CONGA: Distributed Congestion-Aware Load Balancing for Datacenters

Mohammad Alizadeh

Tom Edsall, Sarang Dharmapurikar, Ramanan Vaidyanathan, Kevin Chu, Andy Fingerhut, Vinh The Lam★, Francis Matus, Rong Pan, Navindra Yadav, George Varghese§

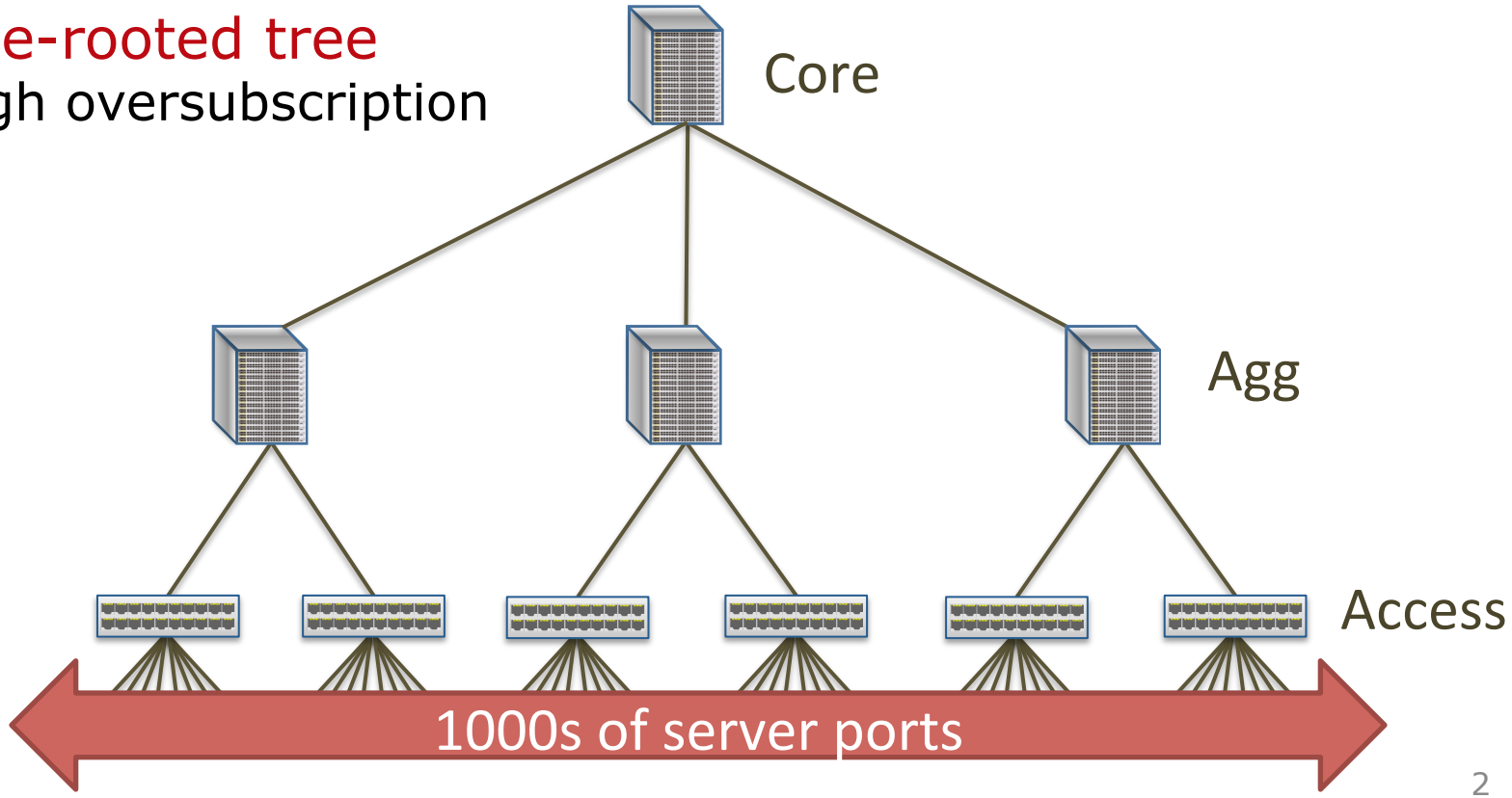
Google★ CISCO Insieme NETWORKS Microsoft §

Motivation

DC networks need large bisection bandwidth for distributed apps (big data, HPC, web services, etc)

Single-rooted tree

- High oversubscription

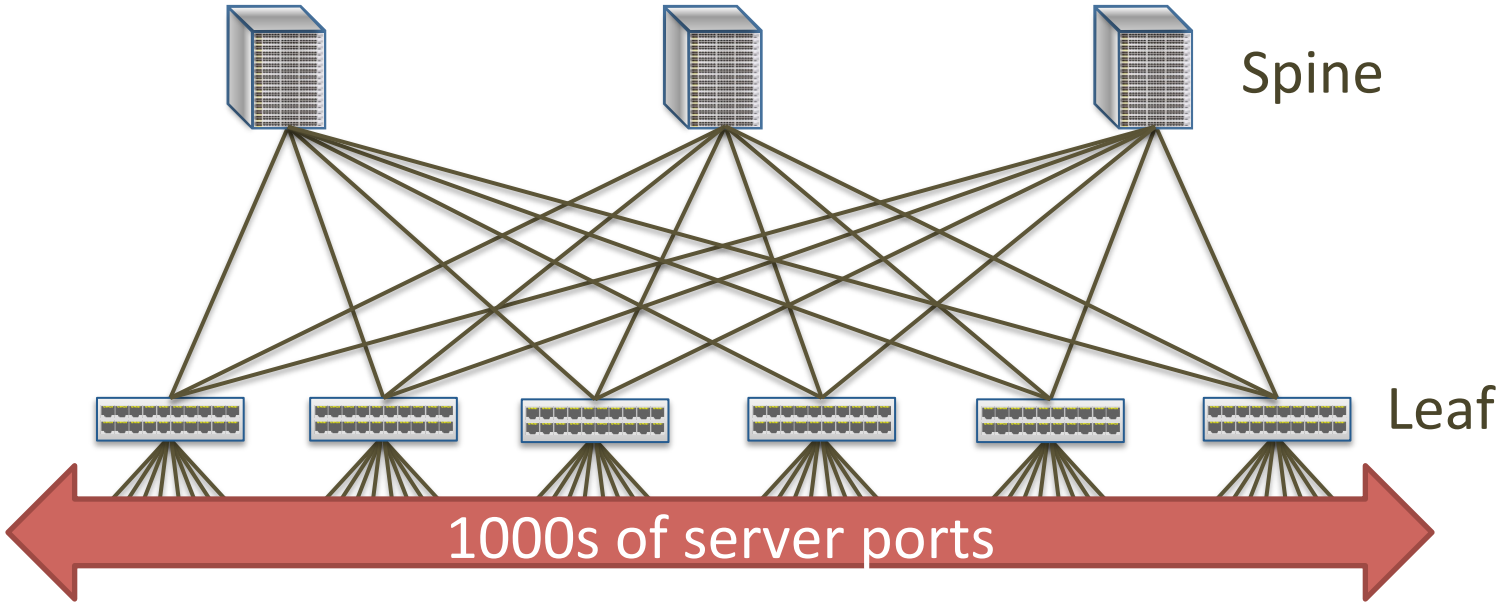


Motivation

DC networks need large bisection bandwidth for distributed apps (big data, HPC, web services, etc)

Multi-rooted tree [Fat-tree, Leaf-Spine, …]

- Full bisection bandwidth, achieved via multipathing

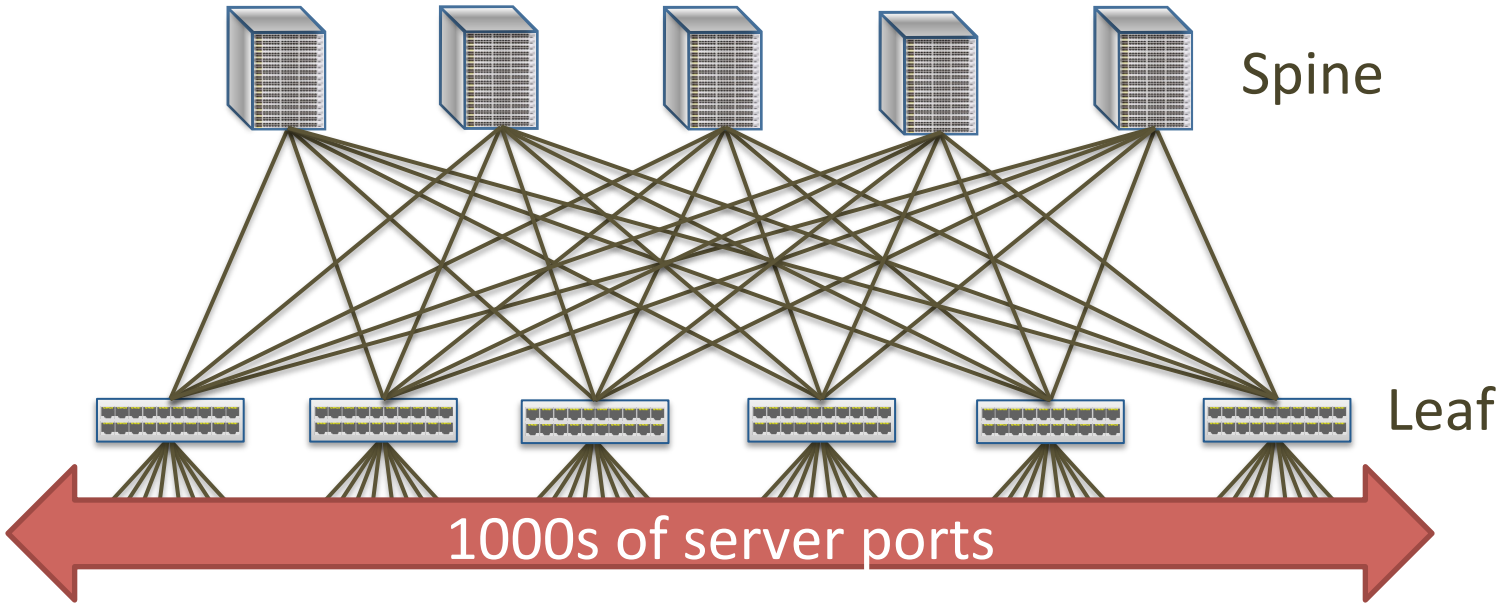


Motivation

DC networks need large bisection bandwidth for distributed apps (big data, HPC, web services, etc)

Multi-rooted tree [Fat-tree, Leaf-Spine, …]

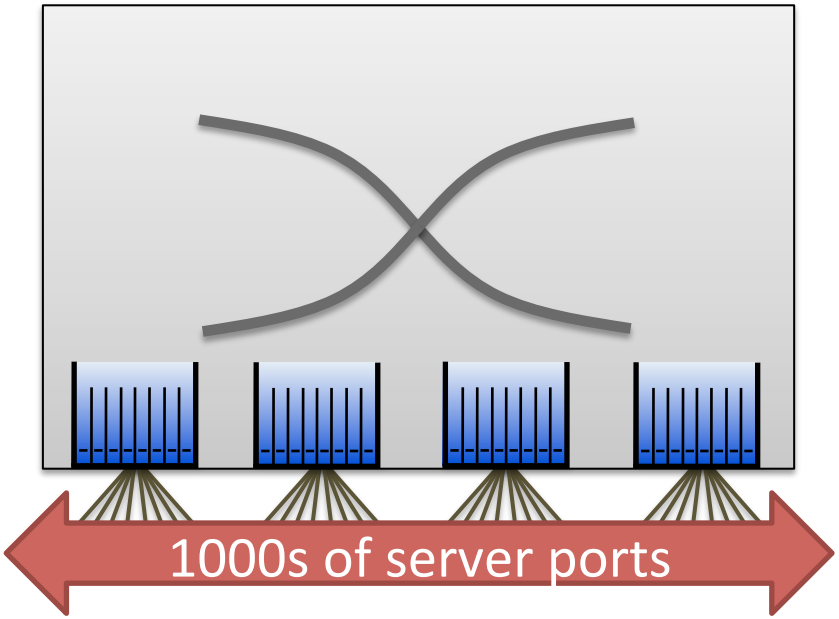
- Full bisection bandwidth, achieved via multipathing



Multi-rooted != Ideal DC Network

Ideal DC network:

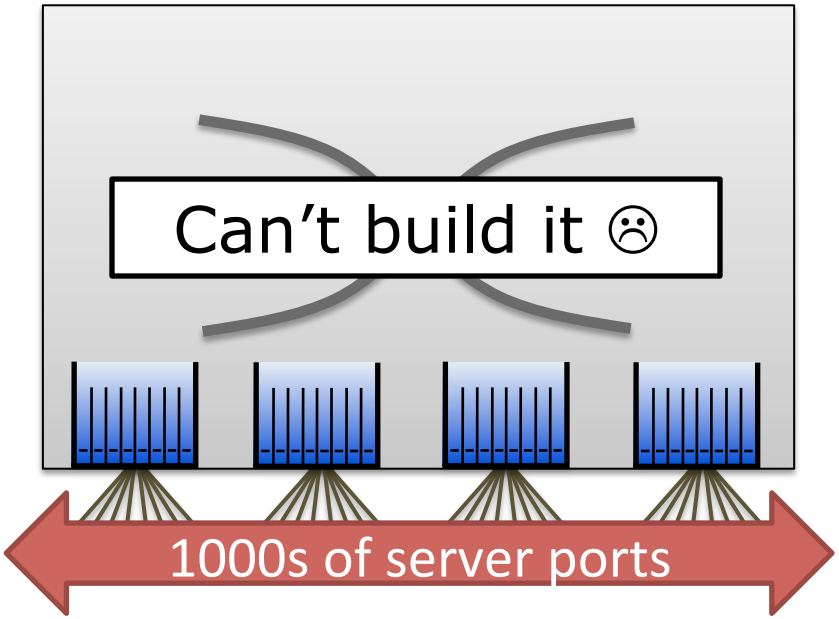
Big output-queued switch



- No internal bottlenecks ➔ predictable
- Simplifies BW management
  [EyeQ, FairCloud, pFabric, Varys, …]

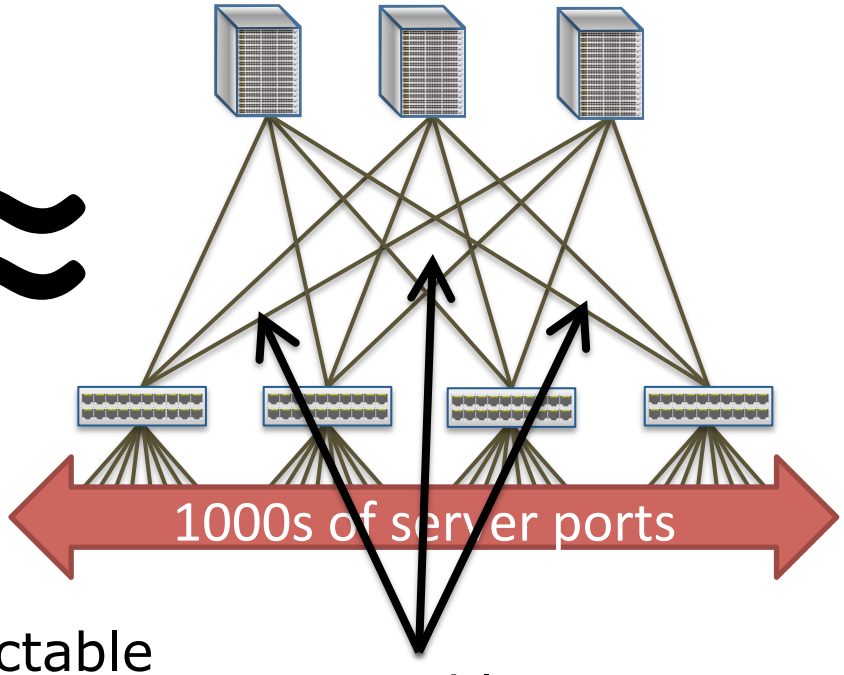
Multi-rooted != Ideal DC Network

Ideal DC network:
Big output-queued switch



≈

Multi-rooted tree

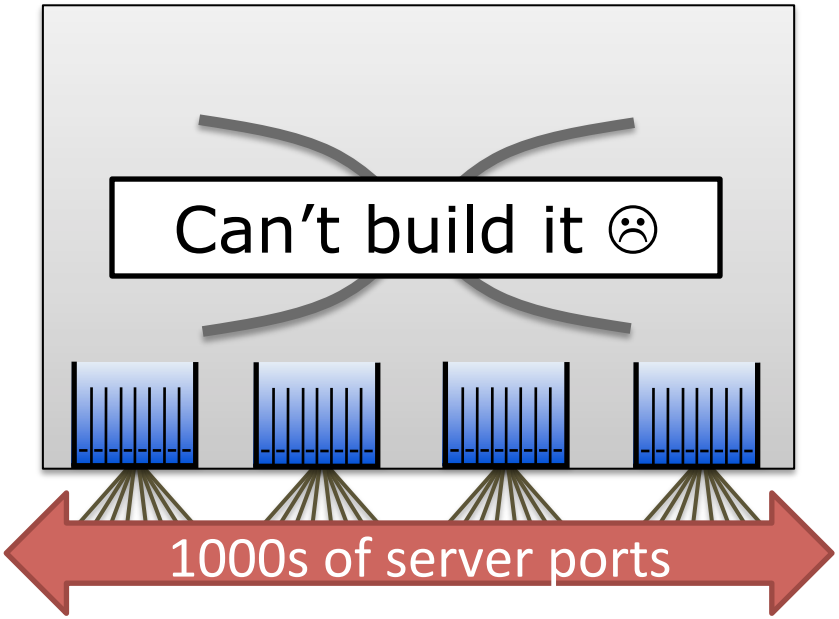


Possible bottlenecks

- No internal bottlenecks ➔ predictable
- Simplifies BW management
  [EyeQ, FairCloud, pFabric, Varys, …]

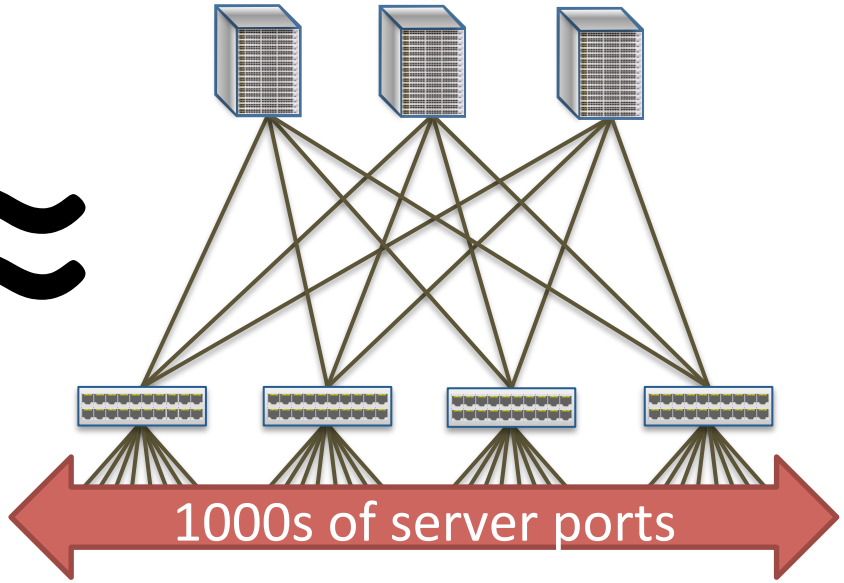
Multi-rooted != Ideal DC Network

Ideal DC network:
Big output-queued switch



≈

Multi-rooted tree

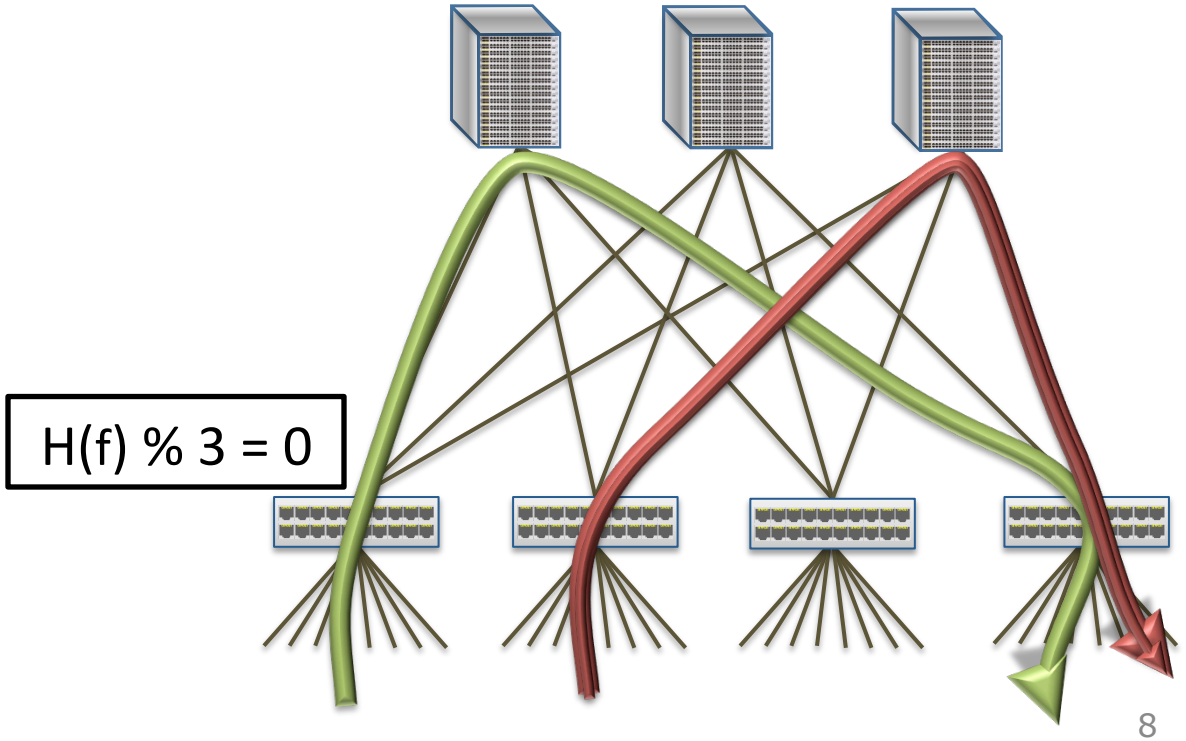


Need precise load balancing

Today: ECMP Load Balancing

Pick among equal-cost paths by a hash of 5-tuple

- Approximates Valiant load balancing
- Preserves packet order



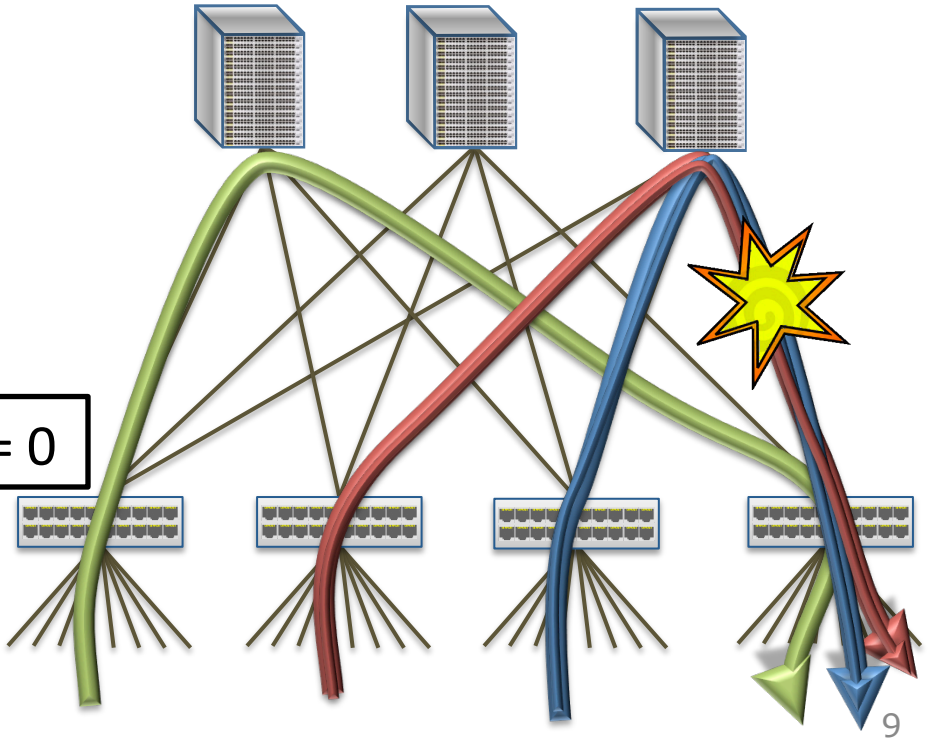
Today: ECMP Load Balancing

Pick among equal-cost paths by a hash of 5-tuple

- Approximates Valiant load balancing
- Preserves packet order

Problems:

– Hash collisions
(coarse granularity)



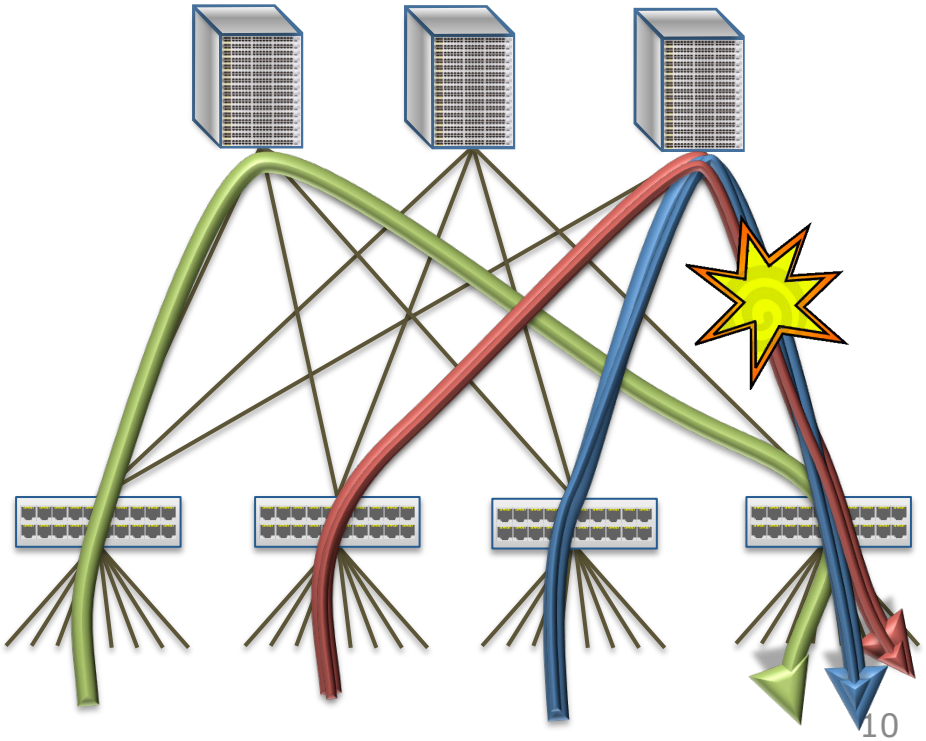
Today: ECMP Load Balancing

Pick among equal-cost paths by a hash of 5-tuple

- Approximates Valiant load balancing
- Preserves packet order

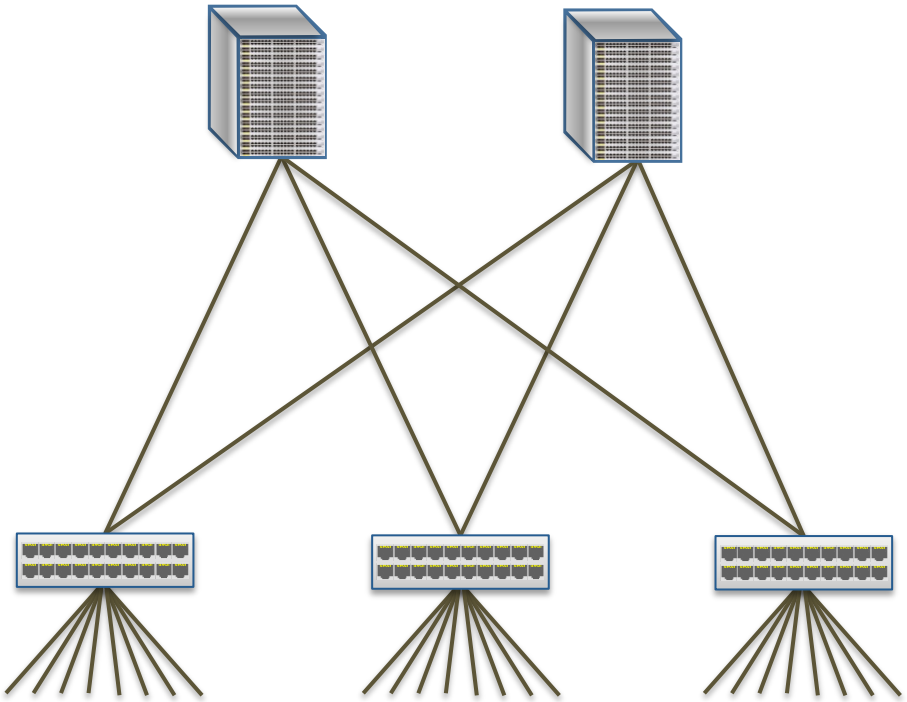
Problems:

- Hash collisions
  (coarse granularity)
- Local & stateless
  (v. bad with asymmetry due to link failures)



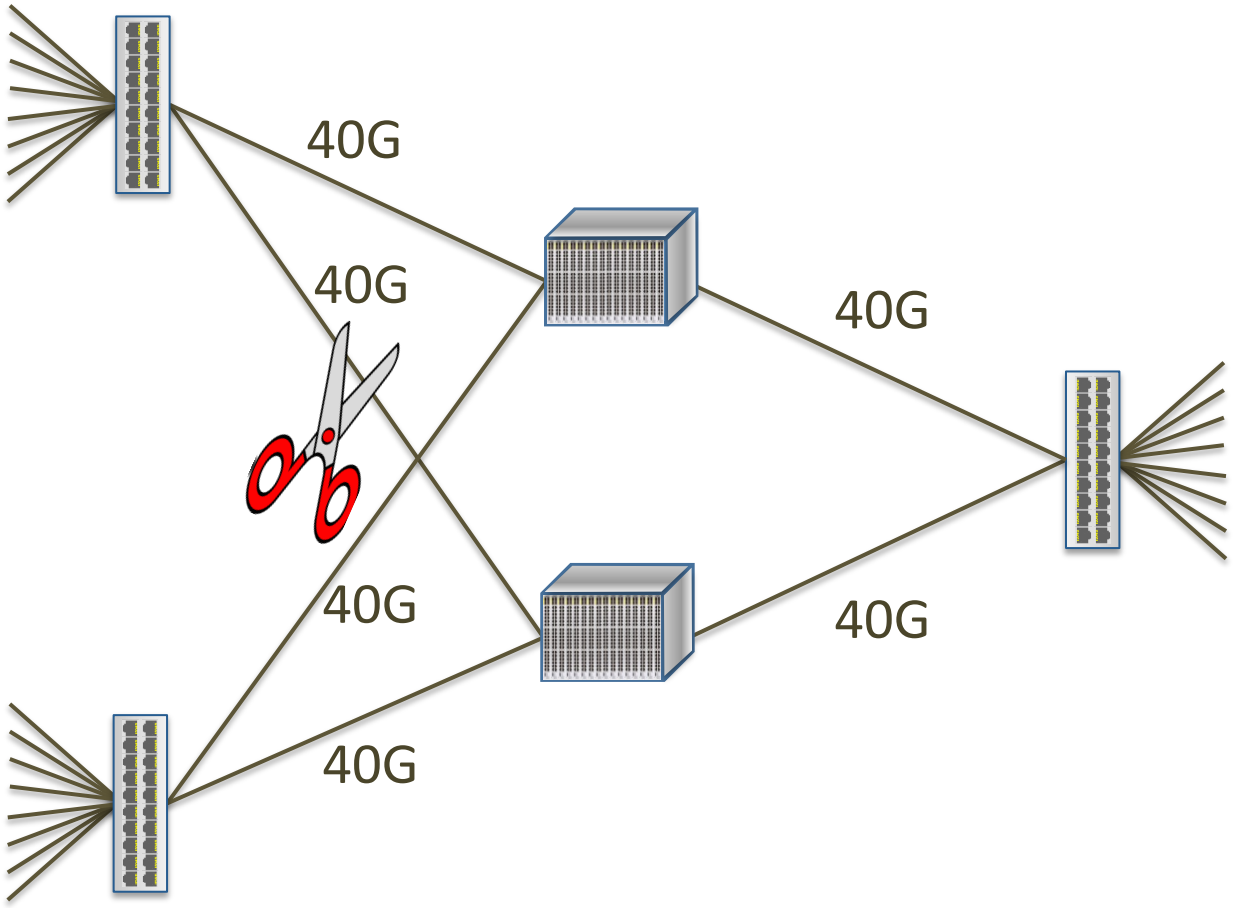
Dealing with Asymmetry

Handling asymmetry needs non-local knowledge



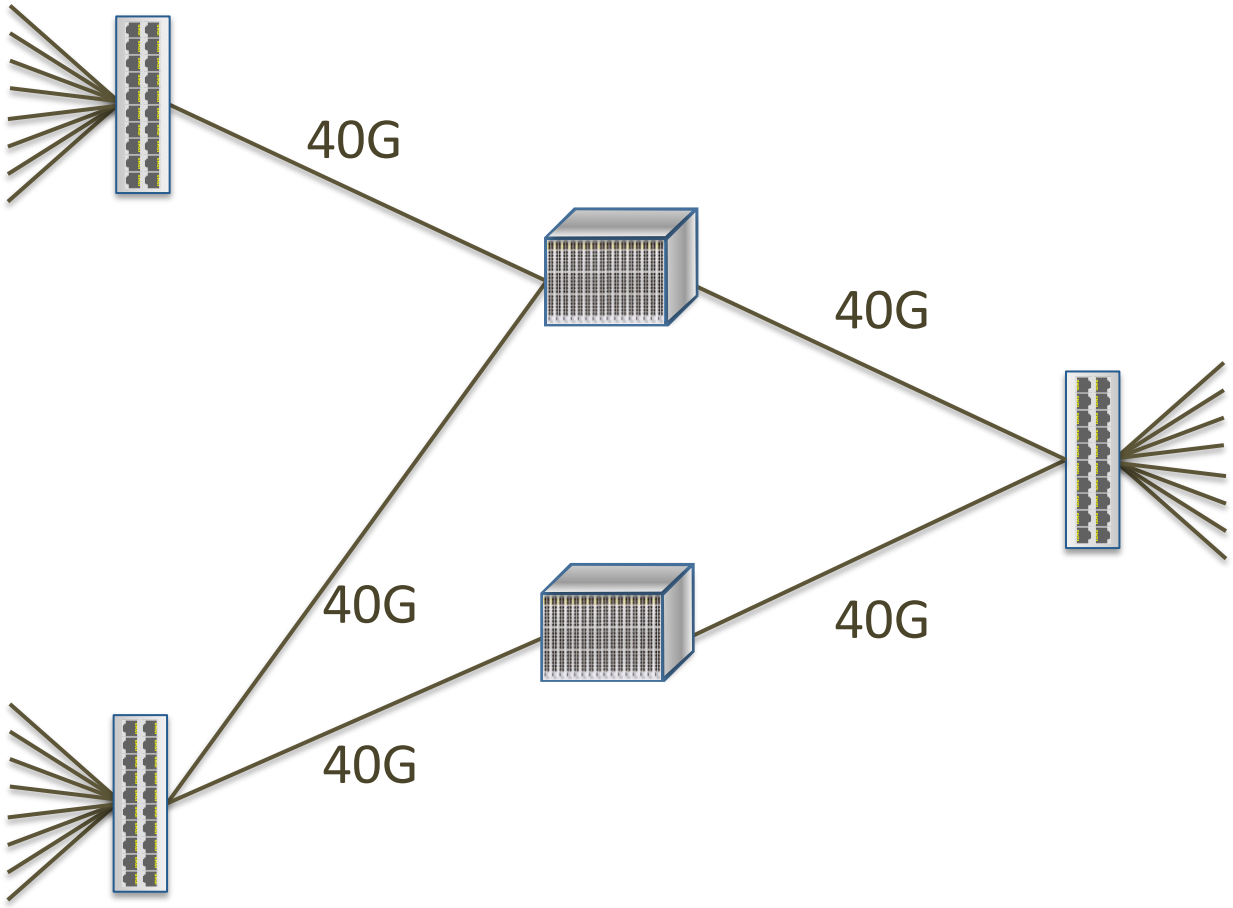
Dealing with Asymmetry

Handling asymmetry needs non-local knowledge

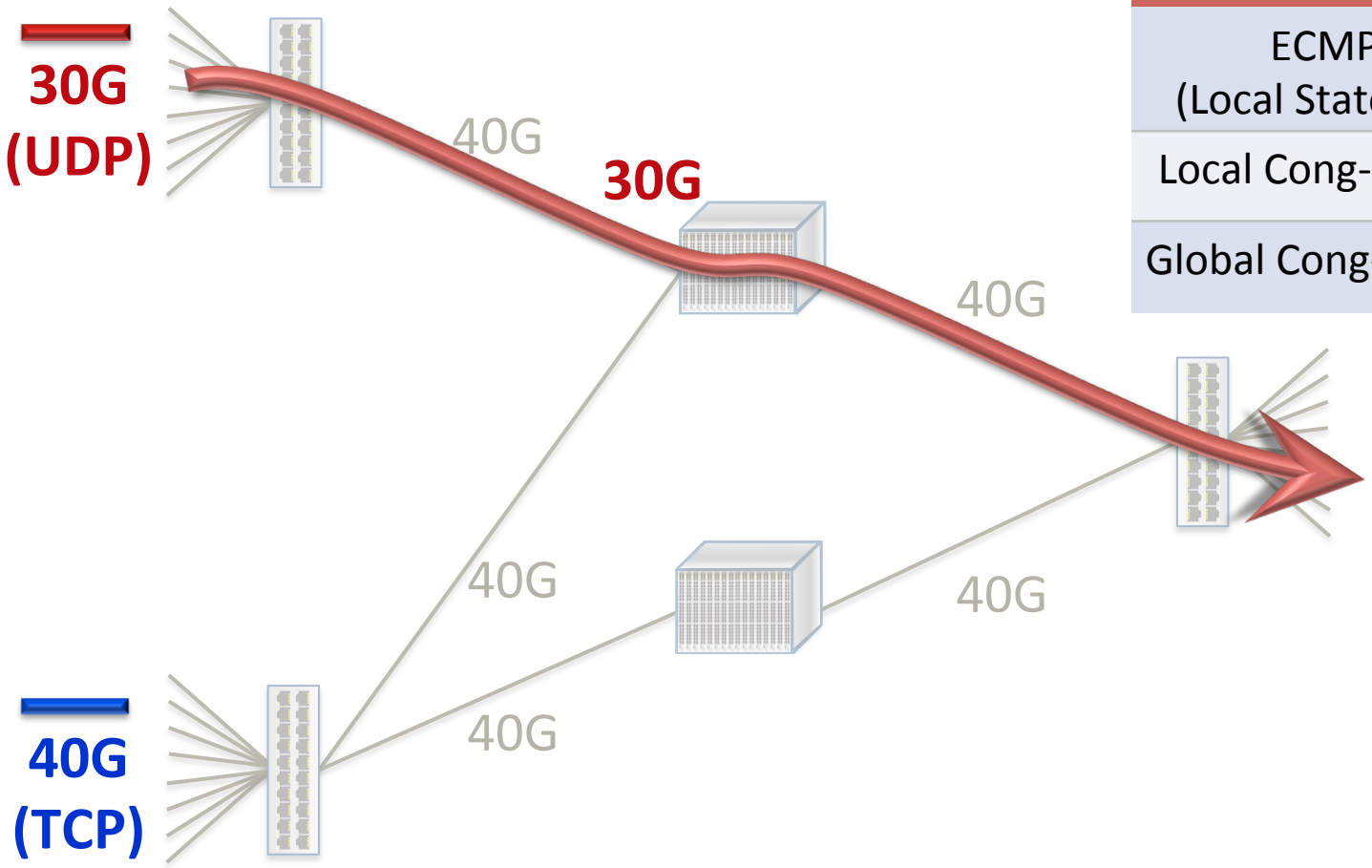


Dealing with Asymmetry

Handling asymmetry needs non-local knowledge

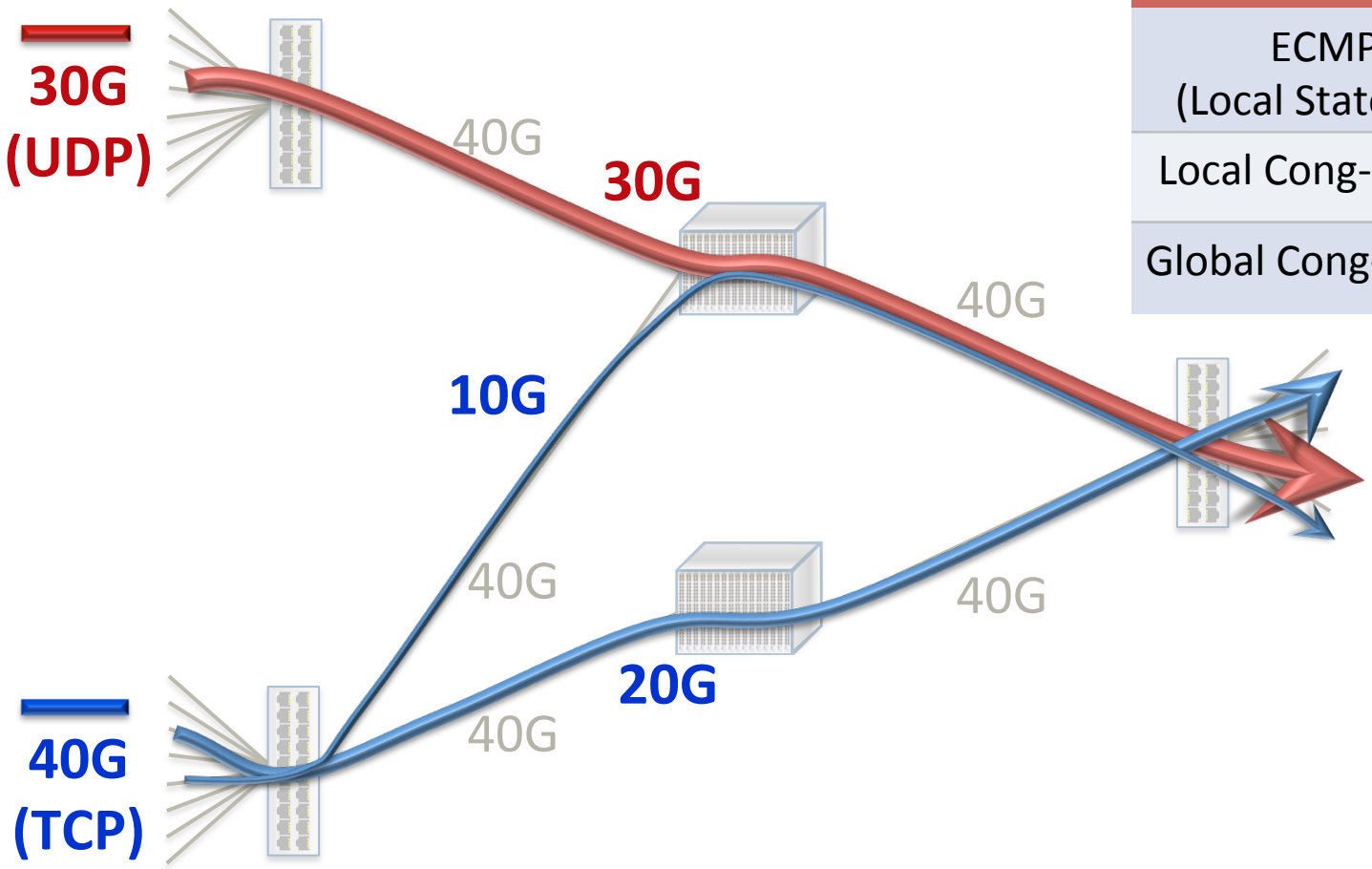


Dealing with Asymmetry



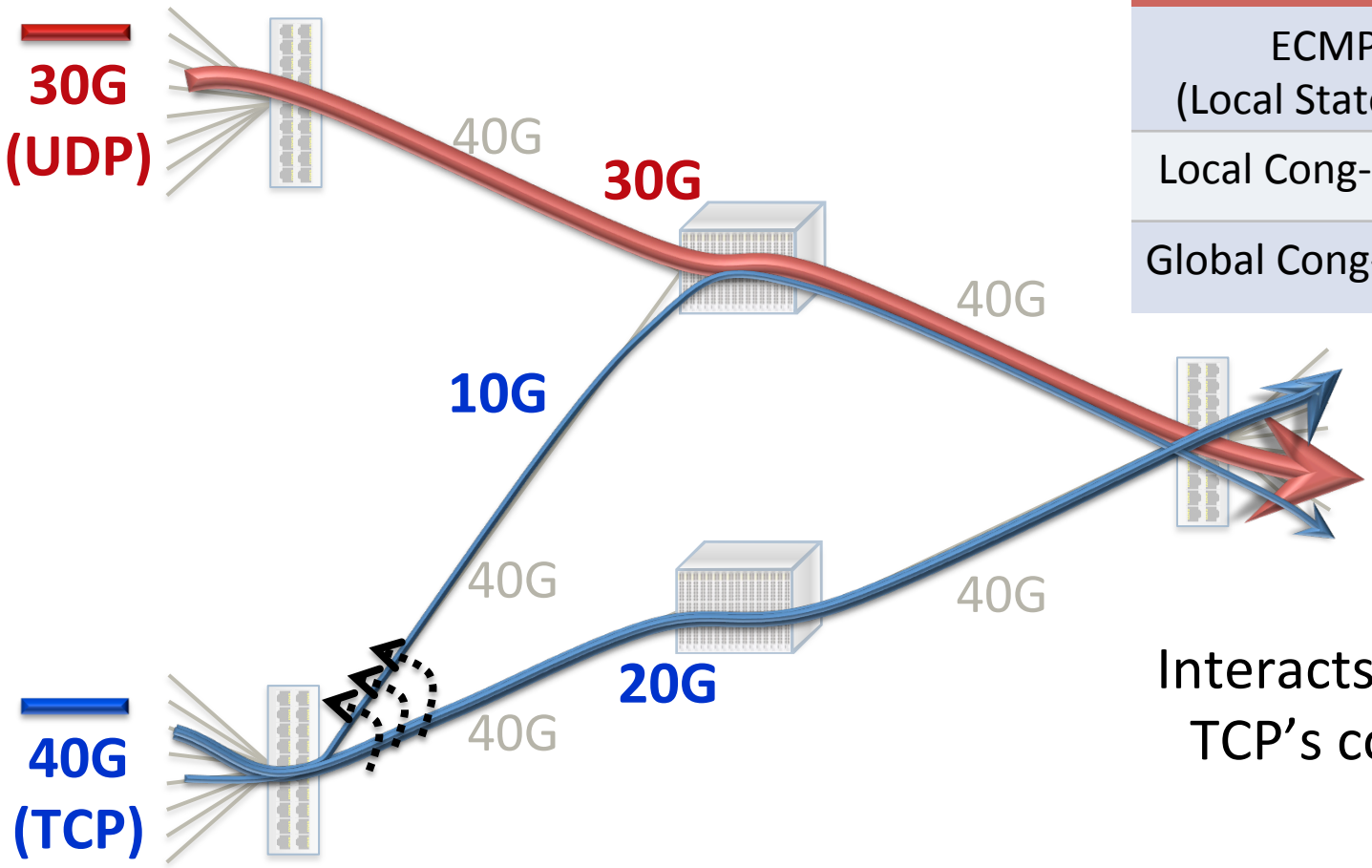
Scheme	Thrput
ECMP (Local Stateless)	
Local Cong-Aware	
Global Cong-Aware	

Dealing with Asymmetry: ECMP



Scheme	Thrput
ECMP (Local Stateless)	**60G**
Local Cong-Aware	
Global Cong-Aware	

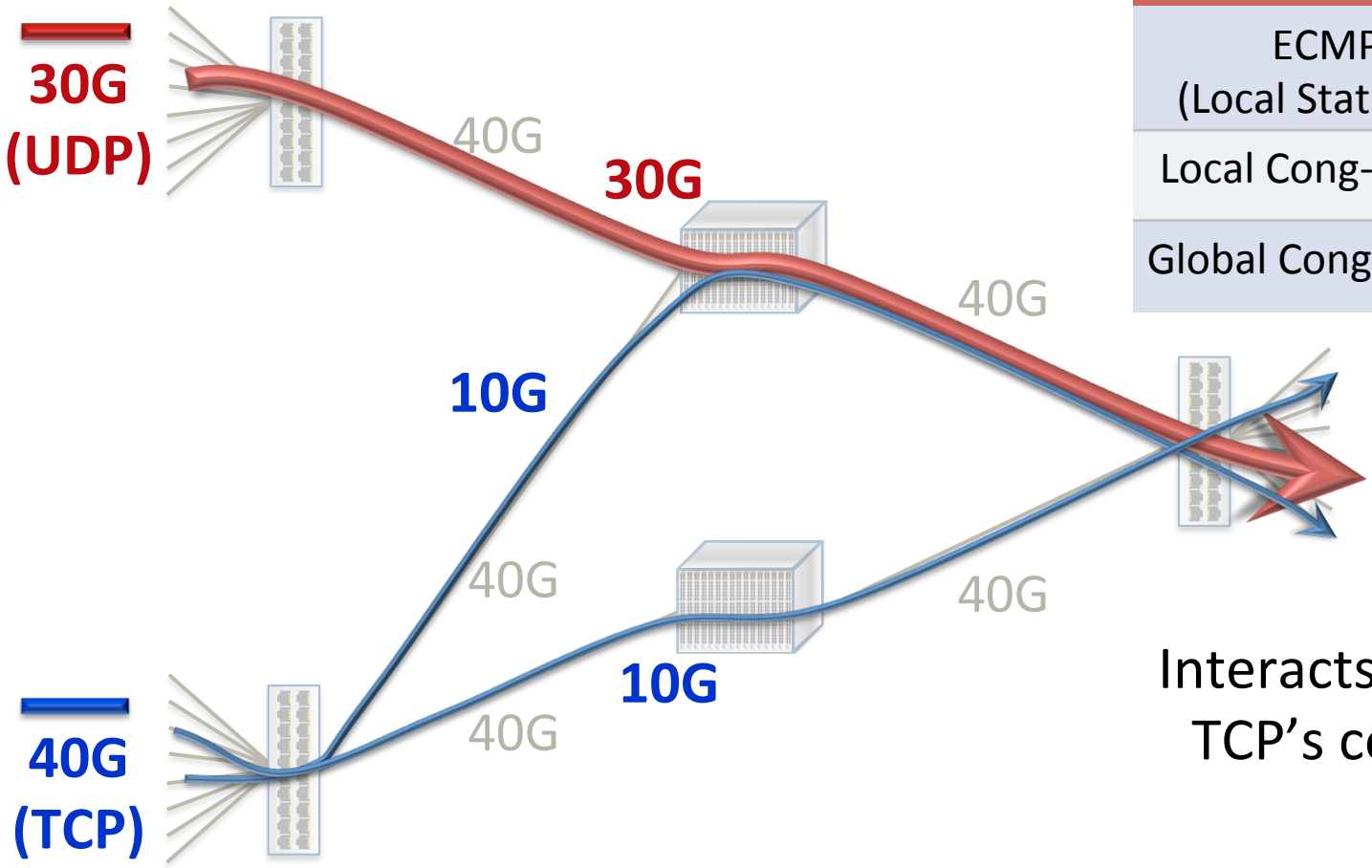
Dealing with Asymmetry: Local Congestion-Aware



Scheme	Thrput
ECMP (Local Stateless)	**60G**
Local Cong-Aware	
Global Cong-Aware	

Interacts poorly with TCP's control loop

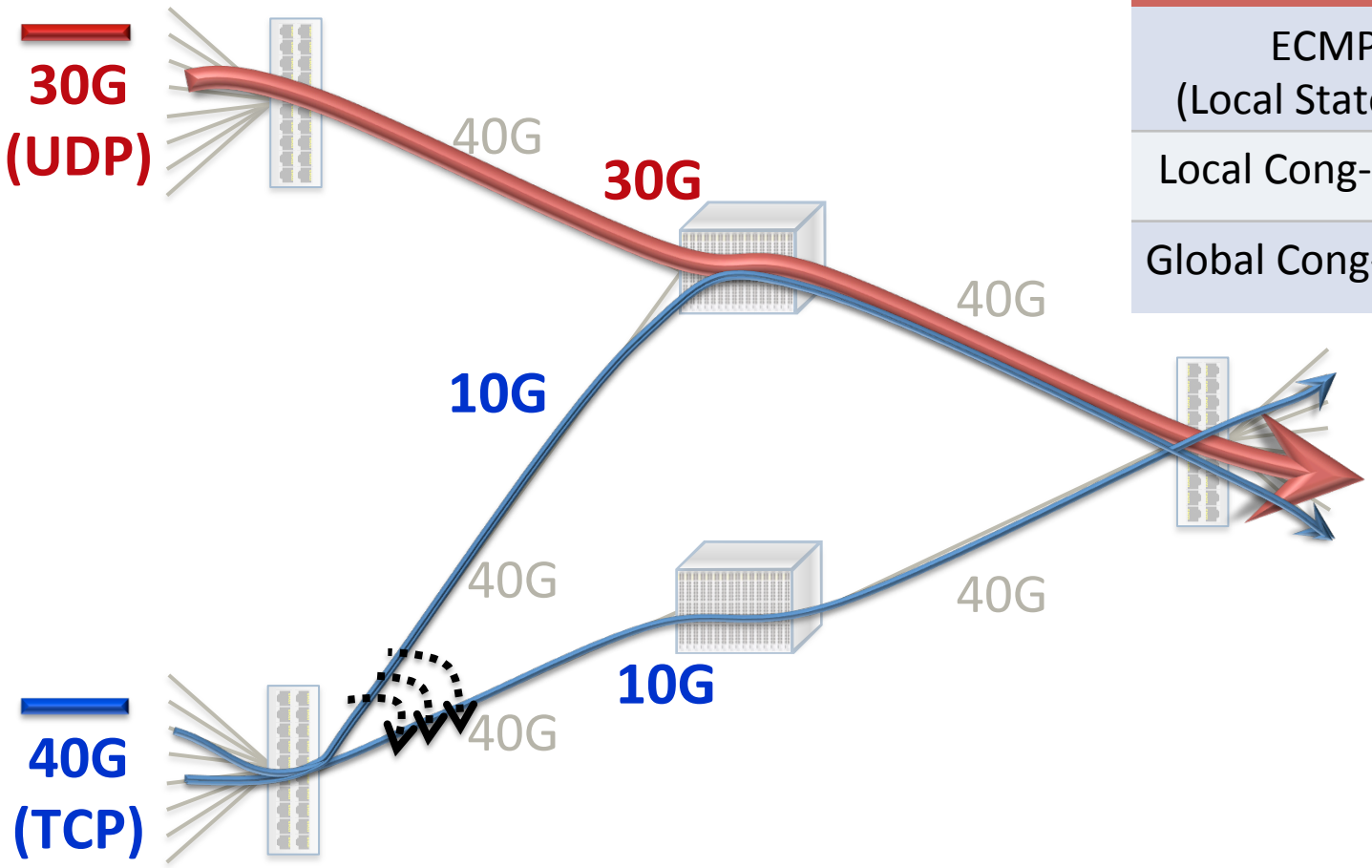
Dealing with Asymmetry: Local Congestion-Aware



Scheme	Thrput
ECMP (Local Stateless)	**60G**
Local Cong-Aware	**50G**
Global Cong-Aware	

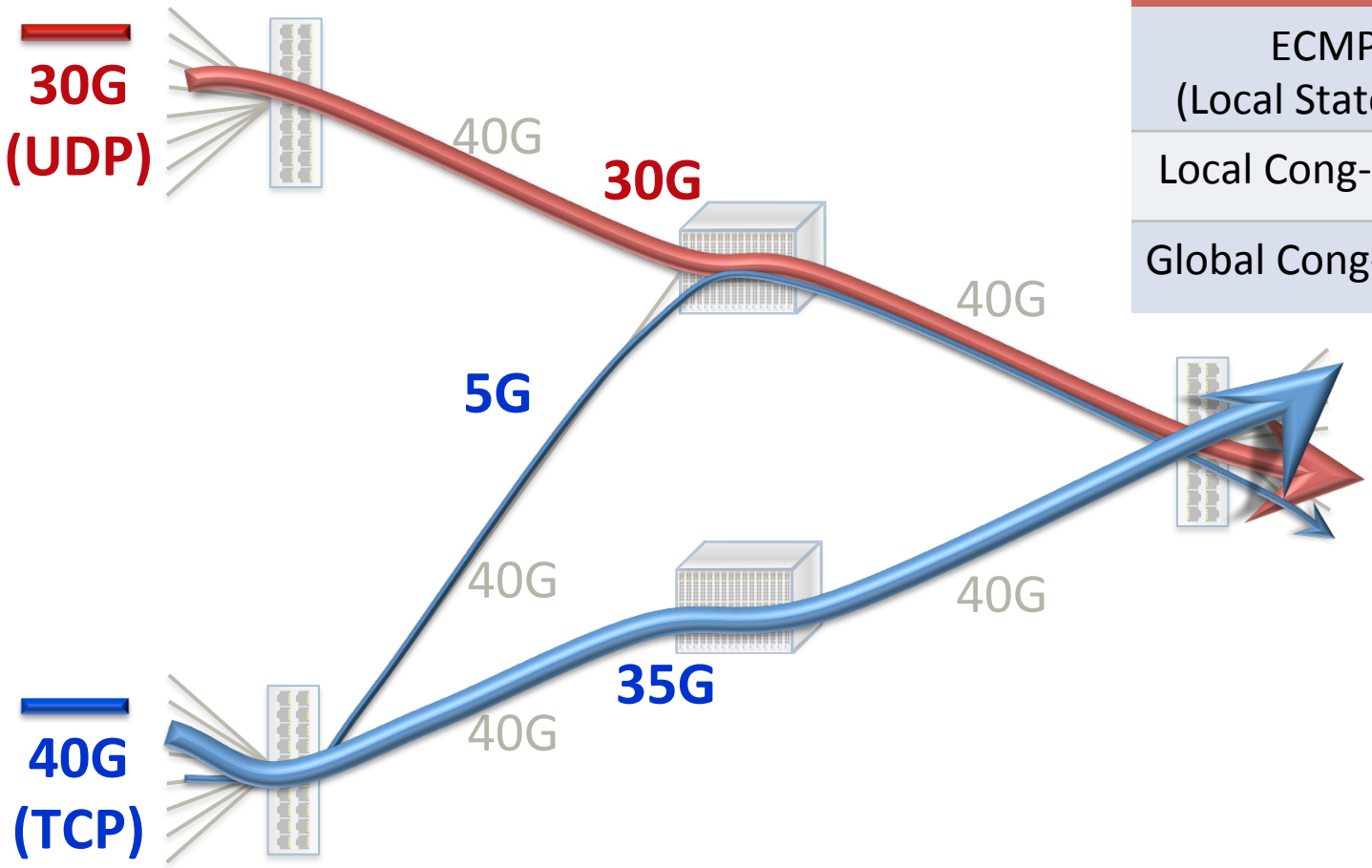
Interacts poorly with TCP's control loop

Dealing with Asymmetry: Global Congestion-Aware



Scheme	Thrput
ECMP (Local Stateless)	**60G**
Local Cong-Aware	**50G**
Global Cong-Aware	

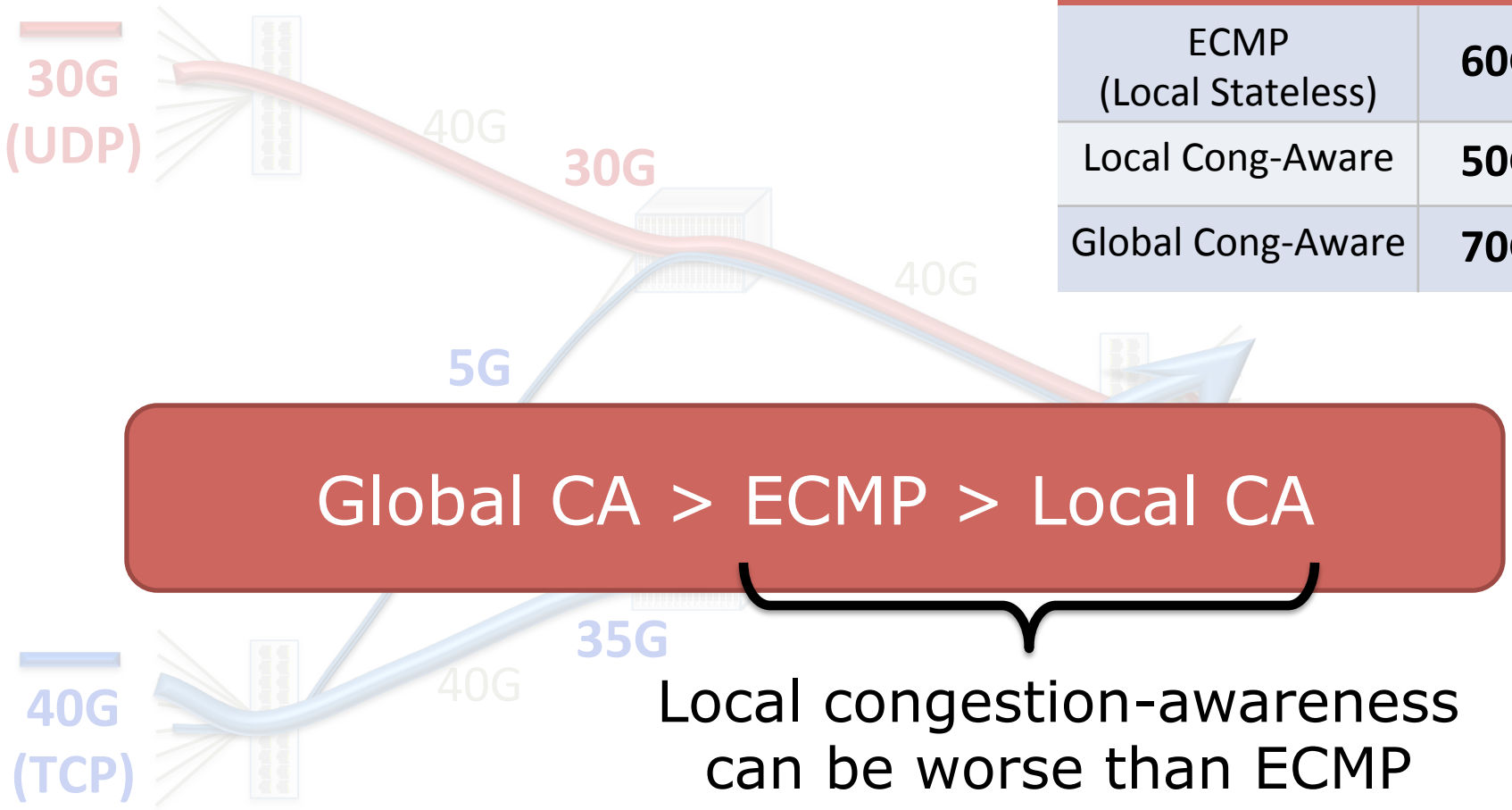
Dealing with Asymmetry: Global Congestion-Aware



Scheme	Thrput
ECMP (Local Stateless)	**60G**
Local Cong-Aware	**50G**
Global Cong-Aware	**70G**

Dealing with Asymmetry: Global Congestion-Aware

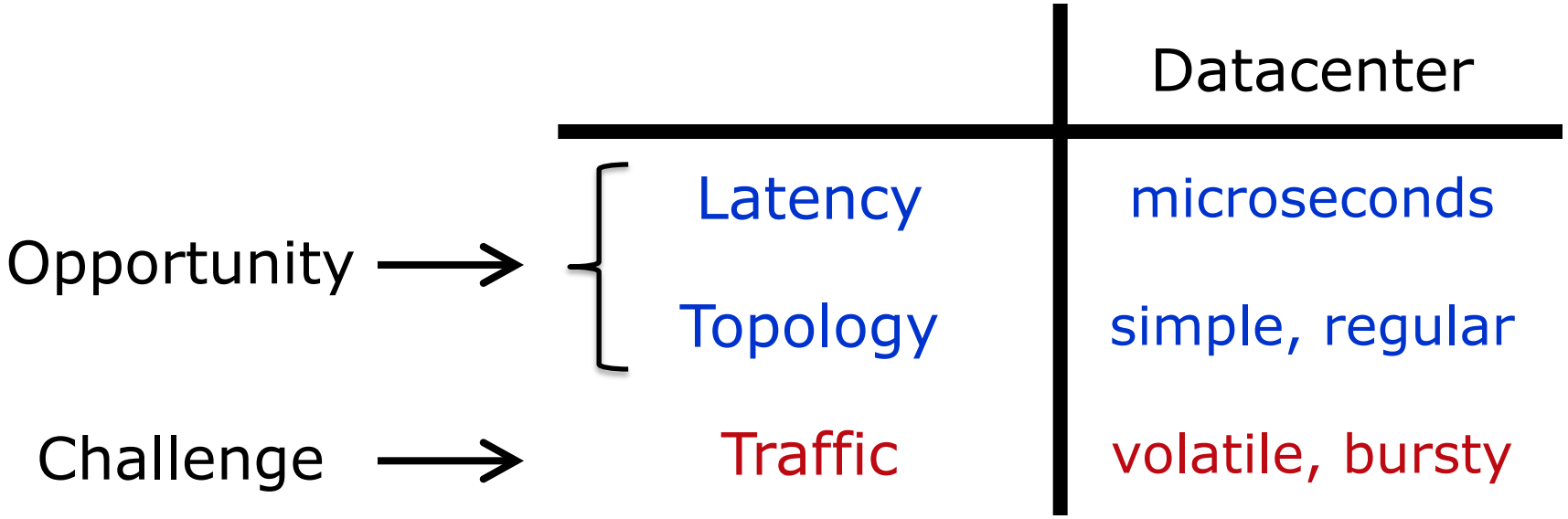
Scheme	Thrput
ECMP (Local Stateless)	**60G**
Local Cong-Aware	**50G**
Global Cong-Aware	**70G**



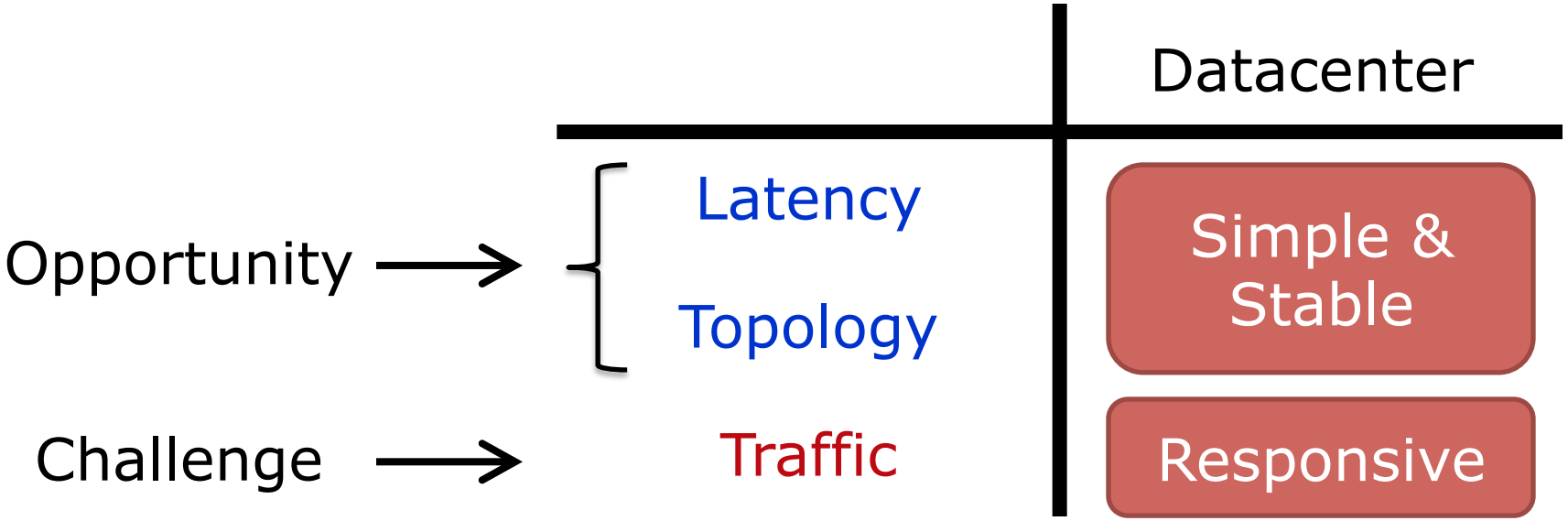
Global CA > ECMP > Local CA

Local congestion-awareness can be worse than ECMP

Global Congestion-Awareness (in Datacenters)



Global Congestion-Awareness (in Datacenters)

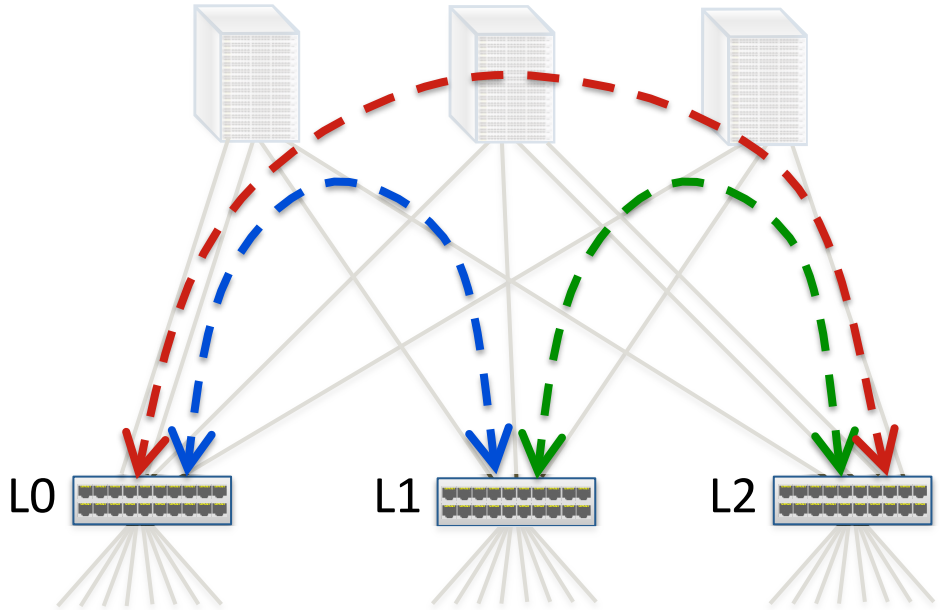


Key Insight:
Use *extremely* fast, low latency distributed control

CONGA in 1 Slide

1. Leaf switches (top-of-rack) track congestion to other leaves on different paths in near real-time

2. Use greedy decisions to minimize bottleneck util



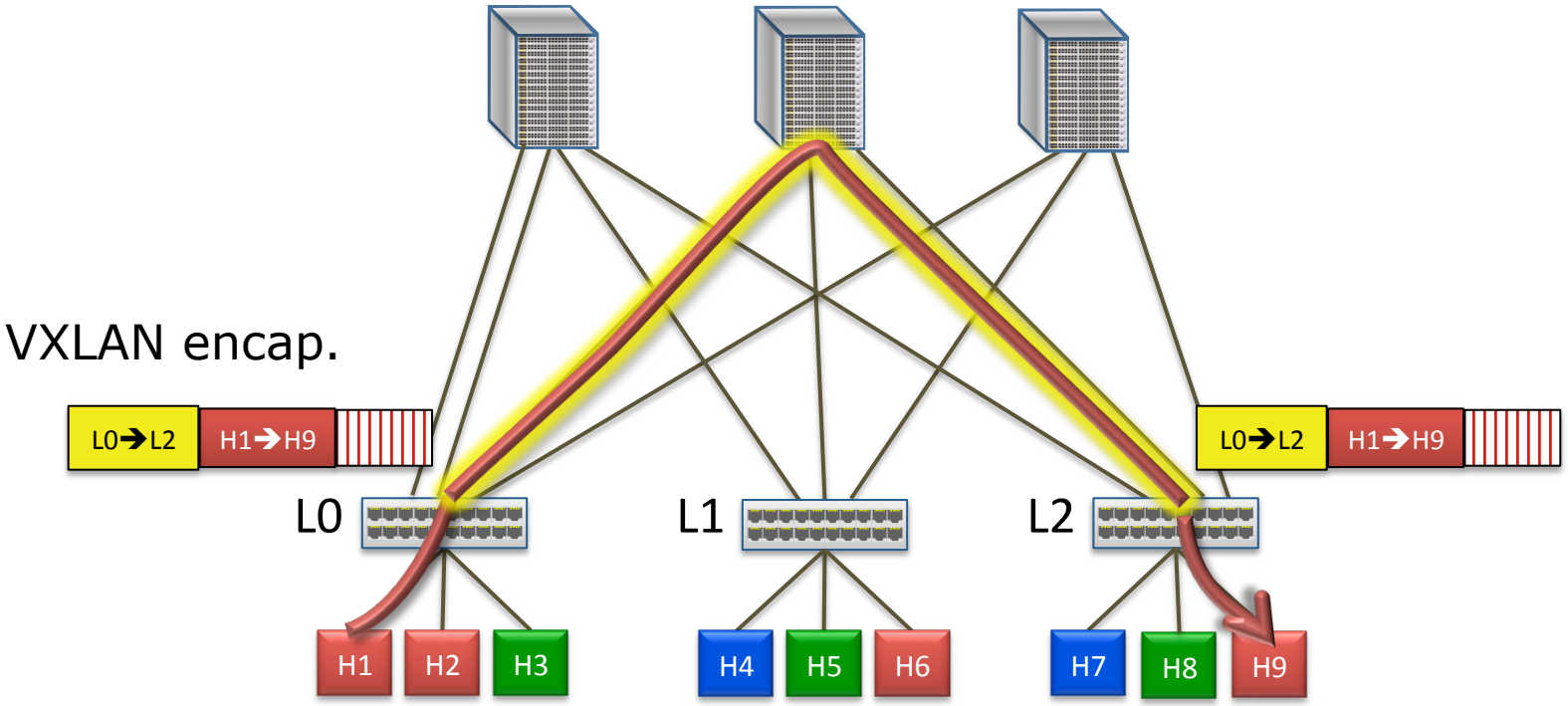
Fast feedback loops between leaf switches, directly in dataplane

CONGA’S DESIGN

Design

CONGA operates over a standard DC overlay (VXLAN)

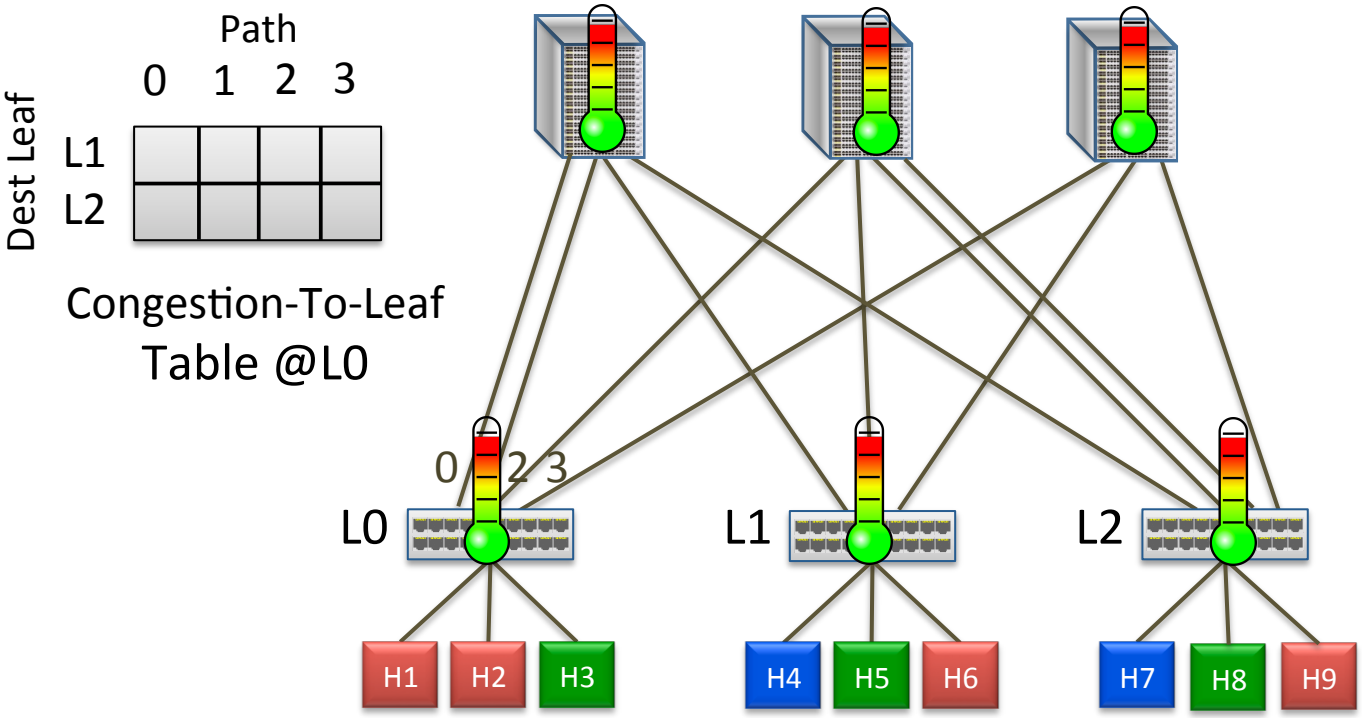
- Already deployed to virtualize the physical network



Design: Leaf-to-Leaf Feedback

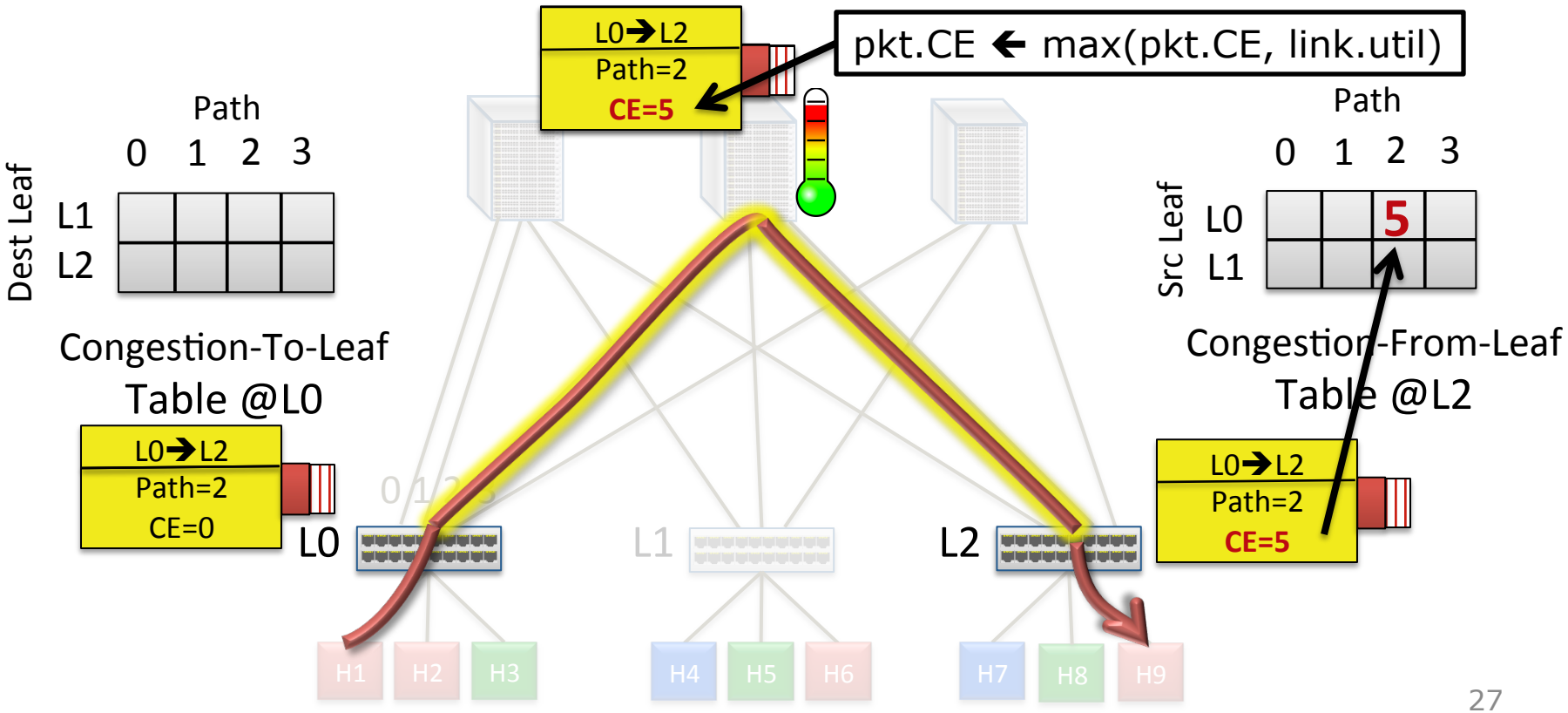
Track path-wise congestion metrics (3 bits) between each pair of leaf switches

Rate Measurement Module
measures link utilization



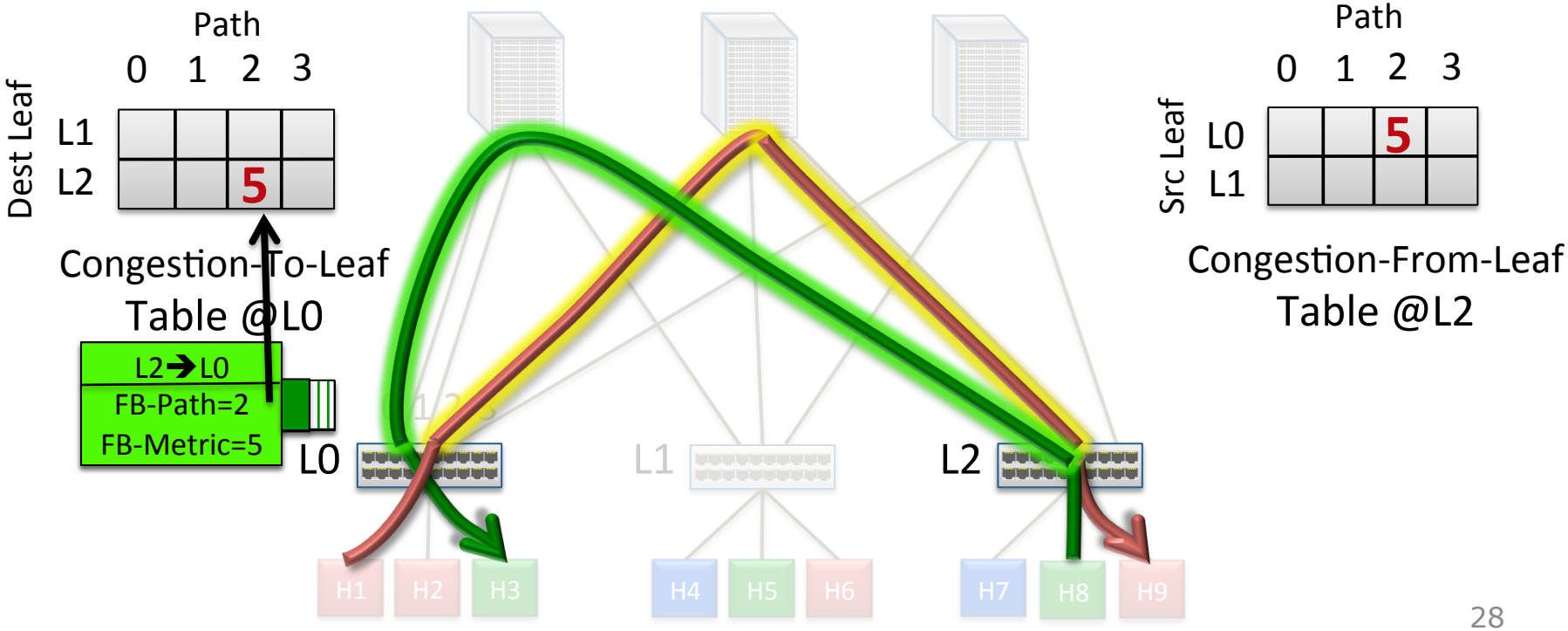
Design: Leaf-to-Leaf Feedback

Track path-wise congestion metrics (3 bits) between each pair of leaf switches



Design: Leaf-to-Leaf Feedback

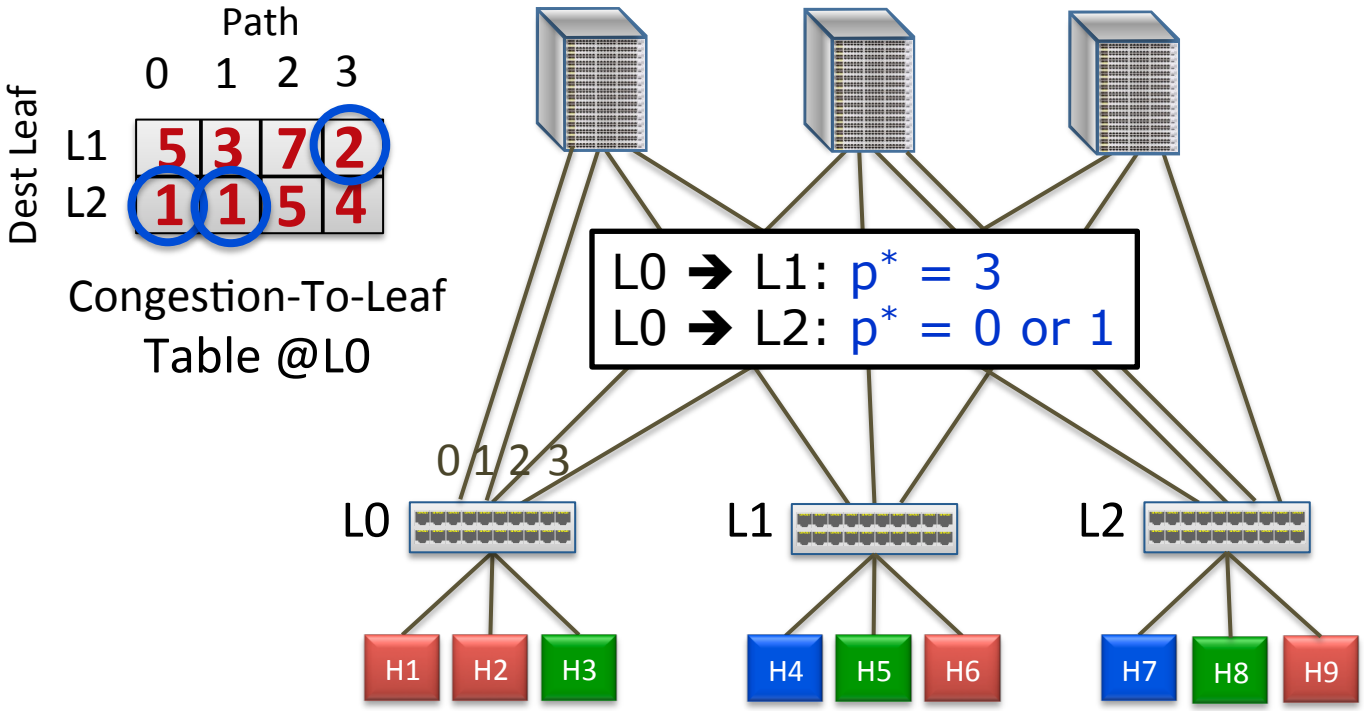
Track path-wise congestion metrics (3 bits) between each pair of leaf switches



Design: LB Decisions

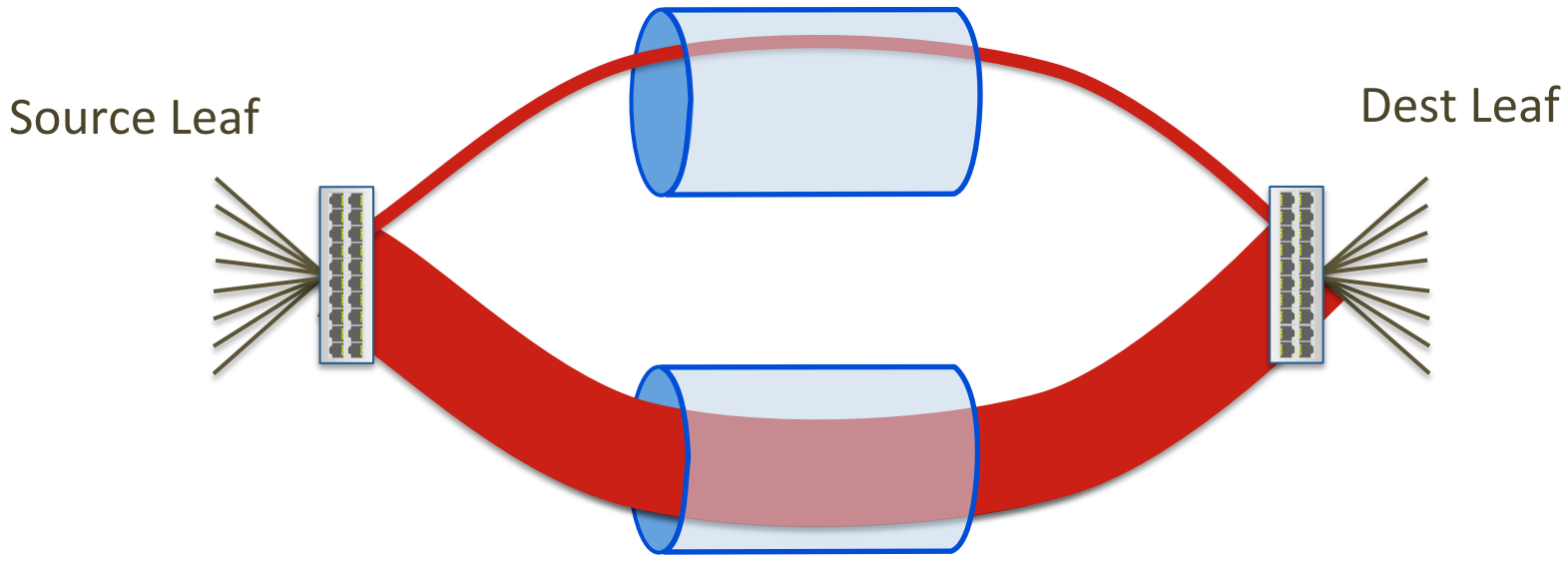
Send each ~~packet~~ on least congested path

flowlet [Kandula et al 2007]



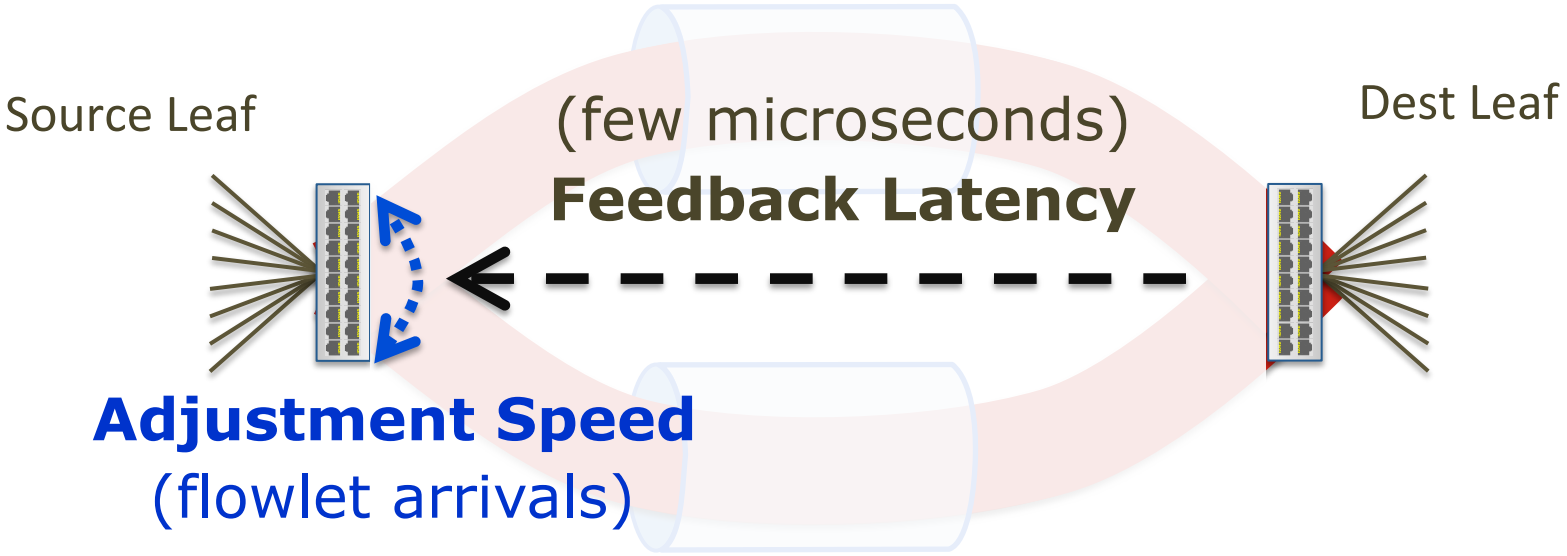
Why is this Stable?

Stability usually requires a sophisticated control law (e.g., TeXCP, MPTCP, etc)



Why is this Stable?

Stability usually requires a sophisticated control law (e.g., TeXCP, MPTCP, etc)



Near-zero latency + flowlets ➔ stable

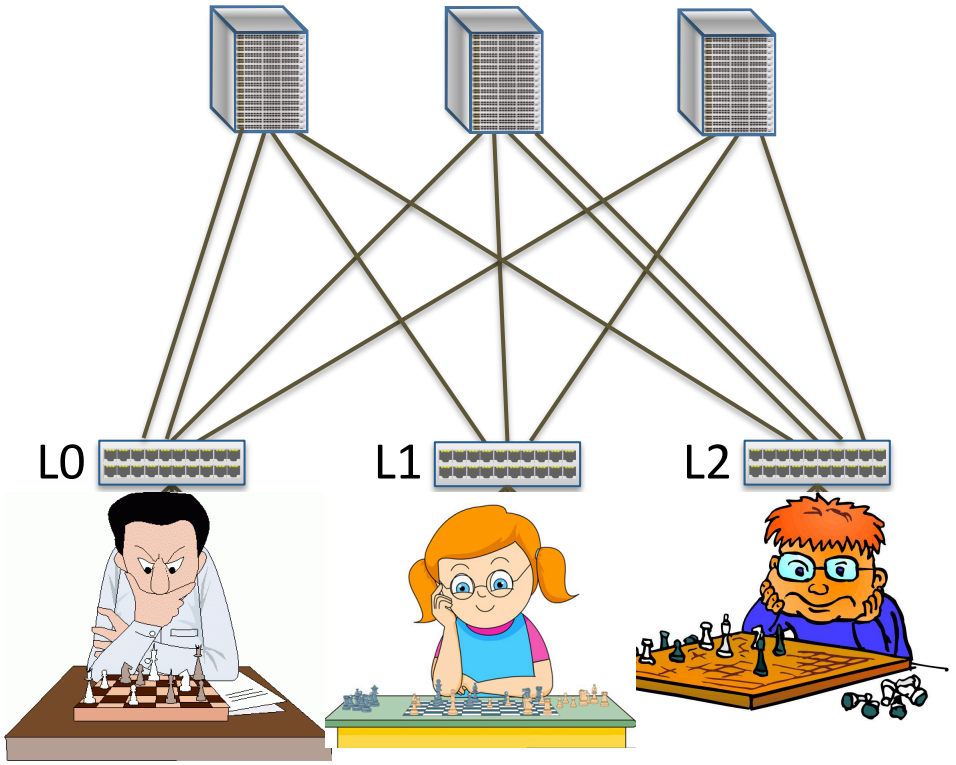
How Far is this from Optimal?

Given traffic demands $[\lambda_{ij}]$:

$$\frac{\max_{l \in \text{Links}} \rho_l \text{ with CONGA}}{\min_{f \in \text{feasible}} \max_{l \in \text{Links}} \rho_l}$$

Price of Anarchy

bottleneck routing game
(Banner & Orda, 2007)



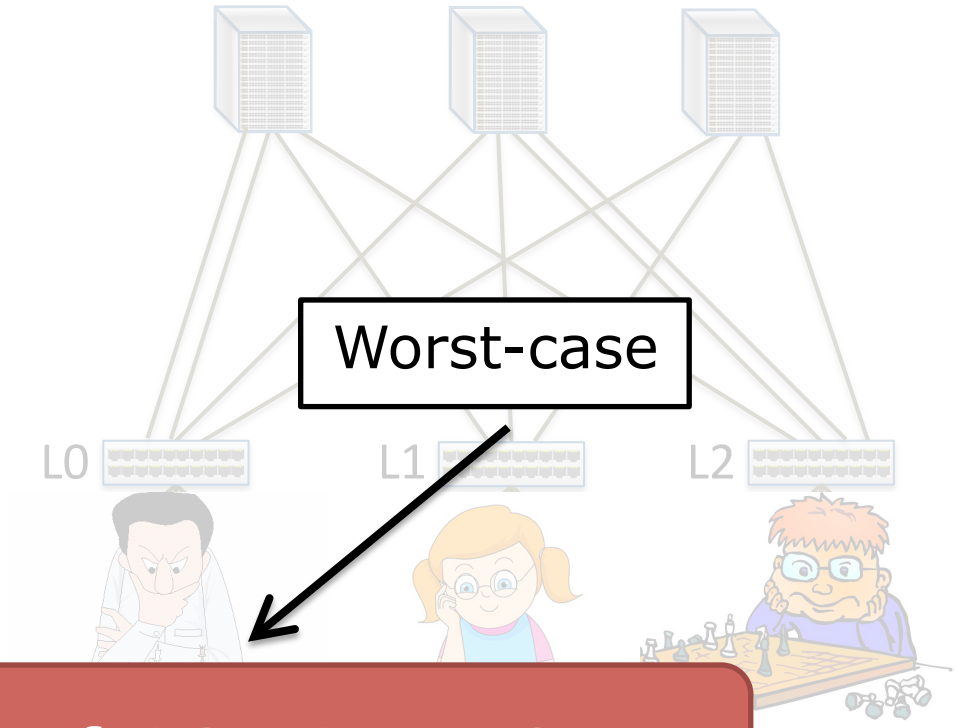
How Far is this from Optimal?

Given traffic demands $[\lambda_{ij}]$:

$$\frac{\max_{l \in \text{Links}} \rho_l \text{ with CONGA}}{\min_{f \in \text{feasible}} \max_{l \in \text{Links}} \rho_l}$$

Price of Anarchy

bottleneck routing game
(Banner & Orda, 2007)

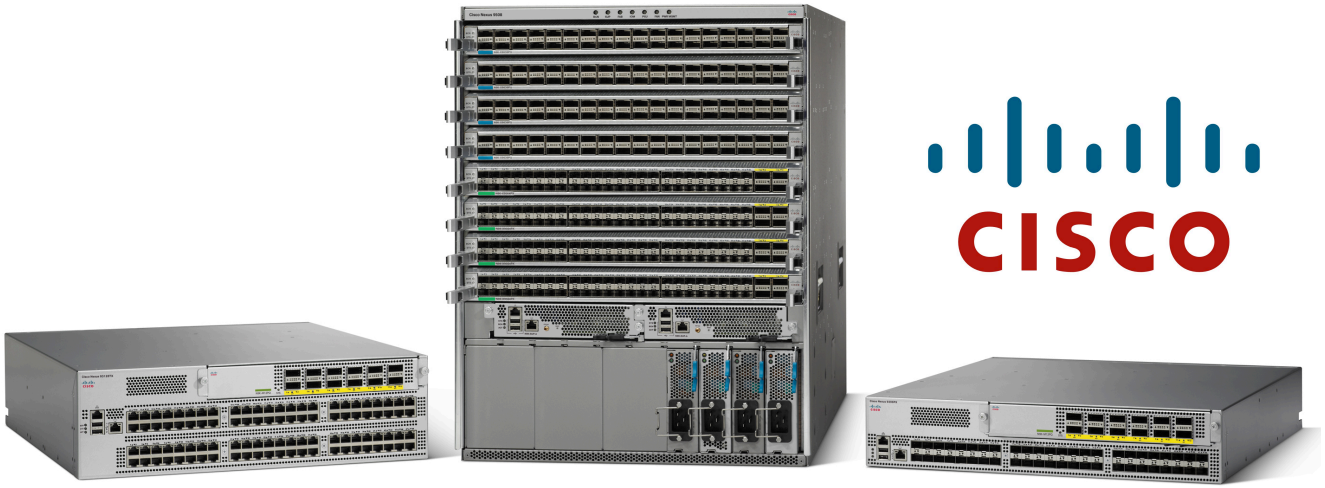


Theorem: PoA of CONGA = 2

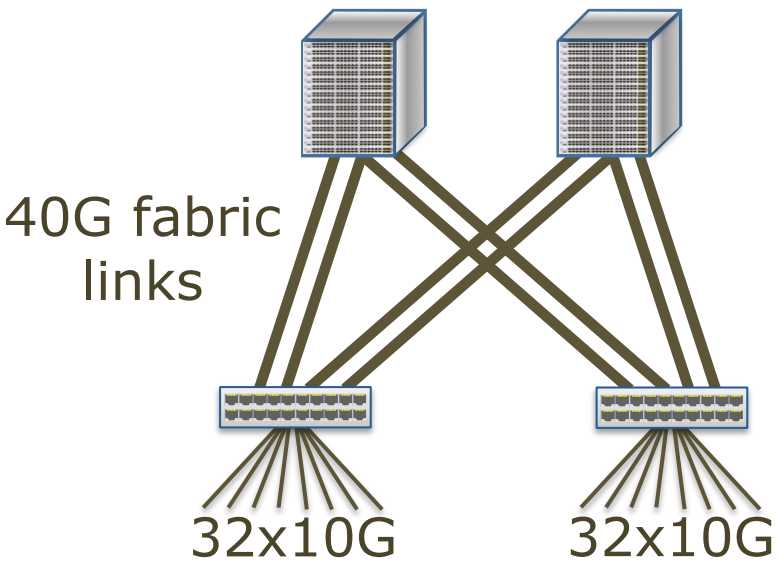
Implementation

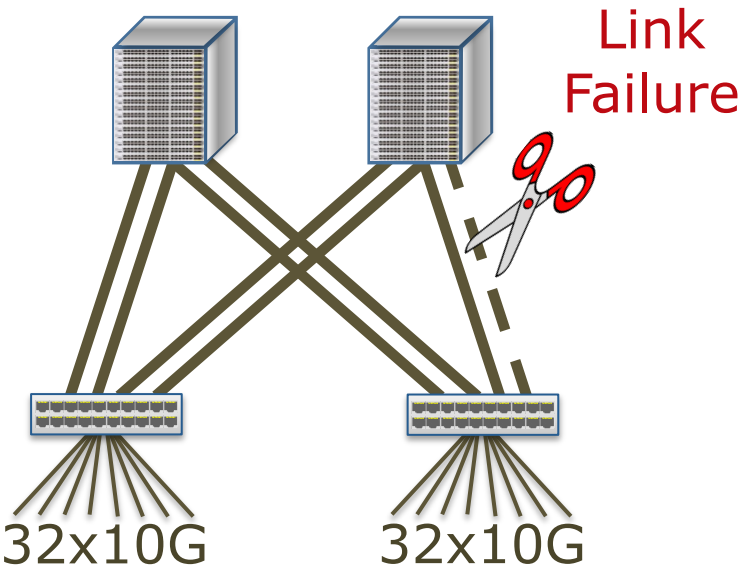
Implemented in silicon for Cisco's new flagship ACI datacenter fabric

- Scales to over 25,000 non-blocking 10G ports (2-tier Leaf-Spine)
- Die area: <2% of chip



Evaluation





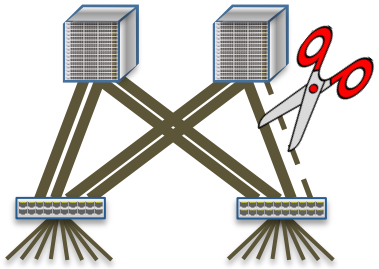
Testbed experiments

- 64 servers, 10/40G switches
- Realistic traffic patterns (enterprise, data-mining)
- HDFS benchmark

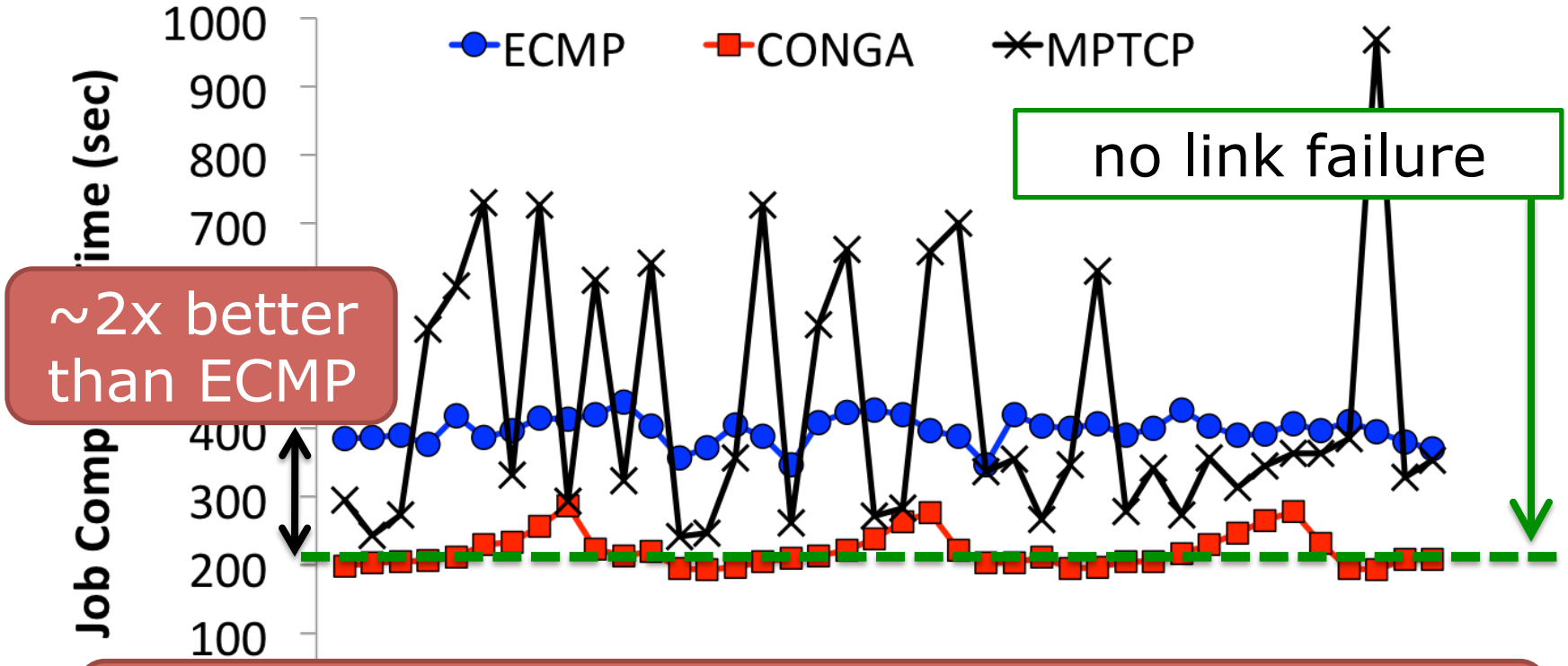
Large-scale simulations

- OMNET++, Linux 2.6.26 TCP
- Varying fabric size, link speed, asymmetry
- Up to 384-port fabric

HDFS Benchmark
1TB Write Test, 40 runs

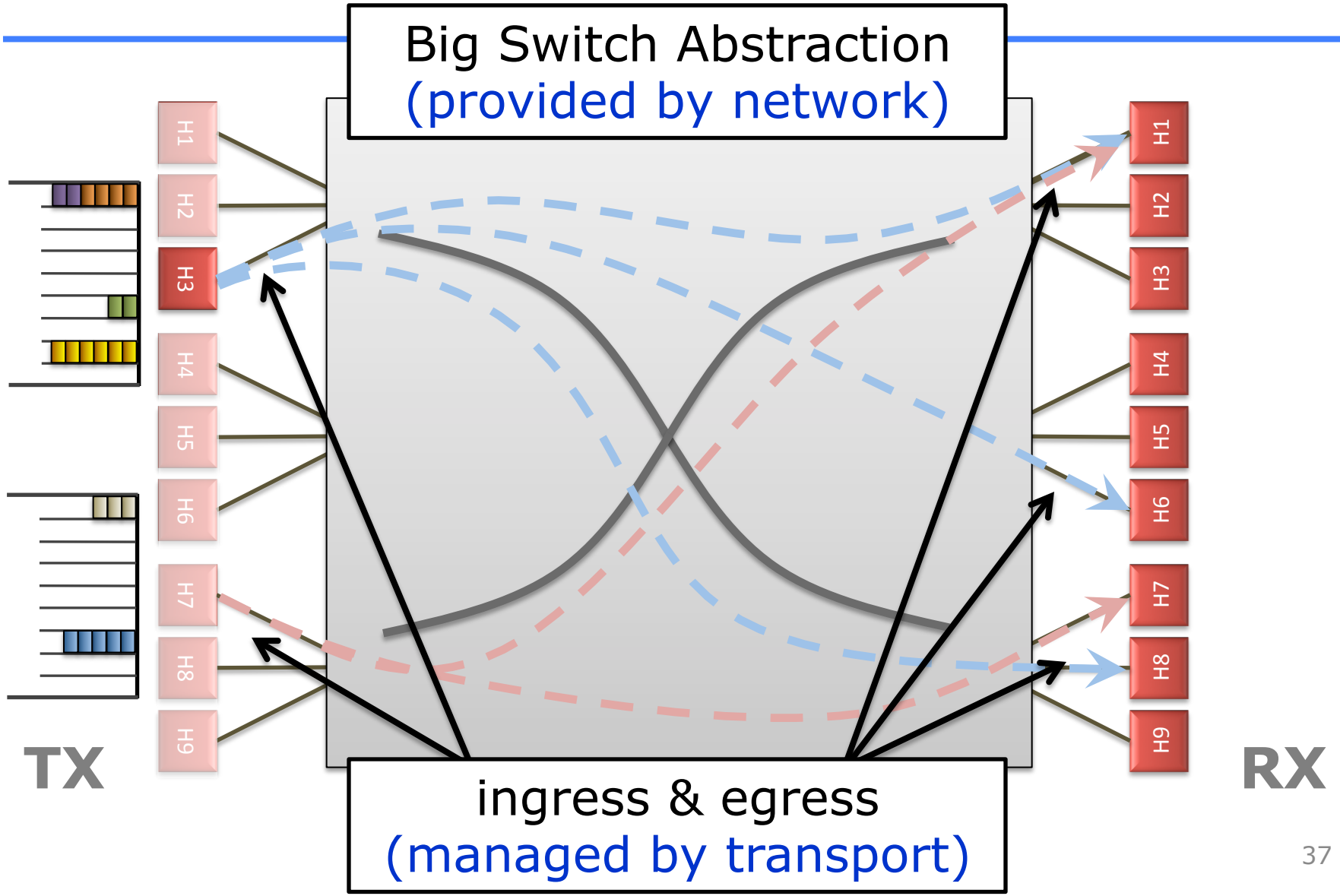


Cloudera hadoop-0.20.2-cdh3u5, 1 NameNode, 63 DataNodes



Link failure has almost no impact with CONGA

Decouple DC LB & Transport



Conclusion

CONGA: Globally congestion-aware LB for DC

… implemented in Cisco ACI datacenter fabric

Key takeaways

1. In-network LB is right for DCs
2. Low latency is your friend; makes feedback control easy

Thank You!