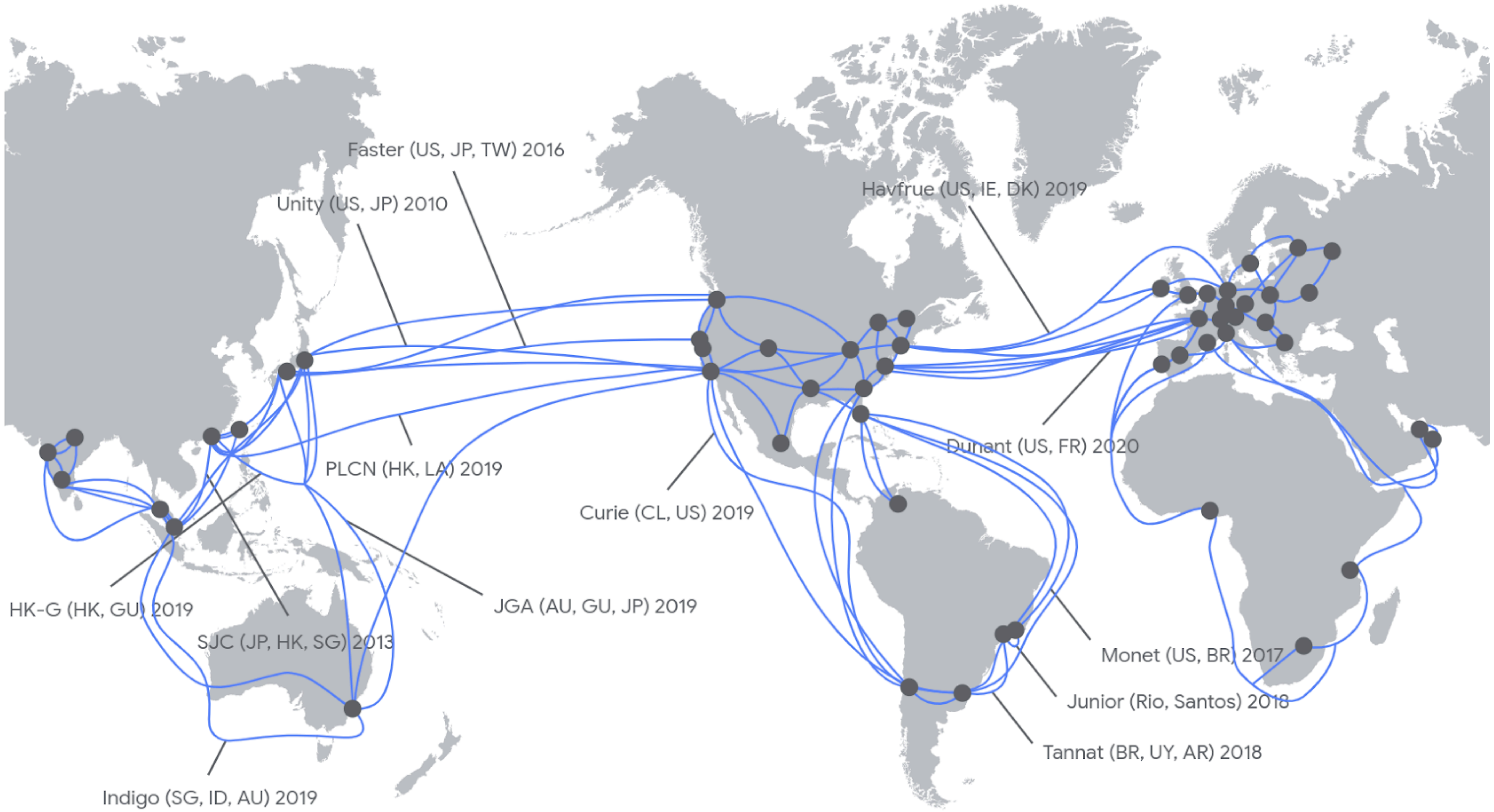
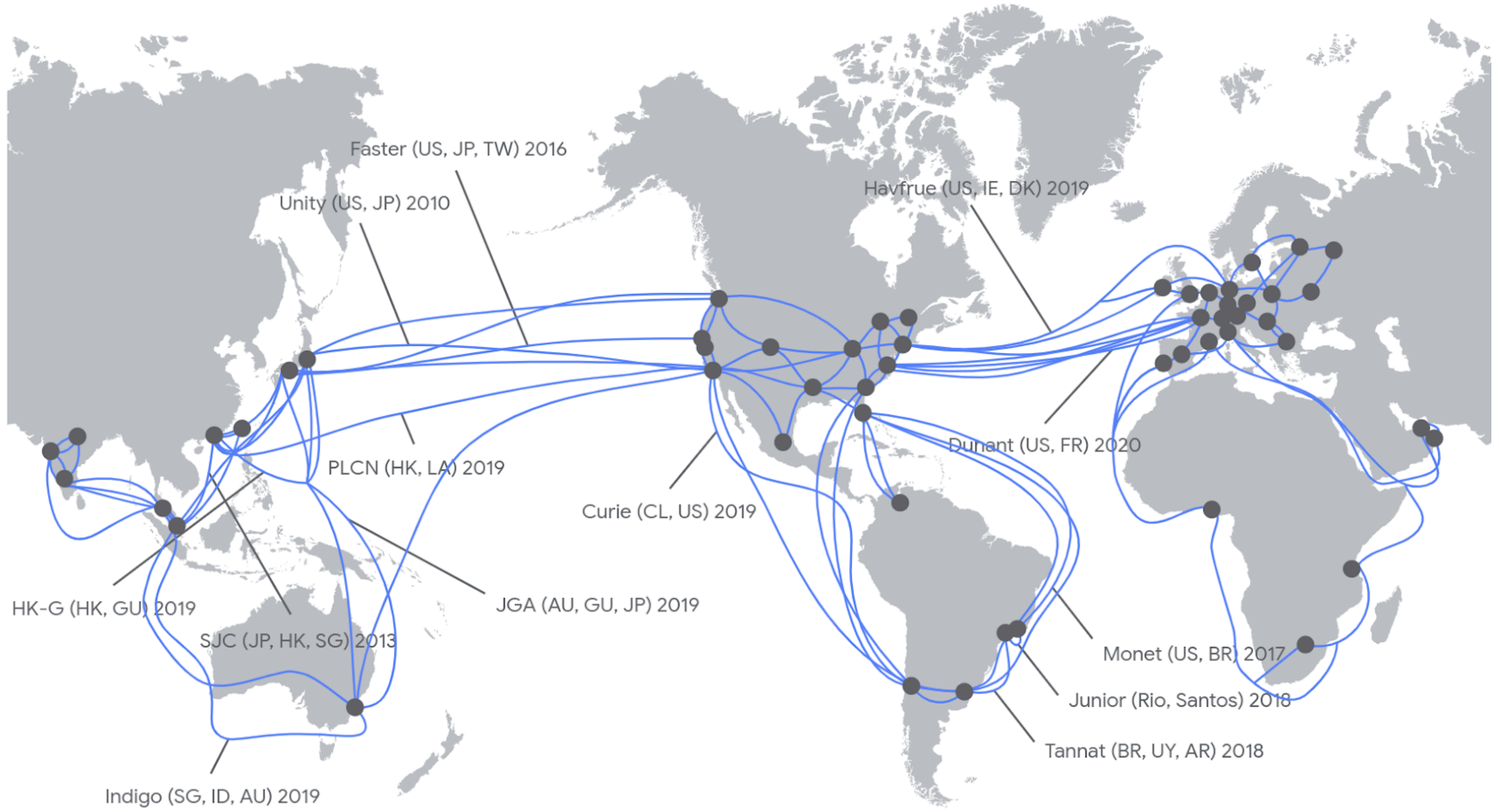
Annulus: A Dual Congestion Control Loop for Datacenter and WAN Traffic Aggregates

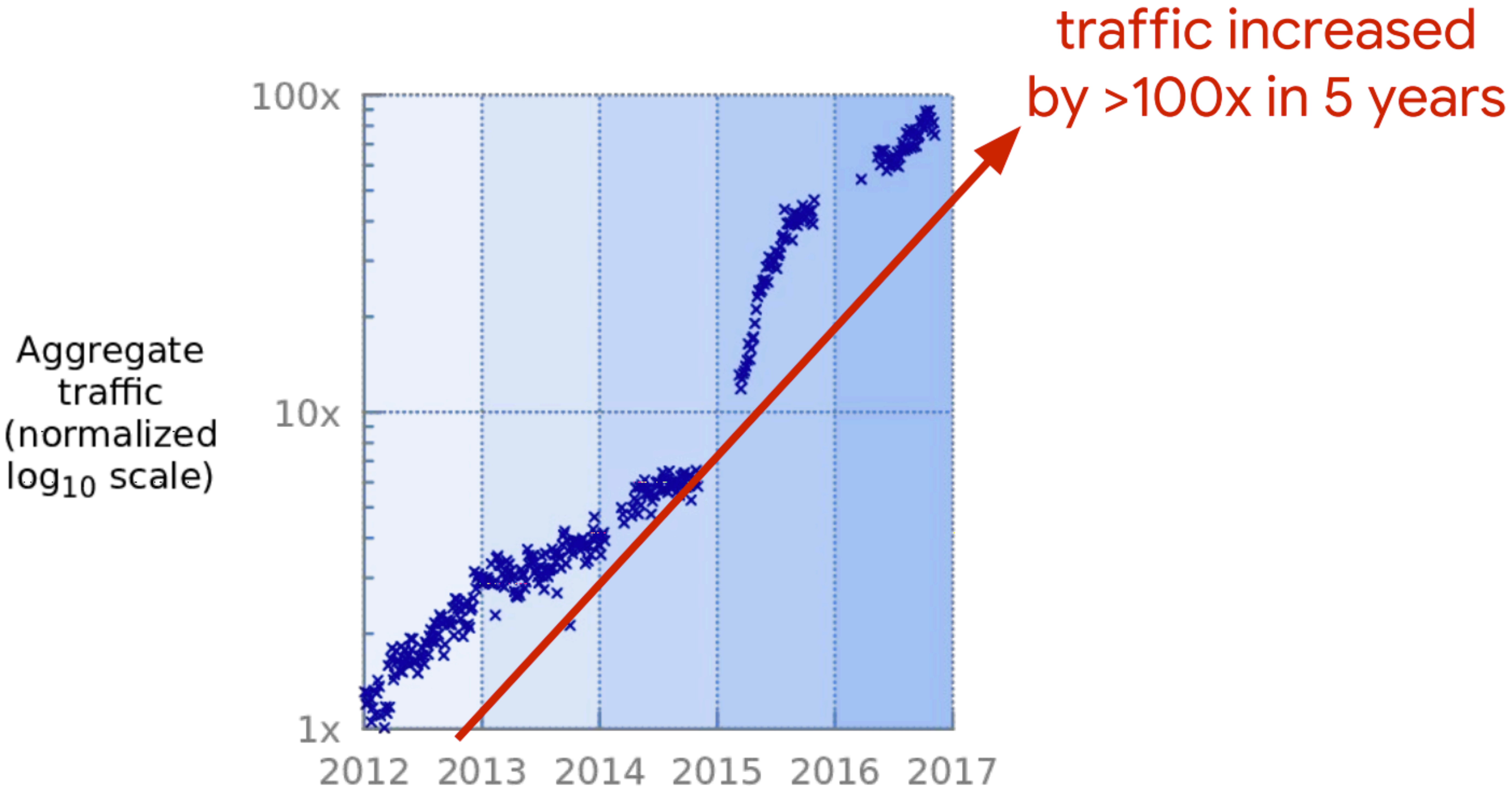
Ahmed Saeed, Varun Gupta, Prateesh Goyal, Milad Sharif, Rong Pan, Mostafa Ammar, Ellen Zegura, Keon Jang, Mohammad Alizadeh, Abdul Kabbani, Amin Vahdat



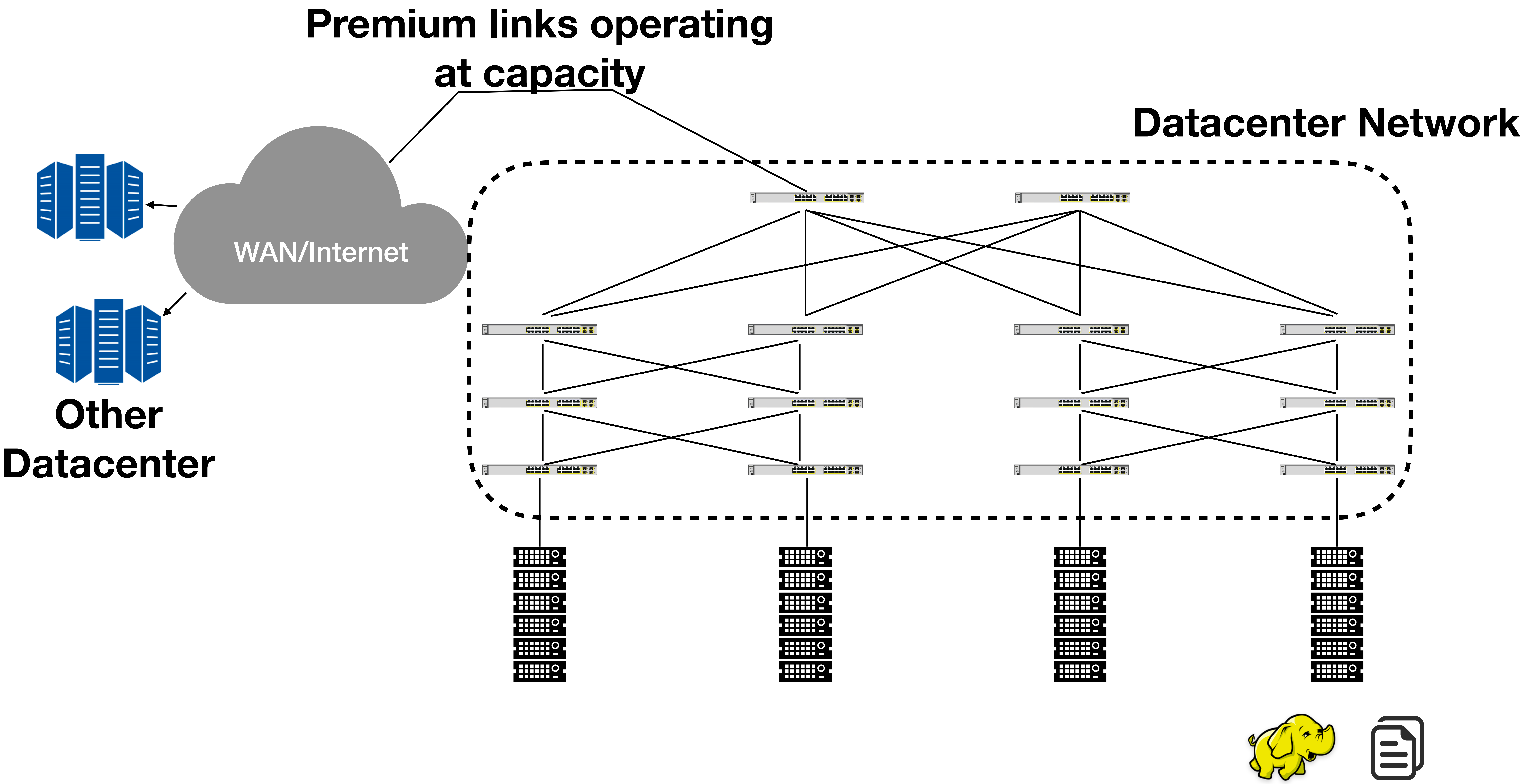
Google's WAN connects different regions through high-capacity links



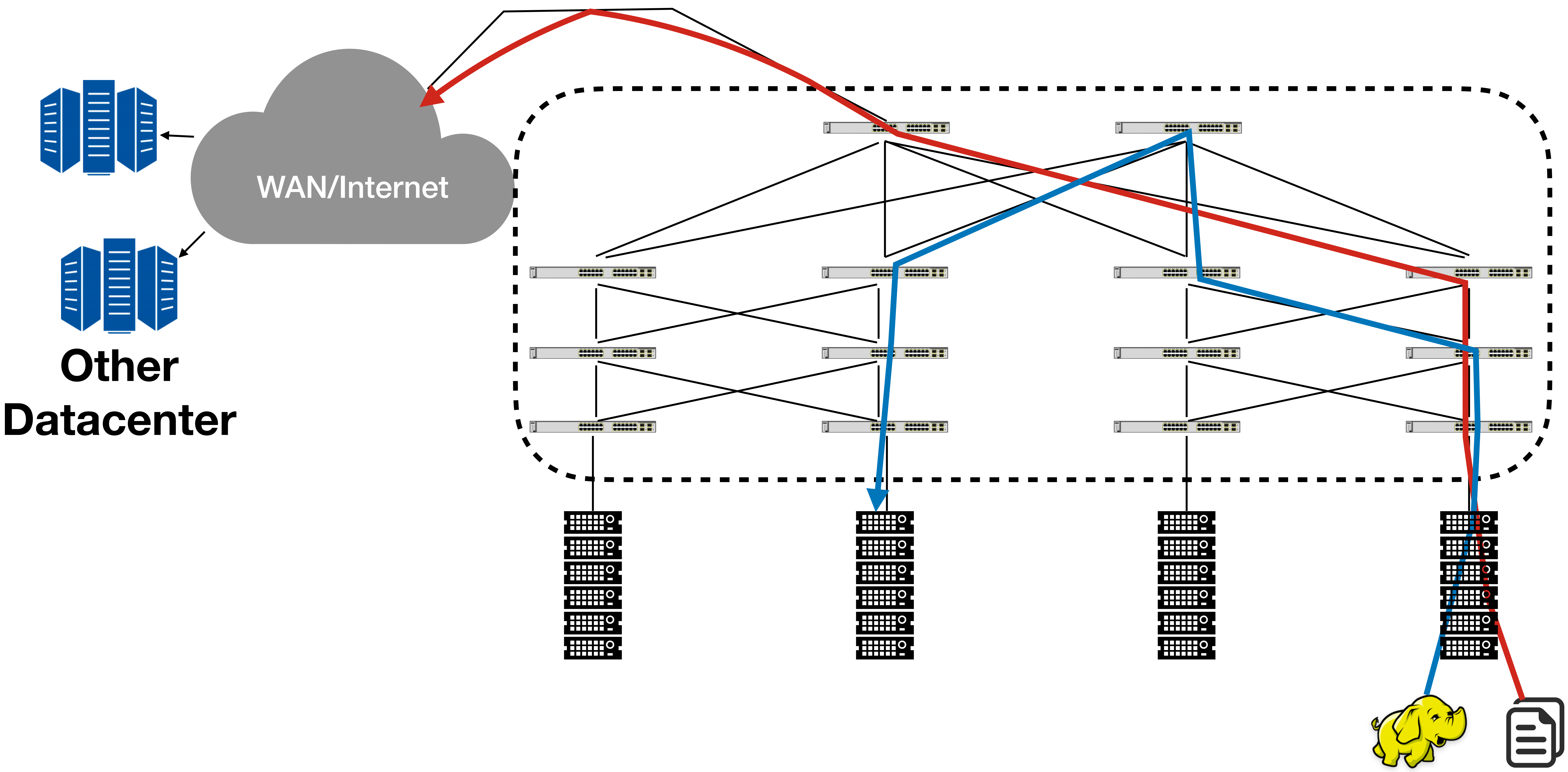
[Chi-Yao Hong et al., SIGCOMM'18]



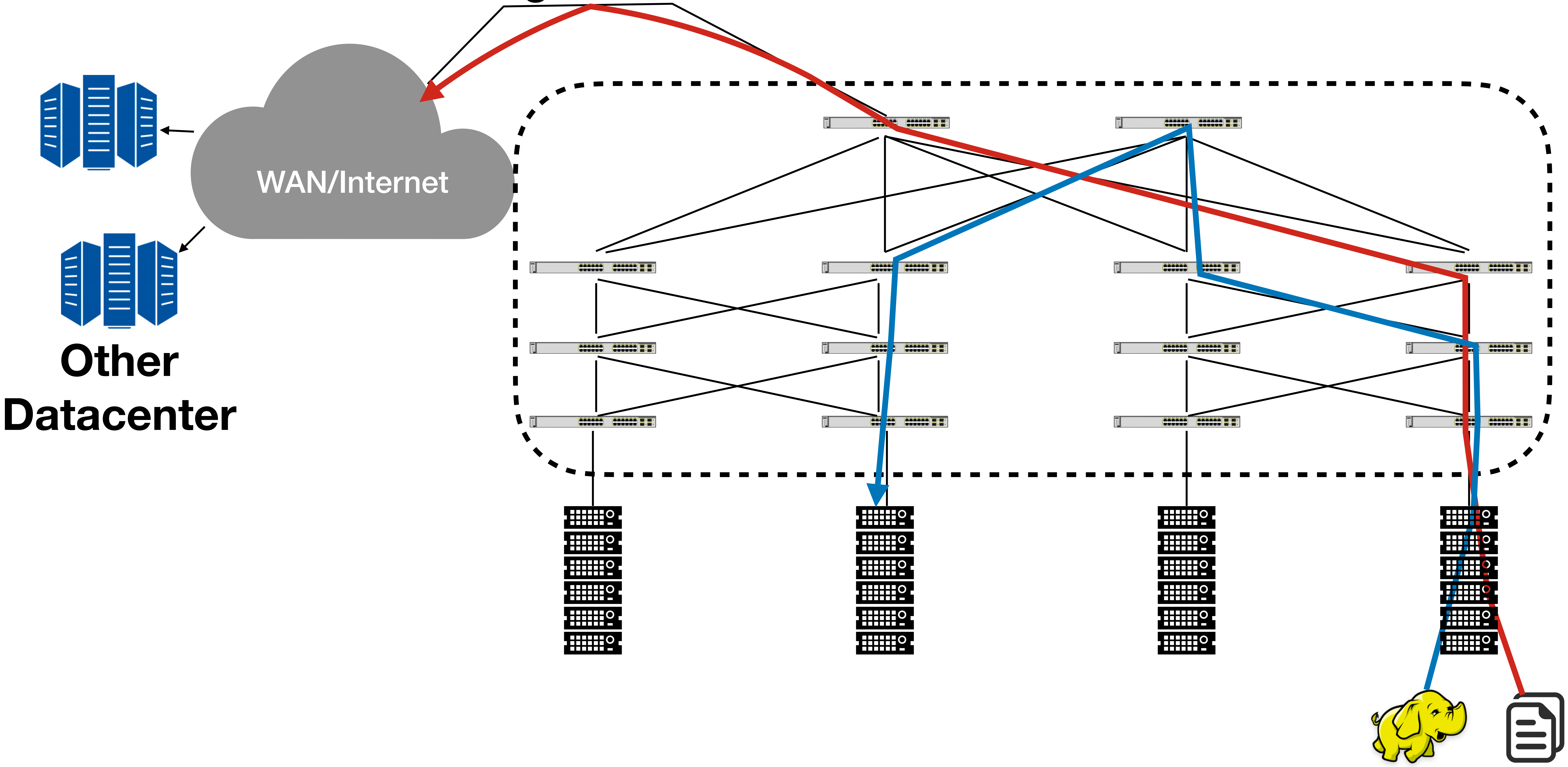
Google's WAN connects different regions through high-capacity links

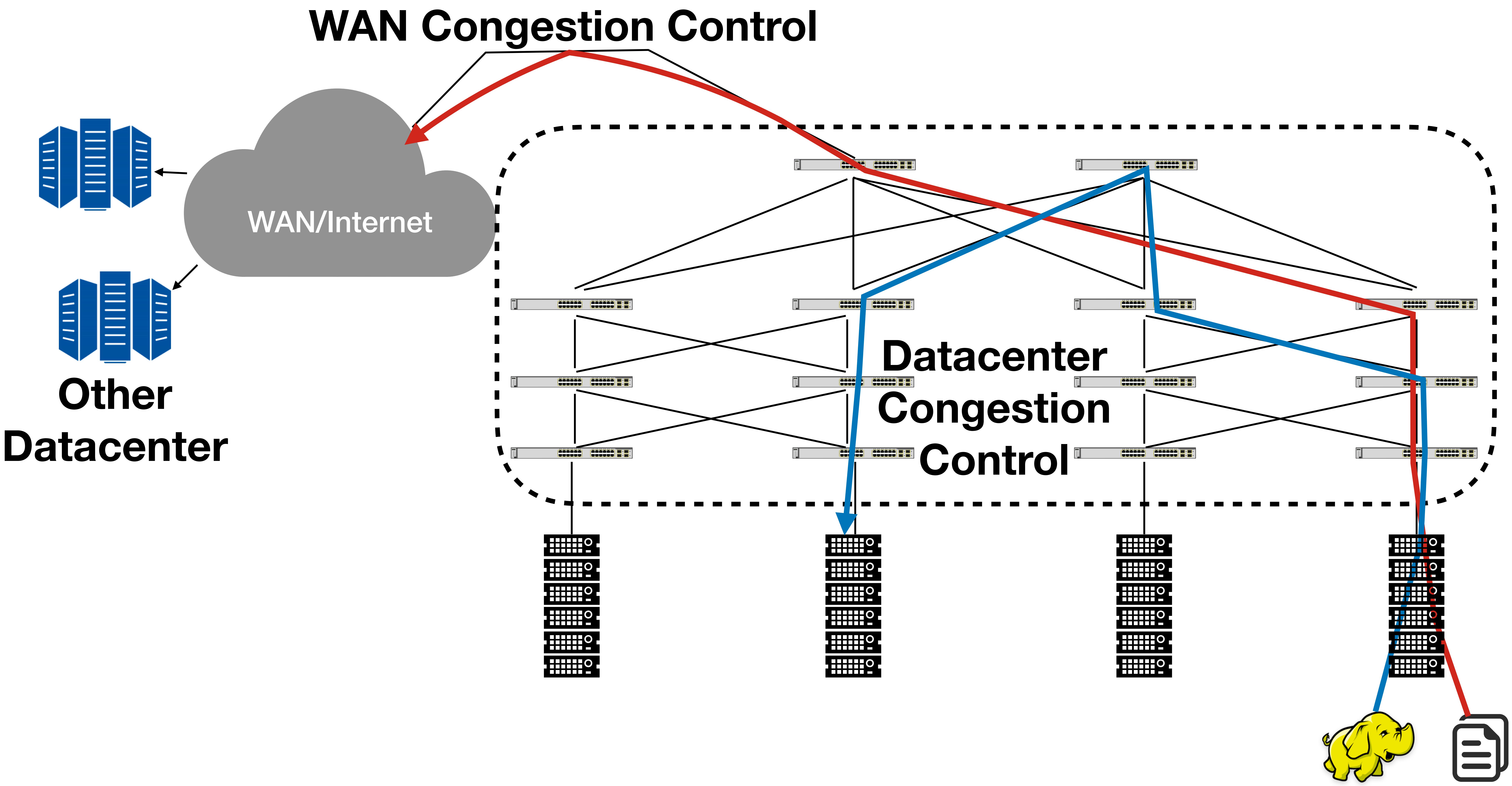


Premium links operating at capacity
Datacenter Network
WAN/Internet
Other Datacenter


WAN/Internet
Other Datacenter

WAN Congestion Control




WAN Congestion Control
WAN/Internet
Other
Datacenter
Datacenter
Congestion
Control

A Tale of Two Networks

A Tale of Two Networks

Datacenter

A Tale of Two Networks

Datacenter

- Short RTTs and shallow buffers

A Tale of Two Networks

Datacenter

- Short RTTs and shallow buffers
- Introduces specific challenges like incast

A Tale of Two Networks

Datacenter

- Short RTTs and shallow buffers
- Introduces specific challenges like incast
- Network falls under a single federation, facilitating deployment and debugging

A Tale of Two Networks

Datacenter

- Short RTTs and shallow buffers
- Introduces specific challenges like incast
- Network falls under a single federation, facilitating deployment and debugging

Data Center TCP (DCTCP)

Mohammad Alizadeh[‡†], Albert Greenberg[†], David A. Maltz[†], Jitendra Padhye[†],
Parveen Patel[†], Balaji Prabhakar[‡], Sudipta Sengupta[†], Murari Sridharan[†]

[†]Microsoft Research [‡]Stanford University
{albert, dmaltz, padhye, parveenp, sudipta, muraris}@microsoft.com
{alizade, balaji}@stanford.edu

A Tale of Two Networks

Datacenter

- Short RTTs and shallow buffers
- Introduces specific challenges like incast
- Network falls under a single federation, facilitating deployment and debugging

Data Center TCP (DCTCP)

Congestion Control for Large-Scale RDMA Deployments

Yibo Zhu[1,3] Haggai Eran[2] Daniel Firestone[1] Chuanxiong Guo[1] Marina Lipshteyn[1]
Yehonatan Liron[2] Jitendra Padhye[1] Shachar Raindel[2] Mohamad Haj Yahia[2] Ming Zhang[1]

[1]Microsoft [2]Mellanox [3]U. C. Santa Barbara

A Tale of Two Networks

Datacenter

- Short RTTs and shallow buffers
- Introduces specific challenges like incast
- Network falls under a single federation, facilitating deployment and debugging

Data Center TCP (DCTCP)

Congestion Control for Large-Scale RDMA Deployments

TIMELY: RTT-based Congestion Control for the Datacenter

ng[1]

Radhika Mittal*(UC Berkeley), Vinh The Lam, Nandita Dukkipati, Emily Blem, Hassan Wassel, Monia Ghobadi*(Microsoft), Amin Vahdat, Yaogong Wang, David Wetherall, David Zats

Google, Inc.

A Tale of Two Networks

Datacenter

- Short RTTs and shallow buffers
- Introduces specific challenges like incast
- Network falls under a single federation, facilitating deployment and debugging

Data Center TCP (DCTCP)

Congestion Control for Large-Scale RDMA Deployments

TIMELY: RTT-based Congestion Control for the Datacenter

ng[1]

HPCC: High Precision Congestion Control

Yuliang Li♠♡, Rui Miao♠, Hongqiang Harry Liu♠, Yan Zhuang♠, Fei Feng♠, Lingbo Tang♠, Zheng Cao♠, Ming Zhang♠, Frank Kelly◇, Mohammad Alizadeh♣, Minlan Yu♡
Alibaba Group♠, Harvard University♡, University of Cambridge◇, Massachusetts Institute of Technology♣

A Tale of Two Networks

Datacenter

Internet/WAN

- Short RTTs and shallow buffers
- Introduces specific challenges like incast
- Network falls under a single federation, facilitating deployment and debugging

Data Center TCP (DCTCP)

Congestion Control for Large-Scale RDMA Deployments

TIMELY: RTT-based Congestion Control for the Datacenter

ng[1]

HPCC: High Precision Congestion Control

Yuliang Li[♠♡], Rui Miao[♠], Hongqiang Harry Liu[♠], Yan Zhuang[♠], Fei Feng[♠], Lingbo Tang[♠], Zheng Cao[♠], Ming Zhang[♠],
Frank Kelly[◇], Mohammad Alizadeh[♣], Minlan Yu[♡]
Alibaba Group[♠], Harvard University[♡], University of Cambridge[◇], Massachusetts Institute of Technology[♣]

A Tale of Two Networks

Datacenter

- Short RTTs and shallow buffers
- Introduces specific challenges like incast
- Network falls under a single federation, facilitating deployment and debugging

Data Center TCP (DCTCP)

Congestion Control for Large-Scale RDMA Deployments

TIMELY: RTT-based Congestion Control for the Datacenter

ng[1]

HPCC: High Precision Congestion Control

Yuliang Li[♠♡], Rui Miao[♠], Hongqiang Harry Liu[♠], Yan Zhuang[♠], Fei Feng[♠], Lingbo Tang[♠], Zheng Cao[♠], Ming Zhang[♠], Frank Kelly[◊], Mohammad Alizadeh[♣], Minlan Yu[♡]
Alibaba Group[♠], Harvard University[♡], University of Cambridge[◊], Massachusetts Institute of Technology[♣]

Internet/WAN

- Long RTTs and high bandwidth

A Tale of Two Networks

Datacenter

- Short RTTs and shallow buffers
- Introduces specific challenges like incast
- Network falls under a single federation, facilitating deployment and debugging

Data Center TCP (DCTCP)

Congestion Control for Large-Scale RDMA Deployments

TIMELY: RTT-based Congestion Control for the Datacenter

ng[1]

HPCC: High Precision Congestion Control

Yuliang Li[♠♡], Rui Miao[♠], Hongqiang Harry Liu[♠], Yan Zhuang[♠], Fei Feng[♠], Lingbo Tang[♠], Zheng Cao[♠], Ming Zhang[♠],
Frank Kelly[◇], Mohammad Alizadeh[♣], Minlan Yu[♡]
Alibaba Group[♠], Harvard University[♡], University of Cambridge[◇], Massachusetts Institute of Technology[♣]

Internet/WAN

- Long RTTs and high bandwidth
- Mix of deep and shallow buffers

A Tale of Two Networks

Datacenter

- Short RTTs and shallow buffers
- Introduces specific challenges like incast
- Network falls under a single federation, facilitating deployment and debugging

Data Center TCP (DCTCP)

Congestion Control for Large-Scale RDMA Deployments

TIMELY: RTT-based Congestion Control for the Datacenter

ng[1]

HPCC: High Precision Congestion Control

Yuliang Li[♠♡], Rui Miao[♠], Hongqiang Harry Liu[♠], Yan Zhuang[♠], Fei Feng[♠], Lingbo Tang[♠], Zheng Cao[♠], Ming Zhang[♠], Frank Kelly[◇], Mohammad Alizadeh[♣], Minlan Yu[♡]
Alibaba Group[♠], Harvard University[♡], University of Cambridge[◇], Massachusetts Institute of Technology[♣]

Internet/WAN

- Long RTTs and high bandwidth
- Mix of deep and shallow buffers
- Network heterogeneity remains a challenge

A Tale of Two Networks

Datacenter

- Short RTTs and shallow buffers
- Introduces specific challenges like incast
- Network falls under a single federation, facilitating deployment and debugging

Data Center TCP (DCTCP)

Congestion Control for Large-Scale RDMA Deployments

TIMELY: RTT-based Congestion Control for the Datacenter

ng[1]

HPCC: High Precision Congestion Control

Yuliang Li♠♡, Rui Miao♠, Hongqiang Harry Liu♠, Yan Zhuang♠, Fei Feng♠, Lingbo Tang♠, Zheng Cao♠, Ming Zhang♠, Frank Kelly◇, Mohammad Alizadeh♣, Minlan Yu♡
Alibaba Group♠, Harvard University♡, University of Cambridge◇, Massachusetts Institute of Technology♣

Internet/WAN

- Long RTTs and high bandwidth
- Mix of deep and shallow buffers
- Network heterogeneity remains a challenge

Measuring bottleneck bandwidth and round-trip propagation time.

BY NEAL CARDWELL, YUCHUNG CHENG, C. STEPHEN GUNN, SOHEIL HASSAS YEGANEH, AND VAN JACOBSON

BBR: Congestion-Based Congestion Control

A Tale of Two Networks

Datacenter

- Short RTTs and shallow buffers
- Introduces specific challenges like incast
- Network falls under a single federation, facilitating deployment and debugging

Data Center TCP (DCTCP)

Congestion Control for Large-Scale RDMA Deployments

TIMELY: RTT-based Congestion Control for the Datacenter

ng[1]

HPCC: High Precision Congestion Control

Yuliang Li, Rui Miao, Hongqiang Harry Liu, Yan Zhuang, Fei Feng, Lingbo Tang, Zheng Cao, Ming Zhang, Frank Kelly, Mohammad Alizadeh, Minlan Yu
Alibaba Group, Harvard University, University of Cambridge, Massachusetts Institute of Technology

Internet/WAN

- Long RTTs and high bandwidth
- Mix of deep and shallow buffers
- Network heterogeneity remains a challenge

Measuring bottleneck bandwidth and round-trip propagation time.

BY NEAL CARDWELL, YUCHUNG CHENG, C. STEPHEN GUNN, SOHEIL HASSAS YEGANEH, AND VAN JACOBSON

BBR: Congestion-Based Congestion Control

PCC: Re-architecting Congestion Control for Consistent High Performance

Mo Dong*, Qingxi Li*, Doron Zarchy**, P. Brighten Godfrey*, and Michael Schapira**

*University of Illinois at Urbana-Champaign
**Hebrew University of Jerusalem

A Tale of Two Networks

Datacenter

- Short RTTs and shallow buffers
- Introduces specific challenges like incast
- Network falls under a single federation, facilitating deployment and debugging

Data Center TCP (DCTCP)

Congestion Control for Large-Scale RDMA Deployments

TIMELY: RTT-based Congestion Control for the Datacenter

ng[1]

HPCC: High Precision Congestion Control

Yuliang Li[♠♡], Rui Miao[♠], Hongqiang Harry Liu[♠], Yan Zhuang[♠], Fei Feng[♠], Lingbo Tang[♠], Zheng Cao[♠], Ming Zhang[♠], Frank Kelly[◊], Mohammad Alizadeh[♣], Minlan Yu[♡]
Alibaba Group[♠], Harvard University[♡], University of Cambridge[◊], Massachusetts Institute of Technology[♣]

Internet/WAN

- Long RTTs and high bandwidth
- Mix of deep and shallow buffers
- Network heterogeneity remains a challenge

Measuring bottleneck bandwidth and round-trip propagation time.

BY NEAL CARDWELL, YUCHUNG CHENG, C. STEPHEN GUNN, SOHEIL HASSAS YEGANEH, AND VAN JACOBSON

BBR: Congestion-Based Congestion Control

PCC: Re-architecting Congestion Control for Consistent High Performance

Copa: Practical Delay-Based Congestion Control for the Internet

Venkat Arun and Hari Balakrishnan
M.I.T. Computer Science and Artificial Intelligence Laboratory
Email: {venkatar,hari}@mit.edu

What about bottlenecks shared between WAN and datacenter traffic?

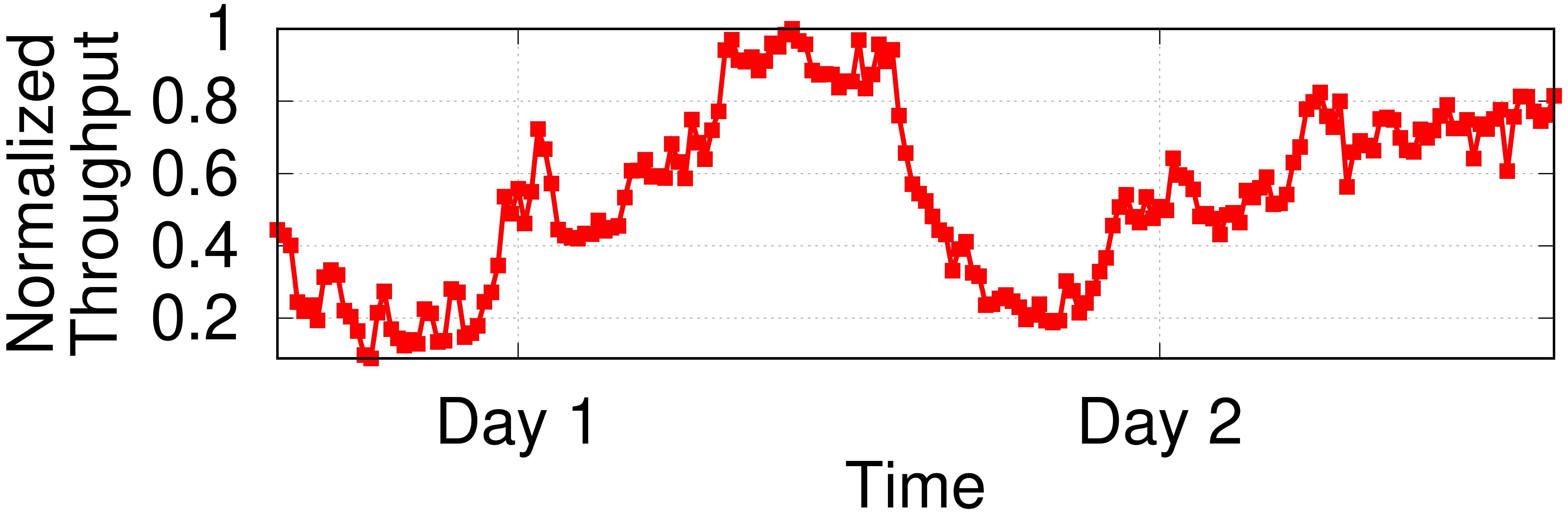
WAN vs LAN in the Wild

WAN vs LAN in the Wild

Data Collected from one of Google's clusters

WAN vs LAN in the Wild

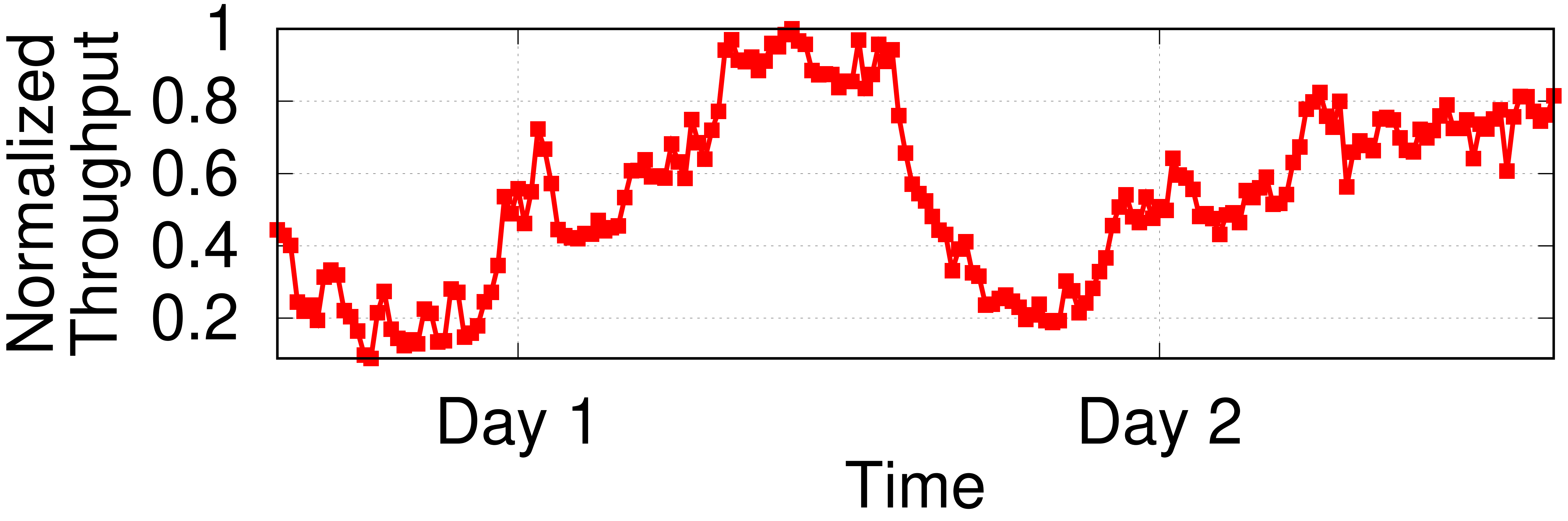
WAN Throughput



Data Collected from one of Google's clusters

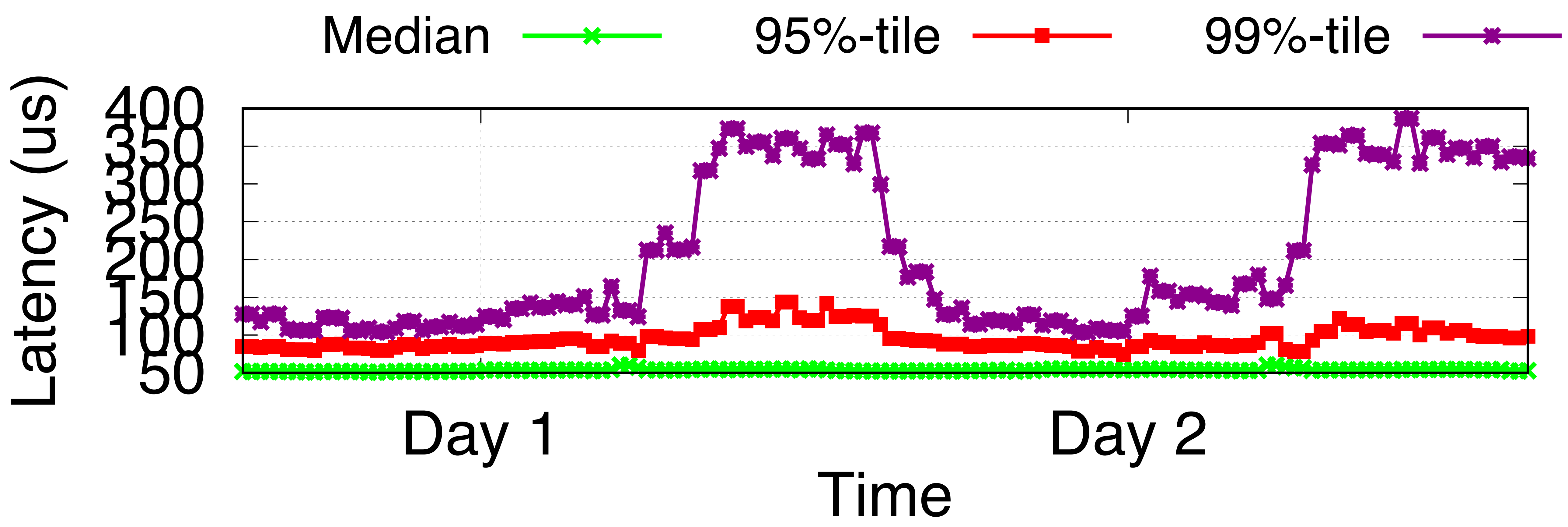
WAN vs LAN in the Wild

WAN Throughput



Data Collected from one of Google's clusters

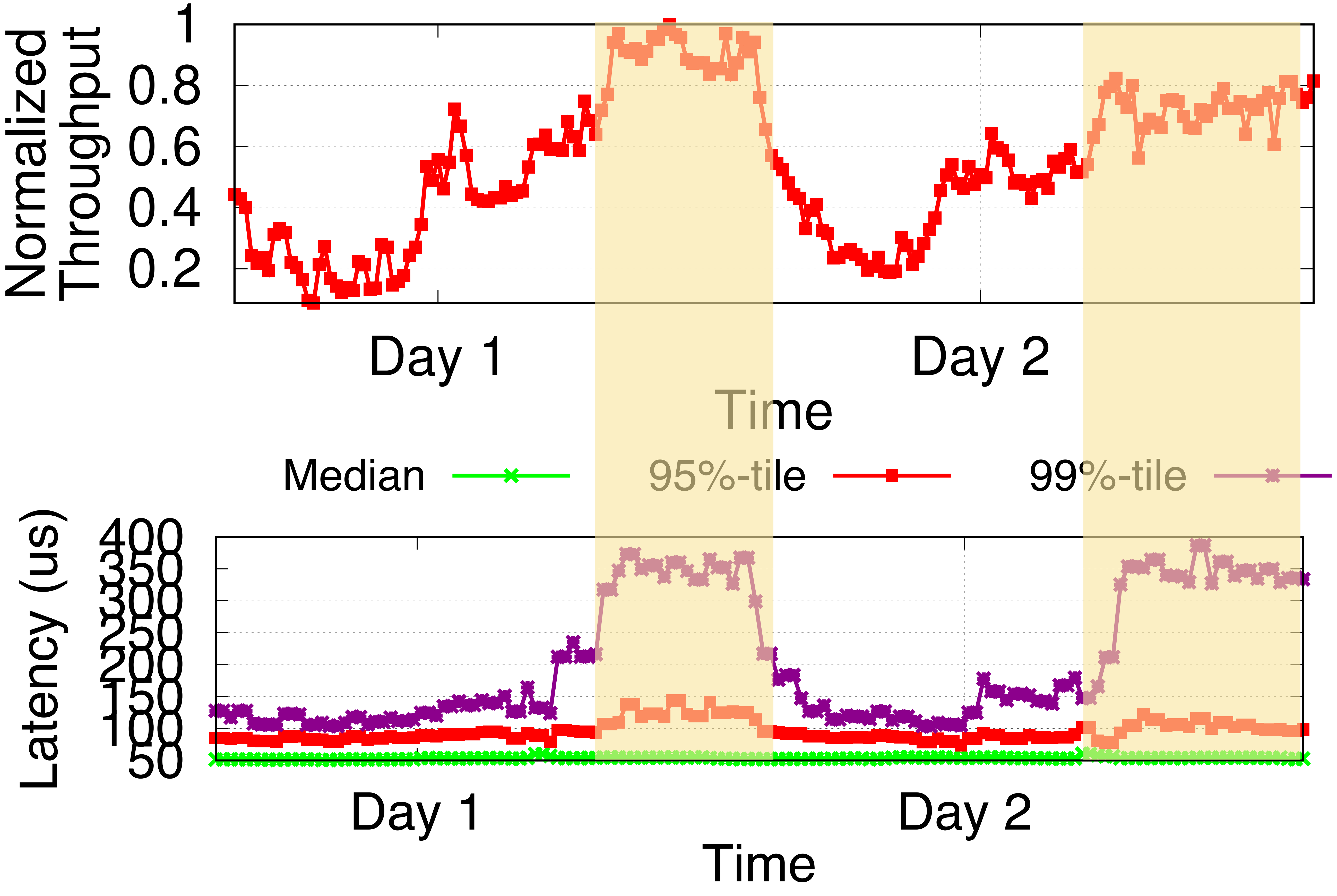
Datacenter Latency



WAN vs LAN in the Wild

WAN Throughput

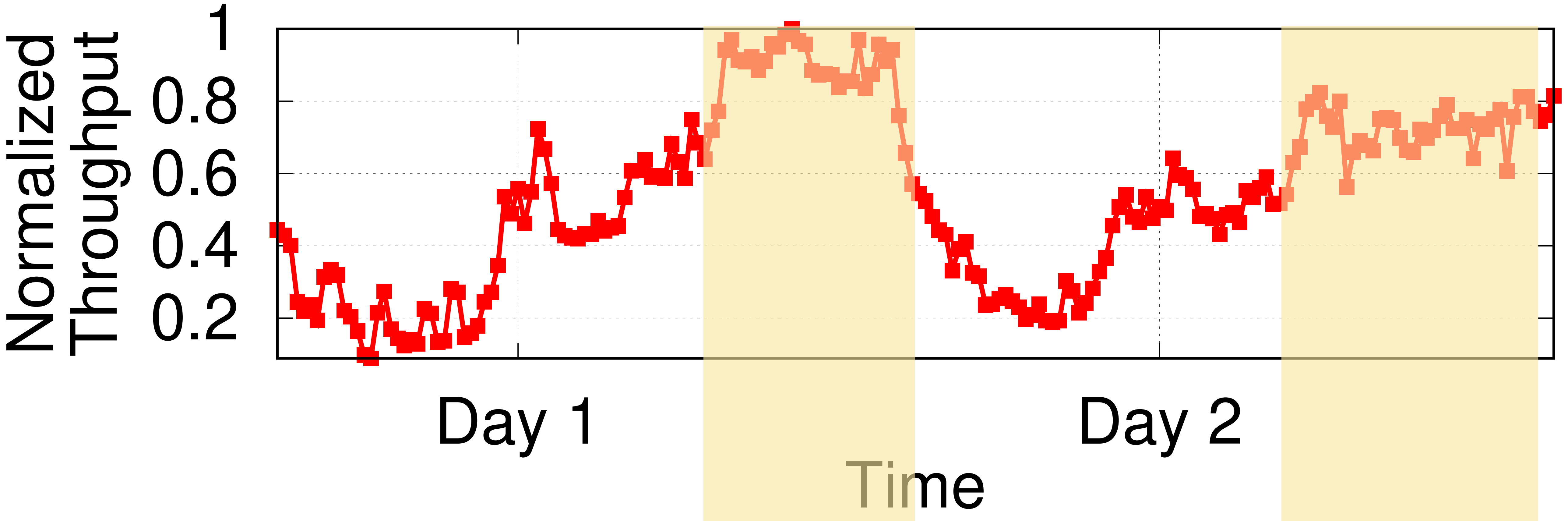
Datacenter Latency



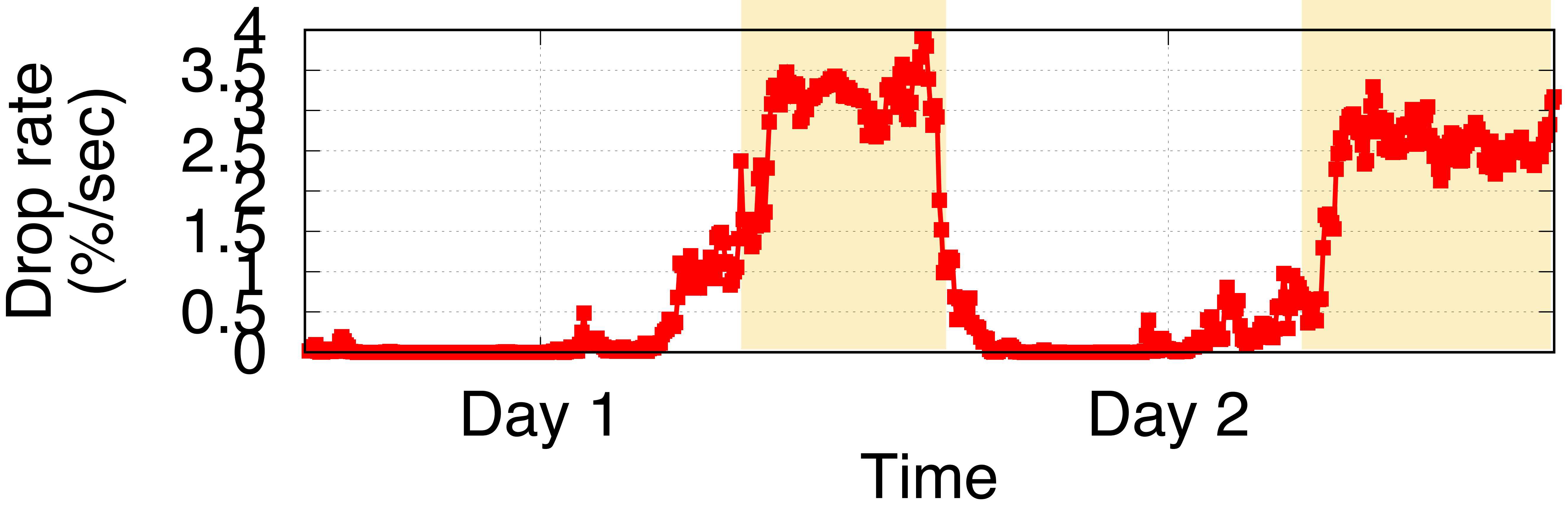
Data Collected from one of Google's clusters

WAN vs LAN in the Wild

WAN Throughput

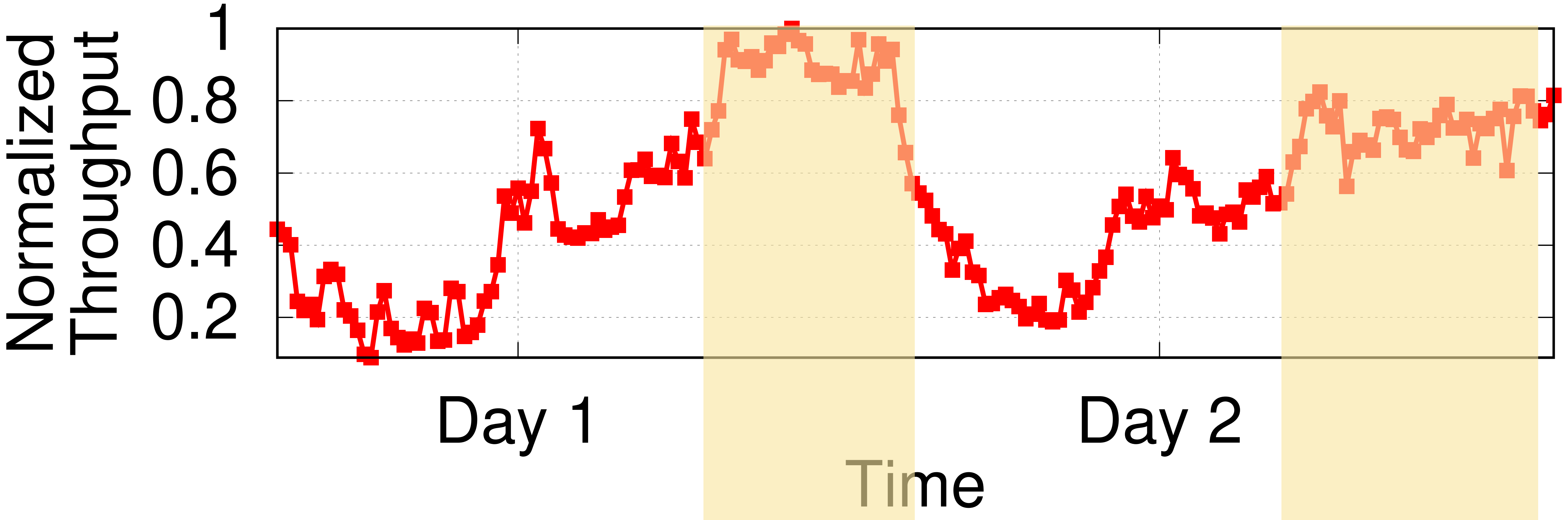


Drop Rate at ToR switches

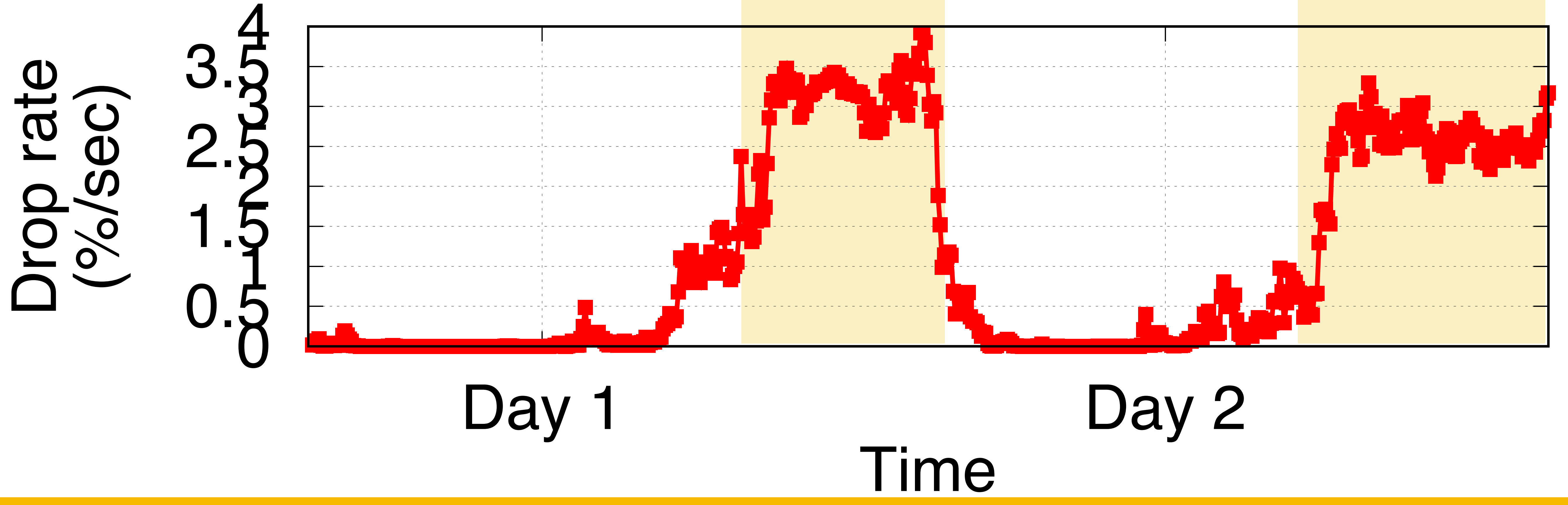


WAN vs LAN in the Wild

WAN Throughput



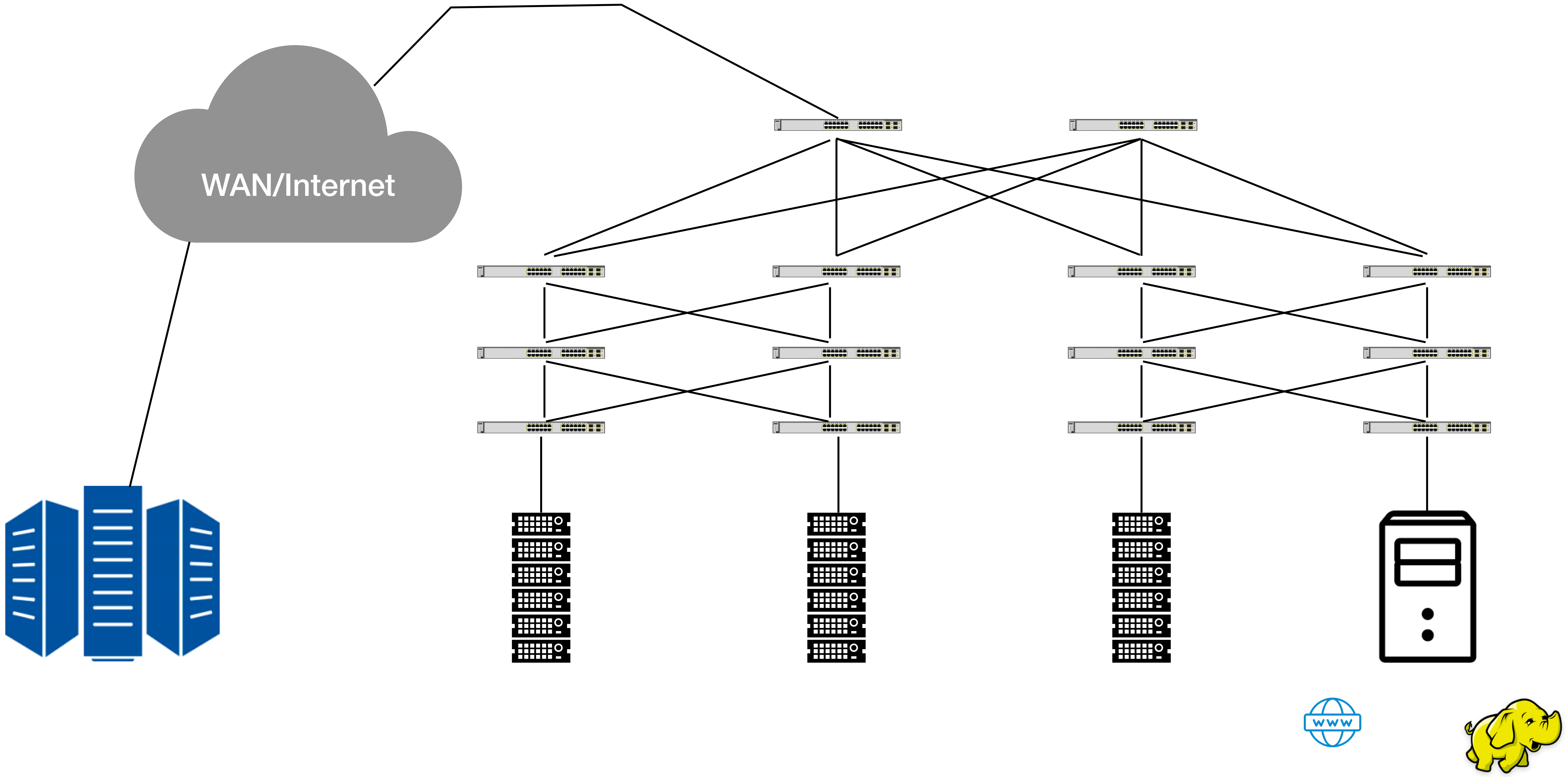
Drop Rate at ToR switches



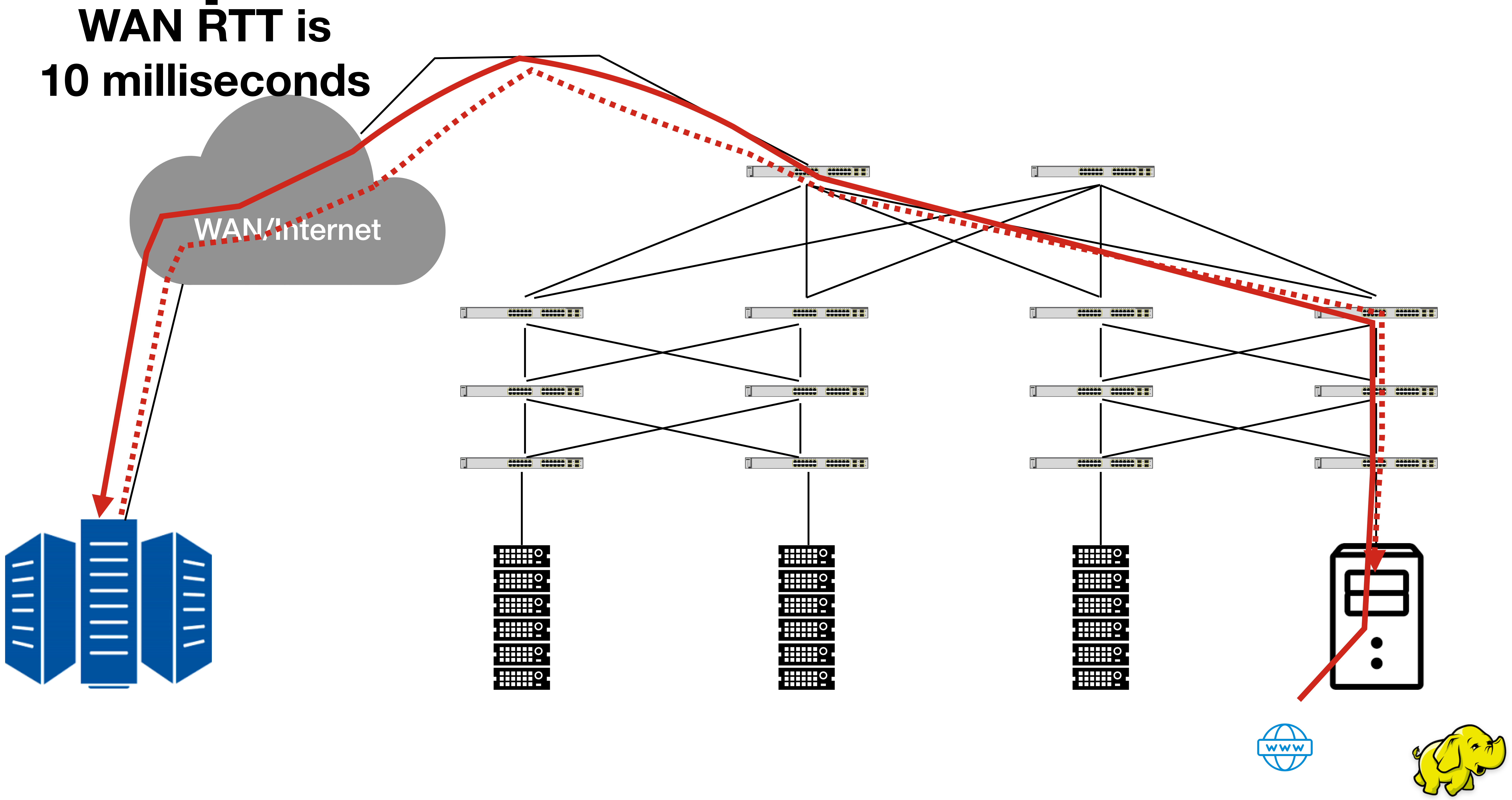
WAN demand significantly impacts the latency and drop rate of datacenter traffic

WAN traffic reaction is too slow to handle the fast dynamics of datacenter traffic

Impact of WAN on Datacenter



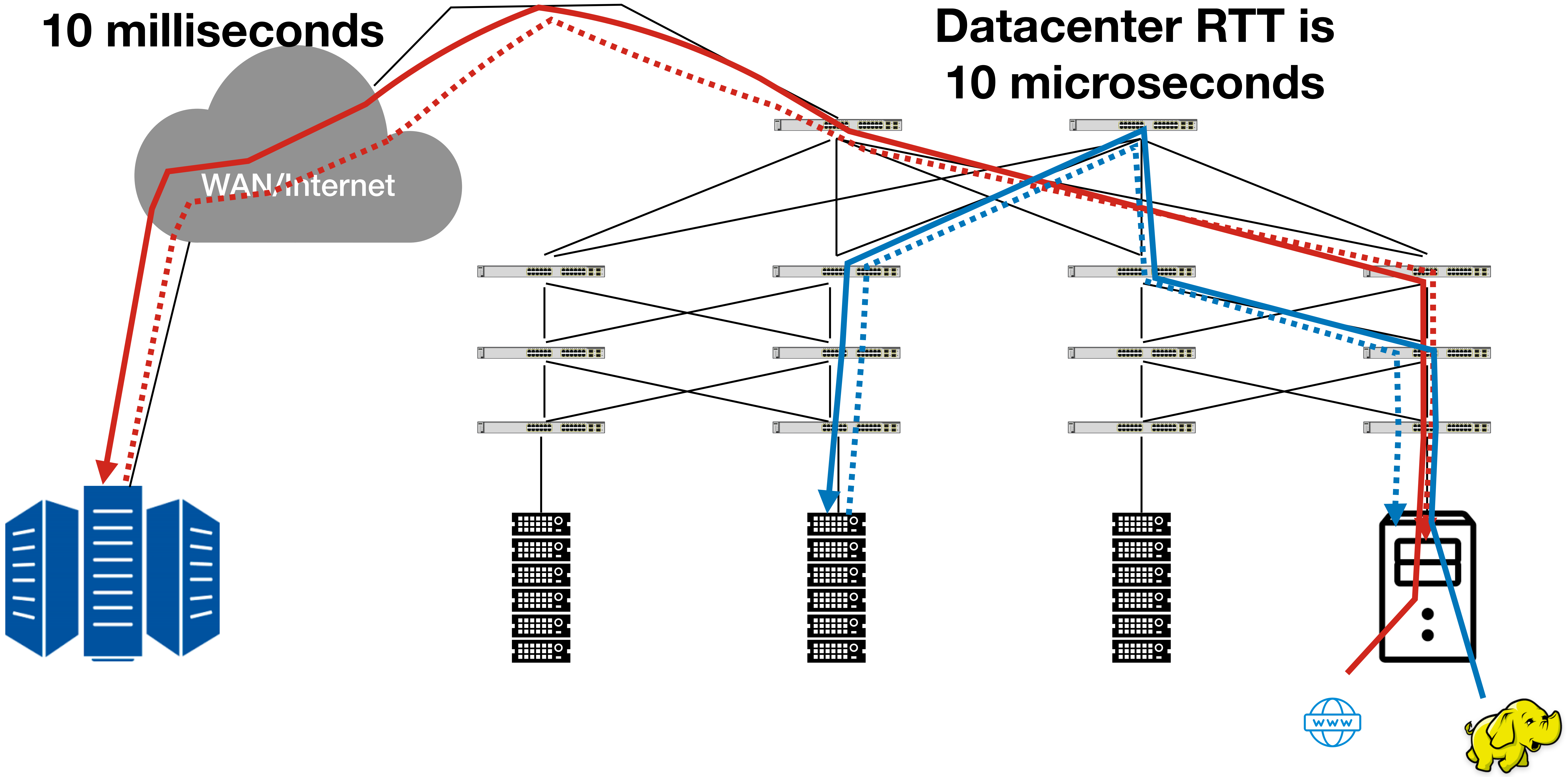
Impact of WAN on Datacenter



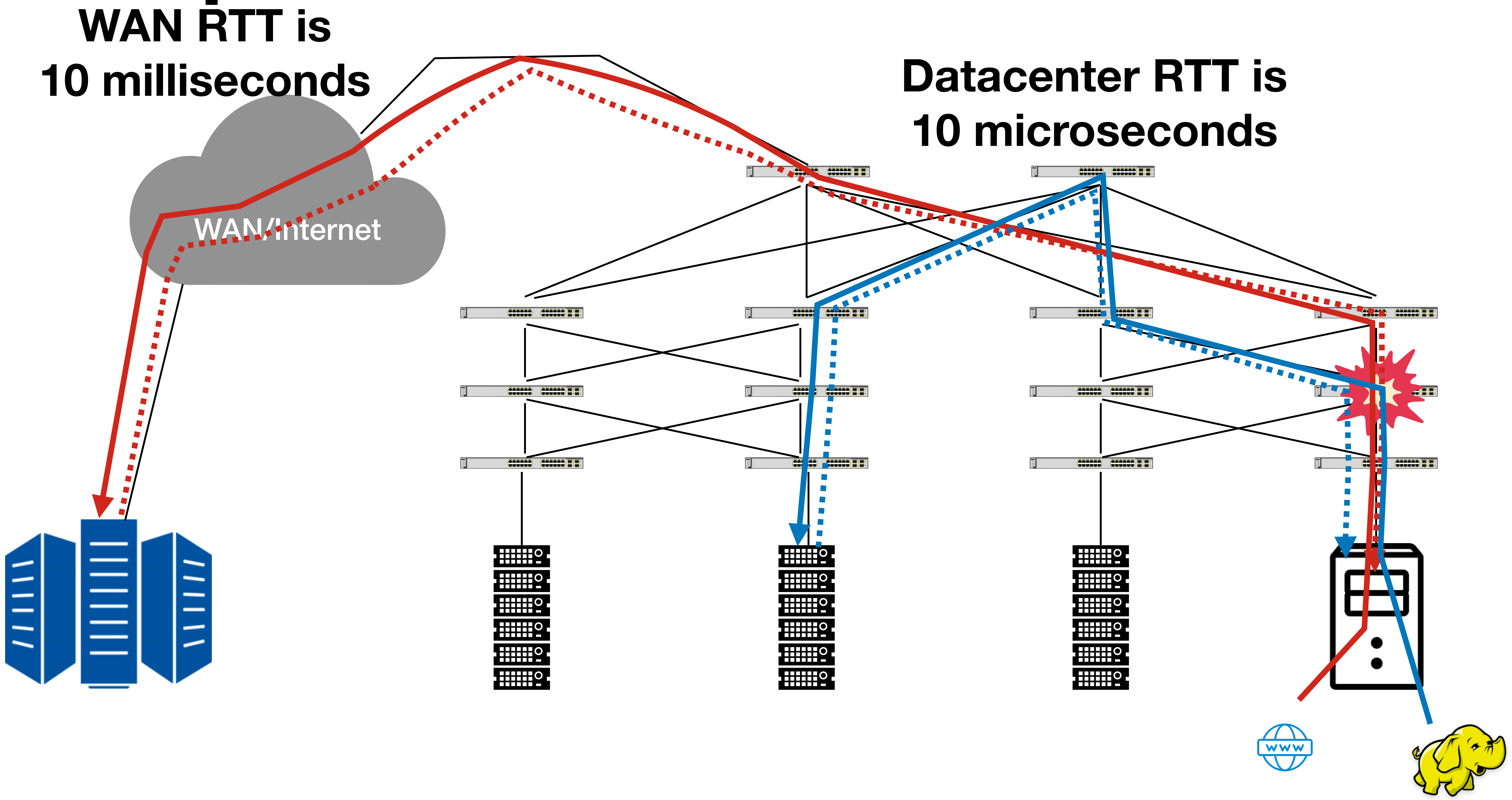
Impact of WAN on Datacenter

WAN RTT is 10 milliseconds

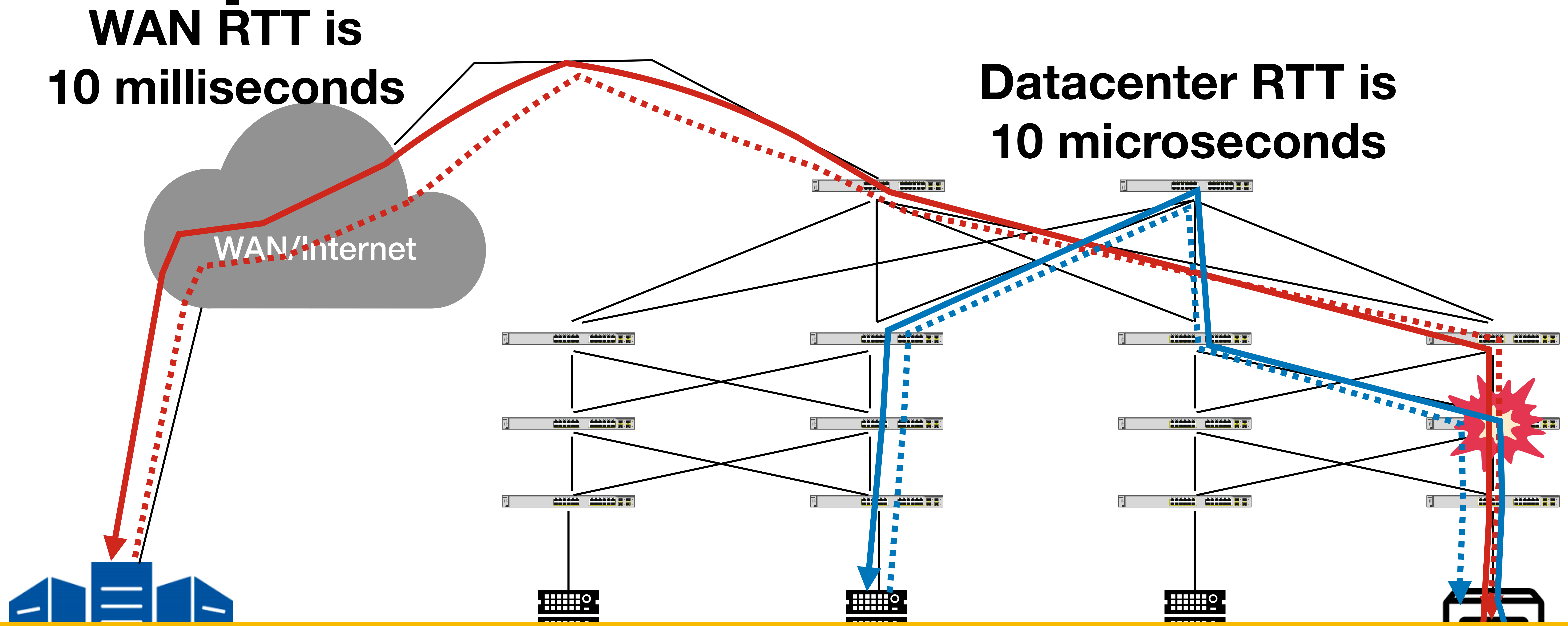
Datacenter RTT is 10 microseconds



Impact of WAN on Datacenter



Impact of WAN on Datacenter



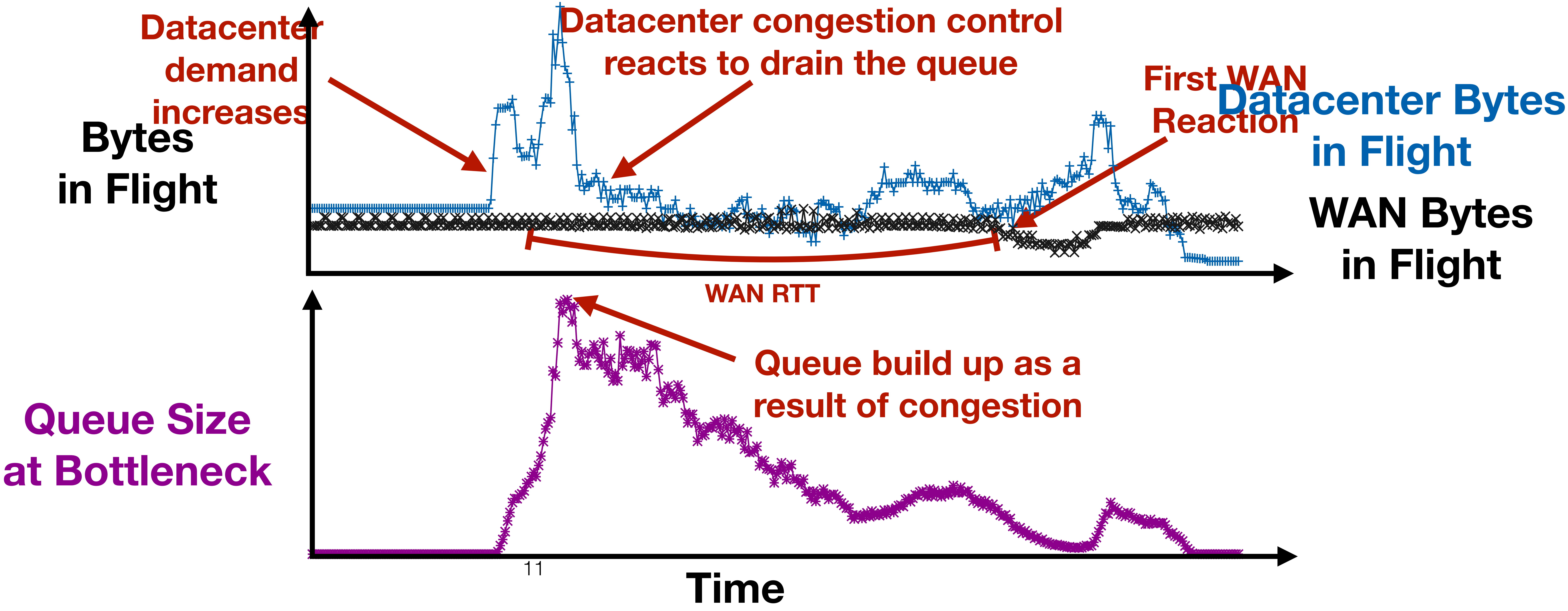
WAN will take a thousand datacenter RTTs to detect the problem, leaving datacenter to solely react to congestion

Example from Simulations

- Datacenter traffic and WAN traffic share a bottleneck

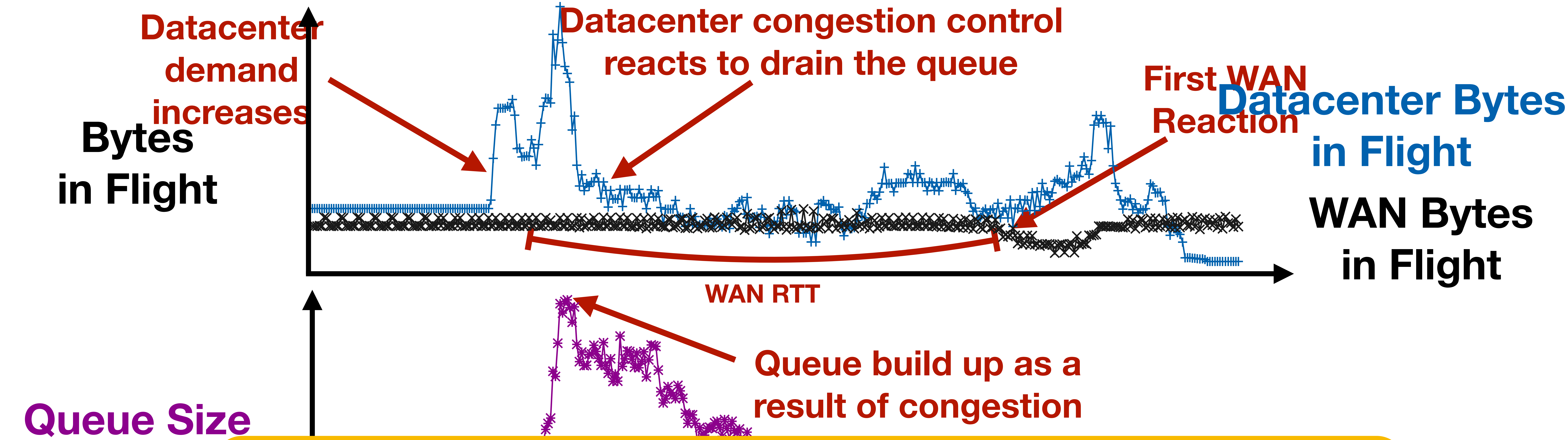
Example from Simulations

- Datacenter traffic and WAN traffic share a bottleneck



Example from Simulations

- Datacenter traffic and WAN traffic share a bottleneck



WAN traffic should react to congestion and drain the queue faster

Impact of Datacenter on WAN

Impact of Datacenter on WAN

- Buffer sizing for WAN flows is proportional to BDP

Impact of Datacenter on WAN

- Buffer sizing for WAN flows is proportional to BDP
  - Short buffers can be problematic

WAN BDP is O(megabytes) per flow

Impact of Datacenter on WAN

- Buffer sizing for WAN flows is proportional to BDP
    - Short buffers can be problematic

WAN BDP is O(megabytes) per flow

- Better algorithms have smaller buffer requirements

BBR or DCTCP

Impact of Datacenter on WAN

- Buffer sizing for WAN flows is proportional to BDP
  - Short buffers can be problematic

WAN BDP is O(megabytes) per flow

- Better algorithms have smaller buffer requirements
  - Assuming available bandwidth is stable

BBR or DCTCP

Impact of Datacenter on WAN

- Buffer sizing for WAN flows is proportional to BDP
  - Short buffers can be problematic

WAN BDP is O(megabytes) per flow

- Better algorithms have smaller buffer requirements
  - Assuming available bandwidth is stable

BBR or DCTCP

- Bandwidth available to WAN flows changes at datacenter RTT timscale

Impact of Datacenter on WAN

- Buffer sizing for WAN flows is proportional to BDP
  - Short buffers can be problematic

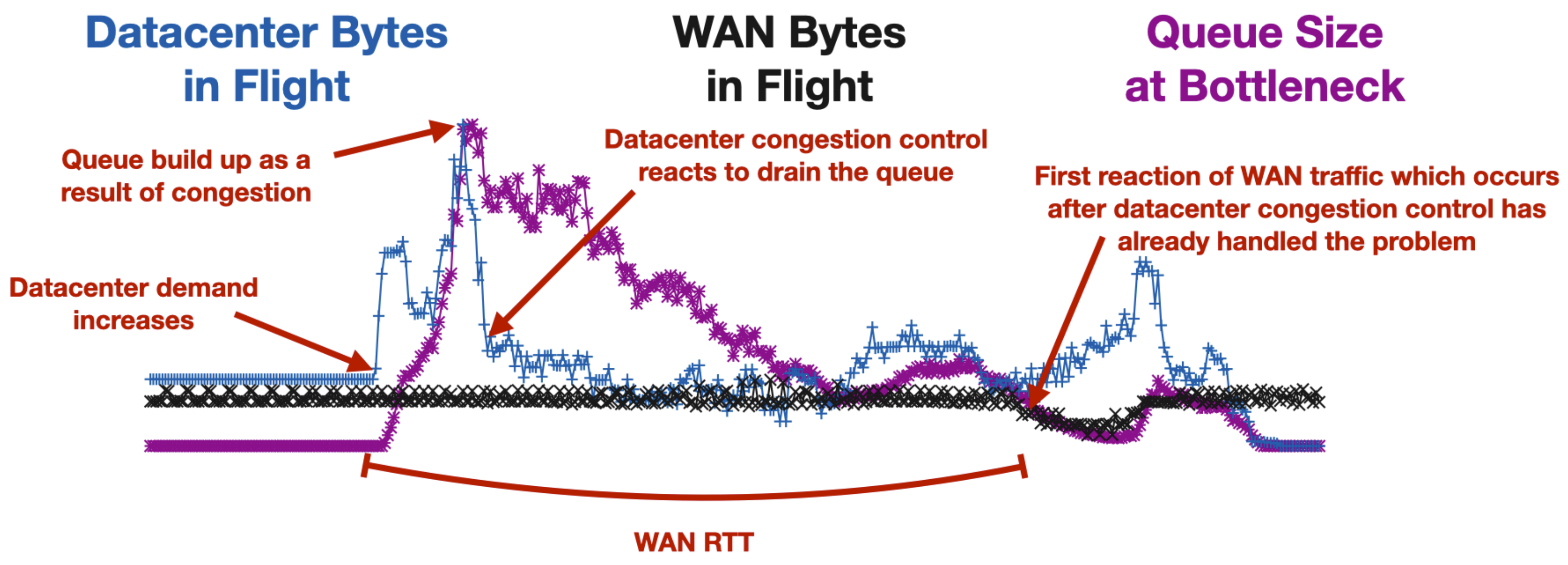
WAN BDP is O(megabytes) per flow

- Better algorithms have smaller buffer requirements
  - Assuming available bandwidth is stable

BBR or DCTCP

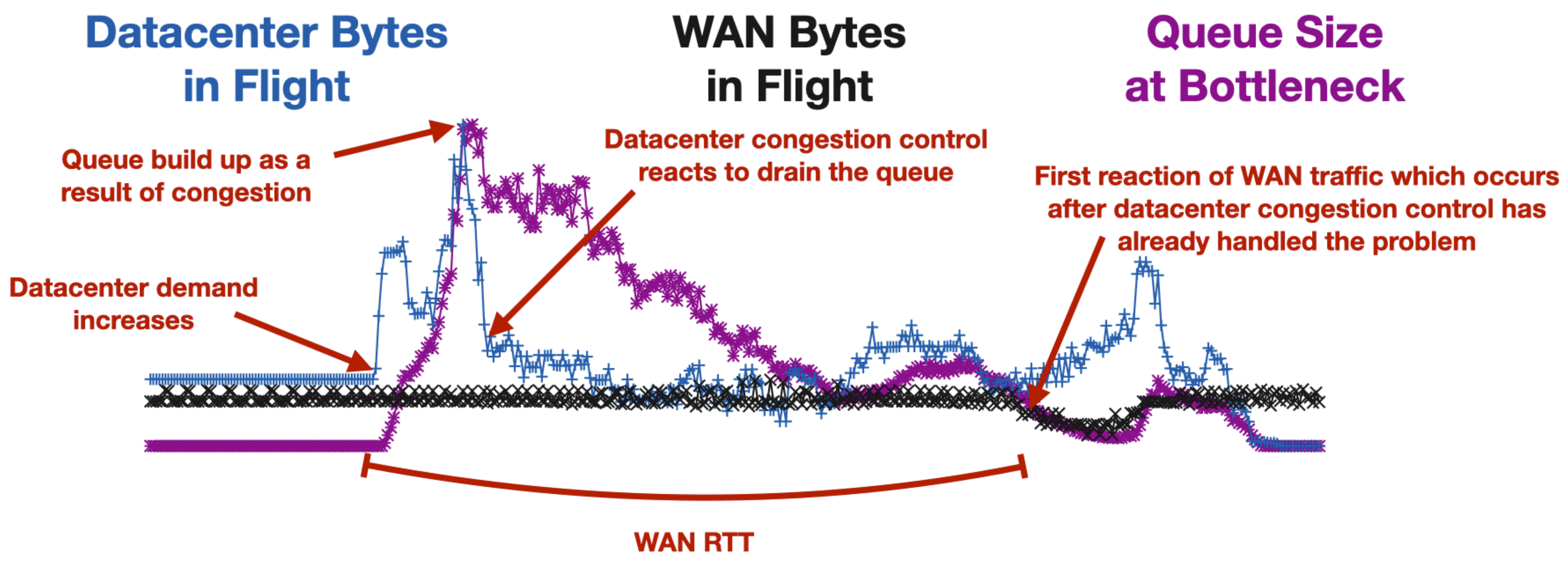
WAN traffic suffers from excessive loss due to lack of buffering and rapid changes in available bandwidth

Summary of Findings



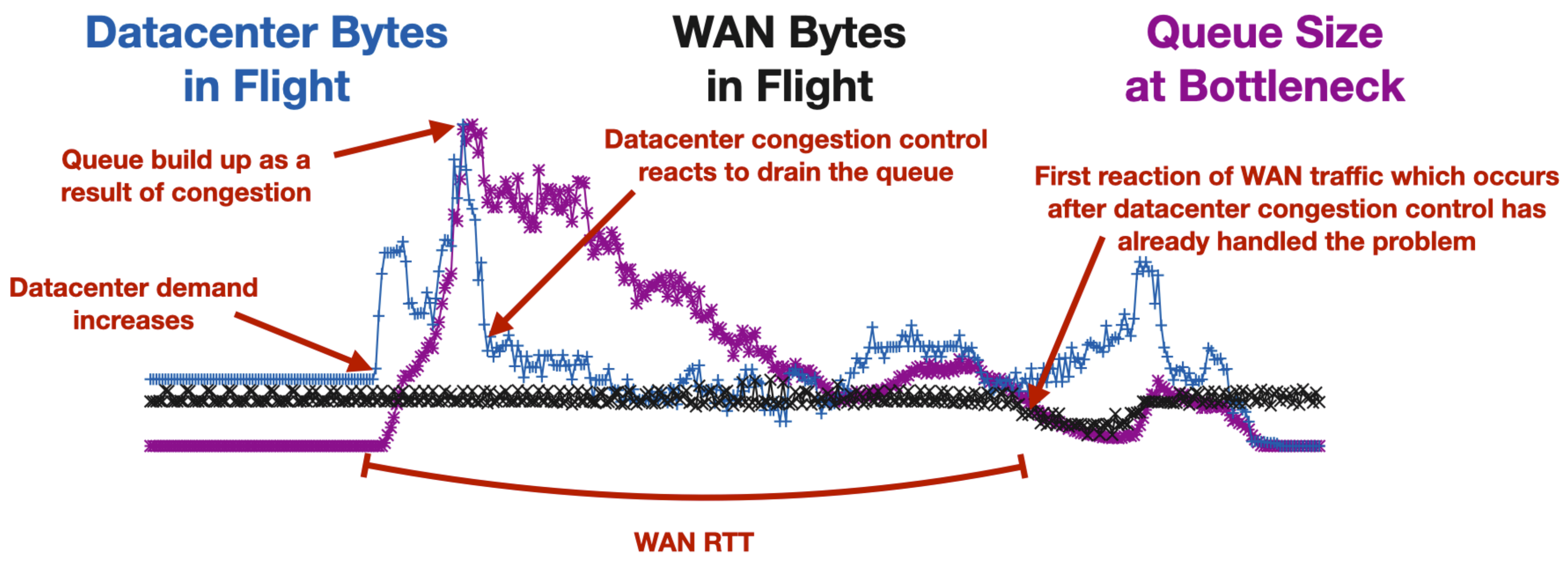
Summary of Findings

- **WAN RTT is too large compared to datacenter dynamics**



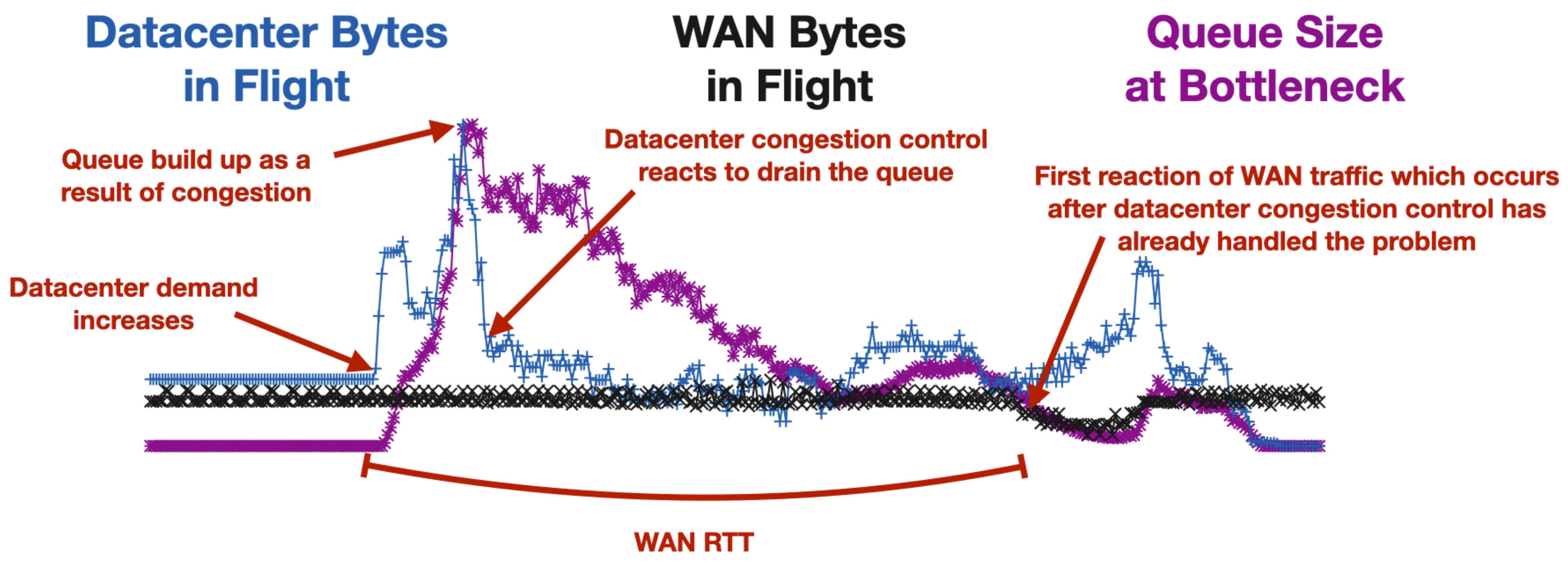
Summary of Findings

- **WAN RTT is too large compared to datacenter dynamics**
  - Datacenter throughput suffers as it solely reacts to congestion



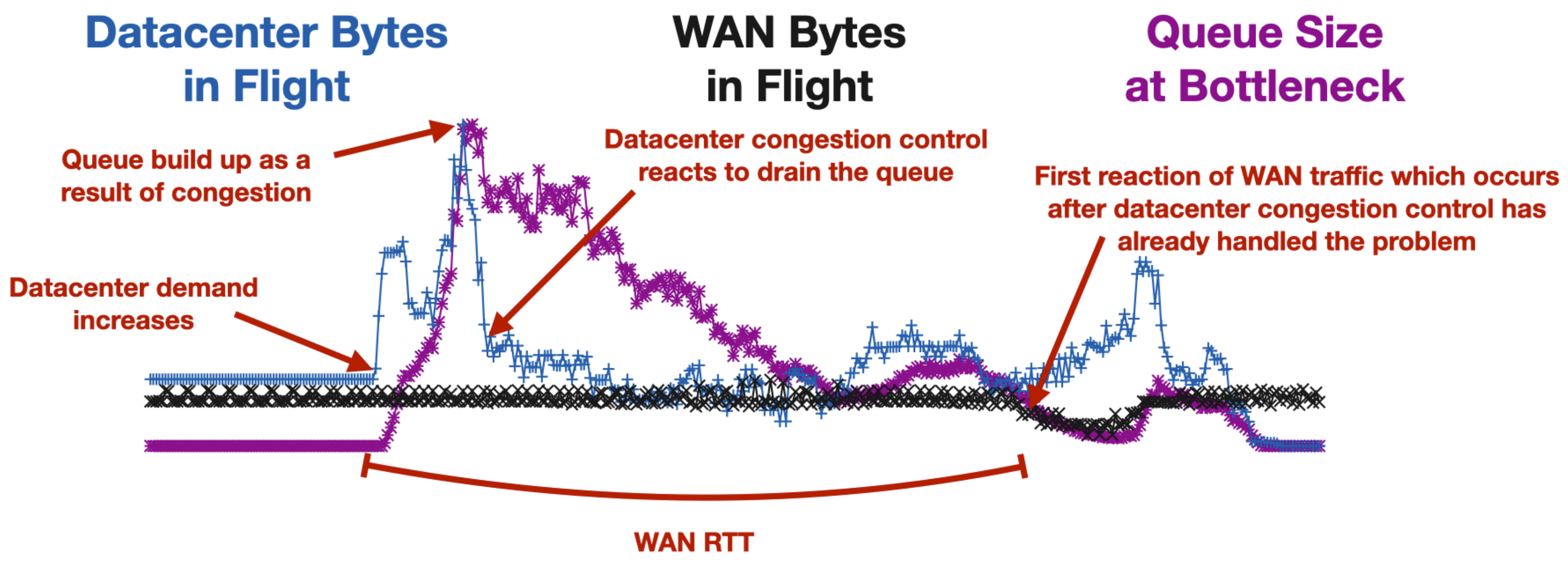
Summary of Findings

- **WAN RTT is too large compared to datacenter dynamics**
  - Datacenter throughput suffers as it solely reacts to congestion
  - WAN creates long queues due to rapid changes in available bandwidth



Summary of Findings

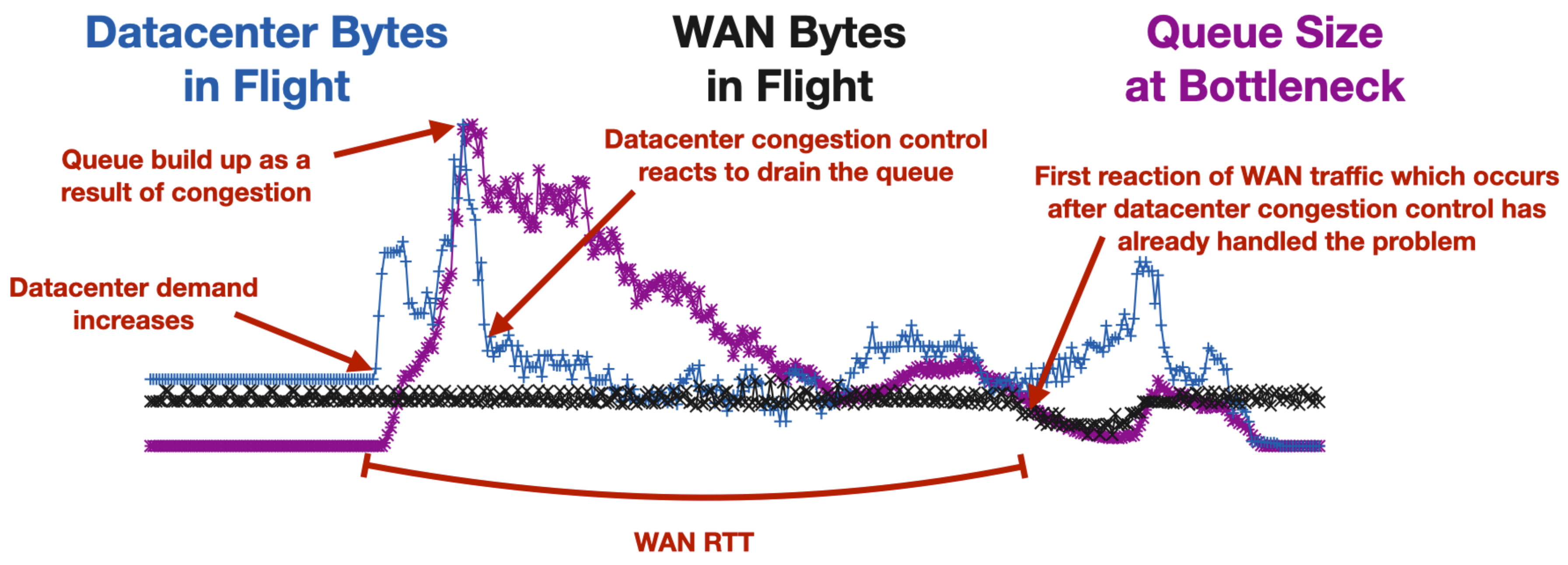
- **WAN RTT is too large compared to datacenter dynamics**
  - Datacenter throughput suffers as it solely reacts to congestion
  - WAN creates long queues due to rapid changes in available bandwidth



- **No buffer space in datacenter switches to absorb WAN bursts**

Summary of Findings

- **WAN RTT is too large compared to datacenter dynamics**
  - Datacenter throughput suffers as it solely reacts to congestion
  - WAN creates long queues due to rapid changes in available bandwidth



- **No buffer space in datacenter switches to absorb WAN bursts**
  - WAN traffic suffers due to excessive drops

What about isolating them at the bottleneck?

What about isolating them at the bottleneck?

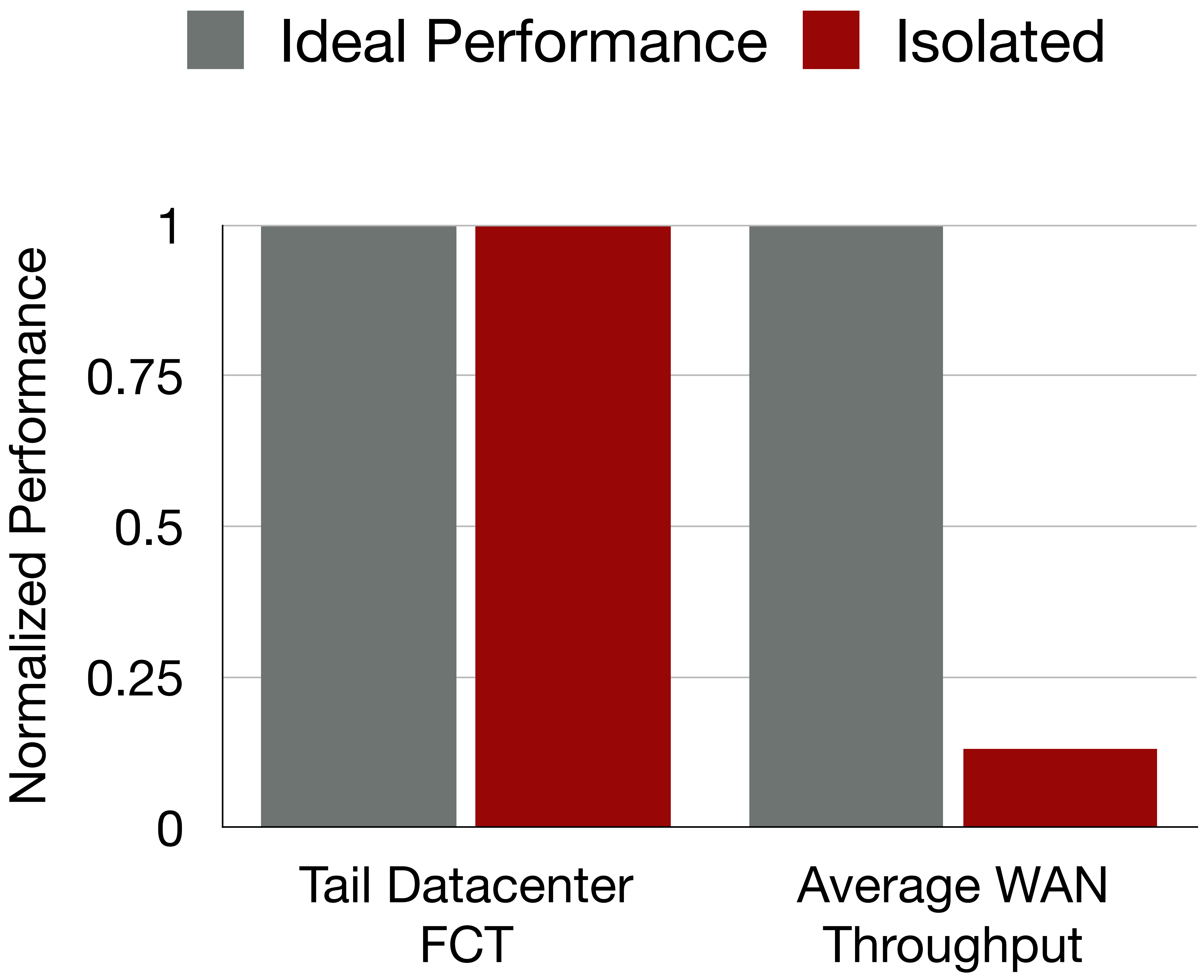
- There is not enough buffer space in datacenter switches to accommodate WAN BDP
  - **WAN traffic will still hurt due to excessive drops**

What about isolating them at the bottleneck?

- There is not enough buffer space in datacenter switches to accommodate WAN BDP
  - **WAN traffic will still hurt due to excessive drops**
- Datacenter and WAN traffic still share bandwidth even if they don’t share buffer space
  - **Exacerbates WAN drop rate**

What about isolating them at the bottleneck?

- There is not enough buffer space in datacenter switches to accommodate WAN BDP
  - **WAN traffic will still hurt due to excessive drops**
- Datacenter and WAN traffic still share bandwidth even if they don't share buffer space
  - **Exacerbates WAN drop rate**



Annulus

How should we handle bottlenecks shared between WAN and datacenter traffic?

How should we handle bottlenecks shared between WAN and datacenter traffic?

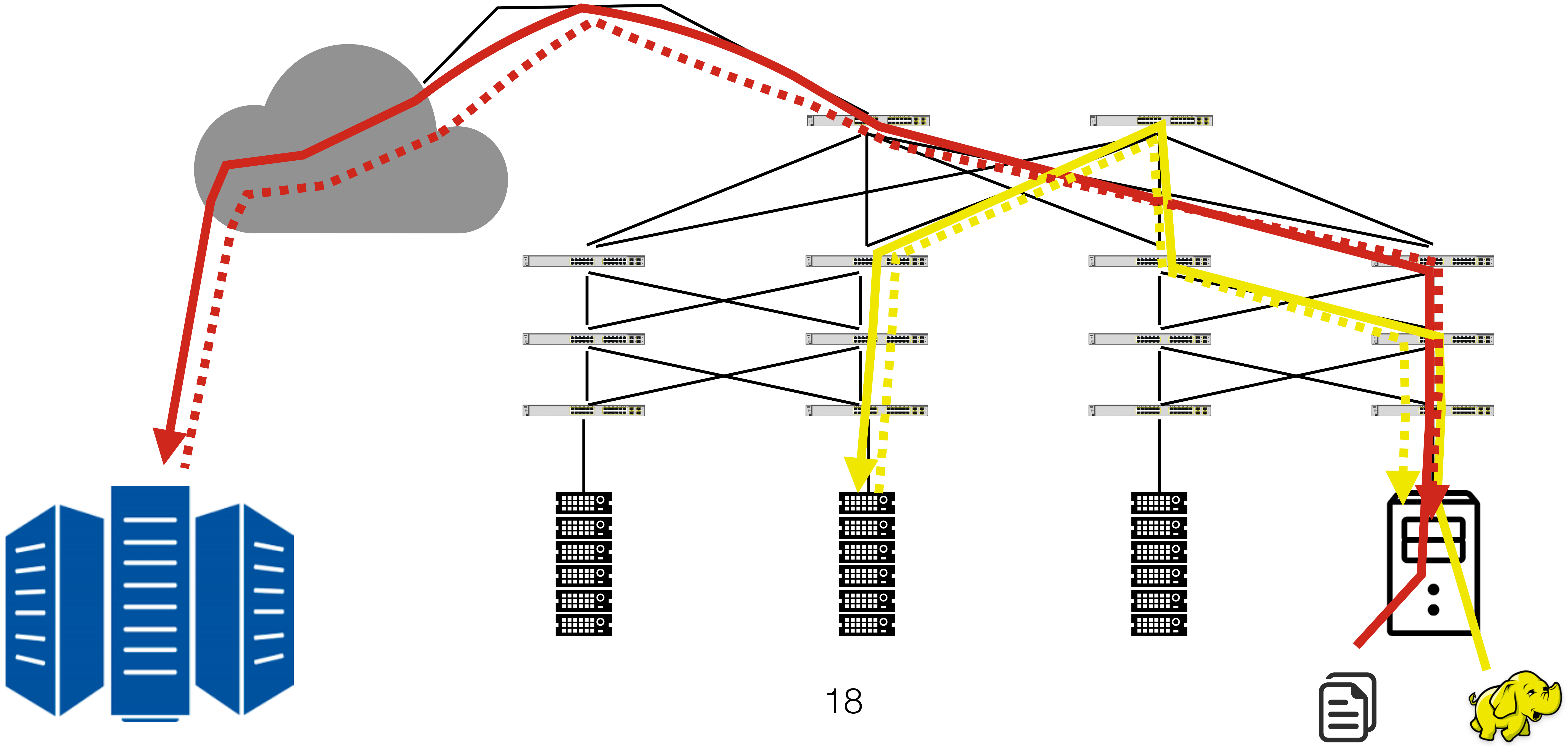
How should we handle the rest of the bottlenecks?

Main Idea

Reduce WAN feedback delay

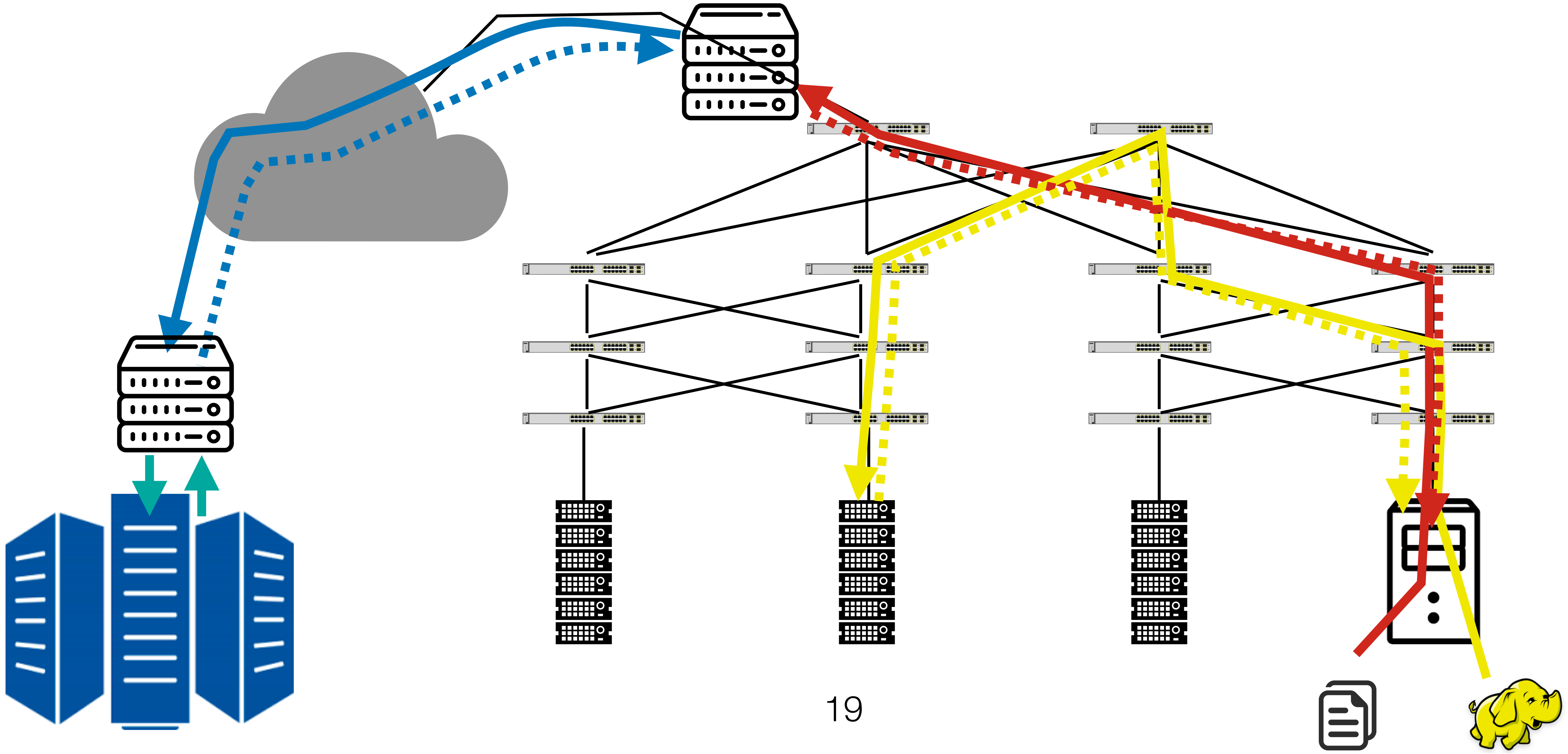
Cutting Feedback Delay

- **Connection termination at the border of the datacenter**



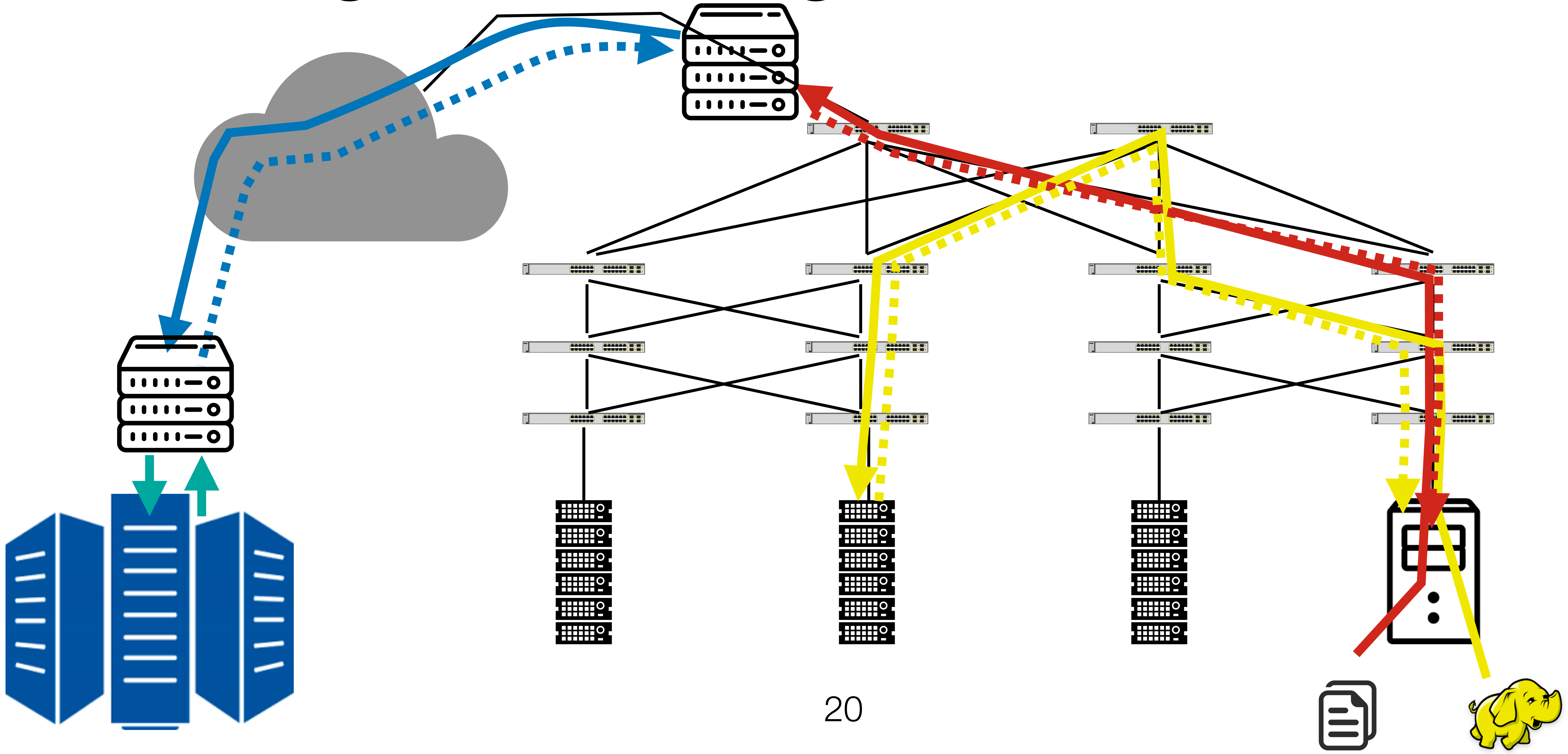
Cutting Feedback Delay

- **Connection termination at the border of the datacenter**



Cutting Feedback Delay

- **Connection termination at the border of the datacenter**
  - Requires middleboxes that handles the state of all WAN traffic entering and exiting datacenter



Cutting Feedback Delay

- **Connection termination at the border of the datacenter**
  - Requires middleboxes that handles the state of all WAN traffic entering and exiting datacenter

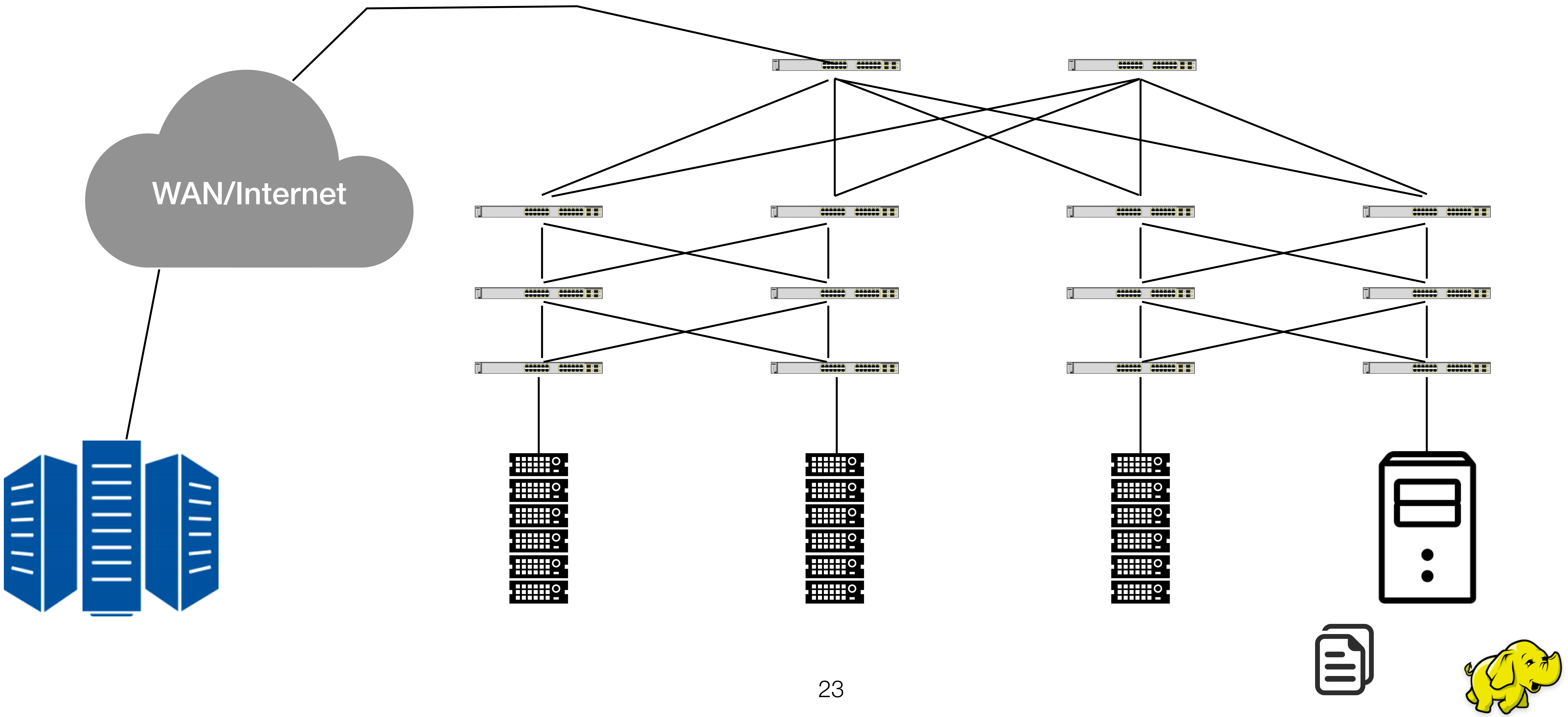
- **Direct signal from the bottleneck**

Cutting Feedback Delay

- **Connection termination at the border of the datacenter**
  - Requires middleboxes that handles the state of all WAN traffic entering and exiting datacenter
- **Direct signal from the bottleneck**
  - Requires switches that support direct congestion feedback

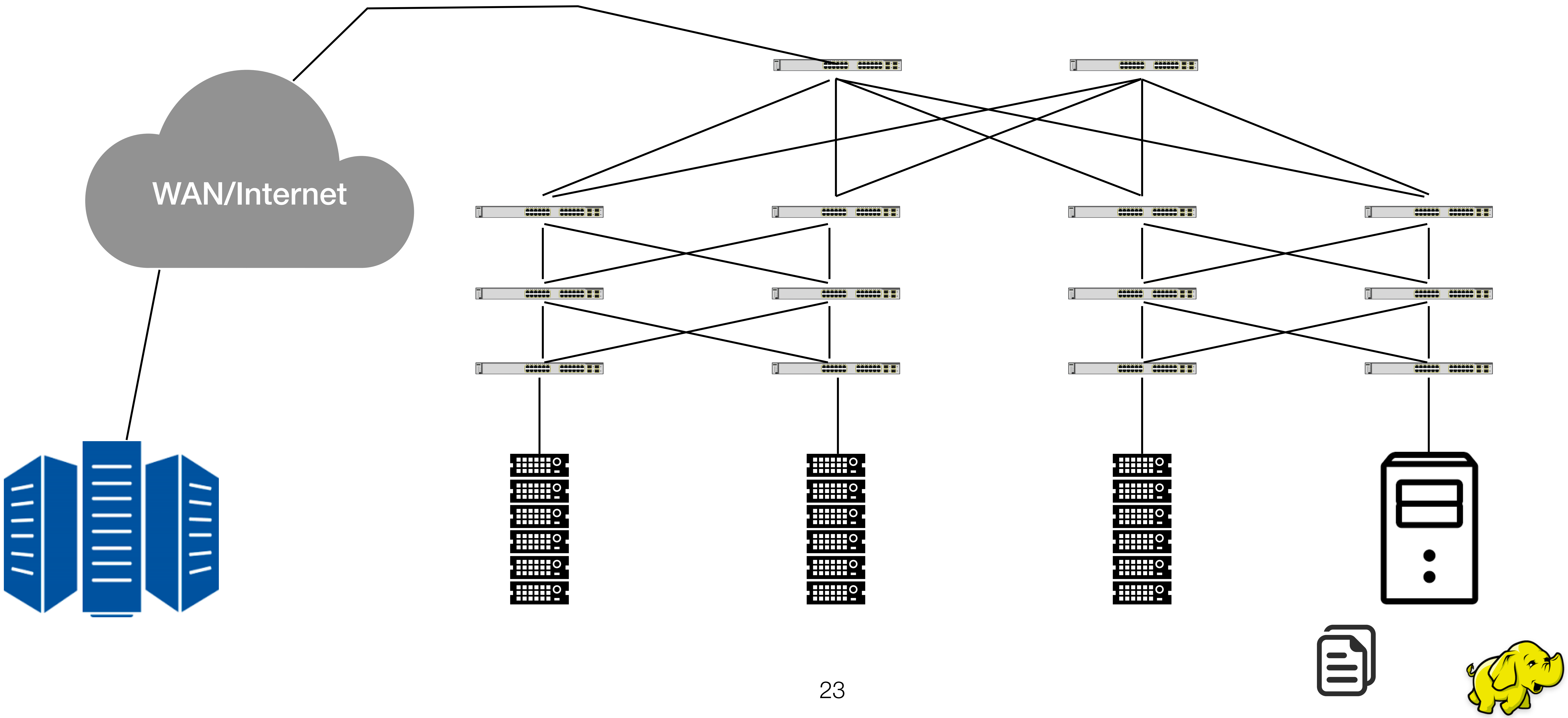
What about bottlenecks that can’t generate a direct signal?

Introducing Annulus



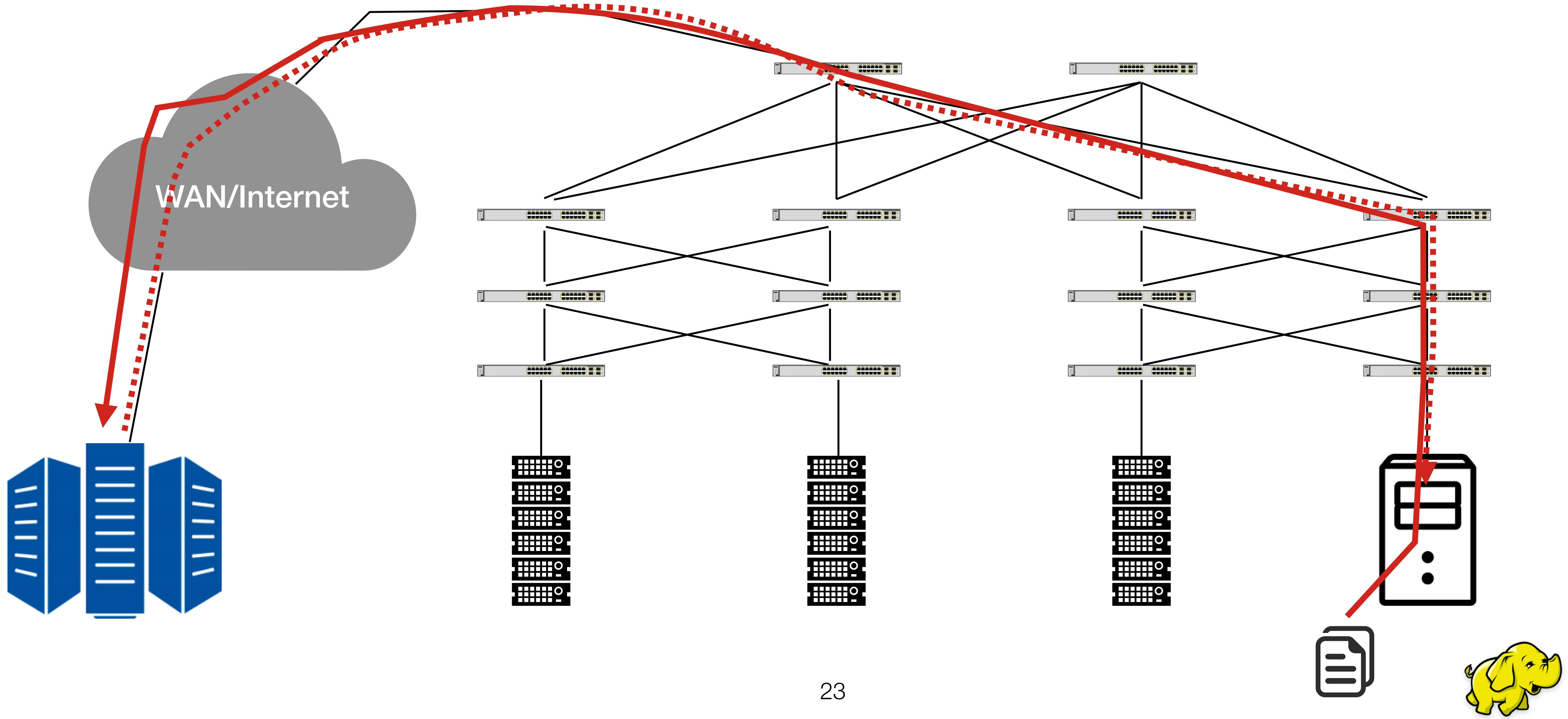
Introducing Annulus

- **Existing congestion control for WAN and datacenter**



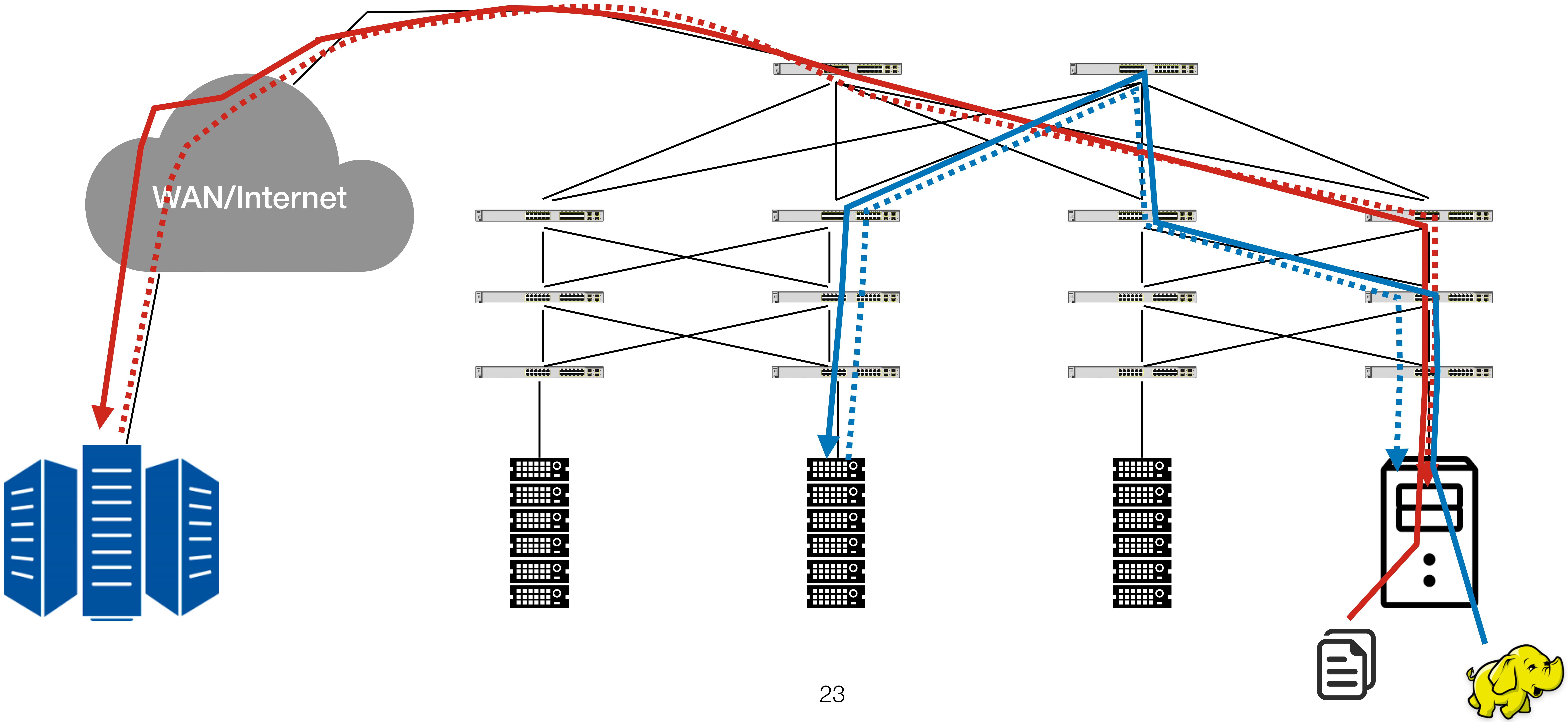
Introducing Annulus

- **Existing congestion control for WAN and datacenter**



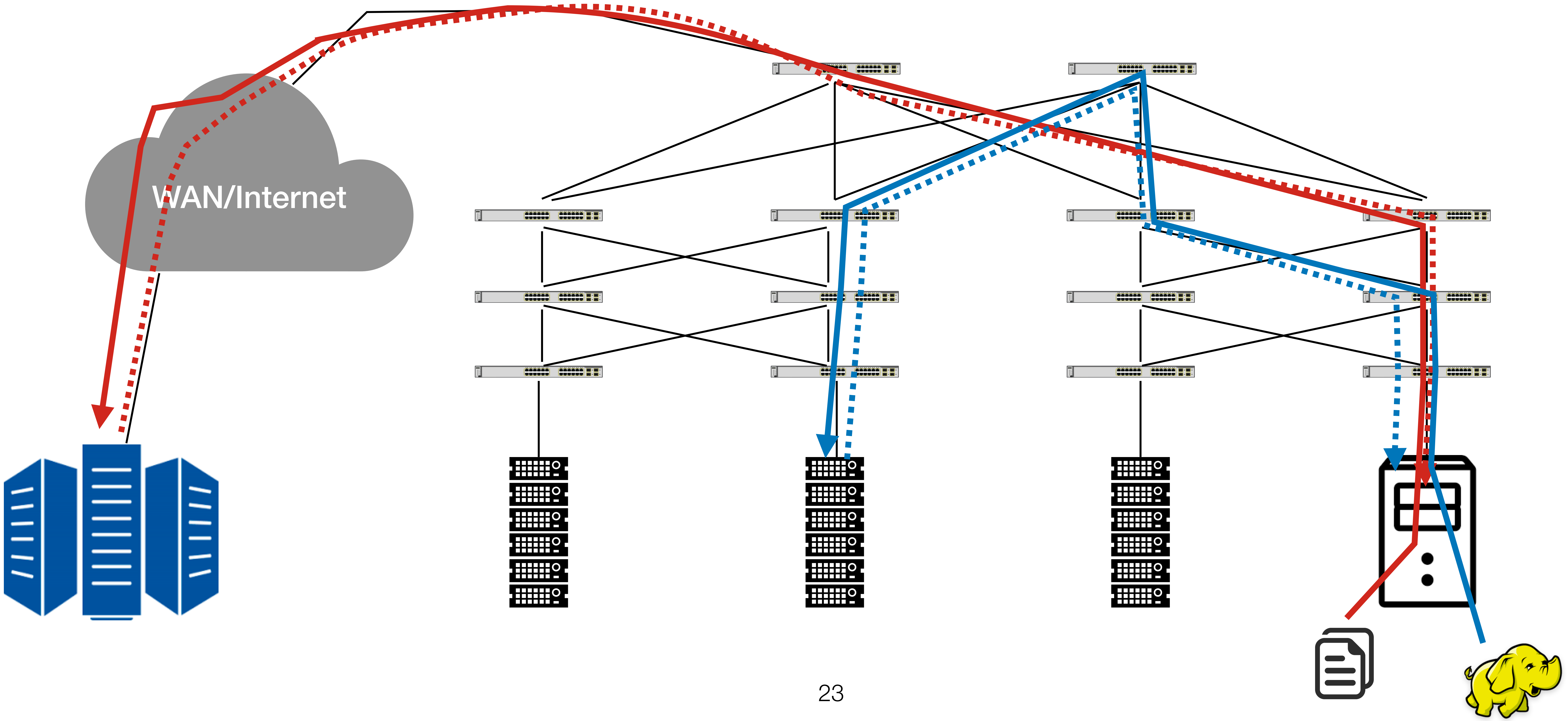
Introducing Annulus

- **Existing congestion control for WAN and datacenter**



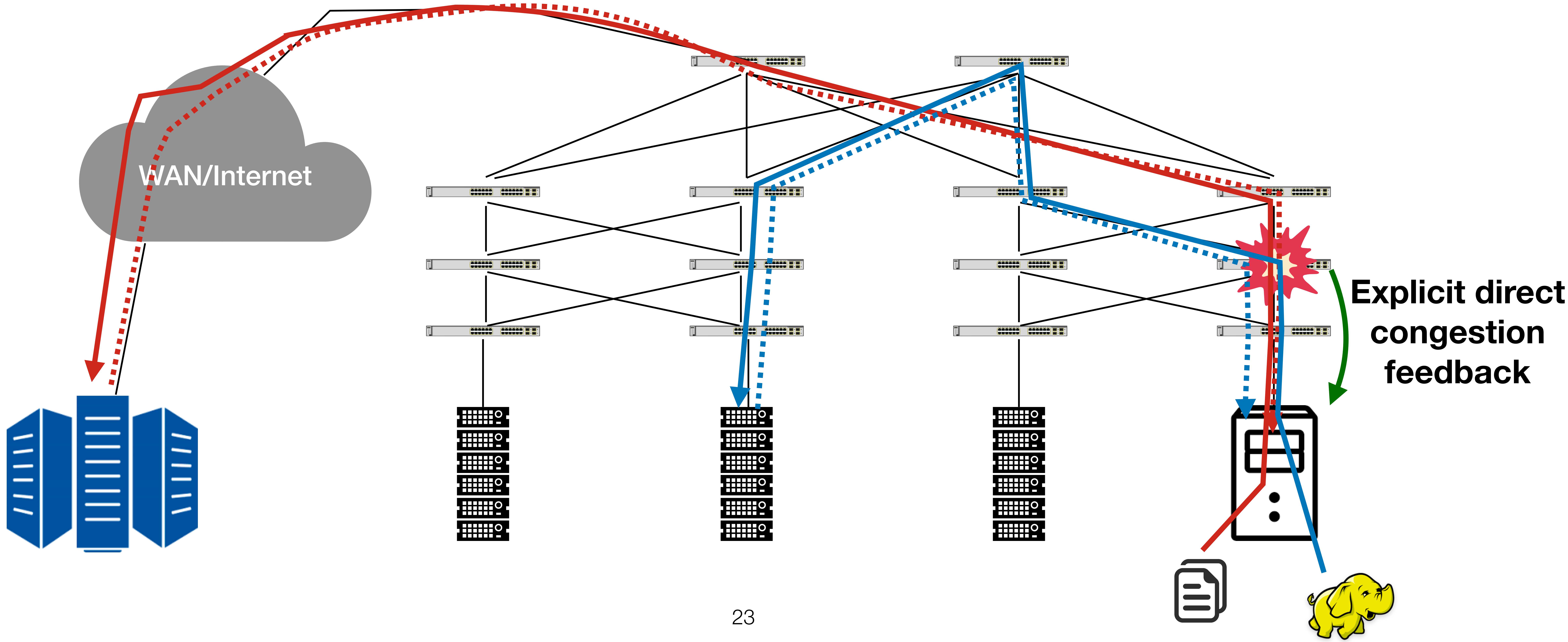
Introducing Annulus

- **Existing congestion control for WAN and datacenter**
- **Near-source control loop that relies on explicit direct feedback**

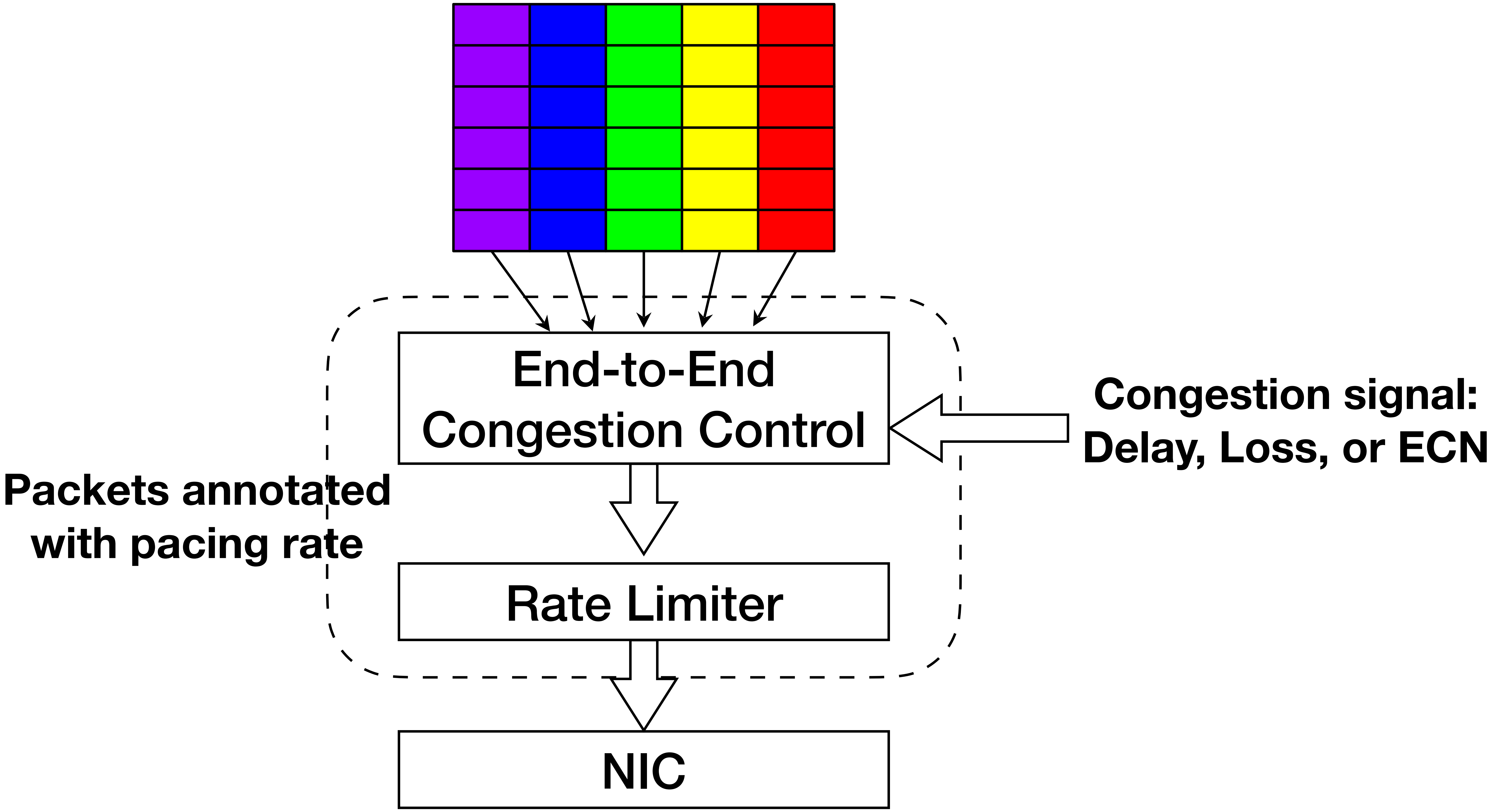


Introducing Annulus

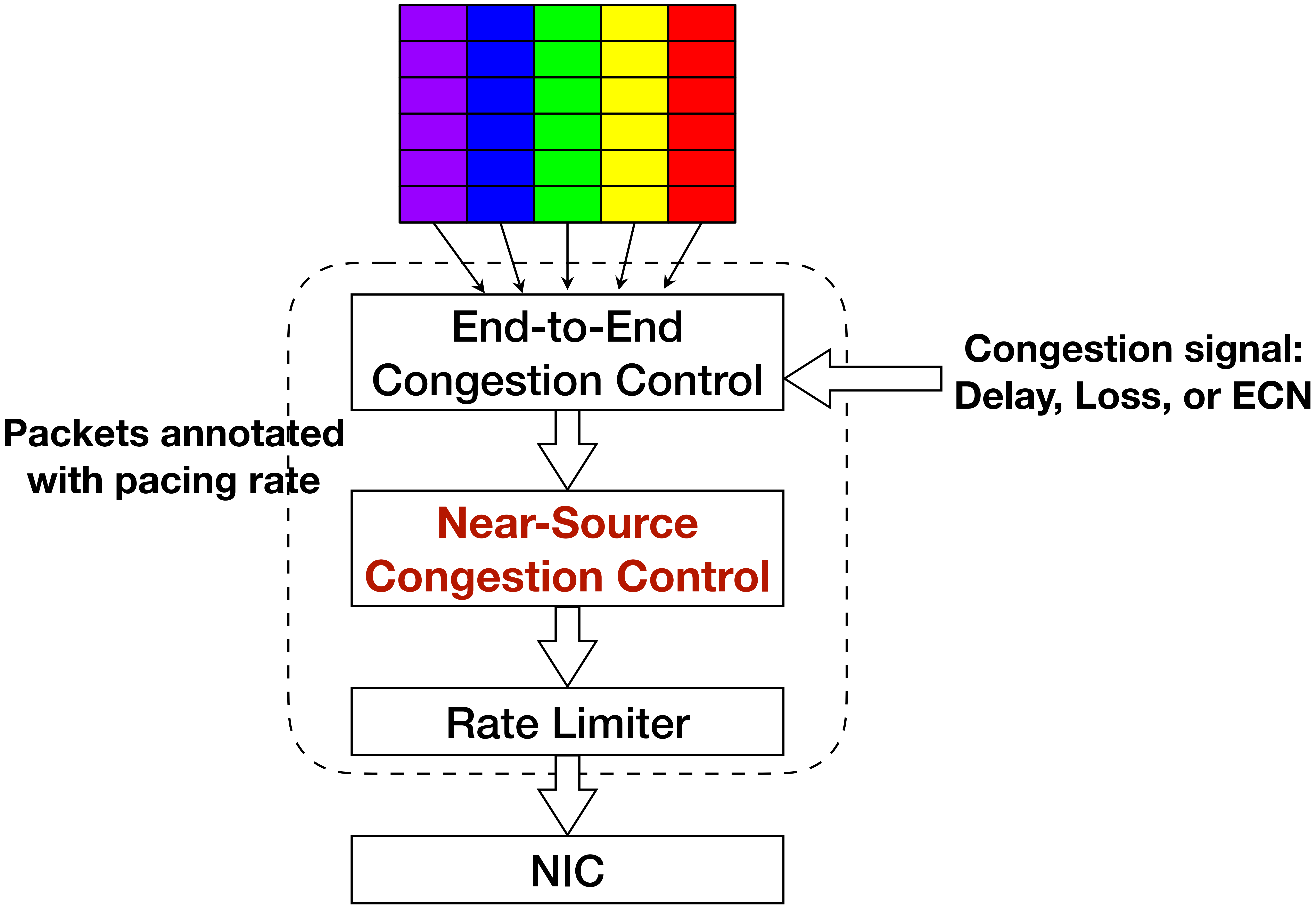
- **Existing congestion control for WAN and datacenter**
- **Near-source control loop that relies on explicit direct feedback**



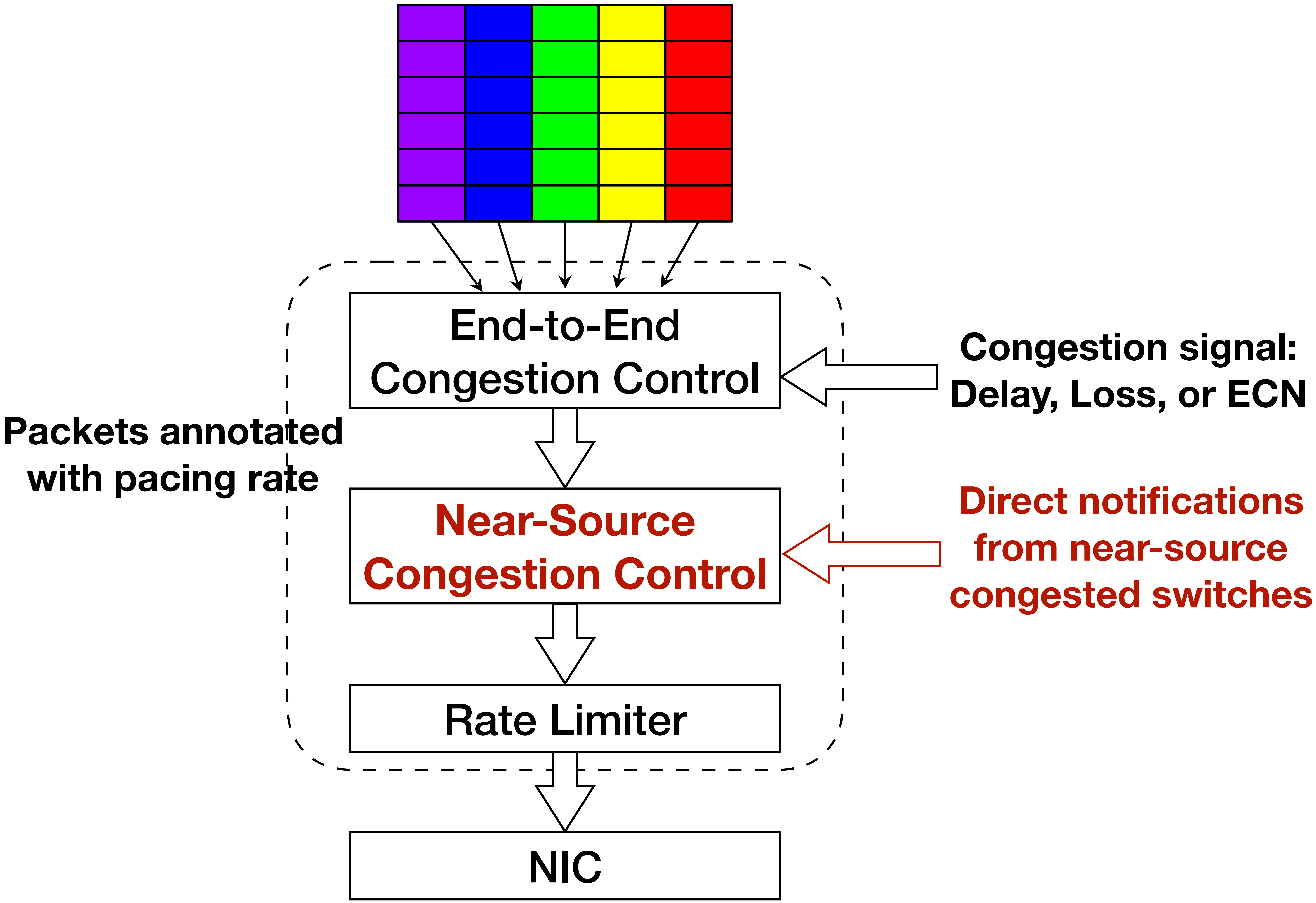
Annulus Overview



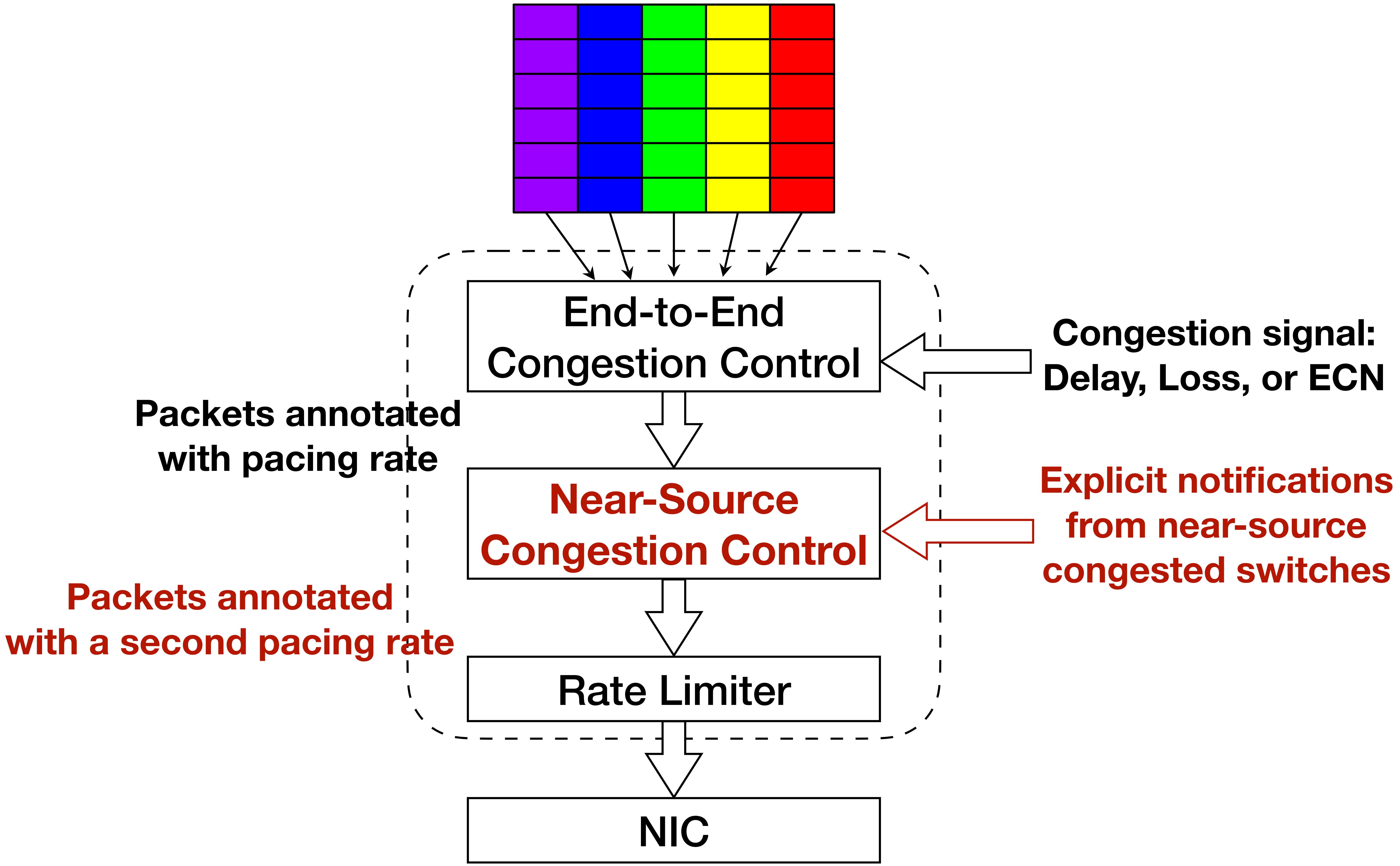
Annulus Overview



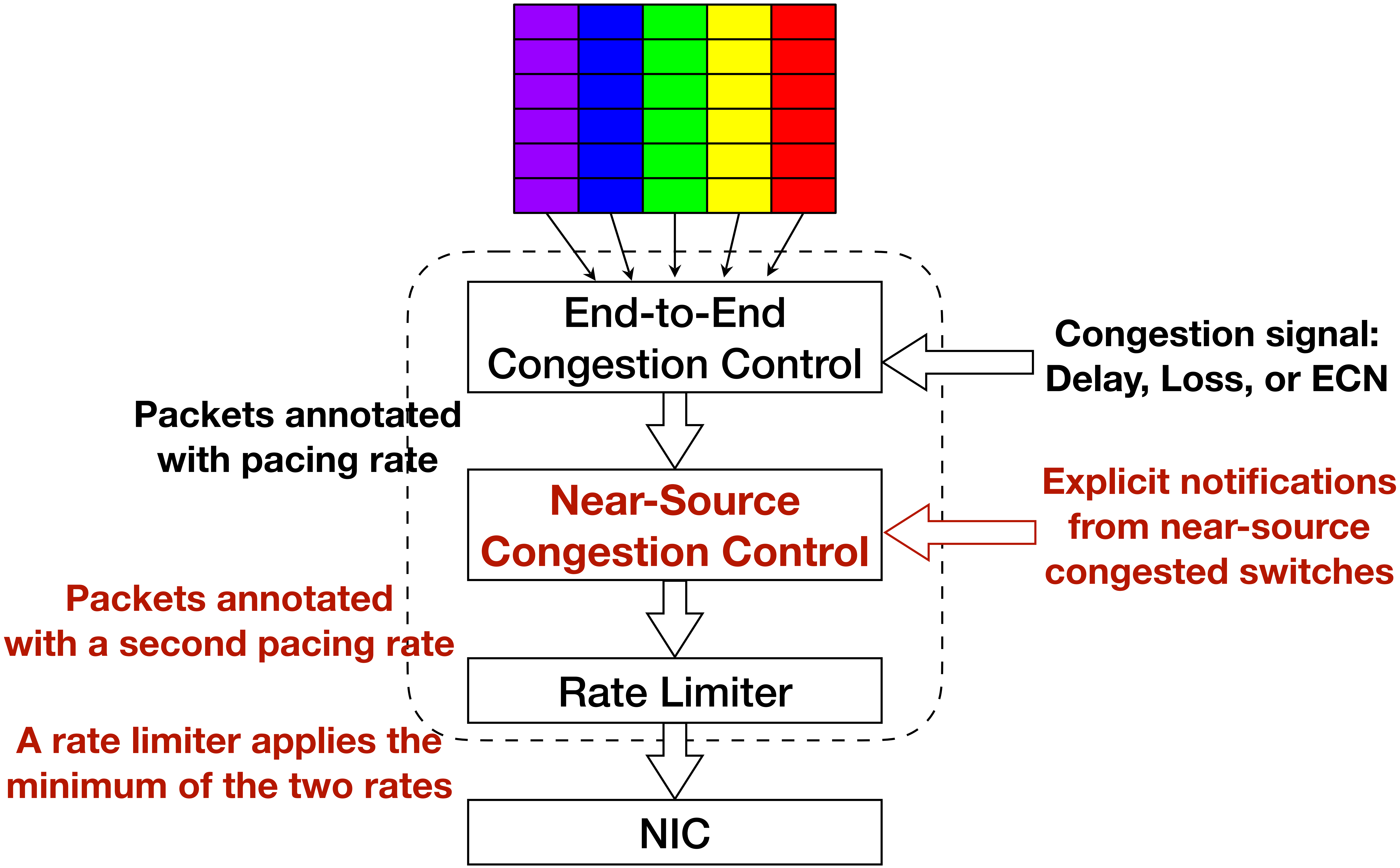
Annulus Overview



Annulus Overview

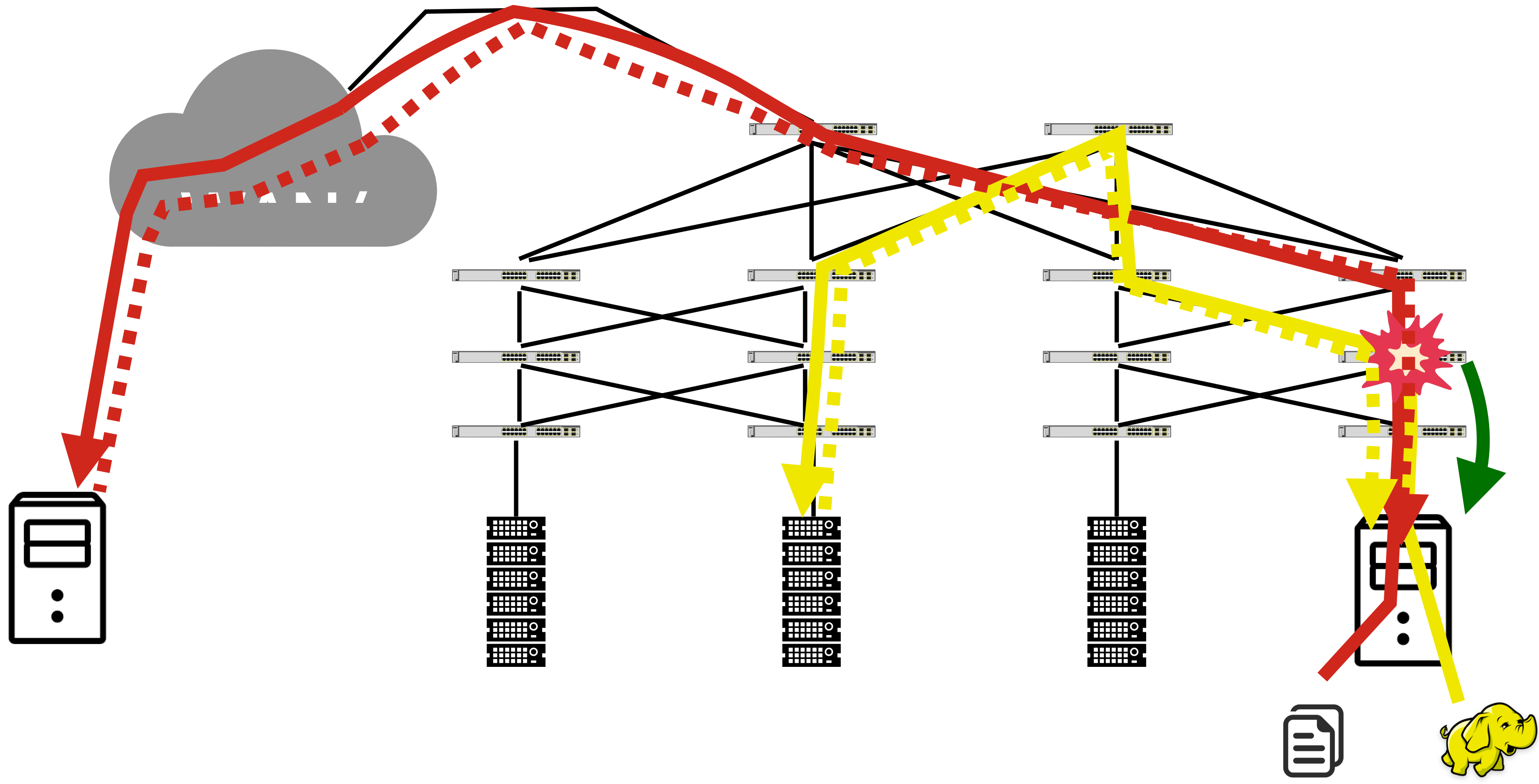


Annulus Overview



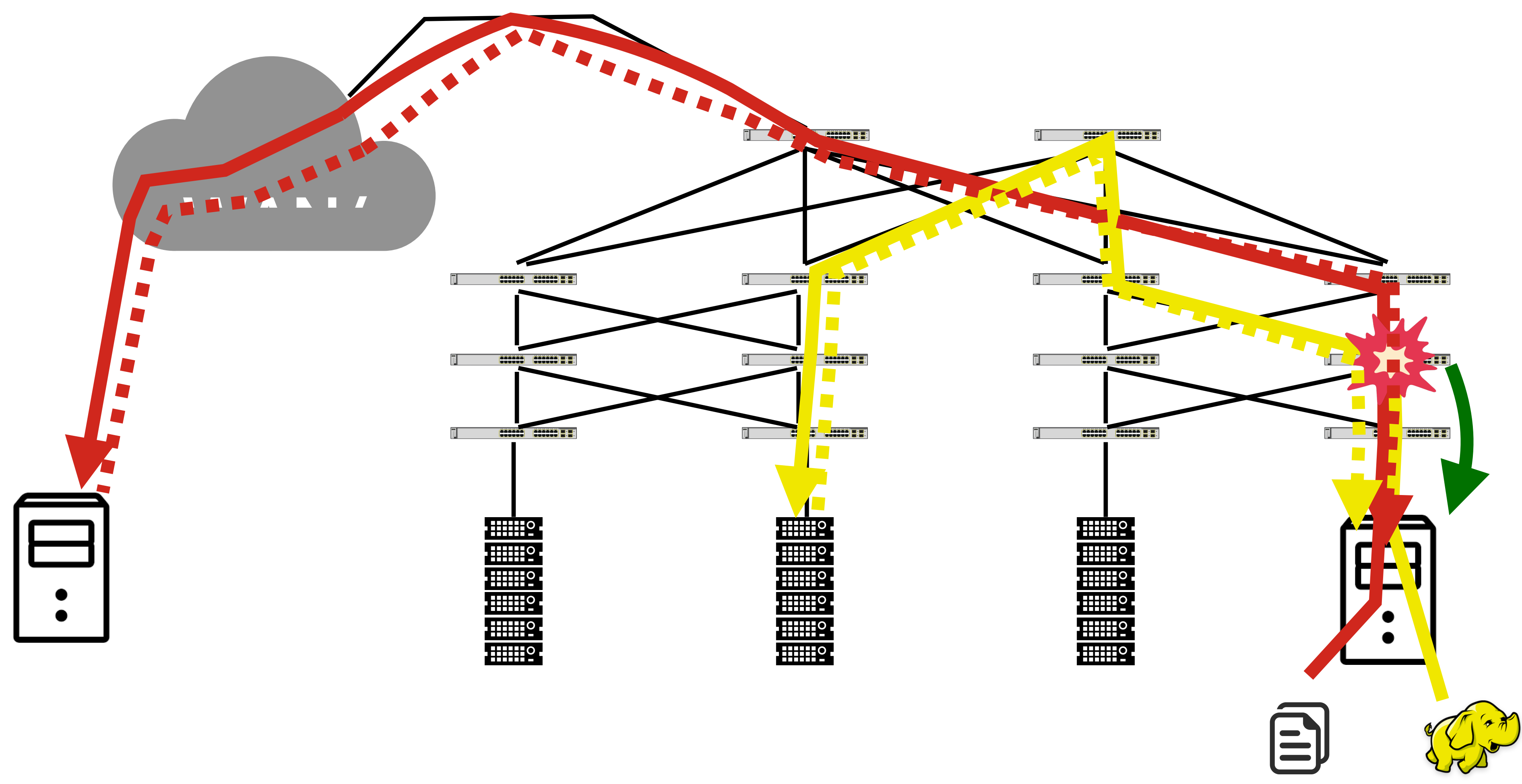
Near-Source
Congestion Control Loop

Near-Source Control Loop



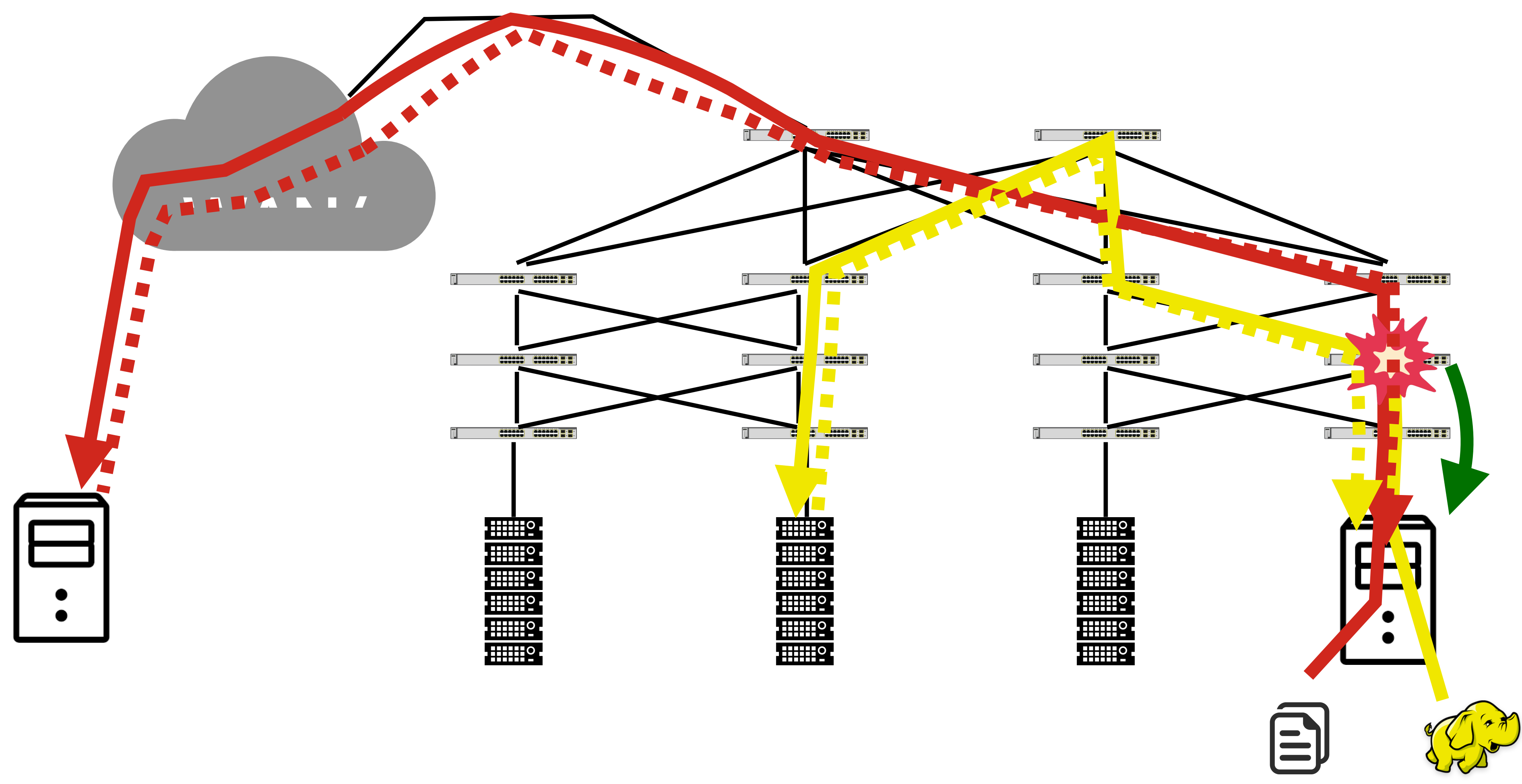
Near-Source Control Loop

- Switches generates direct congestion notification message



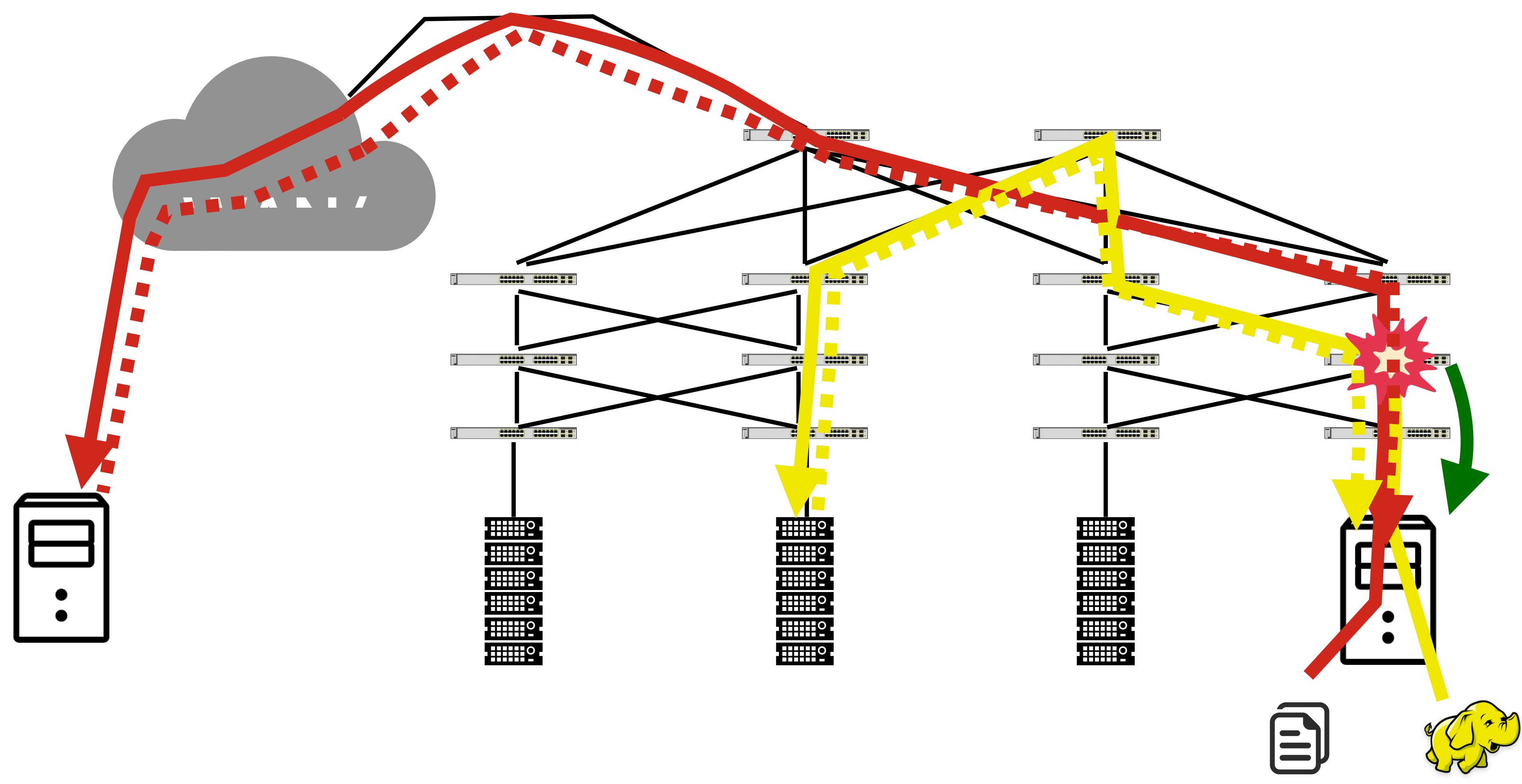
Near-Source Control Loop

- Switches generates direct congestion notification message
- Message indicates the problematic flow and the extent of the congestion



Near-Source Control Loop

- Switches generates direct congestion notification message
- Message indicates the problematic flow and the extent of the congestion
- Sender modulates transmission rate based on congestion level



Challenges

Challenges

- How to implement the direct signal in switches?

Challenges

- How to implement the direct signal in switches?
- How should the two control loops interact?

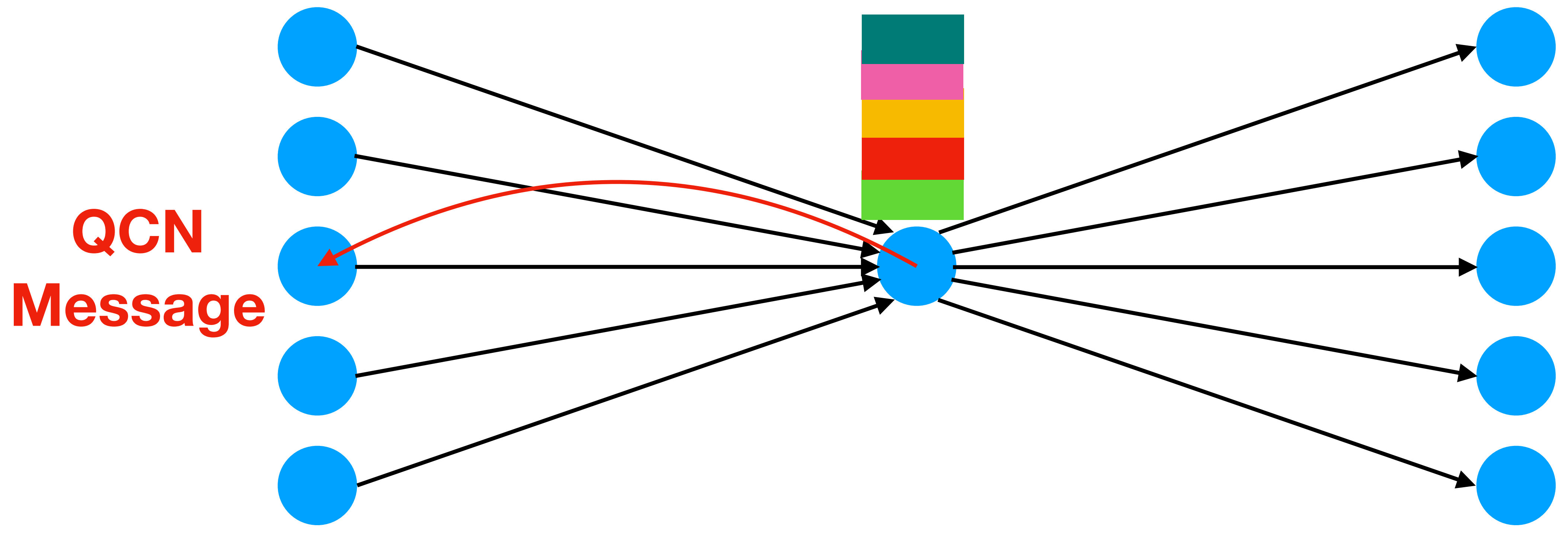
Quantized Congestion Notification (QCN)

- An IEEE standardized L2 congestion control algorithm (IEEE Std 802.1Qau-2010)

- QCN relies on explicit control messages from the point of congestion sent to traffic sources indicating congestion severity

Quantized Congestion Notification (QCN)

- An IEEE standardized L2 congestion control algorithm (IEEE Std 802.1Qau-2010)
- QCN relies on explicit control messages from the point of congestion sent to traffic sources indicating congestion severity



QCN from L2 to L4

QCN from L2 to L4

- QCN messages are L2 messages that rely on L2 routing

QCN from L2 to L4

- QCN messages are L2 messages that rely on L2 routing

QCN messages are routed in L3-routed an data center network by enabling “L2 Learning” feature available in modern switches

QCN from L2 to L4

- QCN messages are L2 messages that rely on L2 routing

  QCN messages are routed in L3-routed an data center network by enabling “L2 Learning” feature available in modern switches

- QCN control logic relies on accurate timers and counters implemented in hardware

QCN from L2 to L4

- QCN messages are L2 messages that rely on L2 routing

  QCN messages are routed in L3-routed an data center network by enabling “L2 Learning” feature available in modern switches

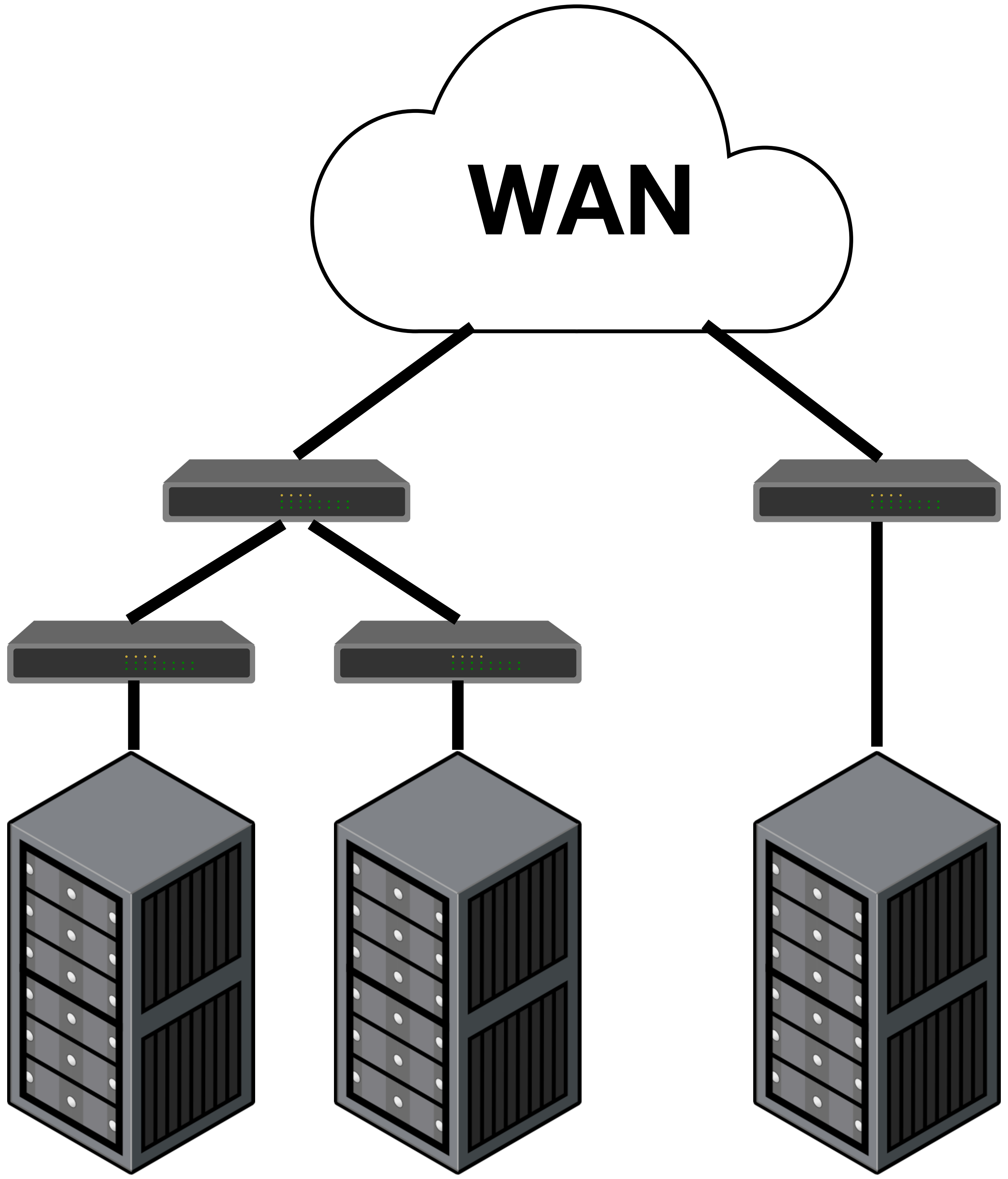
- QCN control logic relies on accurate timers and counters implemented in hardware

  A QCN-based congestion control logic is implemented in a software NIC or in the hypervisor

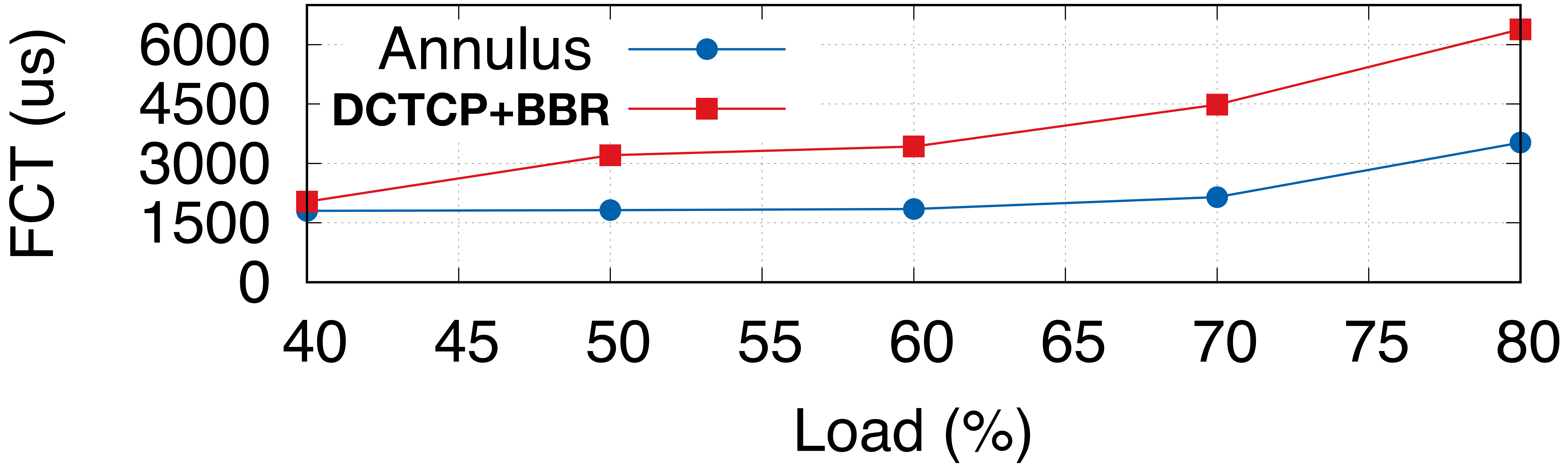
Evaluation

Evaluation Setup

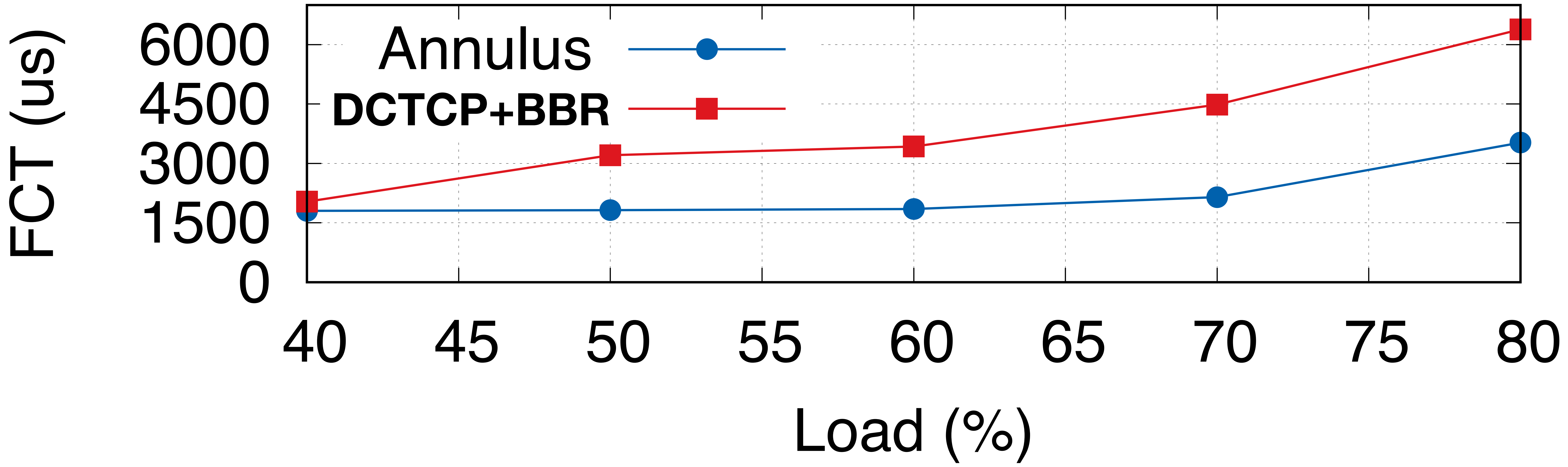
- Annulus is evaluated on three racks:
  Two racks in the same LAN and one connected to them through WAN
- WAN latency is 8ms and LAN latency is tens of microseconds
- Synthetic load is generated using an RPC load generator with cross-rack all to all communication
- Datacenter to WAN traffic ratio is 5:1
- DCTCP and BBR are used for end-to-end congestion control for datacenter and WAN traffic



Tail RPC Completion Time

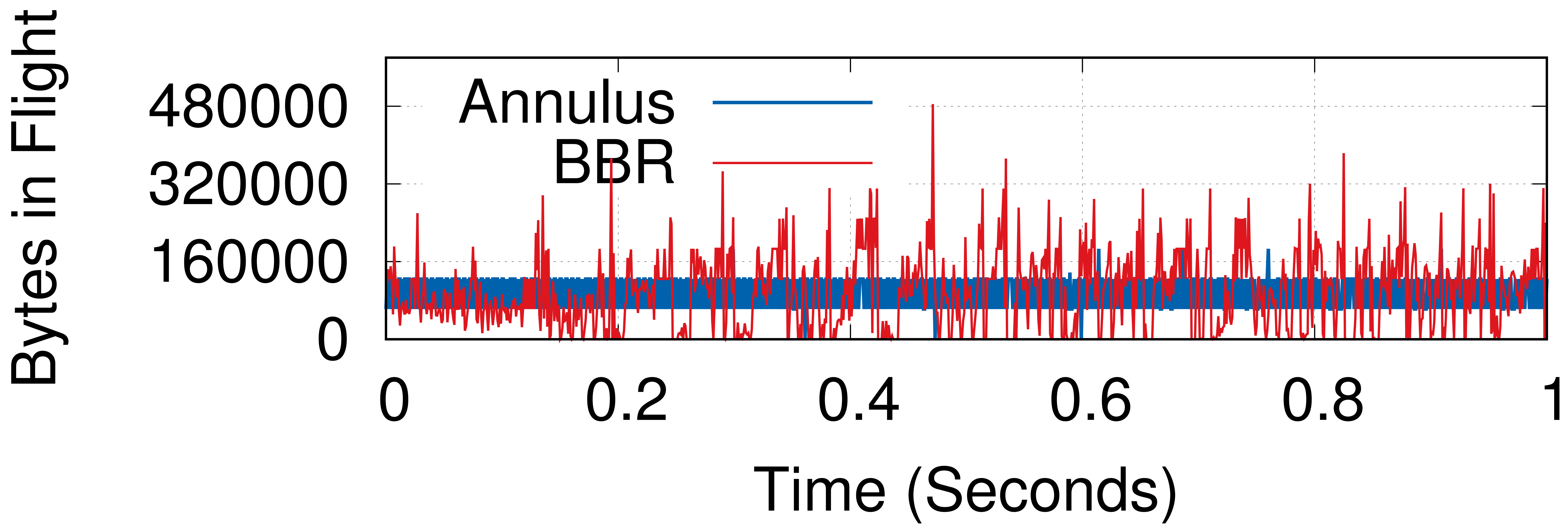


Tail RPC Completion Time

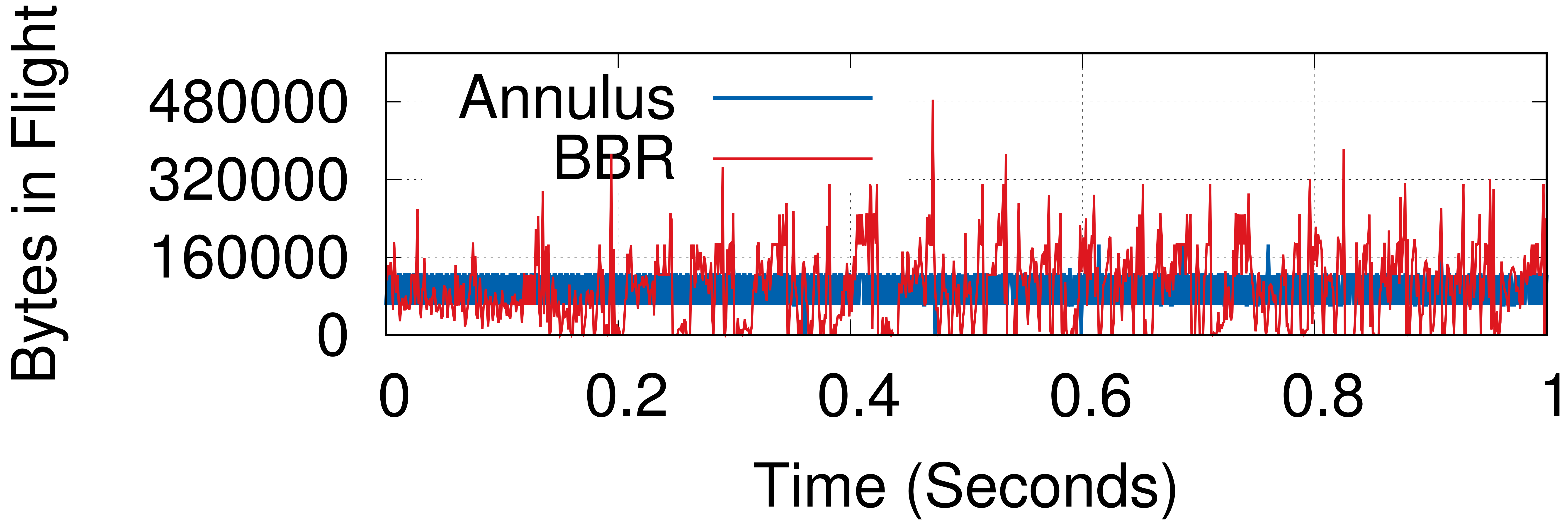


Annulus reduces tail RPC latency by 40% at 50% load

Impact of Annulus on WAN Traffic

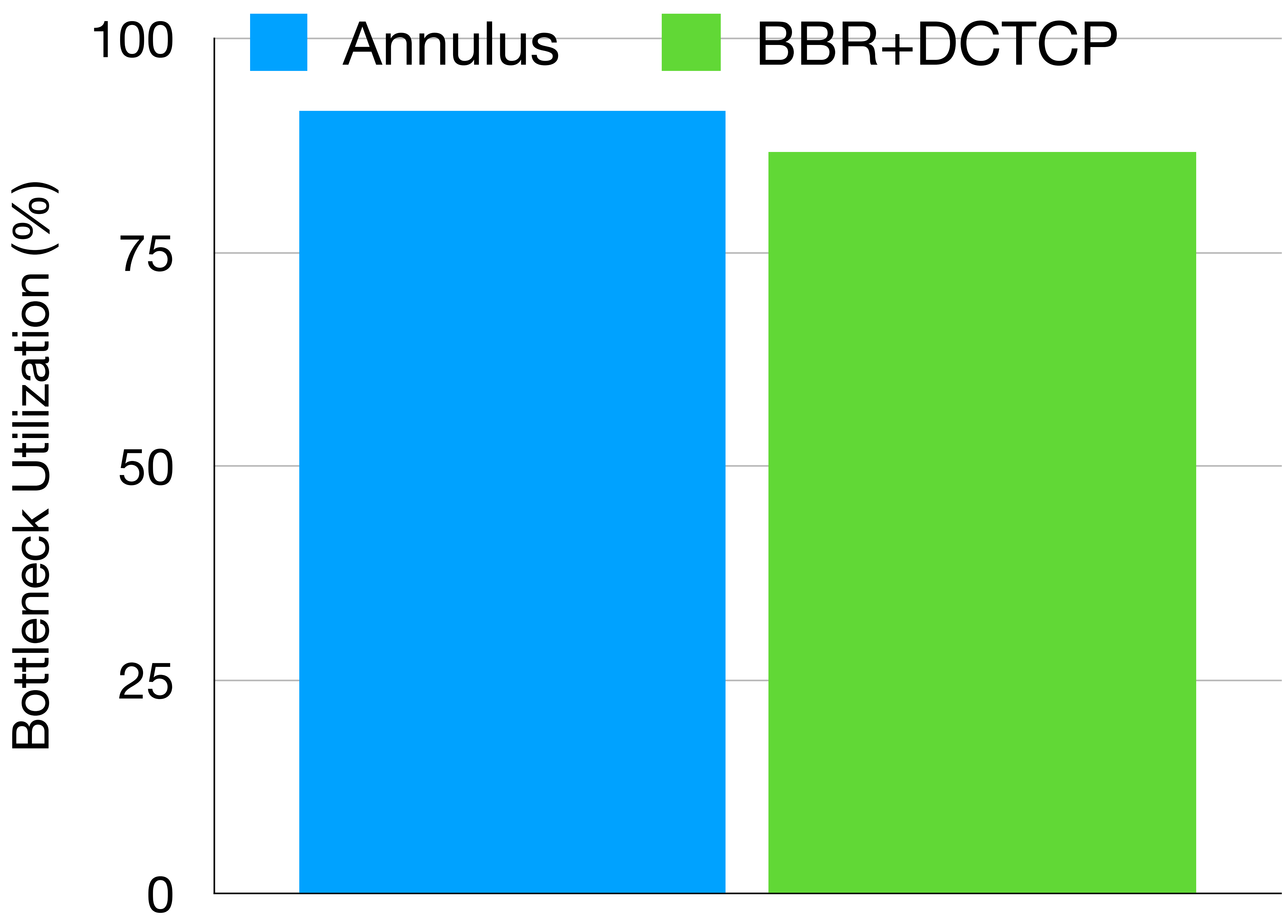


Impact of Annulus on WAN Traffic



Annulus results in less bursty WAN behavior when contending with LAN

Bottleneck Utilization



Stability of Annulus behavior improves utilization by 5%

Conclusions

Conclusions

- A new problem in datacenter congestion control arises when high bandwidth WAN traffic competes with datacenter traffic

Conclusions

- A new problem in datacenter congestion control arises when high bandwidth WAN traffic competes with datacenter traffic
- Annulus makes the case for developing better direct signals that reduce the reaction time and improve the performance of WAN traffic when handling congestion inside the datacenter network

Conclusions

- A new problem in datacenter congestion control arises when high bandwidth WAN traffic competes with datacenter traffic
- Annulus makes the case for developing better direct signals that reduce the reaction time and improve the performance of WAN traffic when handling congestion inside the datacenter network
- Multi-control loop algorithms can help address scenarios where the path has significantly different types of bottlenecks