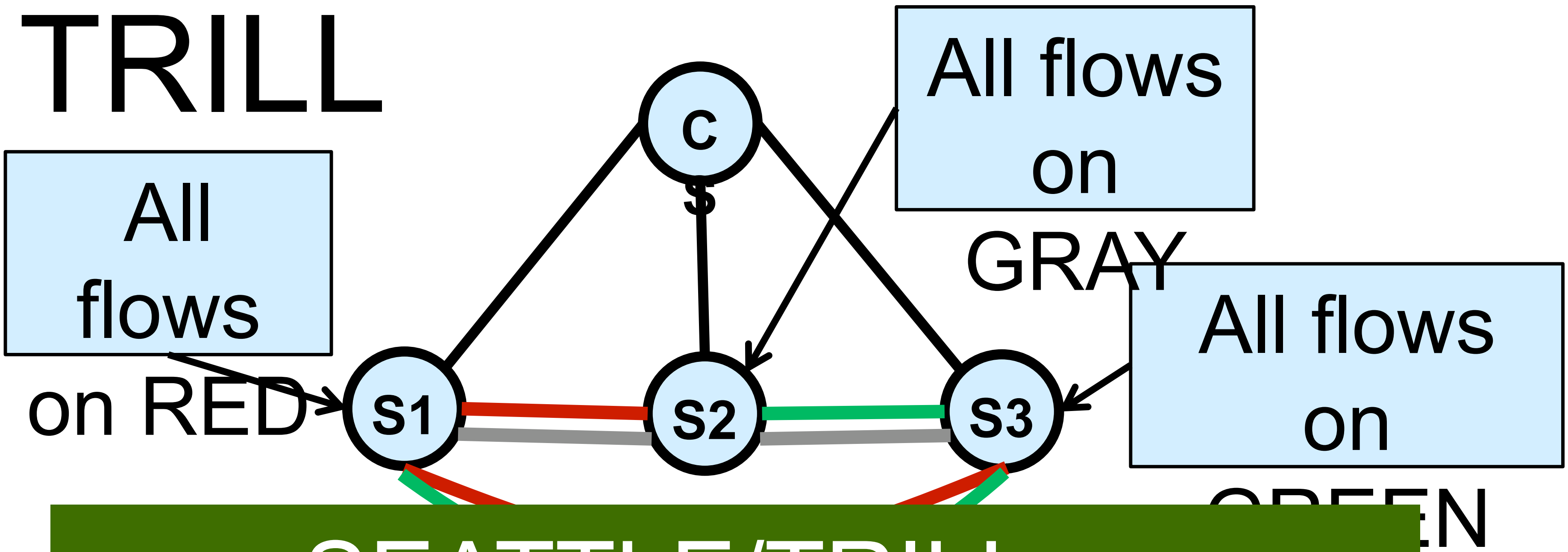
Aggregate Goodput (Gbps)



Aggregate Goodput (Gbps)

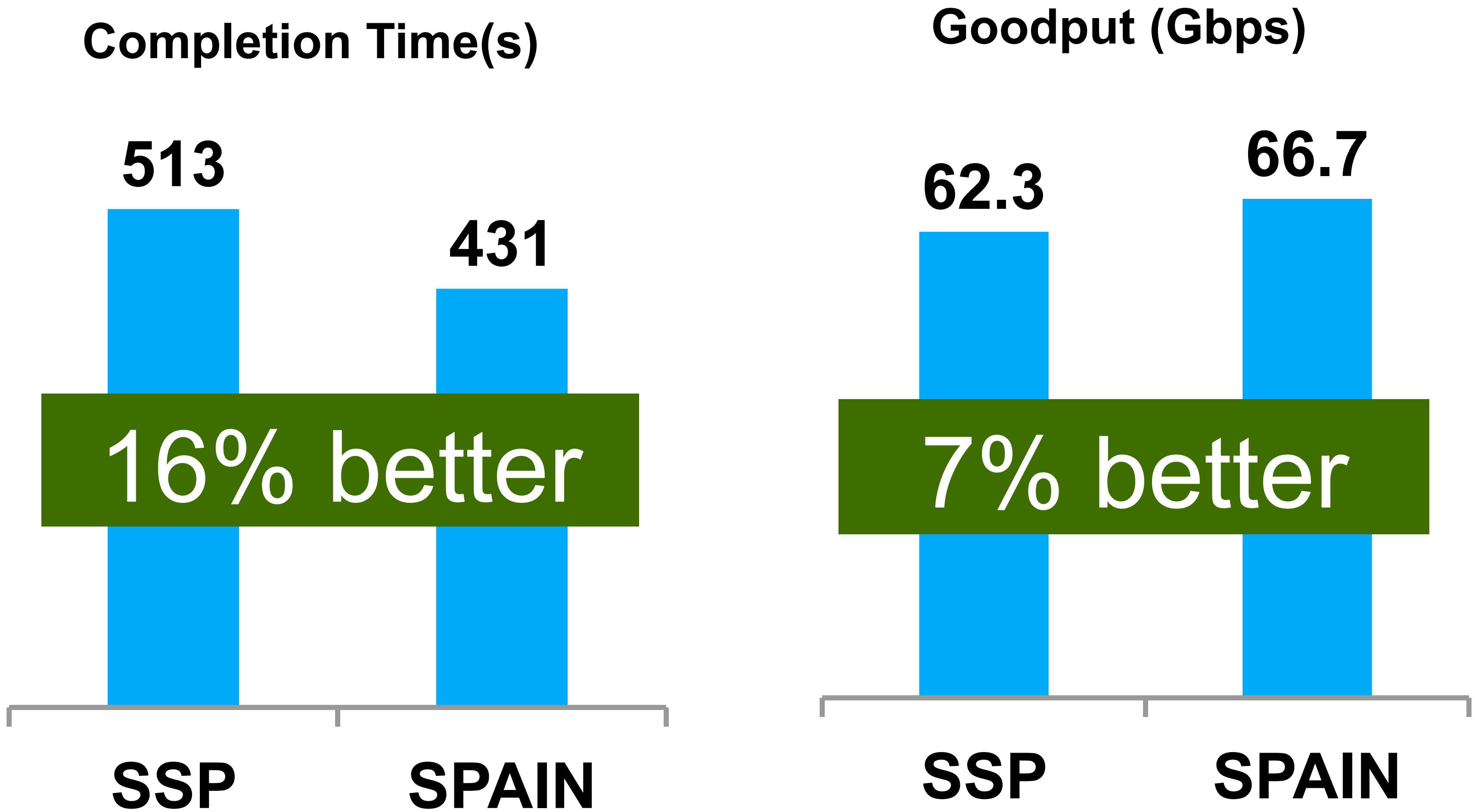


Single Shortest Path(SSP) SEATTLE/ TRILL



SEATTLE/TRILL on unmodified switches with

Comparison with SSP



SPAIN Take-away

Unmodified L2 switches

Multi-pathing via VLANs

Arbitrary topologies

Minor End-host modifs

Low-cost

High-BW

DC Fabric

Today!

Q&A