# Wide-area routing & TE

Ankit Singla

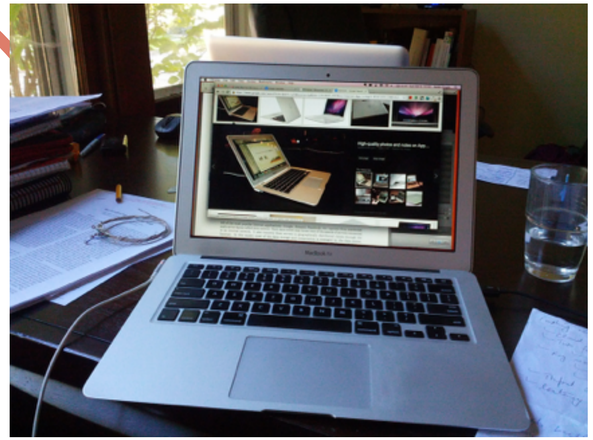
ETH Zürich Spring 2017

# This lecture …

- Zooming out from servers and data centers
- New opportunities in wide-area networking

- Intro: "Network verification & synthesis"
  - *VeriFlow: Verifying Network-Wide Invariants in Real Time*

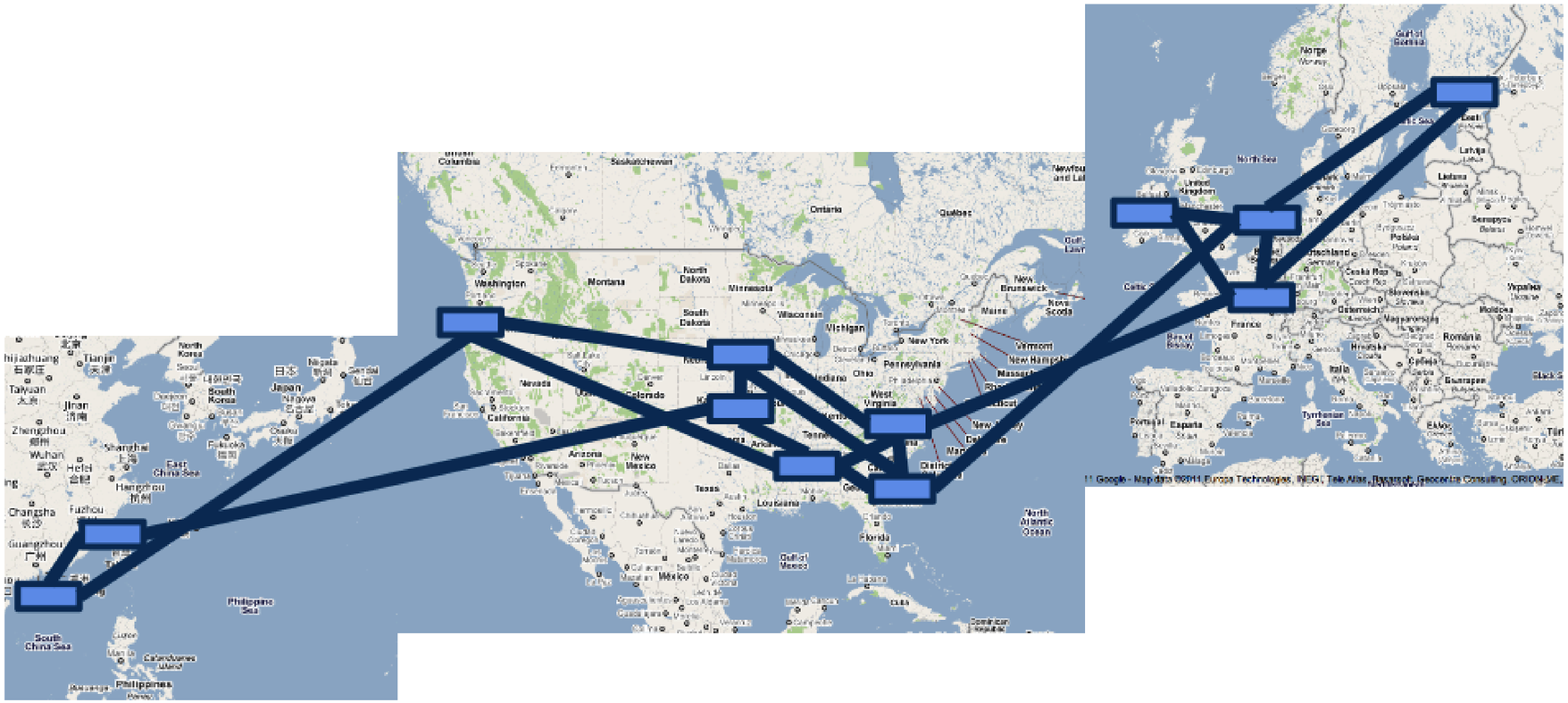
# How large online services work



# How large online services work

The Internet

# Google's WAN (2011)



"B4: Experience with a Globally-Deployed Software Defined WAN"
Jain et al., ACM SIGCOMM 2013

# Why multiple data centers?

- Data availability
- Load balancing
- Latency
- Local data laws
- Hybrid public-private operation

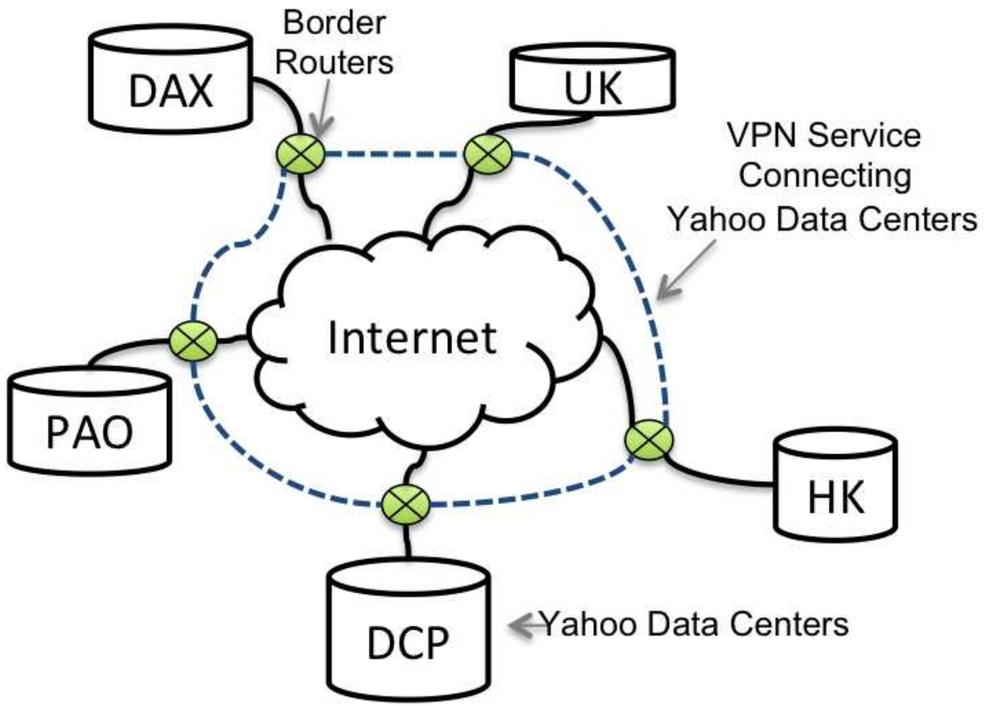
# Inter-data center traffic is significant

**IEEE INFOCOM, 2011**

A First Look at Inter-Data Center Traffic Characteristics via Yahoo! Datasets

Yingying Chen[1], Sourabh Jain[1], Vijay Kumar Adhikari[1], Zhi-Li Zhang[1], and Kuai Xu[2]

[1]University of Minnesota-Twin Cities
[2]Arizona State University



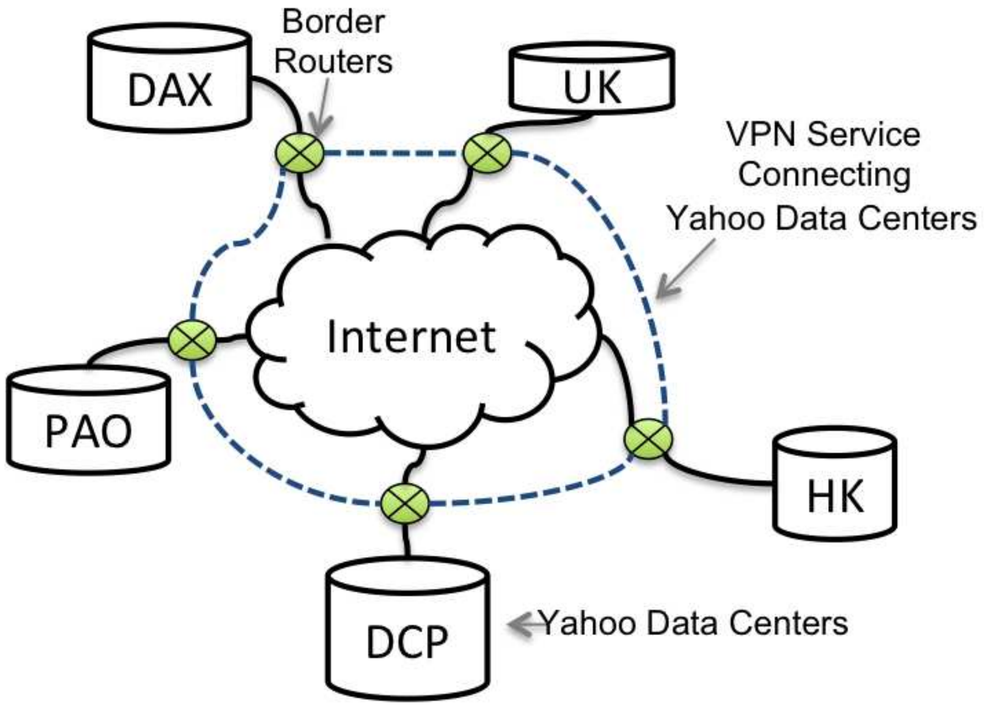
# Inter-data center traffic is significant

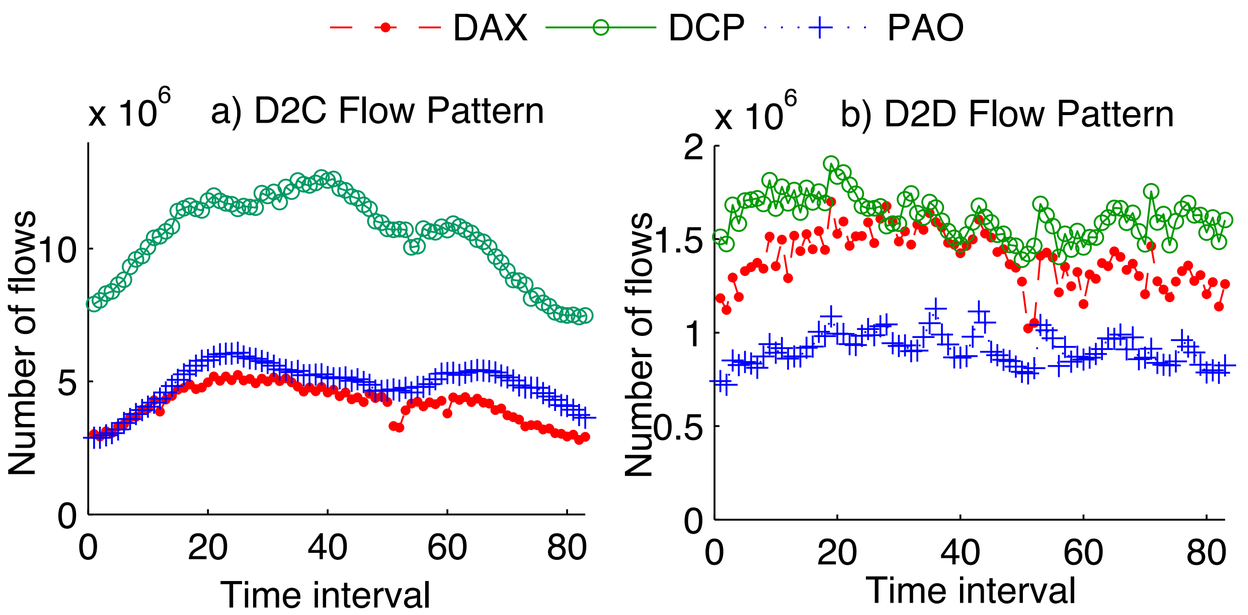
**IEEE INFOCOM, 2011**

A First Look at Inter-Data Center Traffic Characteristics via Yahoo! Datasets

Yingying Chen[1], Sourabh Jain[1], Vijay Kumar Adhikari[1], Zhi-Li Zhang[1], and Kuai Xu[2]

[1]University of Minnesota-Twin Cities
[2]Arizona State University



# Intra-CDN WAN traffic

**ACM IMC, 2014**

## Back-Office Web Traffic on The Internet


Enric Pujol
TU Berlin
enric@inet.tu-berlin.de

Philipp Richter
TU Berlin
prichter@inet.tu-berlin.de

Balakrishnan Chandrasekaran
Duke University
balac@cs.duke.edu

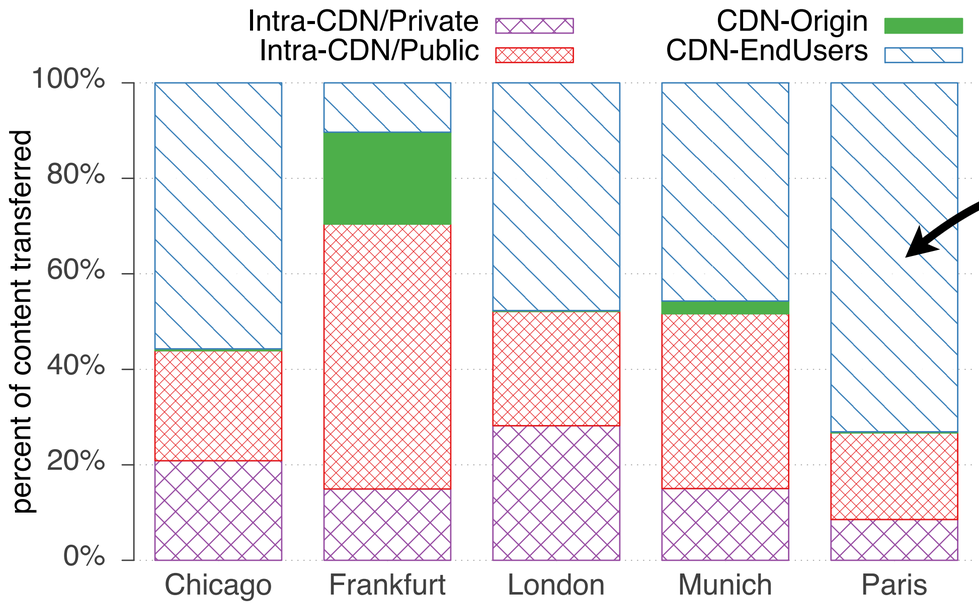
Georgios Smaragdakis
MIT / TU Berlin / Akamai
gsmaragd@csail.mit.edu

Anja Feldmann
TU Berlin
anja@inet.tu-berlin.de

Bruce Maggs
Duke / Akamai
bmm@cs.duke.edu

Keung-Chi Ng
Akamai
kng@akamai.com


*"a major CDN"*




End-user traffic

# Google's WAN (2011)


*"B4 has been in deployment for three years, now carries more traffic than Google's public facing WAN, and has a higher growth rate."*

"B4: Experience with a Globally-Deployed Software Defined WAN"
Jain et al., ACM SIGCOMM 2013

# Why are these networks different?

dedicated, 100s of Gbps

**Persistent connectivity between a (small) set of end-points**

**Internet**
(end-host to app / end-host)

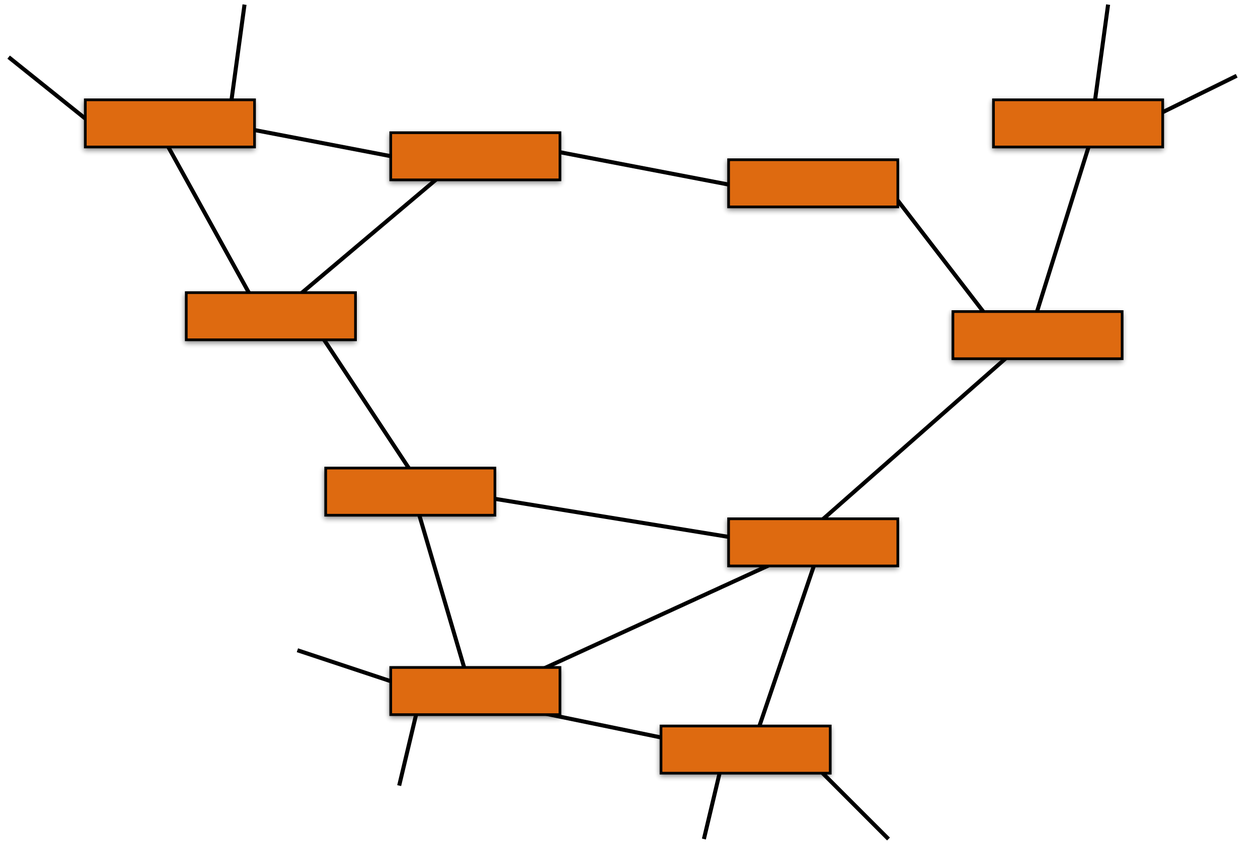
**WAN?**

**Private data center**

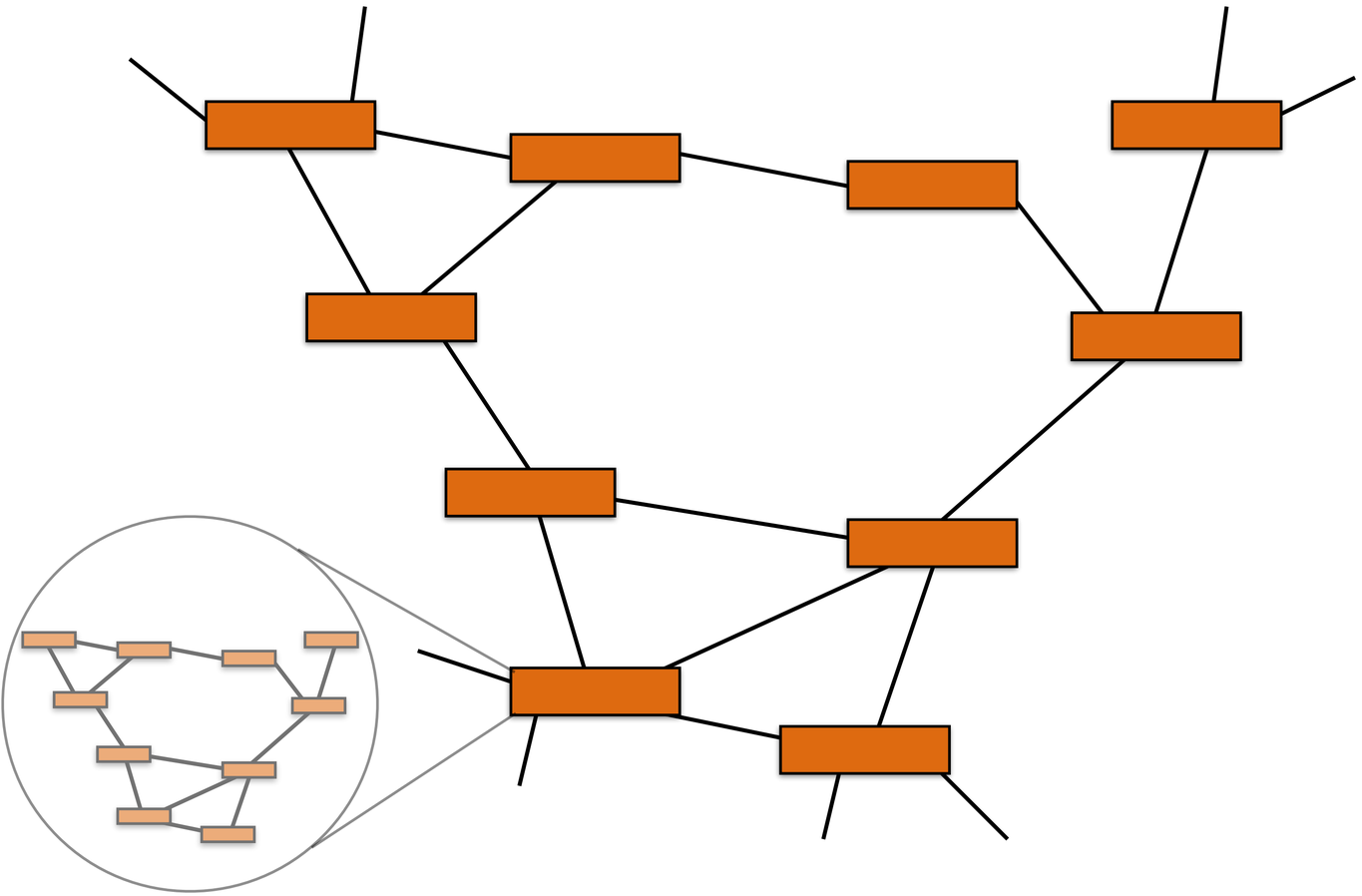

Microsoft: "expensive resource, with amortized annual cost of 100s of millions of dollars"

[Achieving High Utilization with Software-Driven WAN, Hong et al., ACM SIGCOMM'13]

# Traditional WAN approach: MPLS

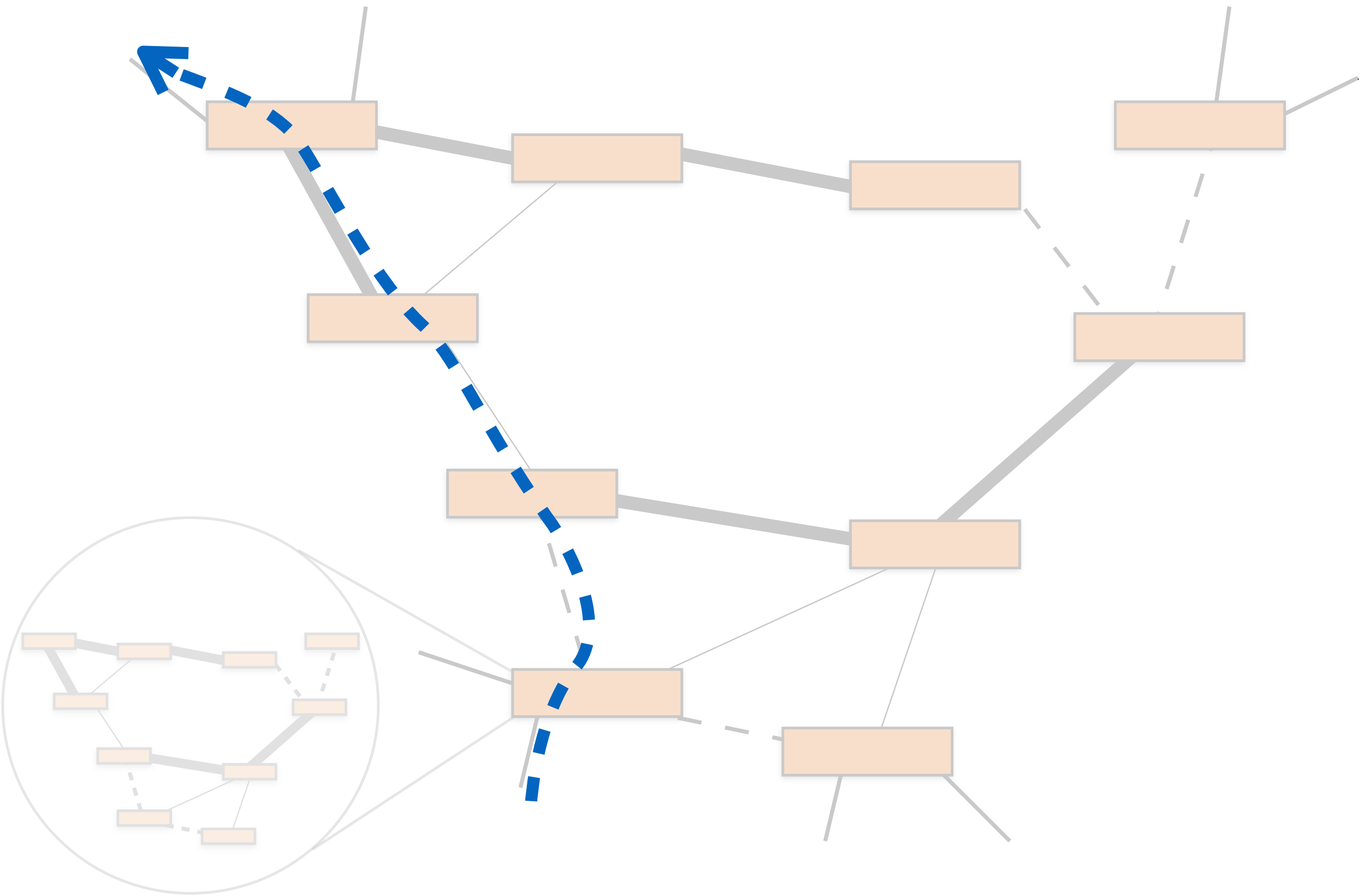


# Traditional WAN approach: MPLS



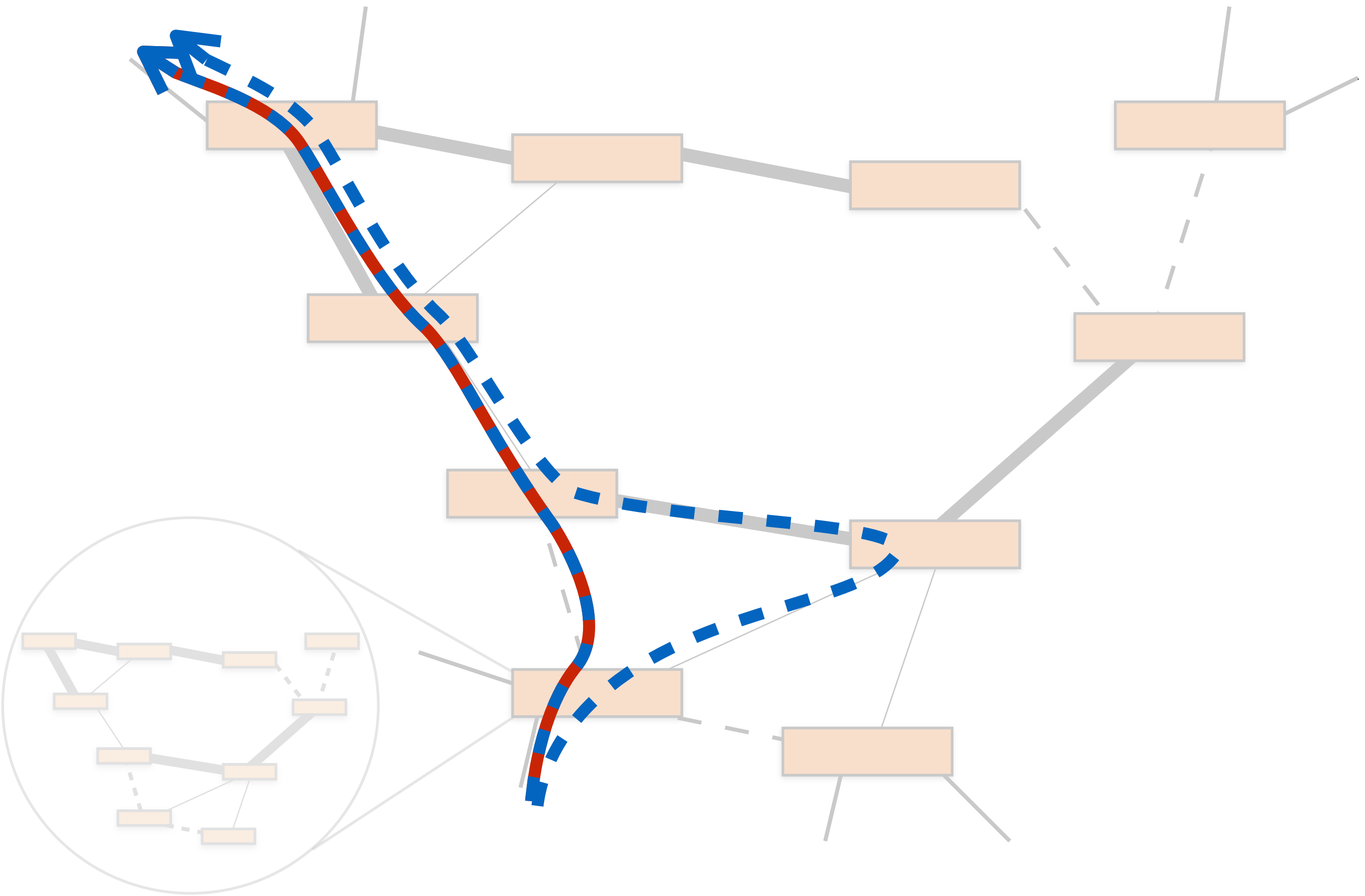
1. Link-state protocol (OSPF / IS-IS)

# Traditional WAN approach: MPLS



1. Link-state protocol (OSPF / IS-IS)
2. Also flood available bandwidth info
3. Fulfill tunnel provisioning requests

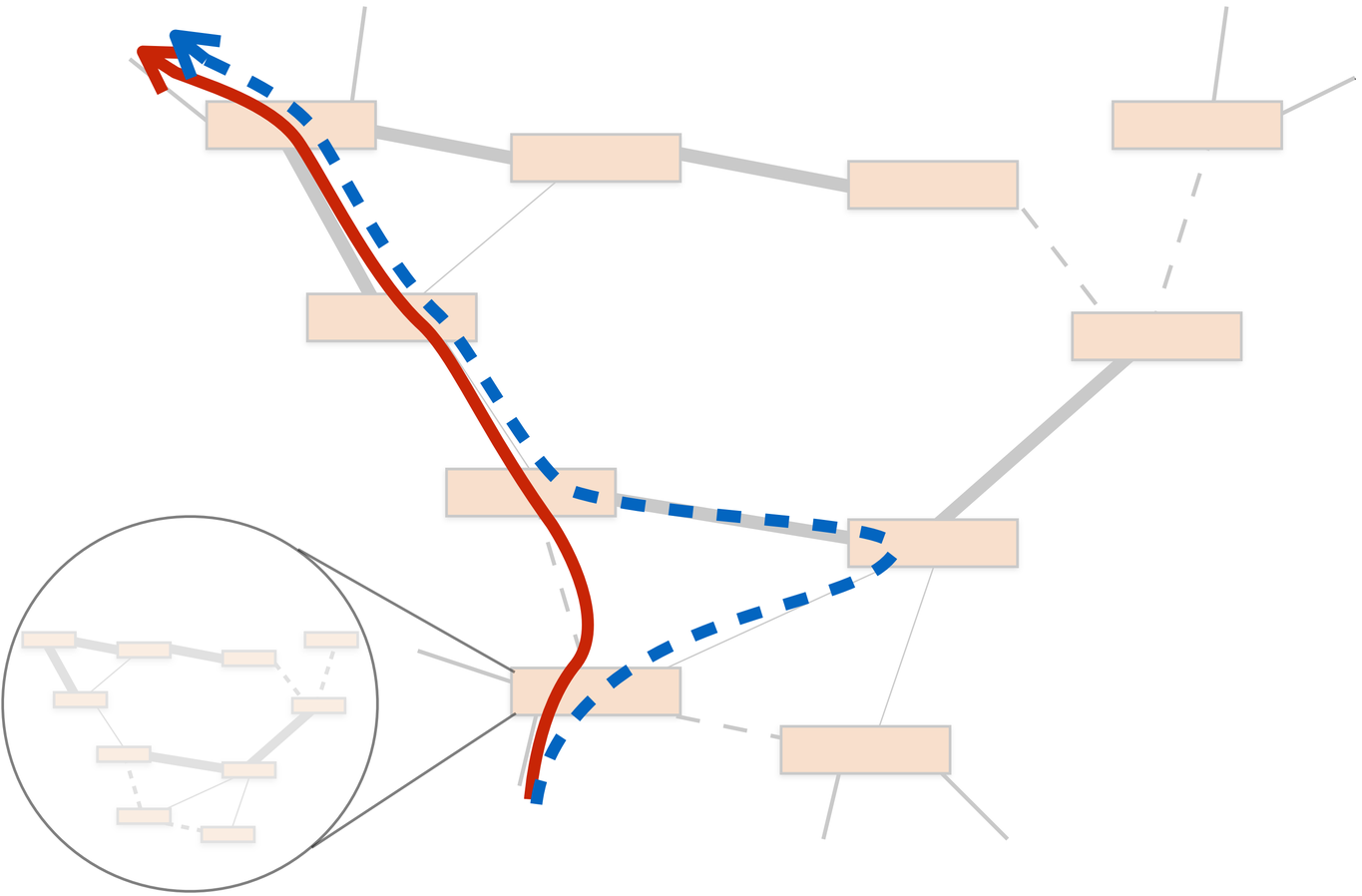
# Traditional WAN approach: MPLS



1. Link-state protocol (OSPF / IS-IS)
2. Also flood available bandwidth info
3. Fulfill tunnel provisioning requests

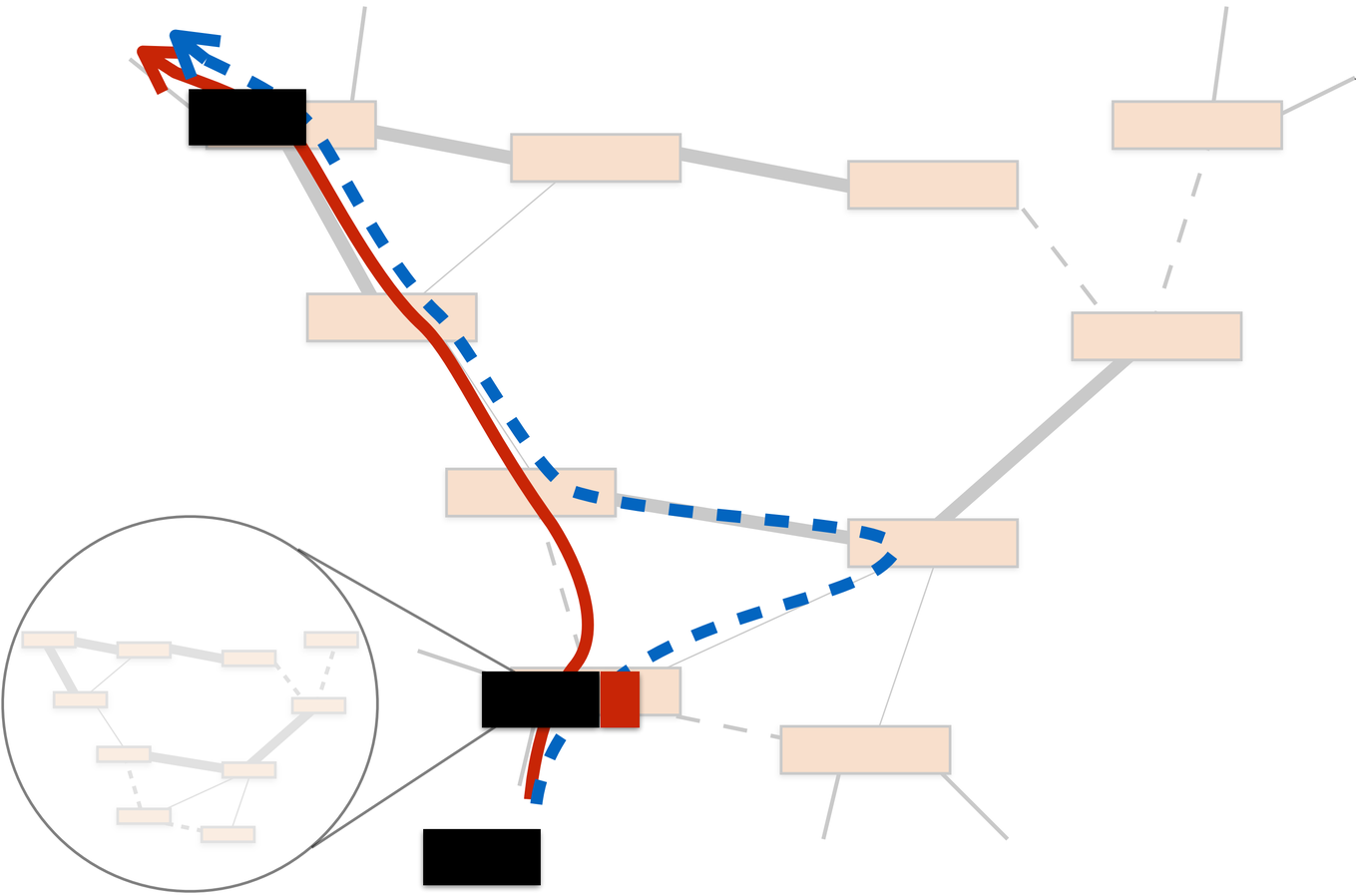
Priorities

# Traditional WAN approach: MPLS



1. Link-state protocol (OSPF / IS-IS)
2. Also flood available bandwidth info
3. Fulfill tunnel provisioning requests
4. Update network state, flood info

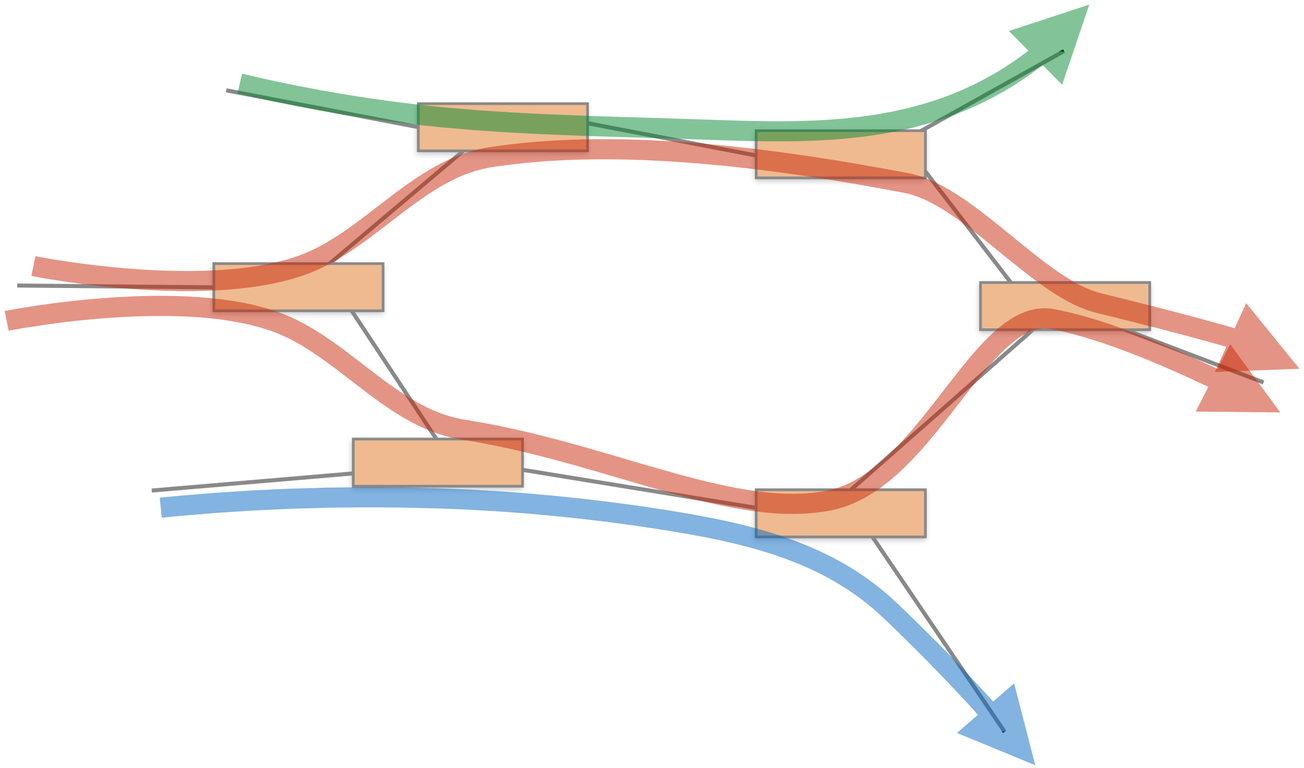
# Traditional WAN approach: MPLS



1. Link-state protocol (OSPF / IS-IS)
2. Also flood available bandwidth info
3. Fulfill tunnel provisioning requests
4. Update network state, flood info

**Only ingress and egress read packet**

# Problem 1:  inflexible sharing



**2x the**

# Problem 2:  complex!

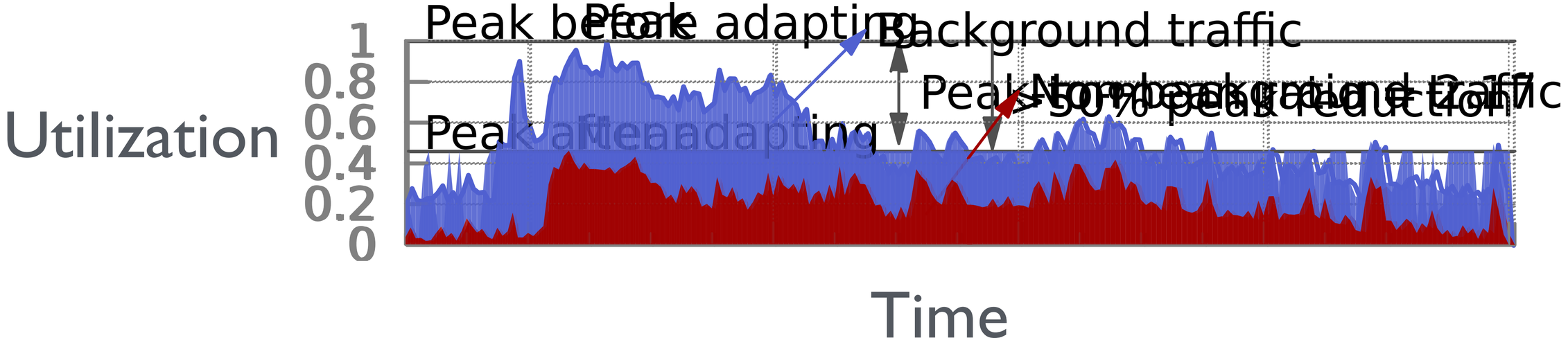
# Problem 3: inefficiency

**ACM SIGCOMM, 2013**

**Achieving High Utilization with Software-Driven WAN**

Chi-Yao Hong (UIUC) Srikanth Kandula Ratul Mahajan Ming Zhang
Vijay Gill Mohan Nanduri Roger Wattenhofer (ETH)

Microsoft



# Cutting-edge WAN TE

## B4

ACM SIGCOMM, 2013

**B4: Experience with a Globally-Deployed Software Defined WAN**

Sushant Jain, Alok Kumar, Subhasree Mandal, Joon Ong, Leon Poutievski, Arjun Singh,
Subbaiah Venkata, Jim Wanderer, Junlan Zhou, Min Zhu, Jonathan Zolla,
Urs Hölzle, Stephen Stuart and Amin Vahdat
Google, Inc.
b4-sigcomm@google.com

## SWAN

ACM SIGCOMM, 2013

**Achieving High Utilization with Software-Driven WAN**

Chi-Yao Hong (UIUC) Srikanth Kandula Ratul Mahajan Ming Zhang
Vijay Gill Mohan Nanduri Roger Wattenhofer (ETH)

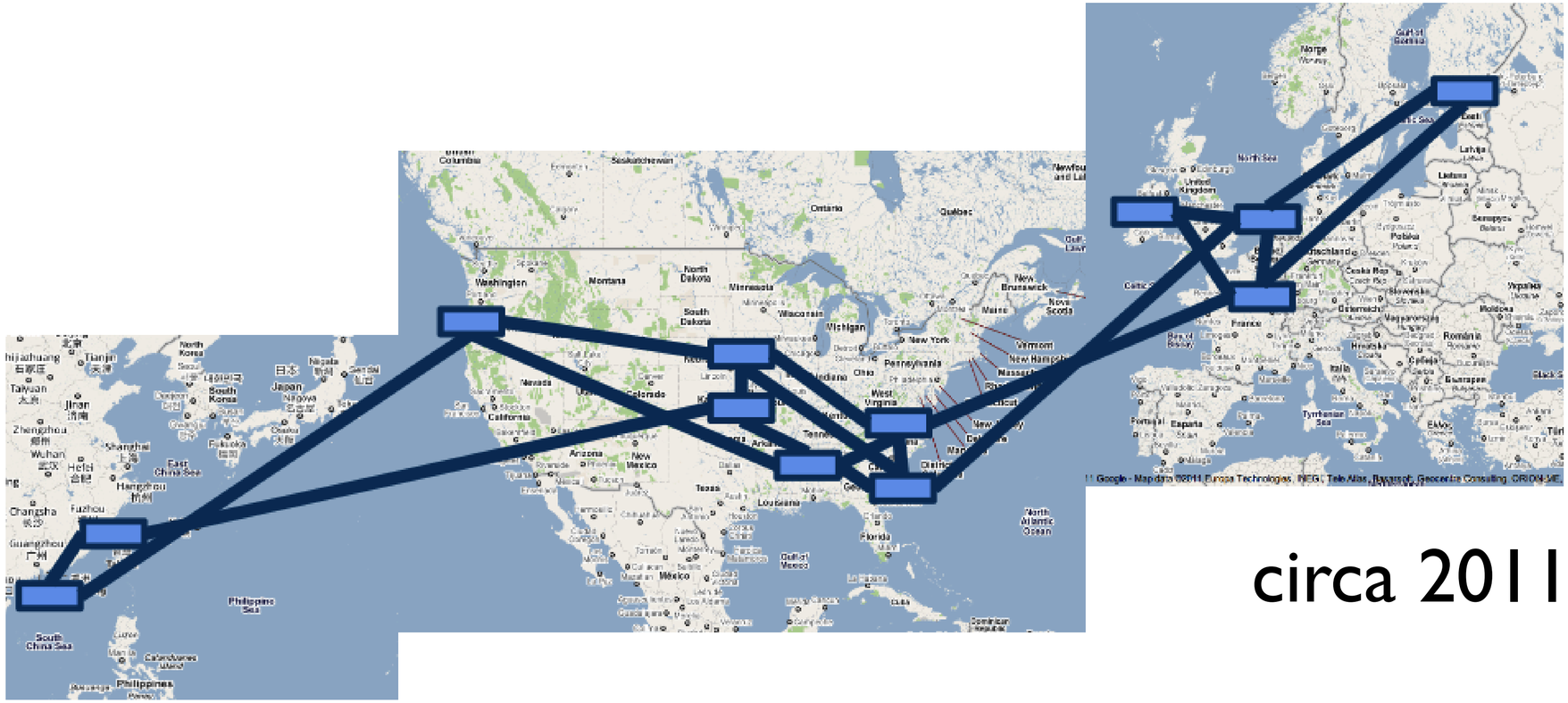
Microsoft

# Common themes

1. Leverage service diversity: some tolerate delay
2. Centralized TE using SDN, OpenFlow
3. Exact linear programming is too slow!
4. Dynamic reallocation of bandwidth
5. Edge rate limiting

# Google's B4

"B4: Experience with a Globally-Deployed Software Defined WAN"
Jain et al., ACM SIGCOMM 2013

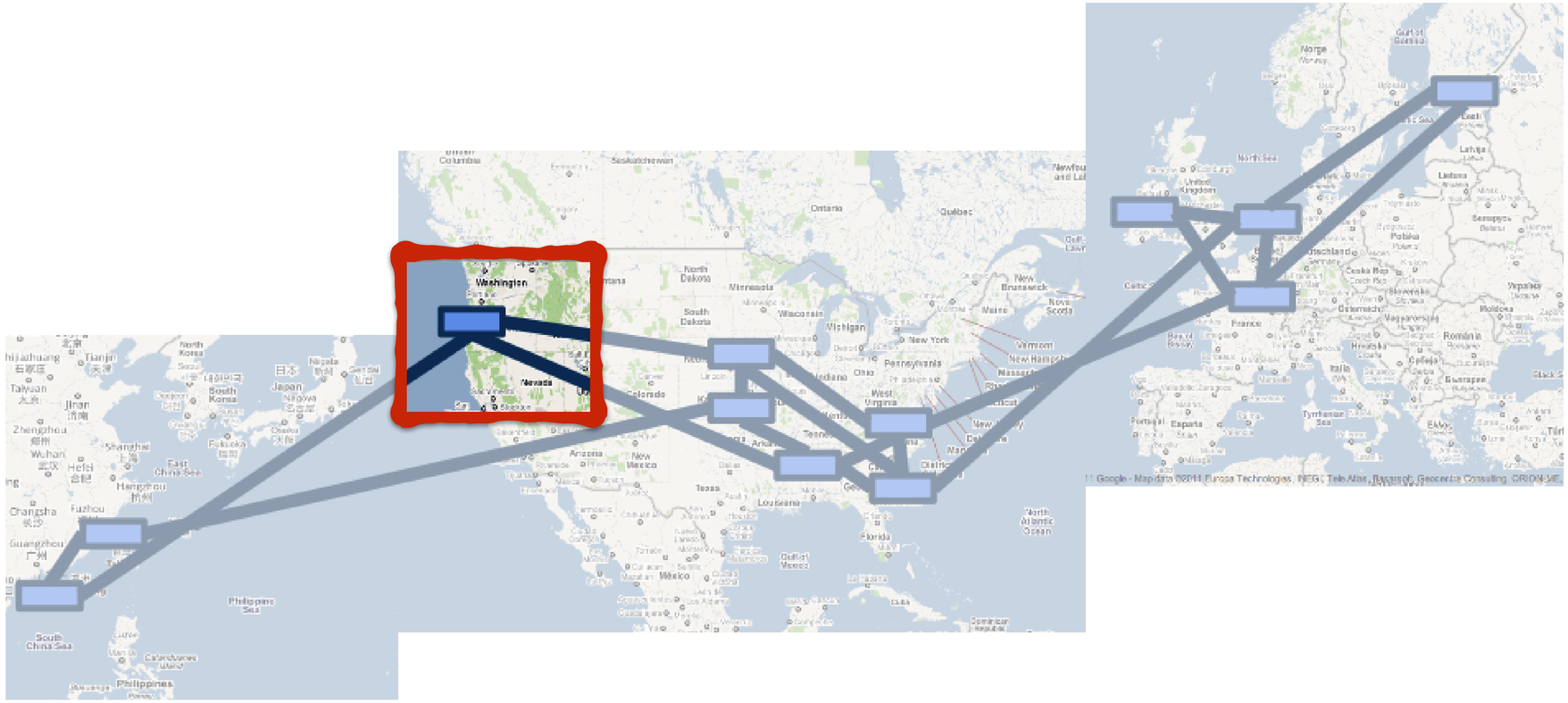


circa 2011

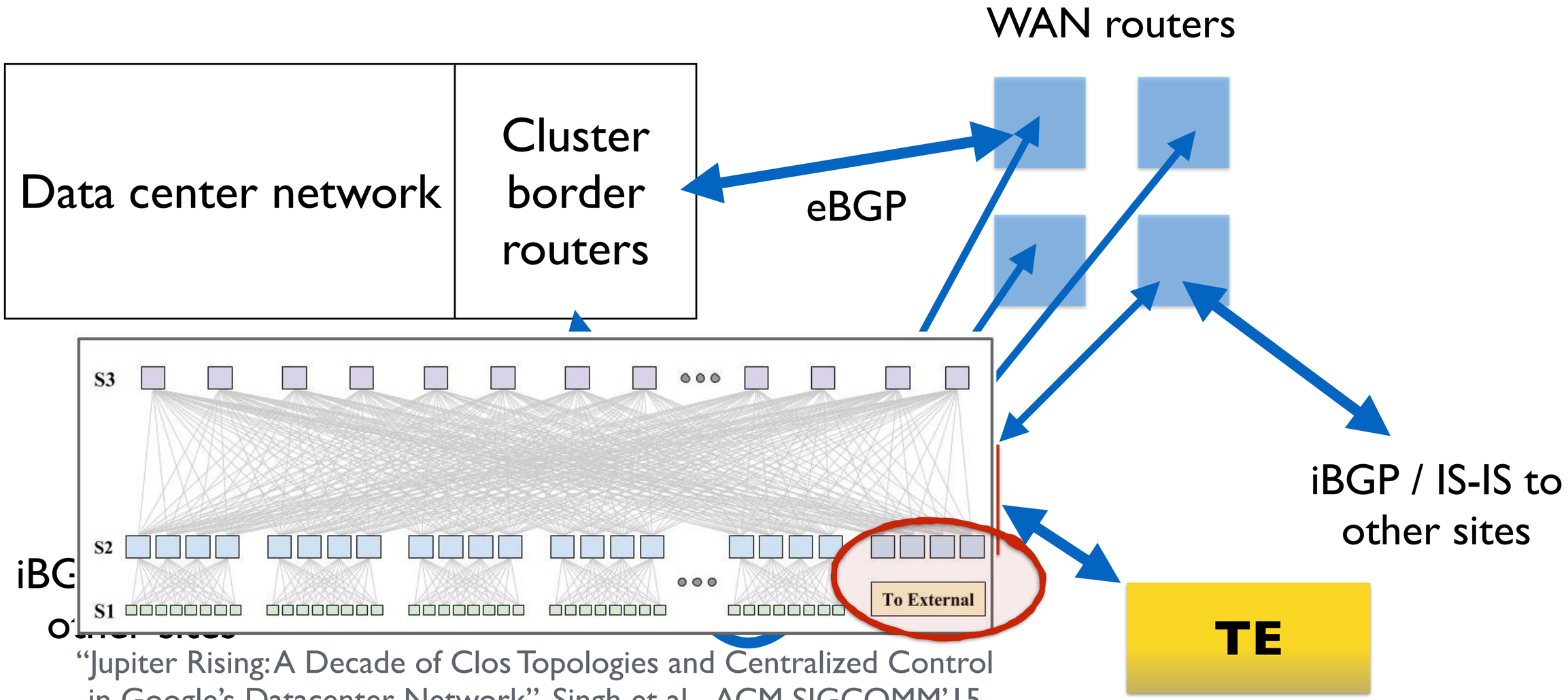
**First highly visible SDN success**

# Google's B4

"B4: Experience with a Globally-Deployed Software Defined WAN"
Jain et al., ACM SIGCOMM 2013

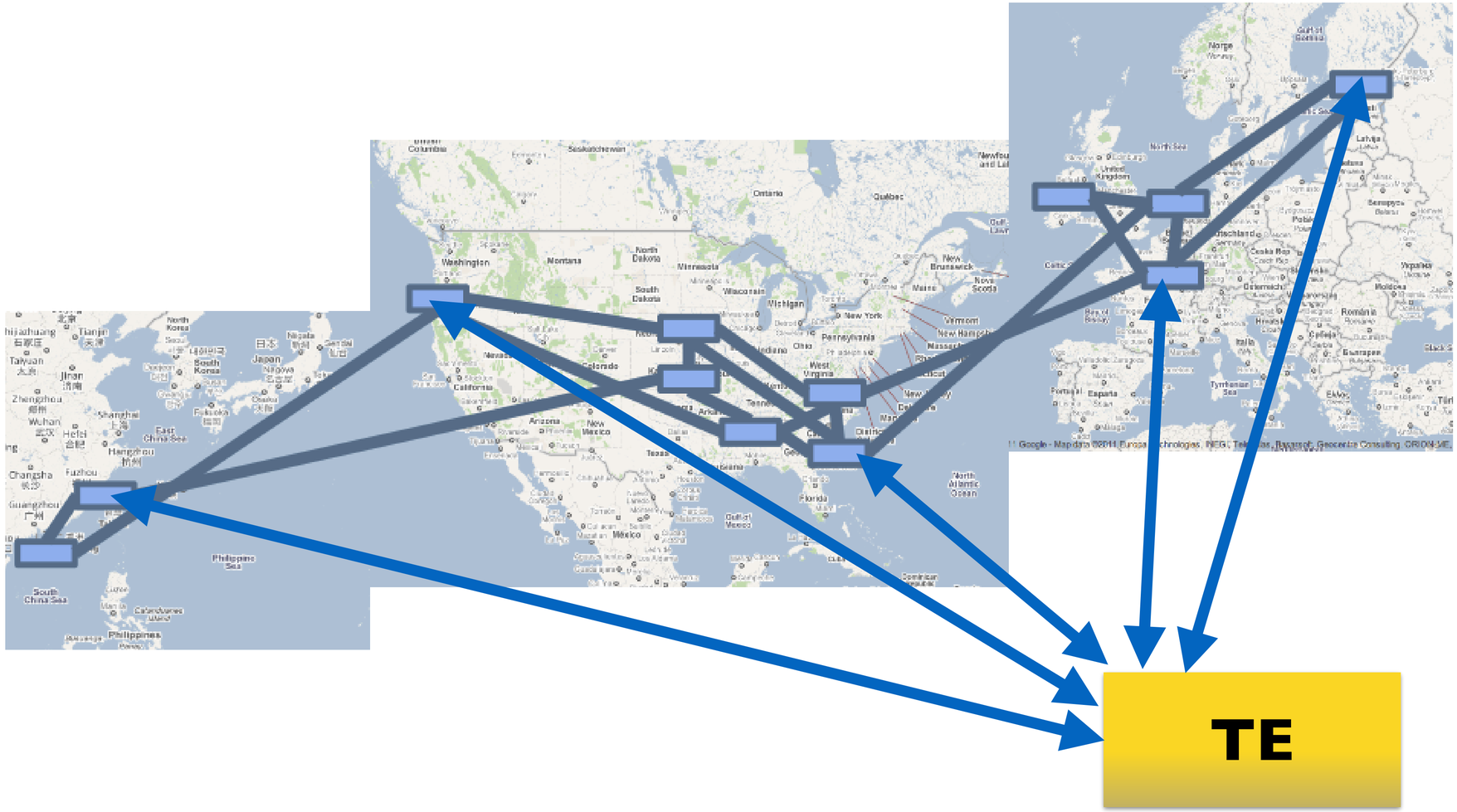


# Google's B4: view at one site




"Jupiter Rising: A Decade of Clos Topologies and Centralized Control in Google's Datacenter Network", Singh et al., ACM SIGCOMM'15

# Google’s B4: Traffic engineering

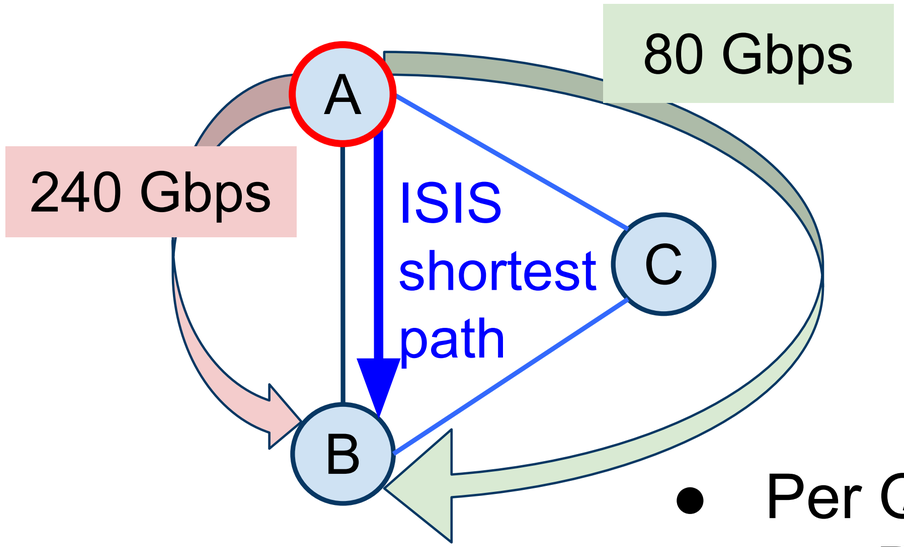


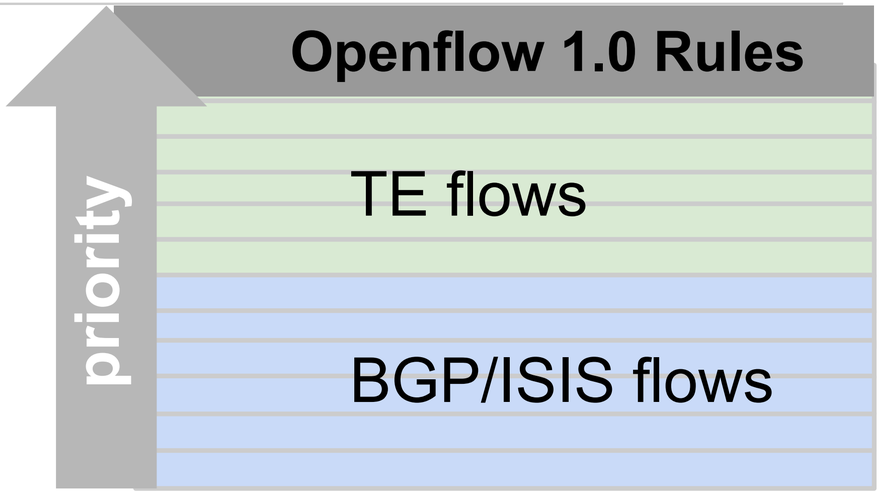
# Google's B4

"B4: Experience with a Globally-Deployed Software Defined WAN"
Jain et al., ACM SIGCOMM 2013

- BGP routing as "big red switch"

# Traffic Engineering Overlay



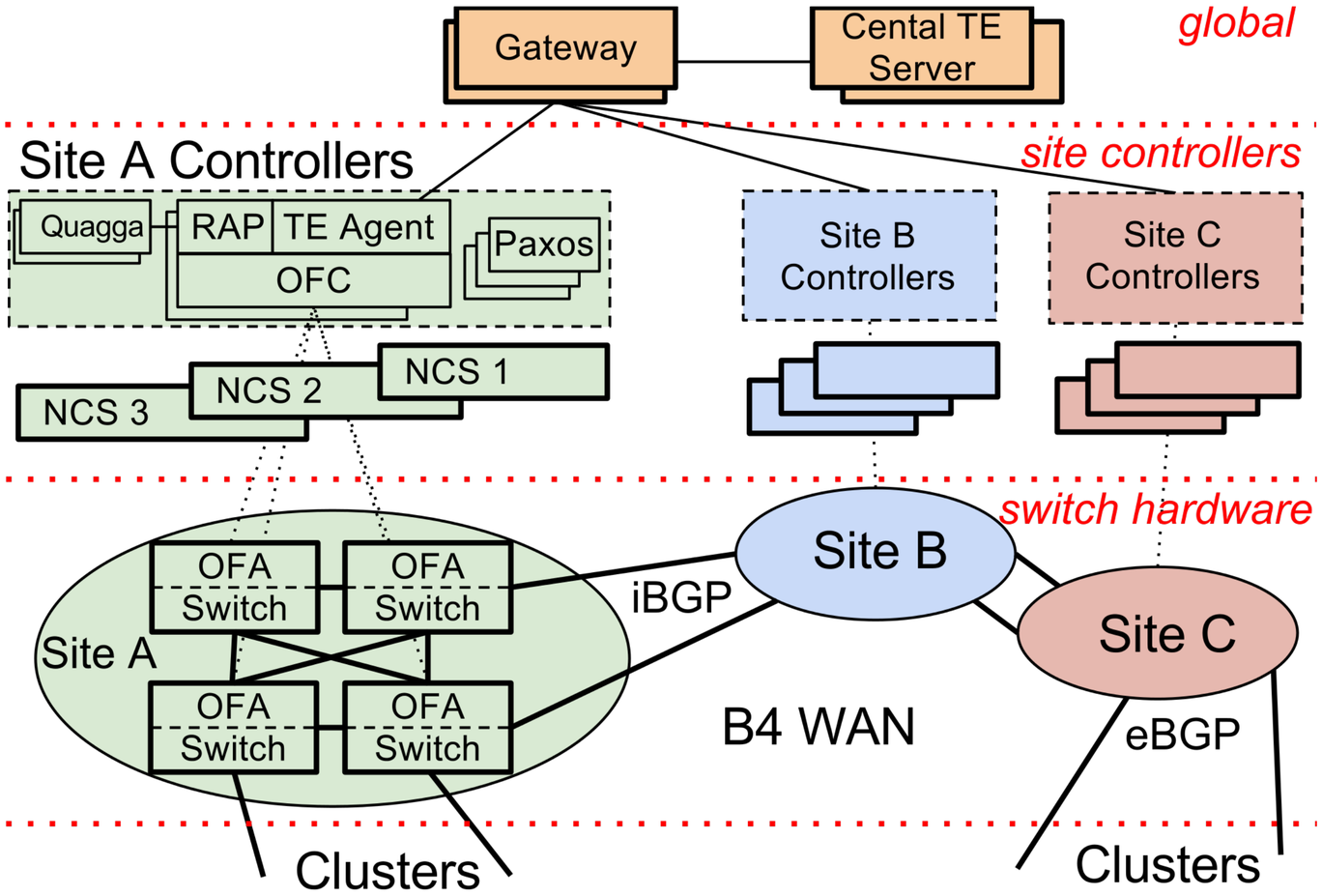




- Per QoS Traffic Engineering (TE)
  - Demand based use of longer paths
  - Max-min fair bandwidth allocation
  - Per app loss/latency/throughput consideration
- TE paths are overlaid on ISIS/BGP routes
  - Higher priority flow rules for TE

"Lessons Learned from B4, Google's SDN WAN"
Subhasree Mandal, talk at USENIX ATC 2015

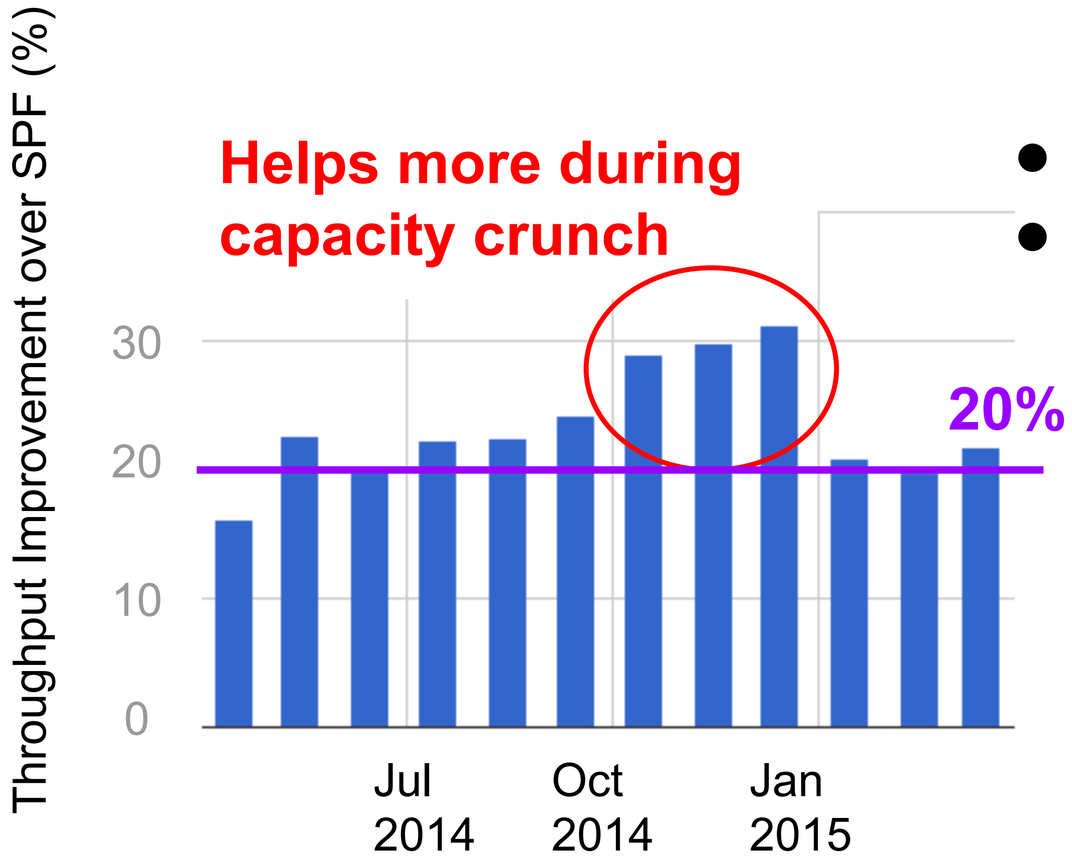
# Google's B4

"B4: Experience with a Globally-Deployed Software Defined WAN"
Jain et al., ACM SIGCOMM 2013




**Nearly 100% utilization, 0.3s TE solution**

# Benefits of TE Over Shortest Path




- ~20% increase in throughput over SPF
- Larger benefits during capacity crunch

Lowers the requirement for bandwidth provisioning

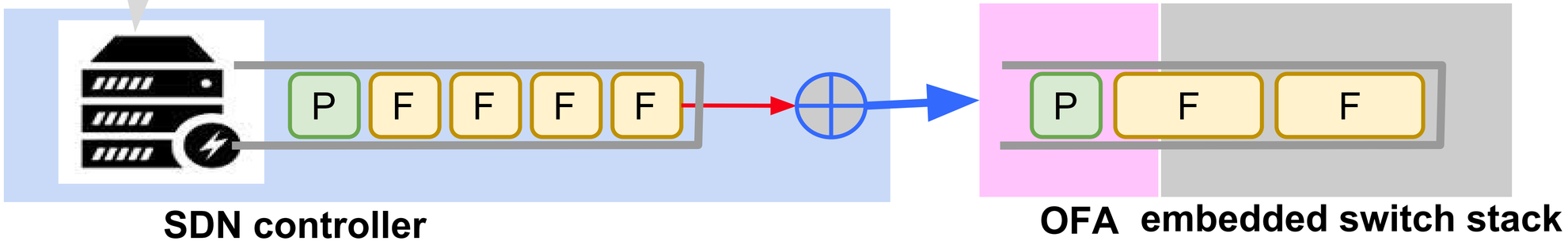
"Lessons Learned from B4, Google's SDN WAN"
Subhasree Mandal, talk at USENIX ATC 2015

# Messages Backlogged and Delayed!

Google

Fast server

Queue build-up on controller and switch due to slow switch CPU



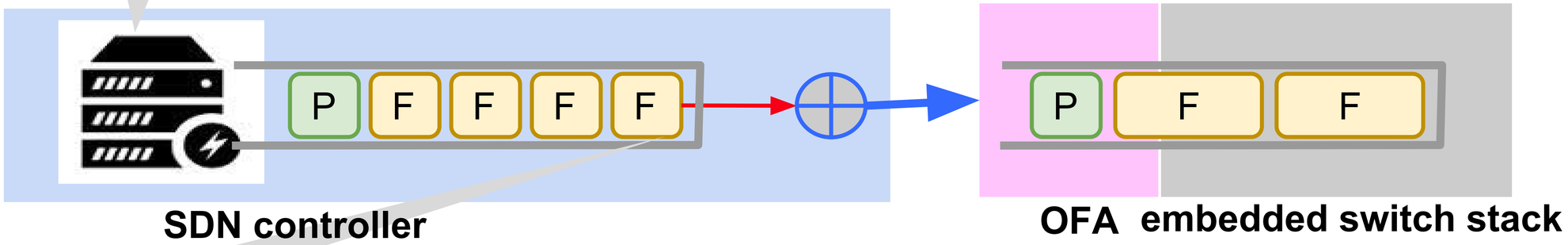

"Lessons Learned from B4, Google's SDN WAN"
Subhasree Mandal, talk at USENIX ATC 2015

# Messages Backlogged and Delayed!

Google

Fast server

Queue build-up on controller and switch due to slow switch CPU




Flow rules generated in bursts

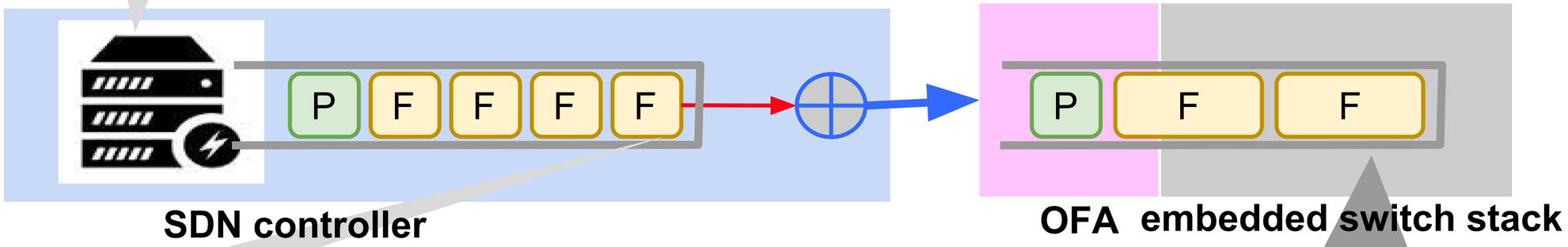
"Lessons Learned from B4, Google's SDN WAN"
Subhasree Mandal, talk at USENIX ATC 2015

# Messages Backlogged and Delayed!

Google

Fast server

Queue build-up on controller and switch due to slow switch CPU




Flow rules generated in bursts

Flow programming in HW is slow

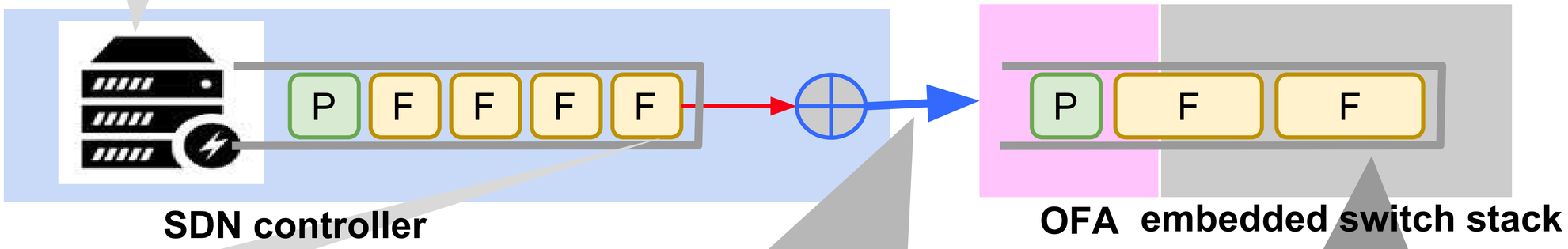
"Lessons Learned from B4, Google's SDN WAN"
Subhasree Mandal, talk at USENIX ATC 2015

# Messages Backlogged and Delayed!

Google

Fast server

Queue build-up on controller and switch due to slow switch CPU




Flow rules generated in bursts

Single OpenFlow connection between controller and OFA

Flow programming in HW is slow

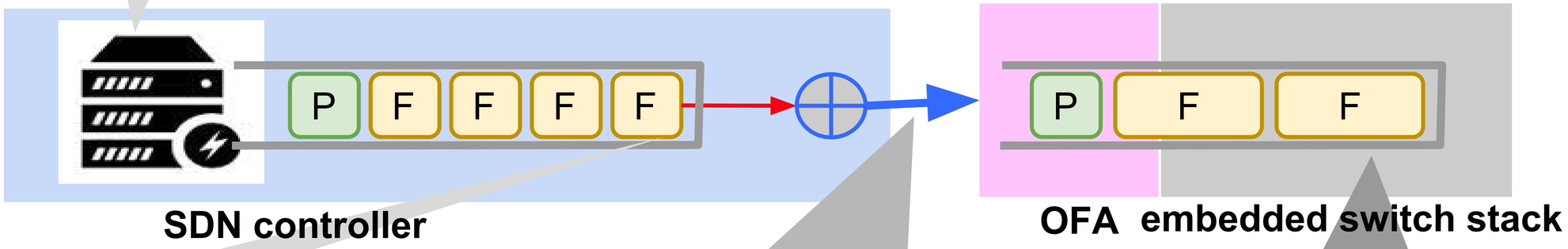
"Lessons Learned from B4, Google's SDN WAN"
Subhasree Mandal, talk at USENIX ATC 2015

# Messages Backlogged and Delayed!

Google

Fast server

Queue build-up on controller and switch due to slow switch CPU




Flow rules generated in bursts

Single OpenFlow connection between controller and OFA

Flow programming in HW is slow

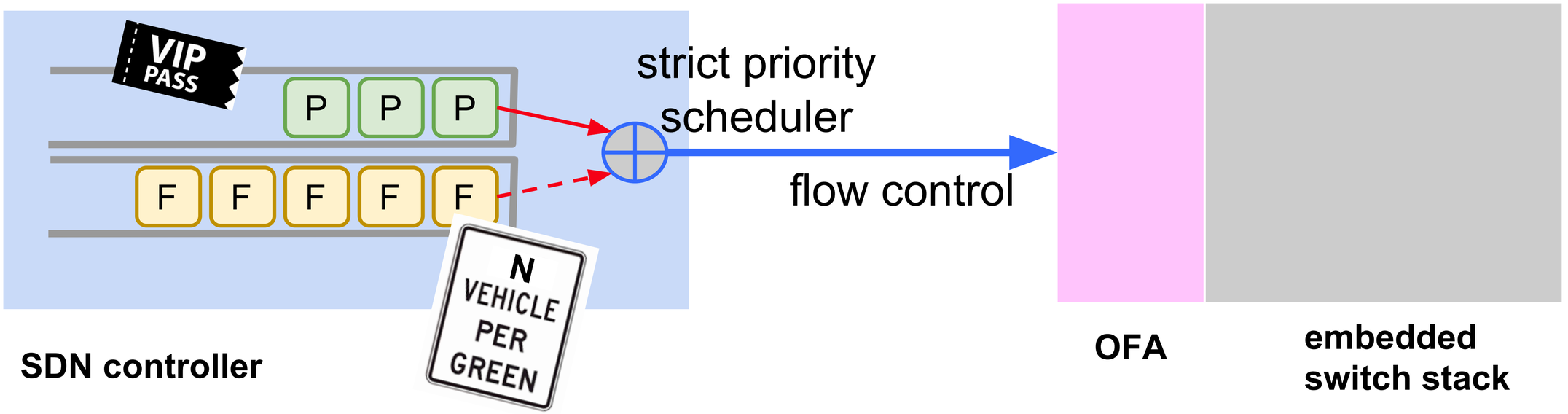
Flow rules cause HOL blocking for packets

packets delayed

protocols timeout
reconvergence produces more flow rules

"Lessons Learned from B4, Google's SDN WAN"
Subhasree Mandal, talk at USENIX ATC 2015

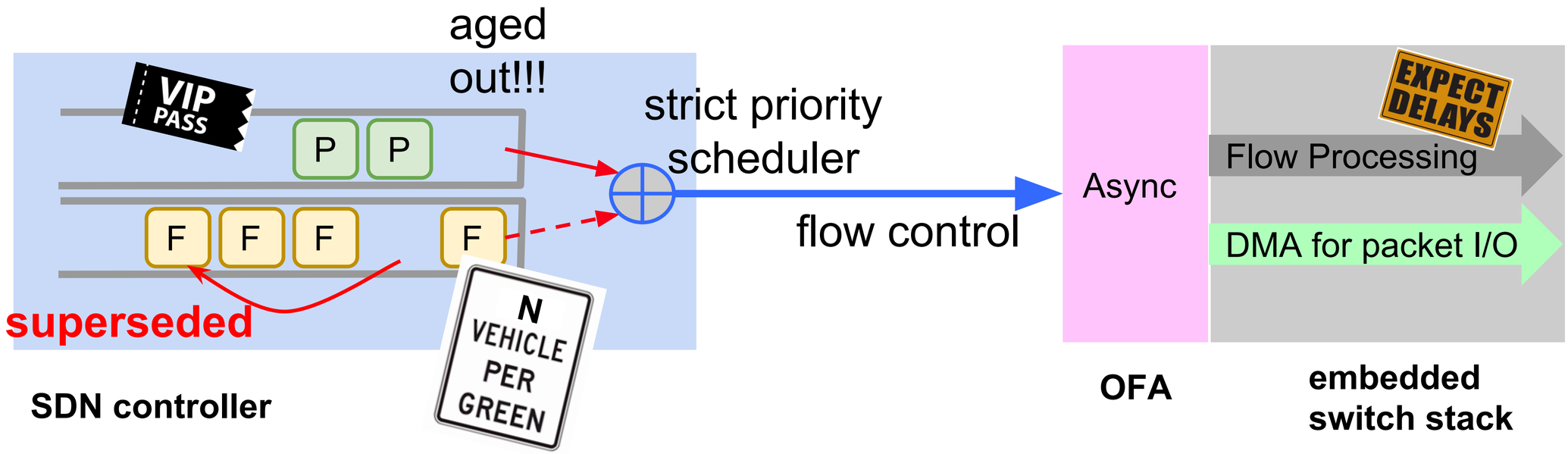
# Lesson: Mitigation with Flow Control




- Separate queue for packet IO and flow request
- Strict priority for packet IO over flow programming
- Limit queue depth in OFA: token based flow control

"Lessons Learned from B4, Google's SDN WAN"
Subhasree Mandal, talk at USENIX ATC 2015

# Lesson: Mitigation with Flow Control




- Separate queue for packet IO and flow request
- Strict priority for packet IO over flow programming
- Limit queue depth in OFA: token based flow control
- Systematics queue drop discipline

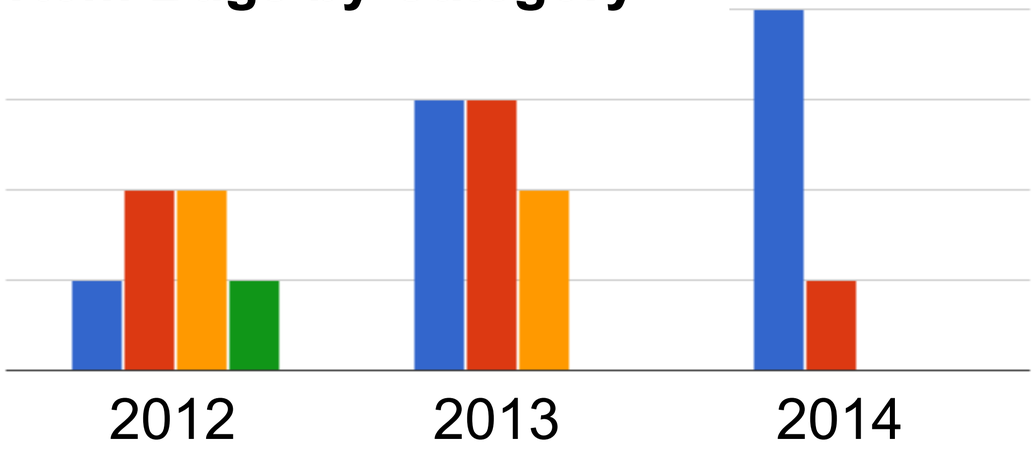
- Asynchronous OFA
- Packet IO out of flow processing pipeline

"Lessons Learned from B4, Google's SDN WAN"
Subhasree Mandal, talk at USENIX ATC 2015

# Outages!!!

Google

**Postmortem Bugs by Category**




Worst Offender

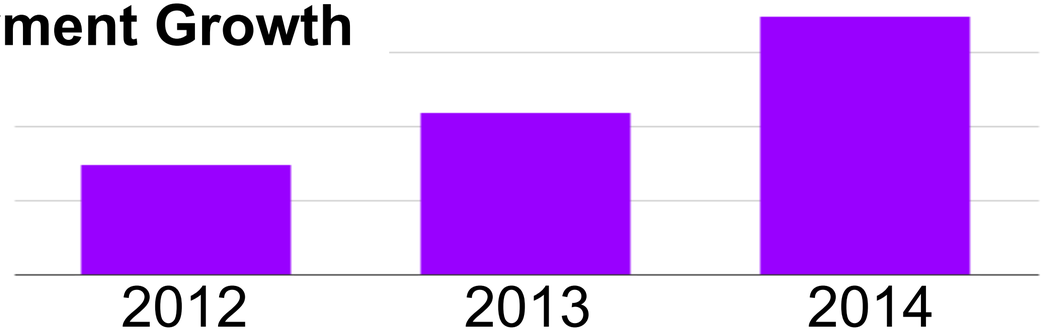
Unstable mastership

Operational Procedure/Tools

Core Software Bugs

Unsupported Software

**Deployment Growth**



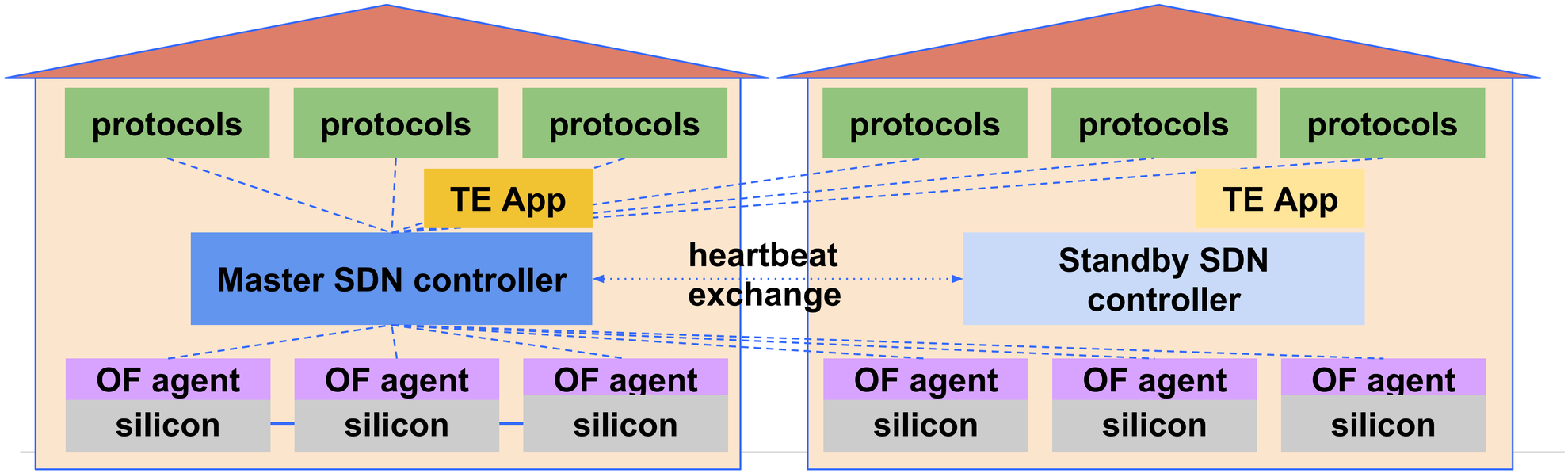

Sites

"Lessons Learned from B4, Google's SDN WAN"
Subhasree Mandal, talk at USENIX ATC 2015

# Control Plane Connectivity: Mastership Google

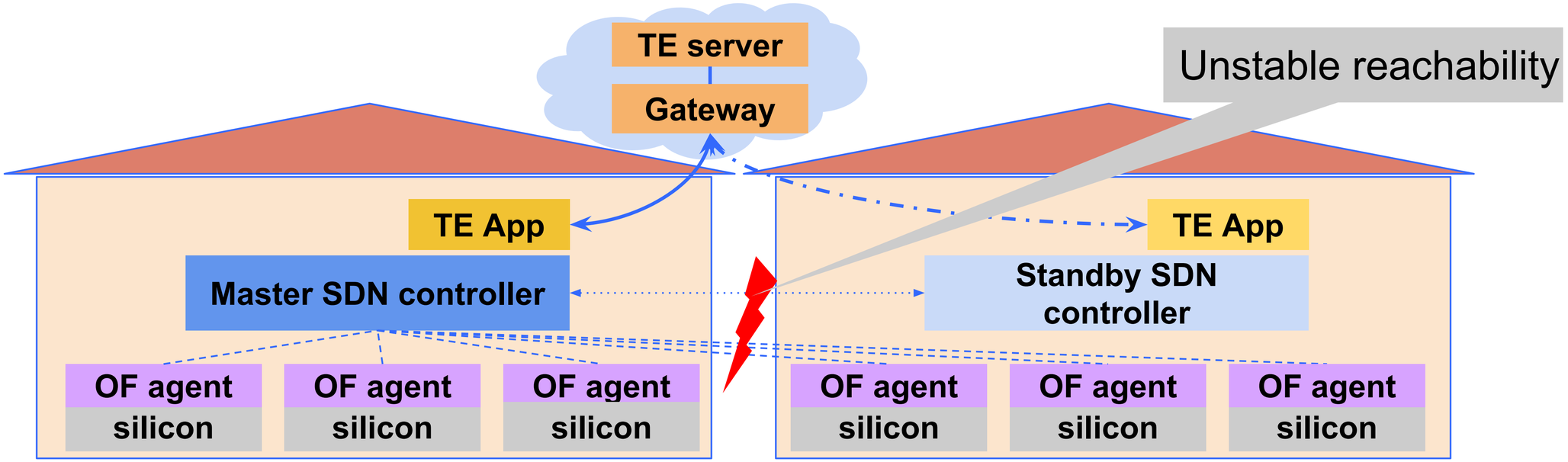
Initial naive design:

- Symmetry between buildings
- Each building can run independently, even if the other one is down
- N+1 controller redundancy sufficient for upgrades, failures etc.




"Lessons Learned from B4, Google's SDN WAN"
Subhasree Mandal, talk at USENIX ATC 2015

# Control Network: Unstable Mastership




- Both controllers declare mastership:
  - Gateway and OFAs can observe mastership flapping frequently
  - Declared master has partial reachability to switches
- Reported topology changes, pathing changes, flow programming fails

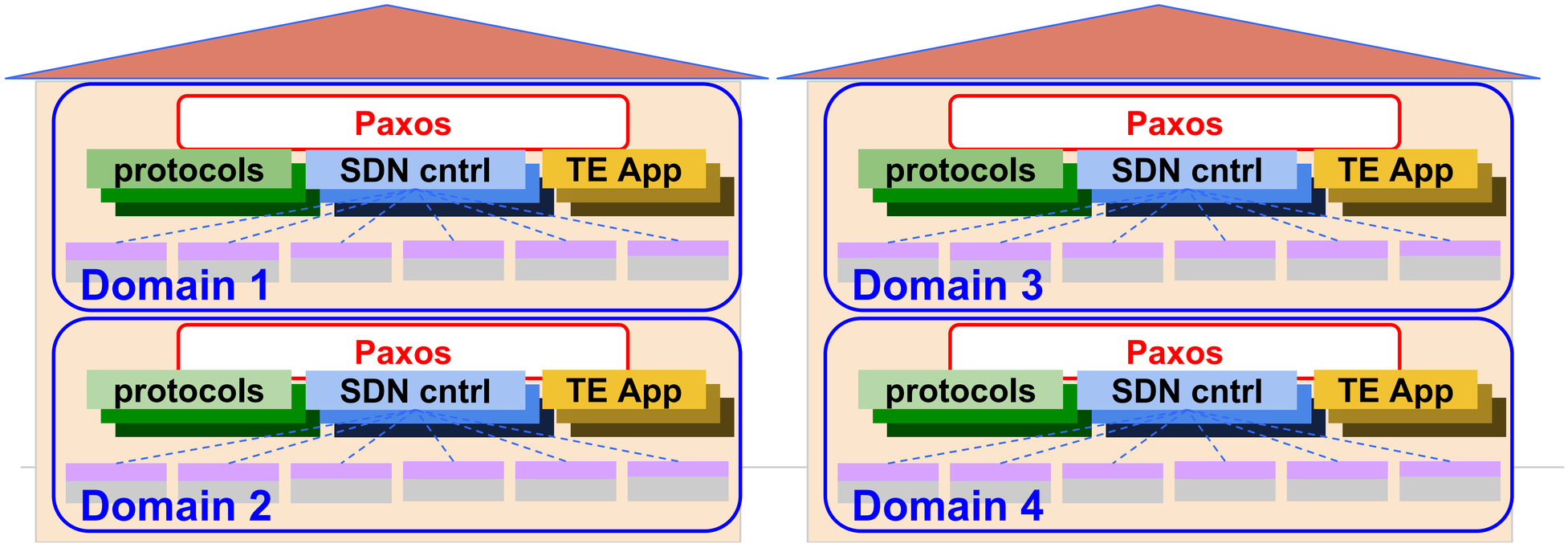
**Non-transitive reachability => Packets dropped!!**

"Lessons Learned from B4, Google's SDN WAN"
Subhasree Mandal, talk at USENIX ATC 2015

# Lesson: Robust Control Reachability

Google

- Multiple independent domains per site: connected only through dataplane
  - Each domain is unit for safe modular upgrade and maintenance
- Paxos: quorum-based robust master election *within* each domain
- Also removes single point of failure in each site




"Lessons Learned from B4, Google's SDN WAN"
Subhasree Mandal, talk at USENIX ATC 2015

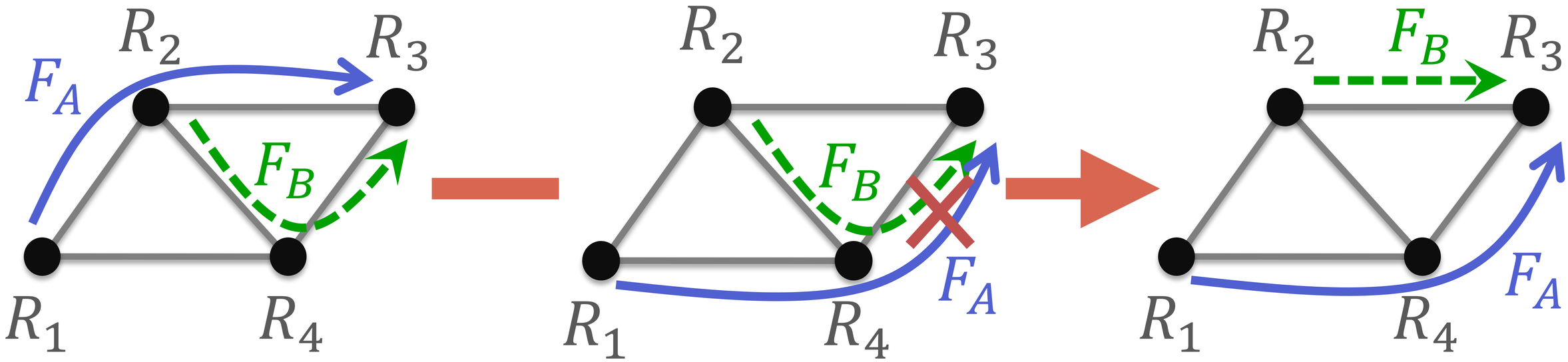
# Microsoft's SWAN

**ACM SIGCOMM, 2013**

**Achieving High Utilization with Software-Driven WAN**

Chi-Yao Hong (UIUC) Srikanth Kandula Ratul Mahajan Ming Zhang
Vijay Gill Mohan Nanduri Roger Wattenhofer (ETH)

Microsoft

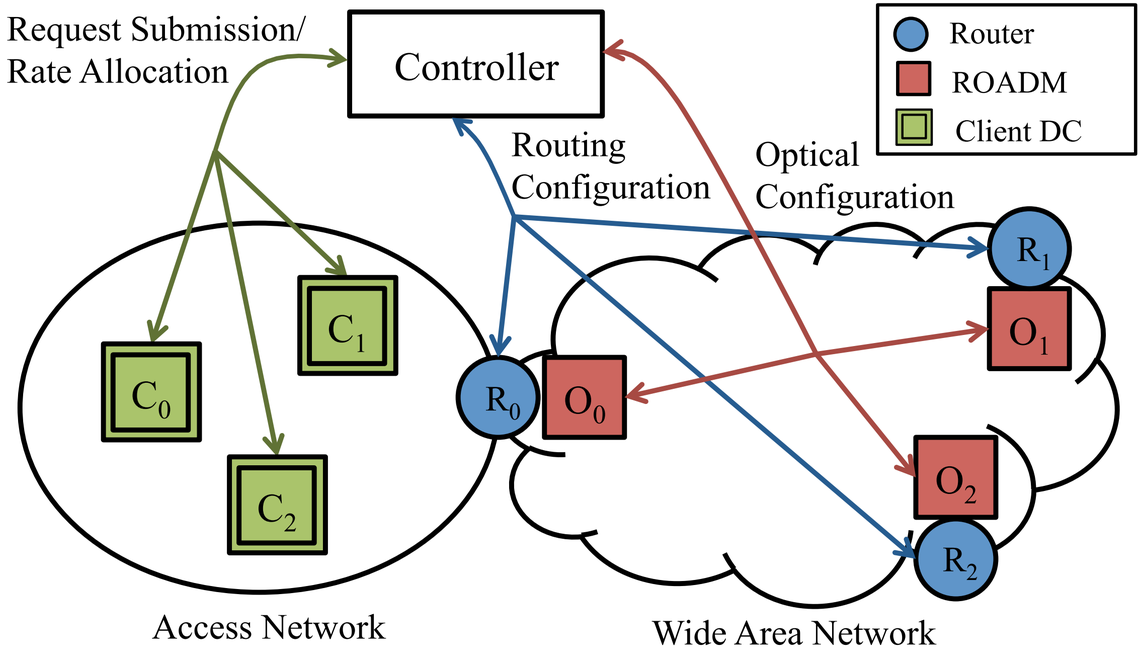


# Joint topology and routing optimization

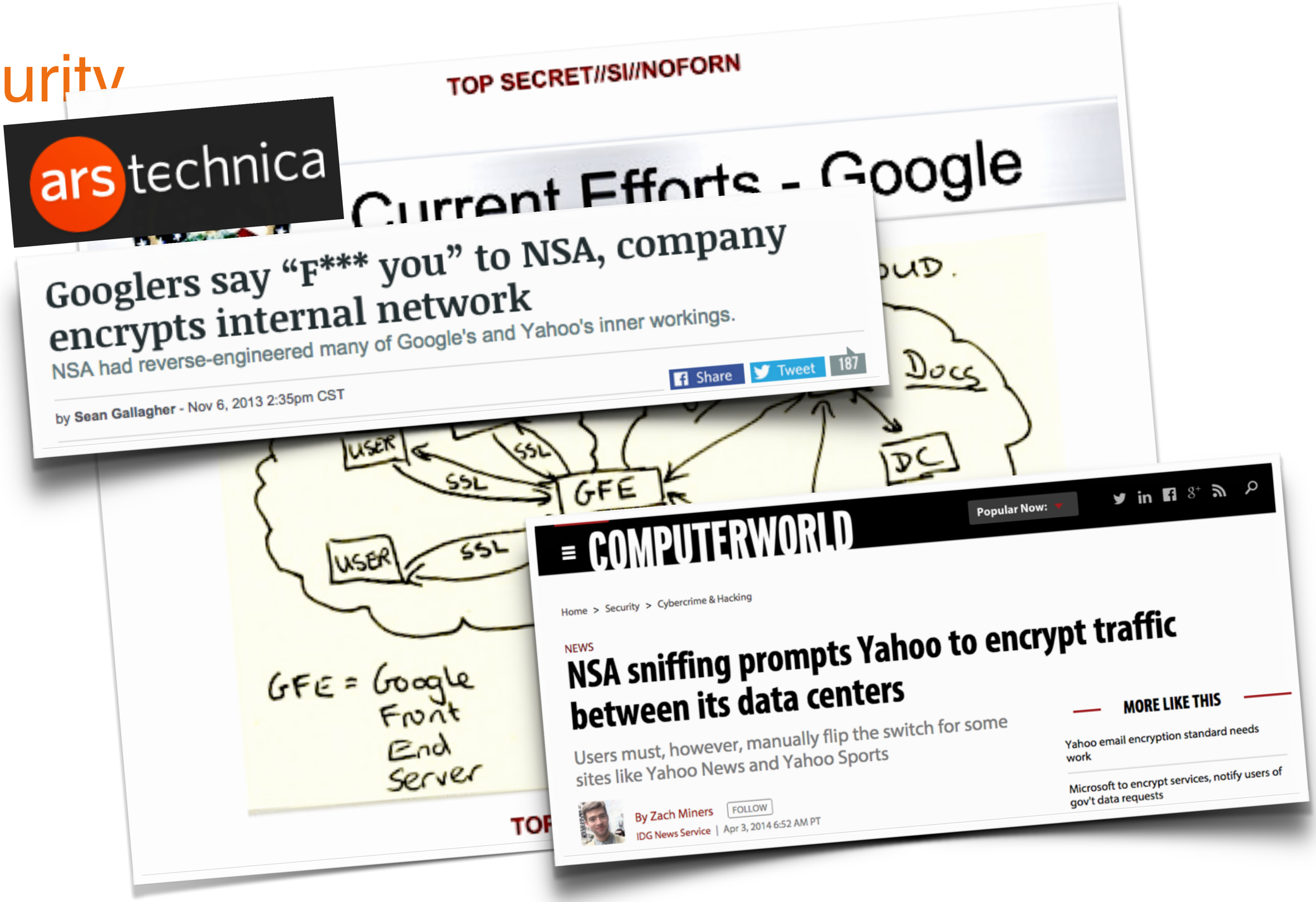
ACM SIGCOMM, 2016

**Optimizing Bulk Transfers with Software-Defined Optical WAN**

Xin Jin†, Yiran Li⋆, Da Wei⋆, Siming Li^, Jie Gao^,
Lei Xu°, Guangzhi Li×, Wei Xu⋆, Jennifer Rexford†



# Security

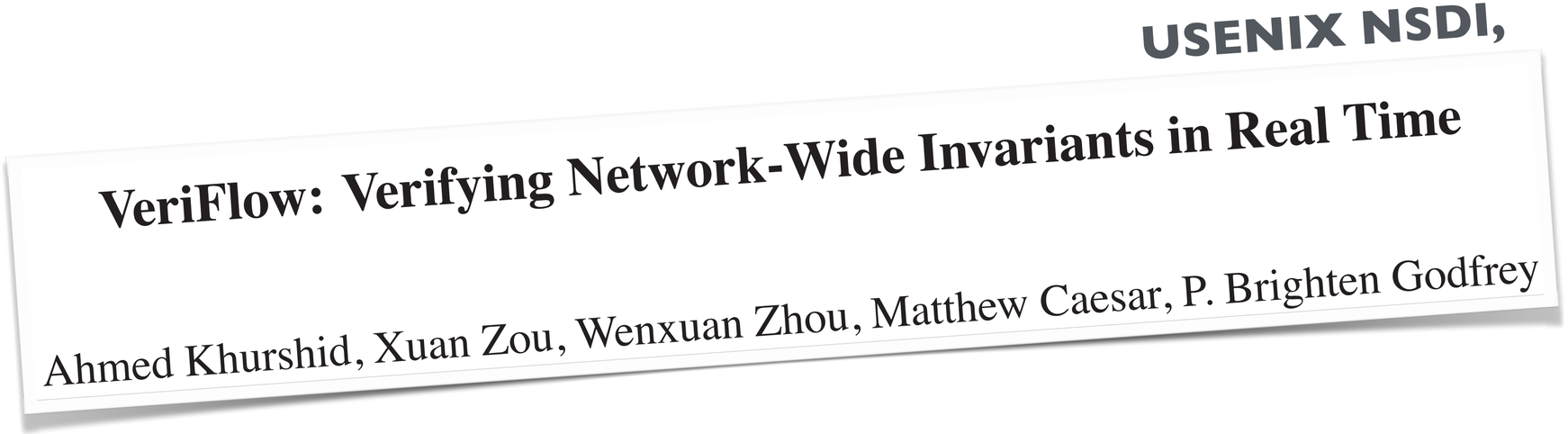

TOP SECRET//SI//NOFORN
Current Efforts - Google
ars technica
Googlers say "F*** you" to NSA, company encrypts internal network
NSA had reverse-engineered many of Google's and Yahoo's inner workings.
by Sean Gallagher - Nov 6, 2013 2:35pm CST
Share
Tweet
187
USER
SSL
SSL
GFE
USER
SSL
Docs
DC
GFE = Google Front End Server
COMPUTERWORLD
Popular Now:
Home > Security > Cybercrime & Hacking
NEWS
NSA sniffing prompts Yahoo to encrypt traffic between its data centers
Users must, however, manually flip the switch for some sites like Yahoo News and Yahoo Sports
By Zach Miners
FOLLOW
IDG News Service | Apr 3, 2014 6:52 AM PT
MORE LIKE THIS
Yahoo email encryption standard needs work
Microsoft to encrypt services, notify users of gov't data requests

# Other possibilities and settings?

- Why let congestion control ruin TE? Per-packet scheduling?
- What can you do when you only control one end?
- Between 1000s of edge locations instead of 12 DC sites?
- What if most traffic is “inelastic”?

# Weekly reading guide

# Network verification




- How do we prevent misconfiguration?
- Guarantee adherence to policy?

# The need for verification

- Complex interactions between elements
- Potential for misconfiguration
- Unforeseen bugs
- Hard to test entire network before deployment

# Errors in network operation can …

- … allow unauthorized traffic into secure zones
- … expose the network to attacks
- … make critical services unavailable
- … affect network performance

# Just verify configurations?


Can still cause unpredictable behavior!

# Does packet-forwarding match policy?

- State policy or "network invariants"
- Verify that *forwarding* behavior matches policy
- Check new changes in **real-time**
- How would you do this?