Network Performance and Measurements

Jorge Crichigno
College of Engineering and Computing
University of South Carolina
http://ce.sc.edu/cyberinfra

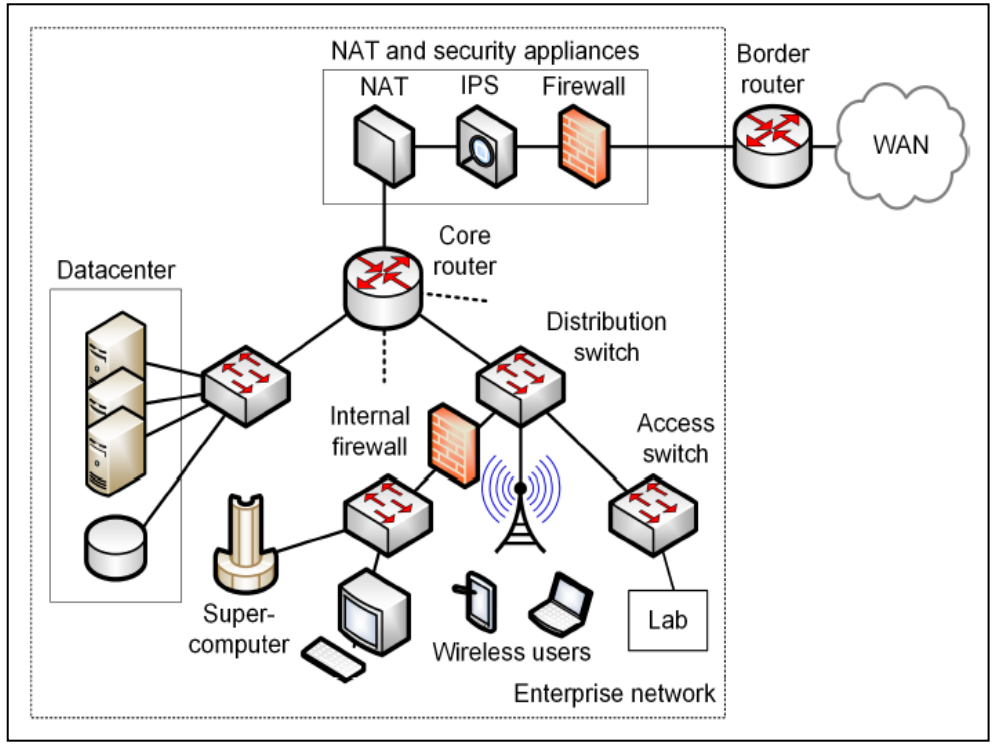
NSF Workshop - Navajo Technical University
Arizona State University
January 30, 2023

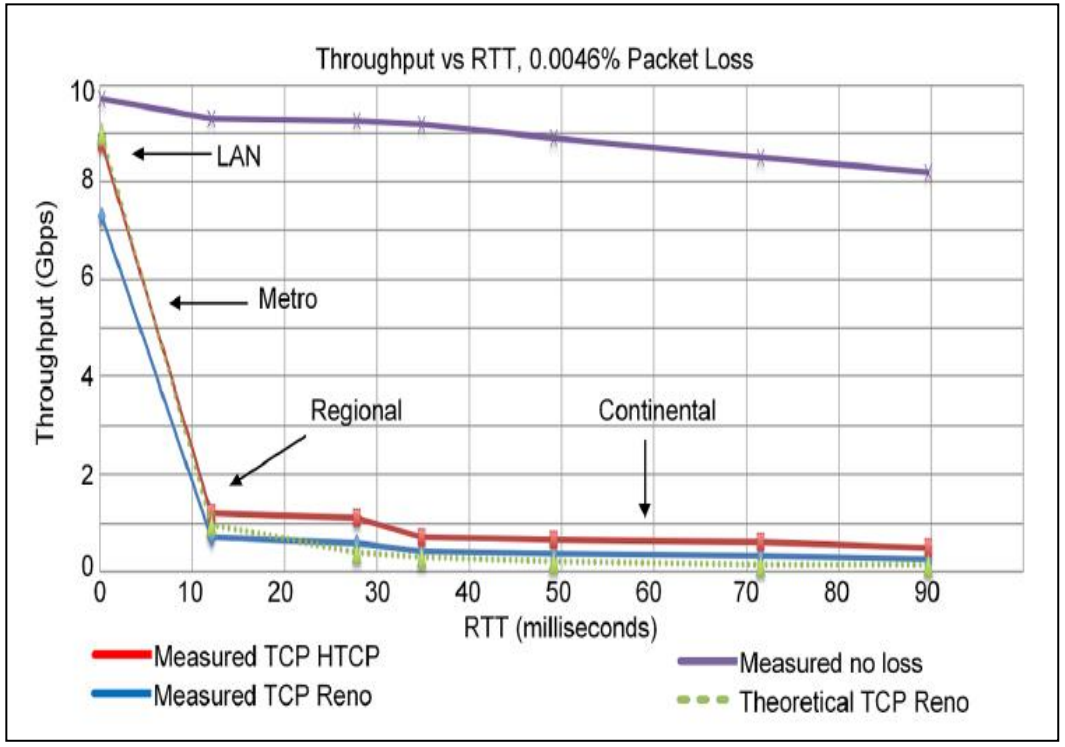
Agenda

- Review – Science DMZ
- Network performance and measurements
  - Importance
  - Examples
- Lessons learned and observations – FABRIC
- Trends
  - Programmable switch ASICs
  - Fine-grained measurements
- Opportunities and challenges

Review – Science DMZ

- Security appliances (IPS, firewalls, etc.) are CPU-intensive
- Small-buffer routers and switches are incapable of absorbing traffic bursts
- Even a small packet loss rate reduces throughput
- Transfers of big science data may last days or even weeks

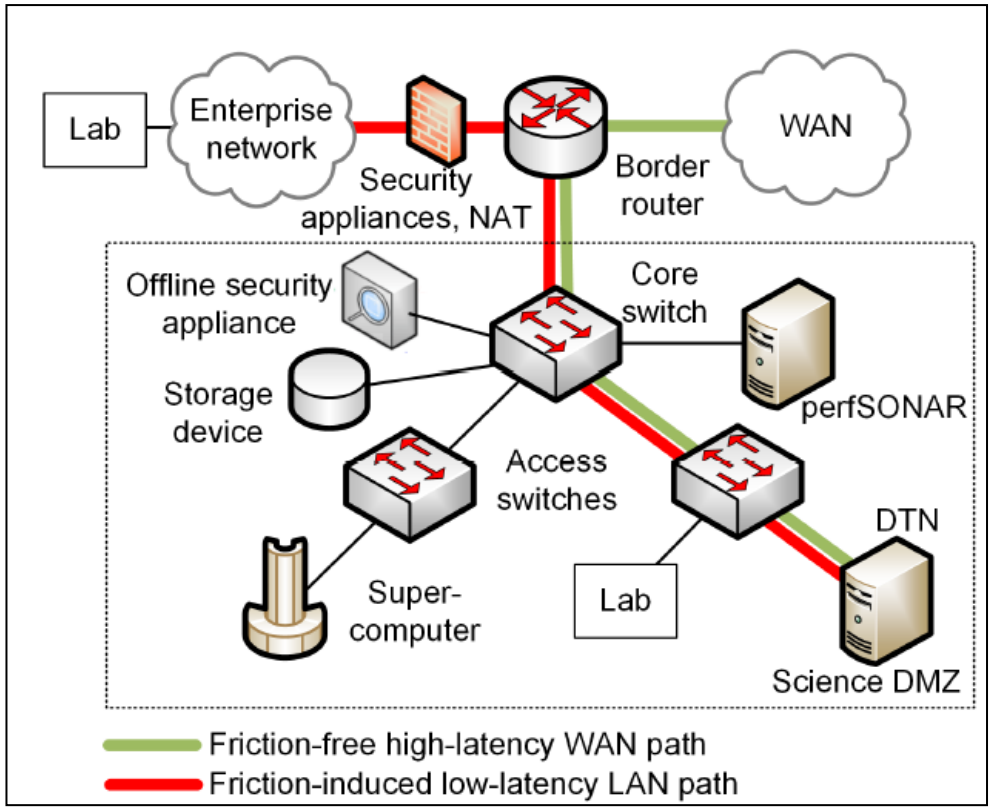




[1]E. Dart, L. Rotman, B. Tierney, M. Hester, J. Zurawski, "The science dmz: a network design pattern for data-intensive science," I*nternational Conference on High Performance Computing, Networking, Storage and Analysis*, Nov. 2013.

Review – Science DMZ

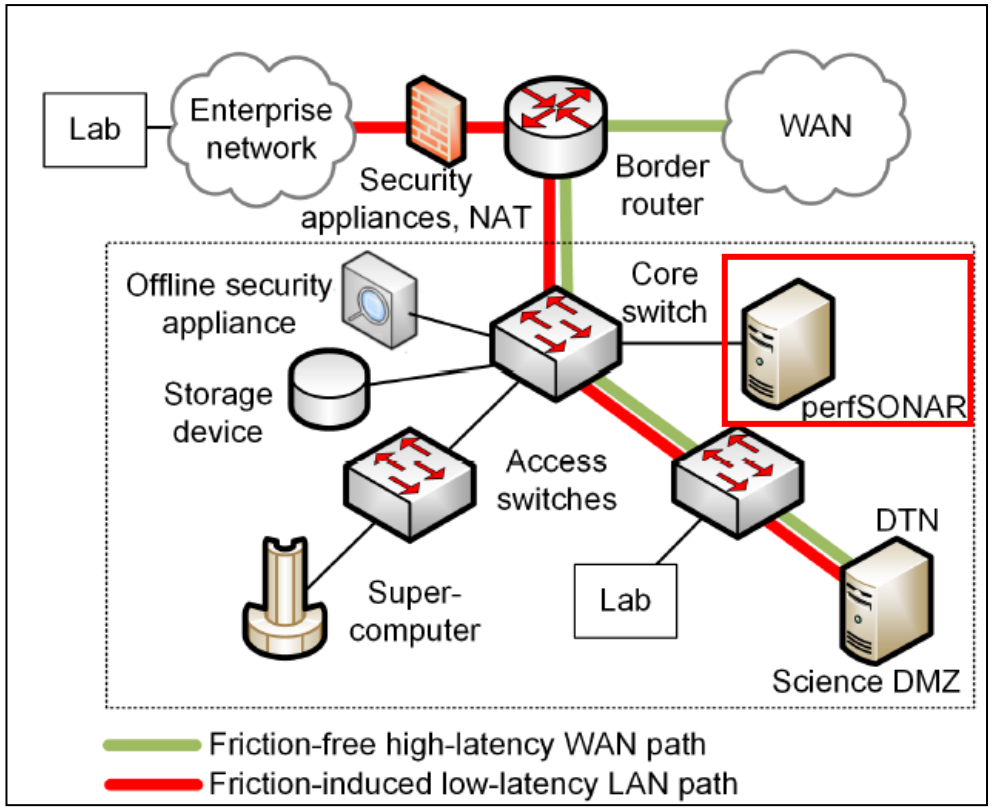
- The Science DMZ is a network designed for big science data[1]
- Main elements
  - High throughput, friction free WAN paths (no inline security appliances)
  - Switches with large buffer size
  - Data Transfer Nodes (DTNs)
  - Security = Access-Control List (ACLs) + offline appliance/s (IDS)
  - End-to-end measurements = perfSONAR



[1]E. Dart, L. Rotman, B. Tierney, M. Hester, J. Zurawski, "The science dmz: a network design pattern for data-intensive science," I*nternational Conference on High Performance Computing, Networking, Storage and Analysis*, Nov. 2013.

Review – Science DMZ

- The Science DMZ is a network designed for big science data[1]
- Main elements
  - High throughput, friction free WAN paths (no inline security appliances)
  - Switches with large buffer size
  - Data Transfer Nodes (DTNs)
  - Security = Access-Control List (ACLs) + offline appliance/s (IDS)
  - End-to-end measurements = perfSONAR



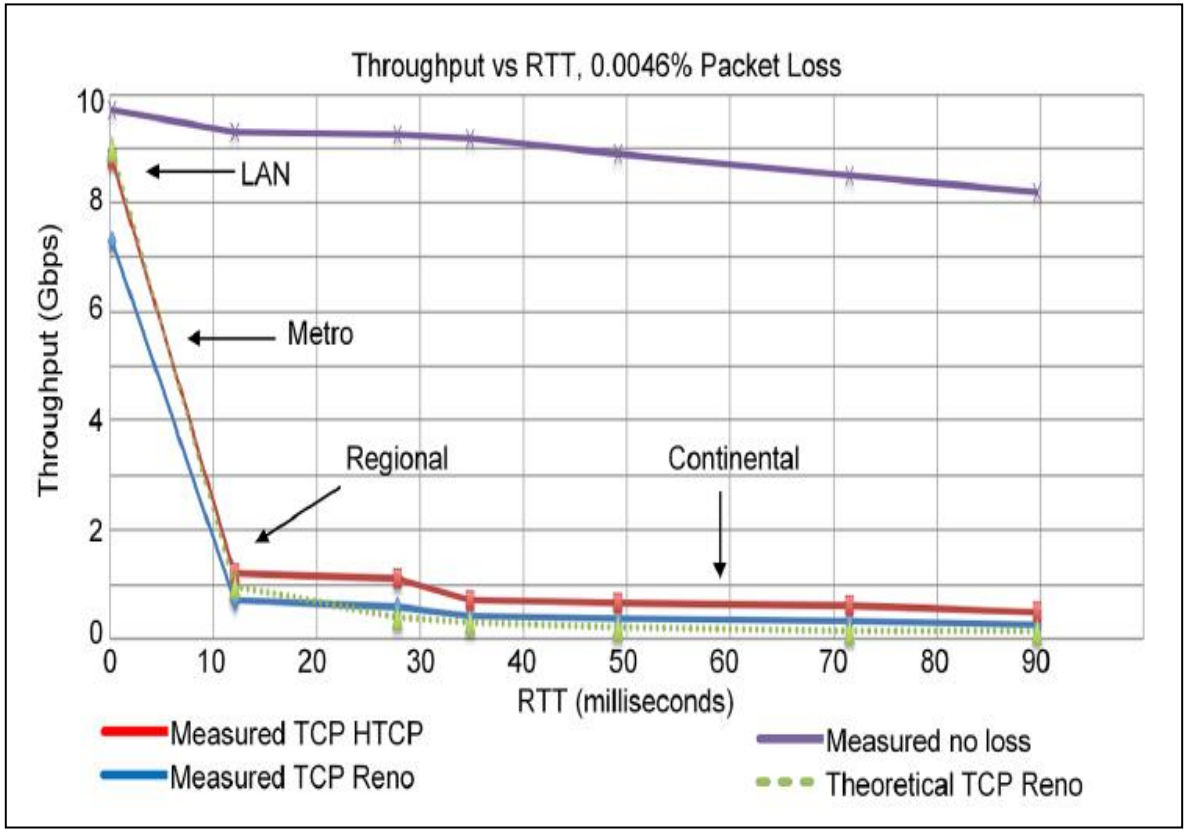
[1]E. Dart, L. Rotman, B. Tierney, M. Hester, J. Zurawski, "The science dmz: a network design pattern for data-intensive science," I*nternational Conference on High Performance Computing, Networking, Storage and Analysis*, Nov. 2013.

Network Performance and Measurements

- The following metrics are often used to measure network performance:
  - Packet loss
  - Throughput (e.g., “how much can I get out of the network”)
  - Latency / round-trip time (RTT)

Network Performance and Measurements

- Soft failures are those failures that do not disrupt connectivity, but may prevent high performance
  - Critical in high-throughput high-latency networks

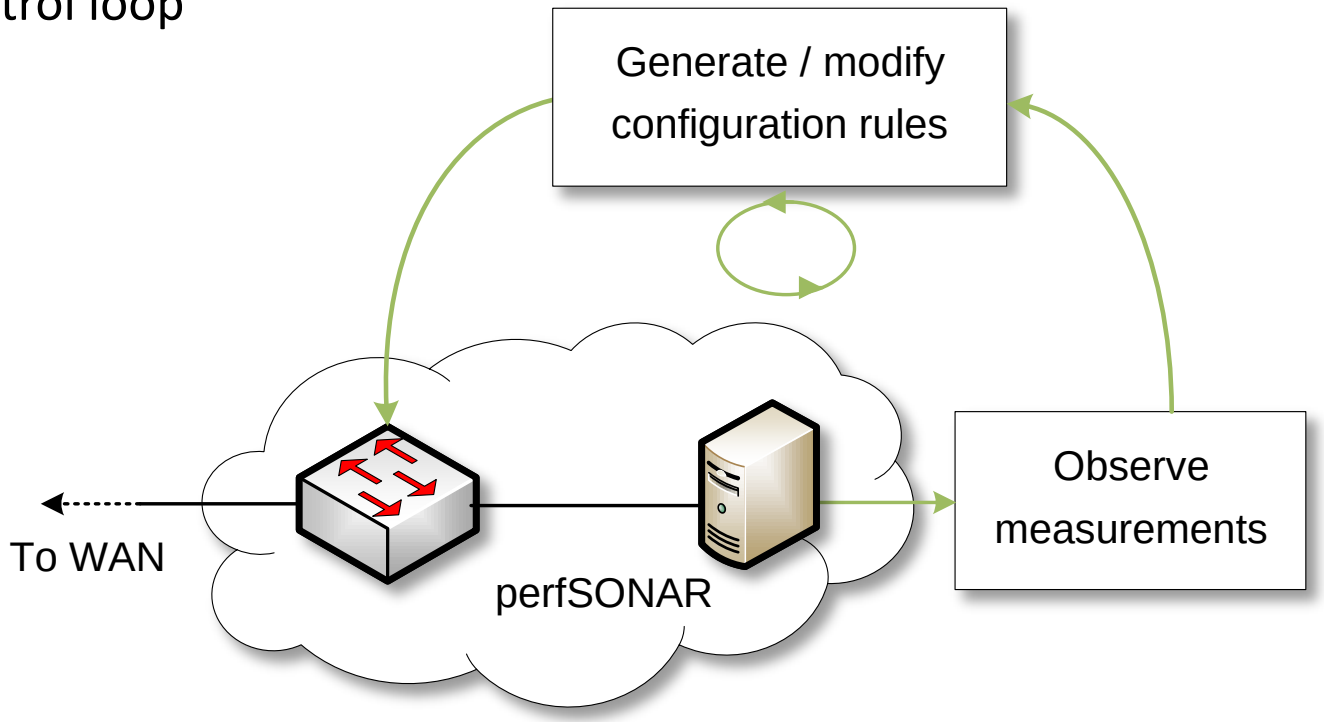


Network Performance and Measurements

- perfSONAR provides information that reflects the state of the network, in a multi-domain basis
- perfSONAR = Measurement Middleware

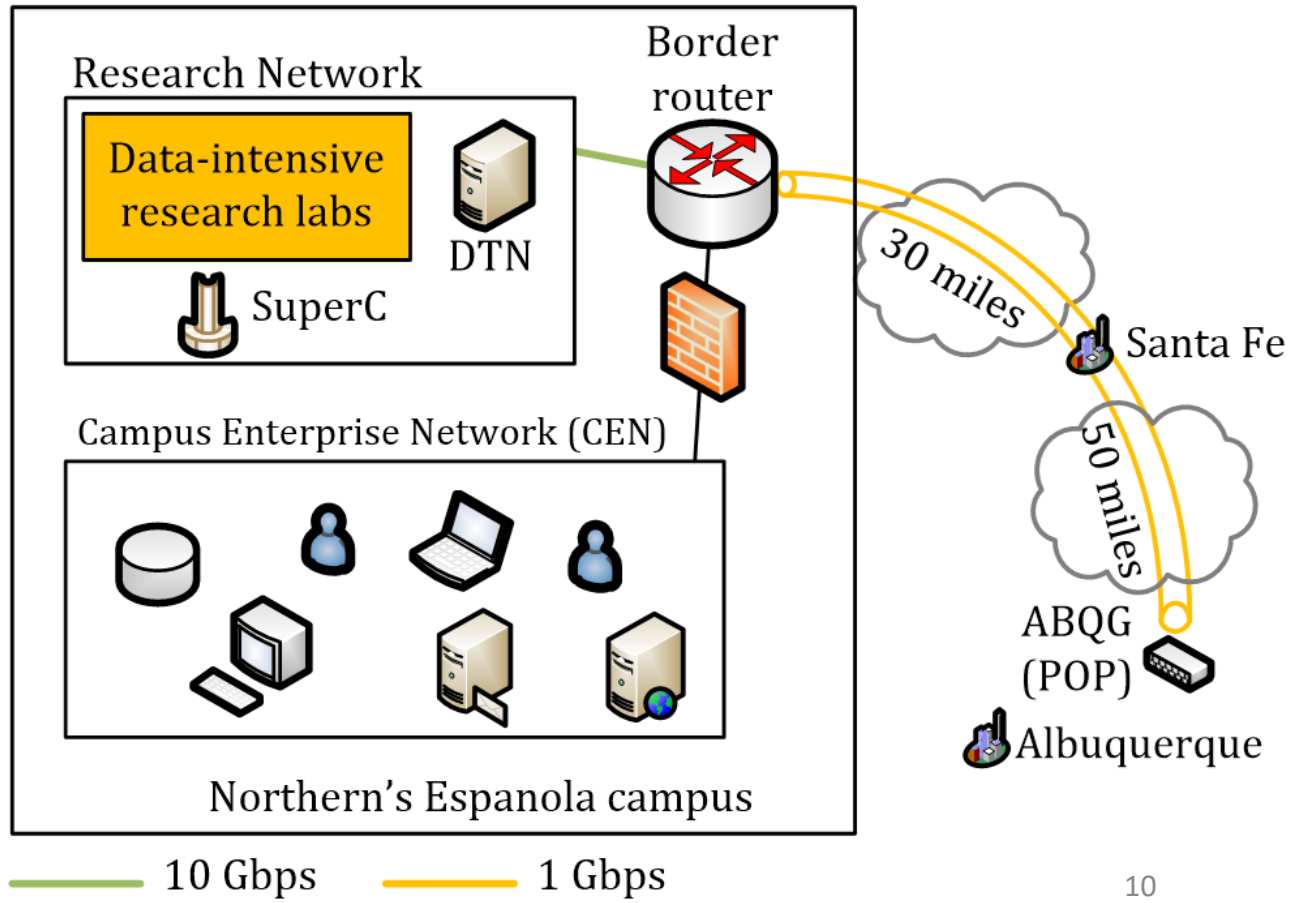
Network Performance and Measurements

- Measurements allow us to:
  - Learn the performance limits (e.g., maximum throughput, minimum delay)
  - Discover the current state of the network
  - Modify configuration rules
  - Implement a control loop



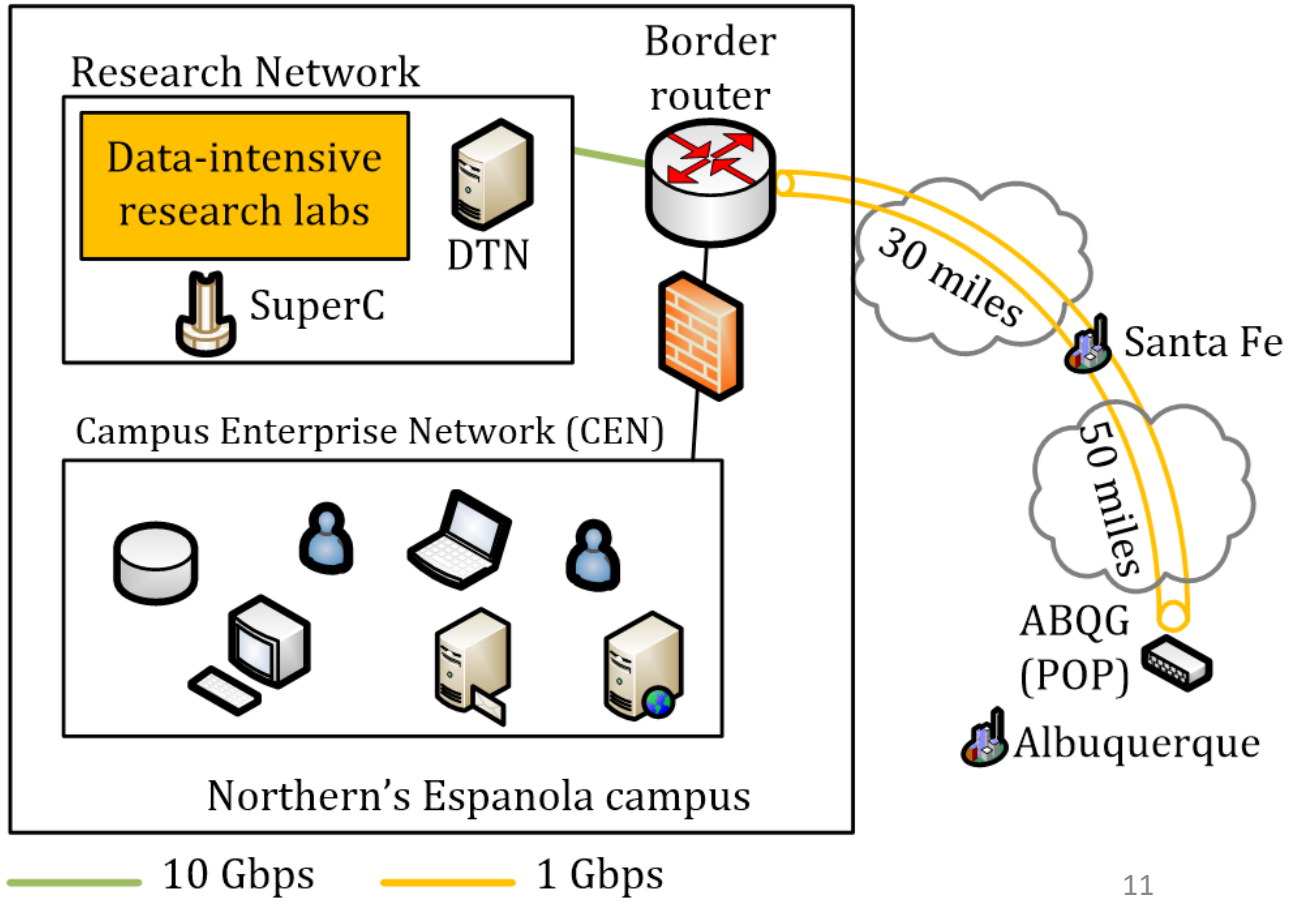
Network Performance and Measurements

- Example 1: connection from Northern New Mexico College (Espanola, NM) to Albuquerque GigaPop (Albuquerque, NM)



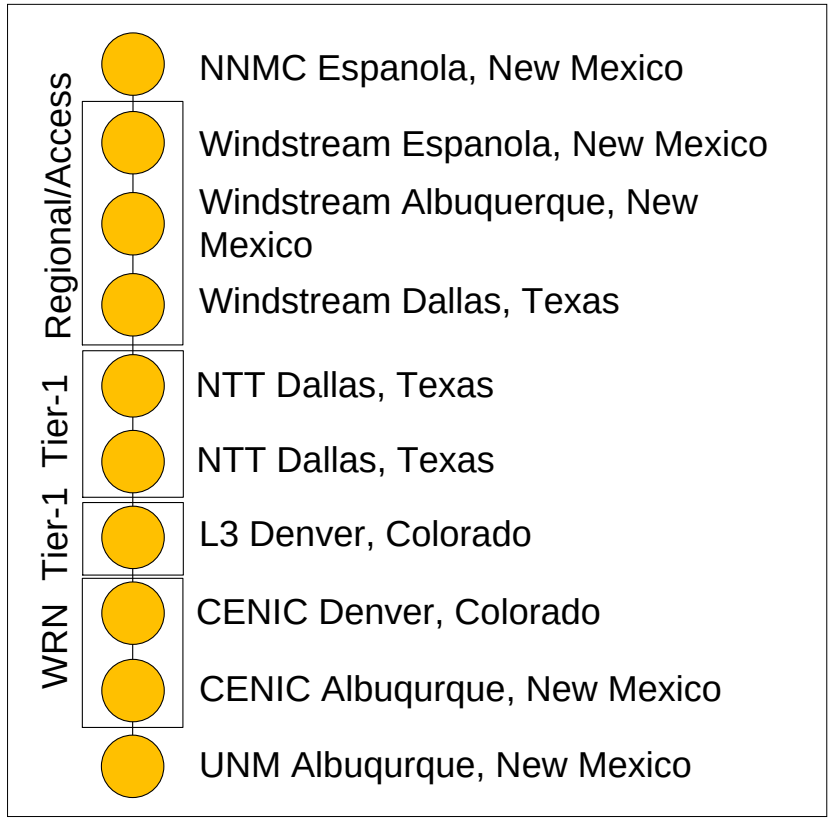
Network Performance and Measurements

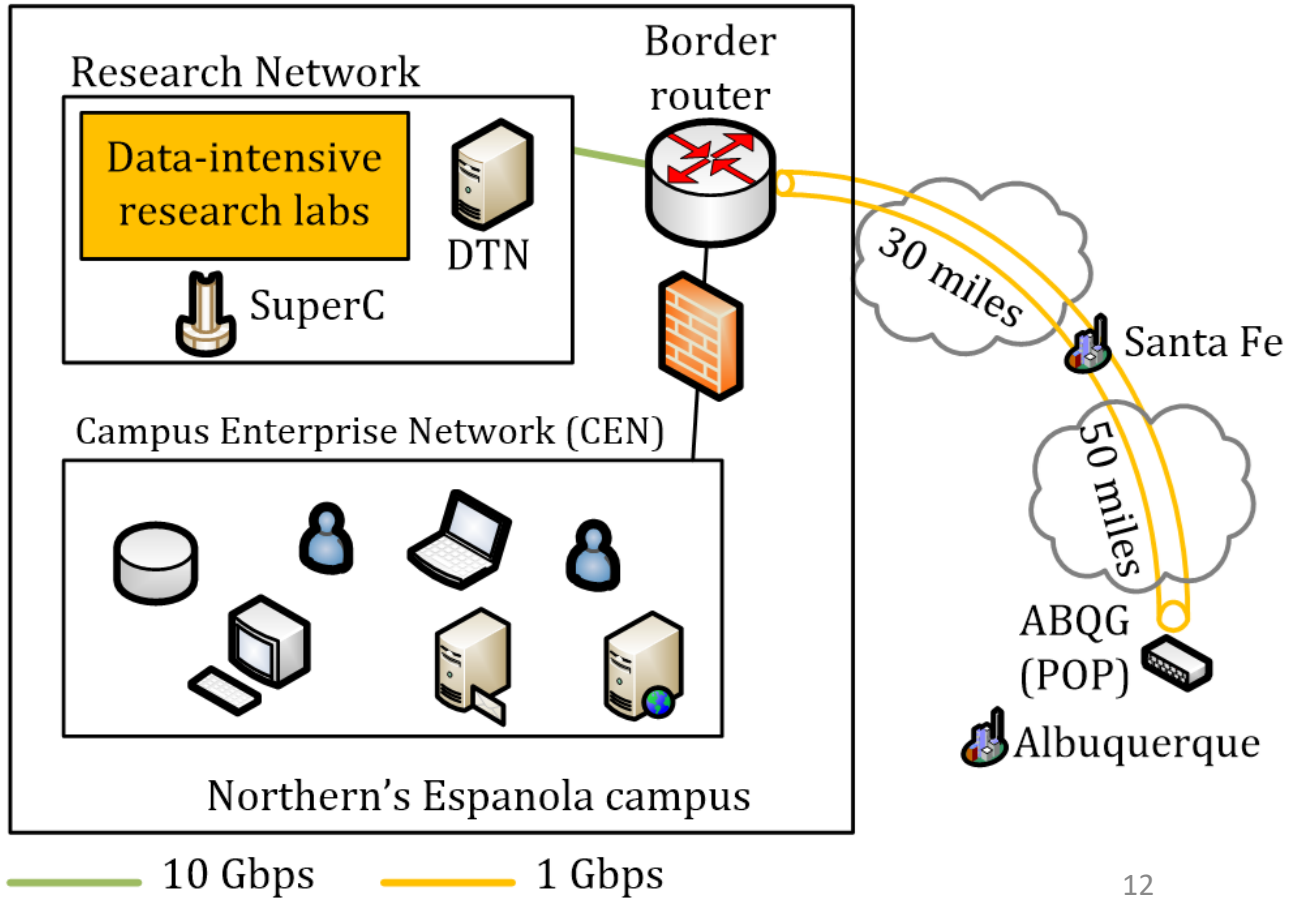
- Example 1: connection from Northern New Mexico College (Espanola, NM) to Albuquerque GigaPop (Albuquerque, NM)
- perfSONAR measured a throughput of ~50 Mbps, latency > 30ms



Network Performance and Measurements

- Example 1: connection from Northern New Mexico College (Espanola, NM) to Albuquerque GigaPop (Albuquerque, NM)
- perfSONAR measured a throughput of ~50 Mbps, latency > 30ms





Network Performance and Measurements

- Example 2: 10 Gbps WAN, 50ms RTT, TCP CUBIC (default TCP algorithm in Linux)

Network Performance and Measurements

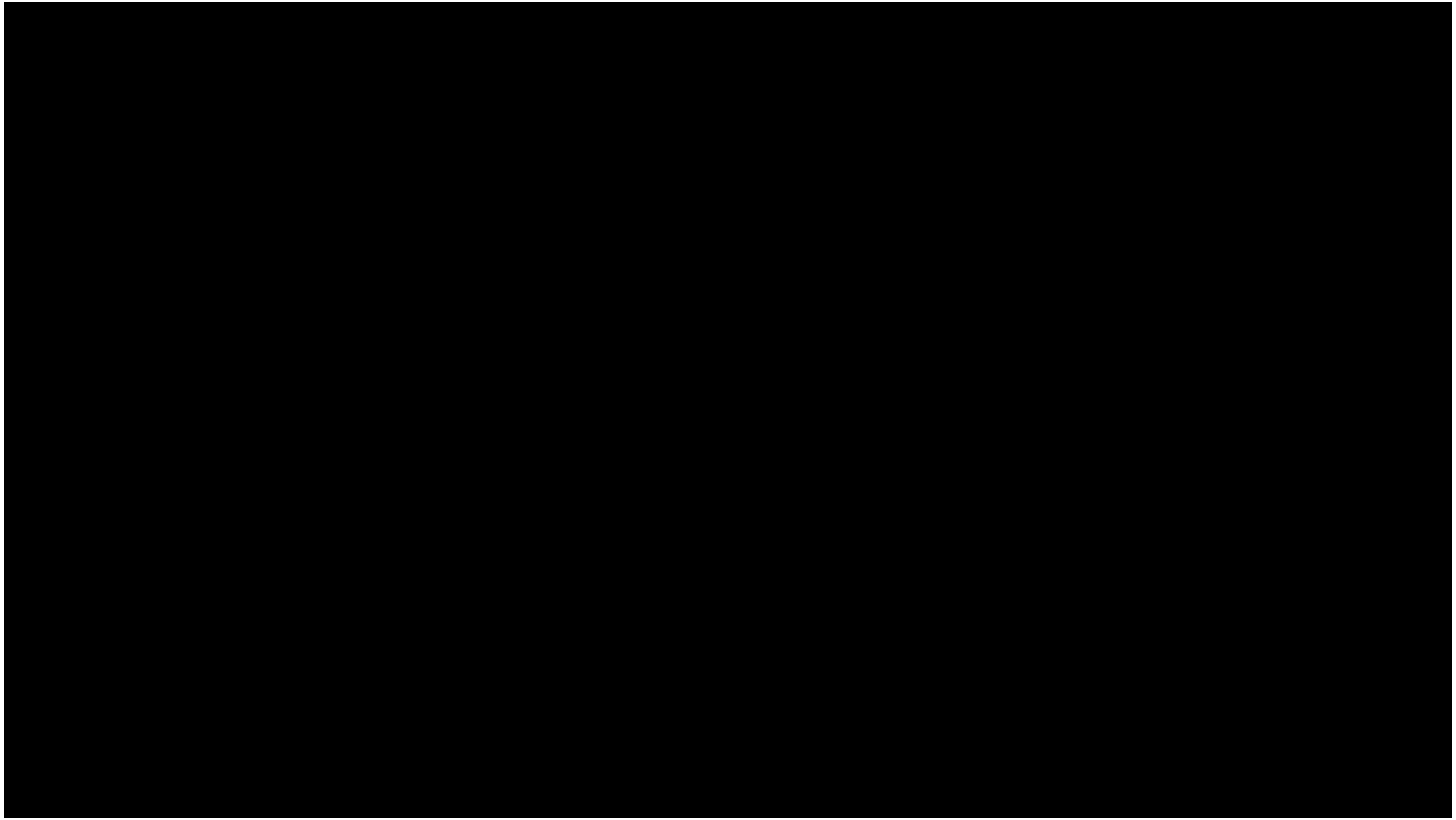
- Example 2: 10 Gbps WAN, 50ms RTT, TCP CUBIC (default TCP algorithm in Linux)

Network Performance and Measurements

- Example 2: 10 Gbps WAN, 50ms RTT, TCP BBR

Network Performance and Measurements

- Example 2: 10 Gbps WAN, 50ms RTT, TCP BBR



Network Performance and Measurements

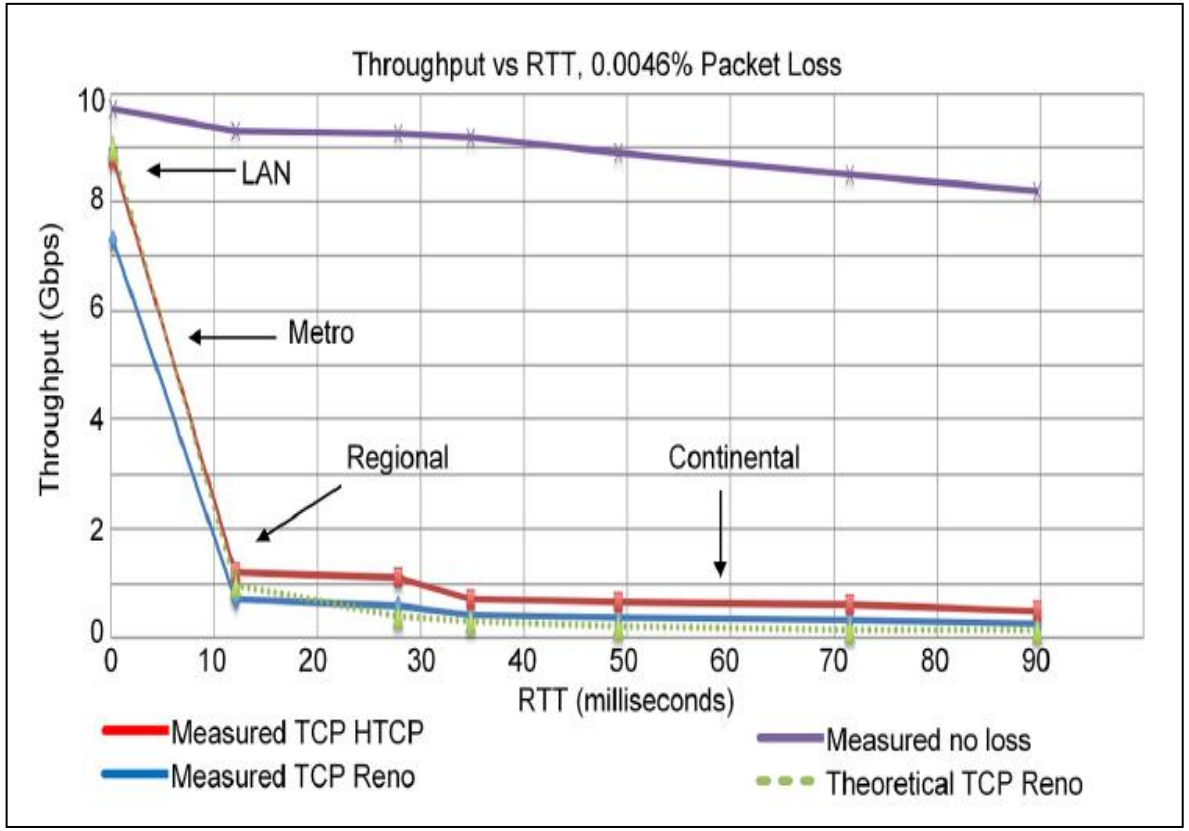
- Example 2: 60ms RTT path from Columbia (SC) to Los Angeles (CA)

Lessons Learned

- Network measurements are essential for performance
  - Classic monitoring systems are good at alerting hard failures (e.g., SNMP, NOC tools indicate when hardware ceases to function, a power failure occurs)
  - perfSONAR helps identify soft failures
- There is training periodically offered by the CI community
  - Are there new alternatives to the Science DMZ model?

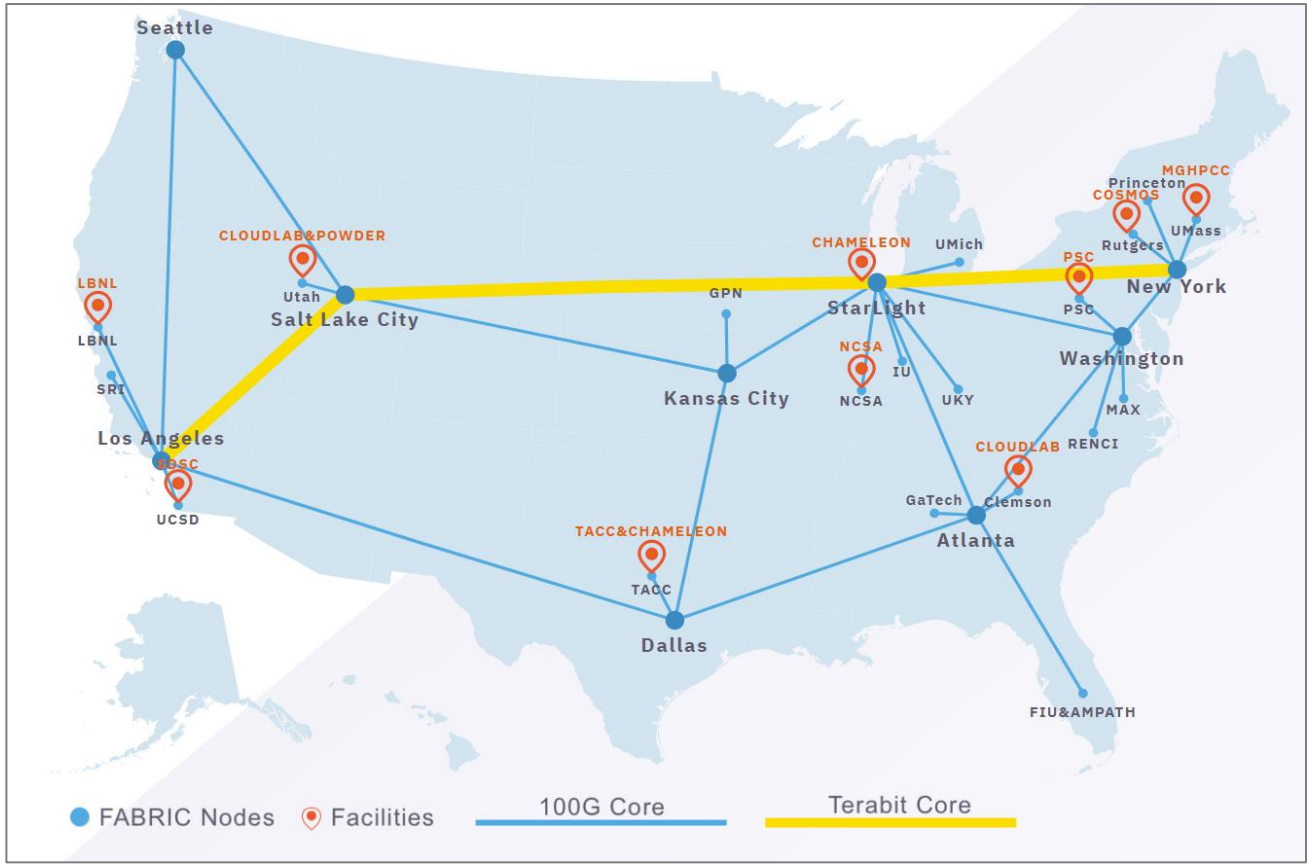
Lessons Learned

- Measurements illustrated in the original Science DMZ paper



Lessons Learned

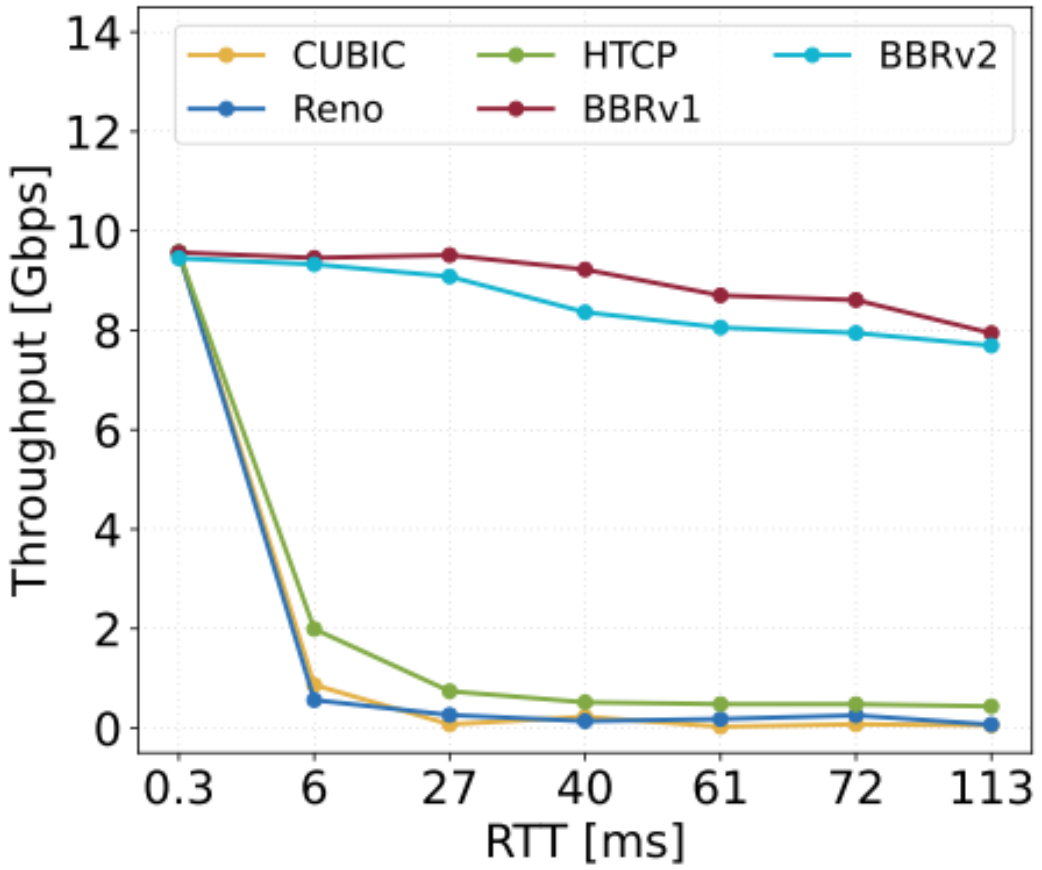
- Some measurements on FABRIC

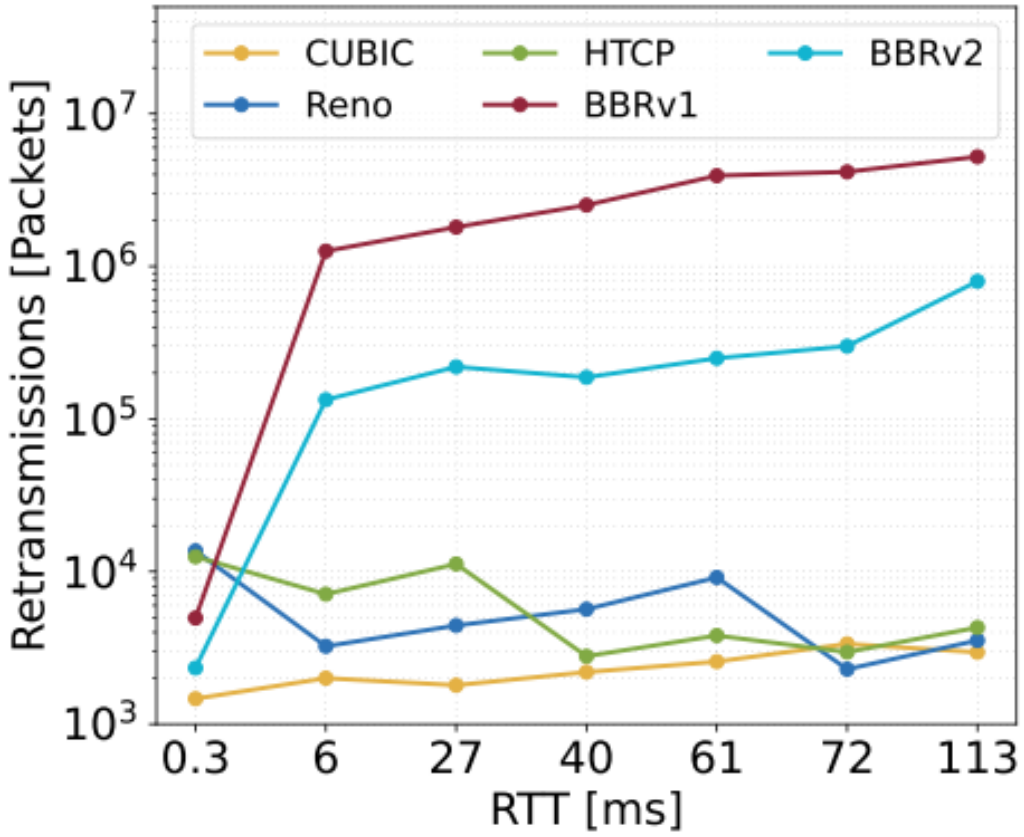


Site 1	Site 2	RTT
TACC (TX)	TACC (TX)	0.3ms
DALL (TX)	TACC (TX)	6ms
DALL (TX)	WASH (DC)	27ms
SALT (UT)	FIU (FL)	44ms
GPN (MO)	DALL (TX)	61ms
UTAH (UT)	WASH (DC)	72ms
GPN (MO)	FIU (FL)	113ms

Lessons Learned

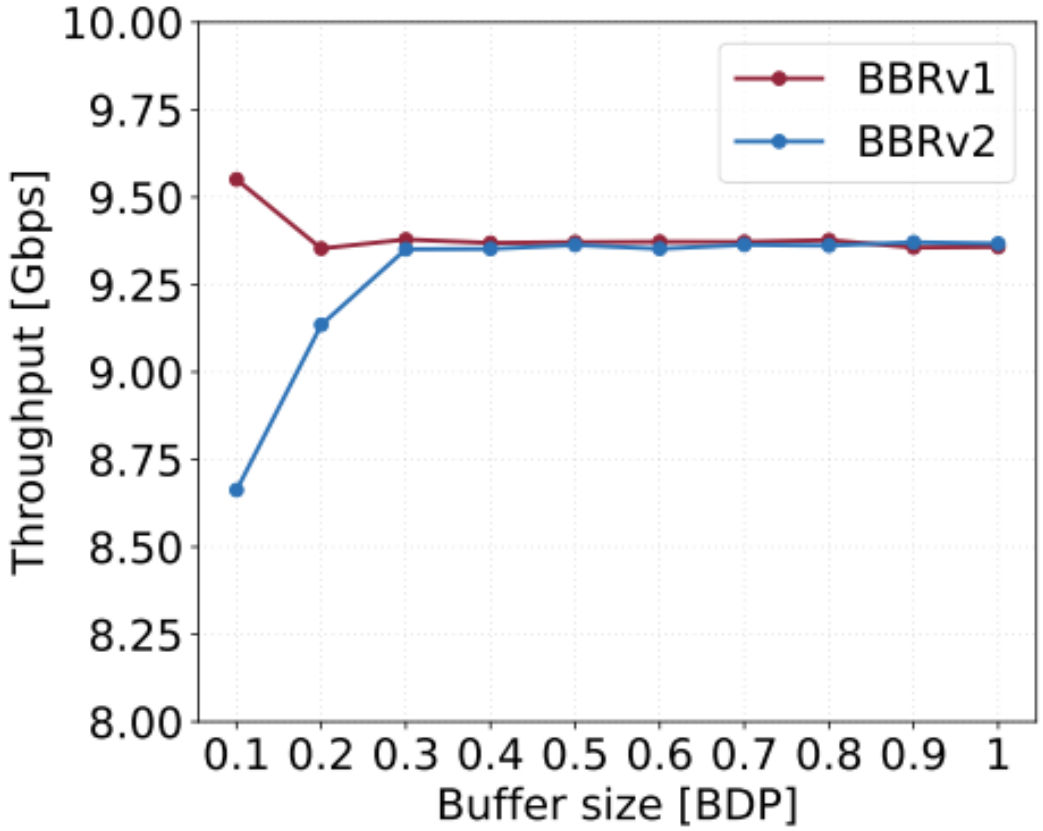
- Performance measurements for a single flow, 0.0046% packet loss rate

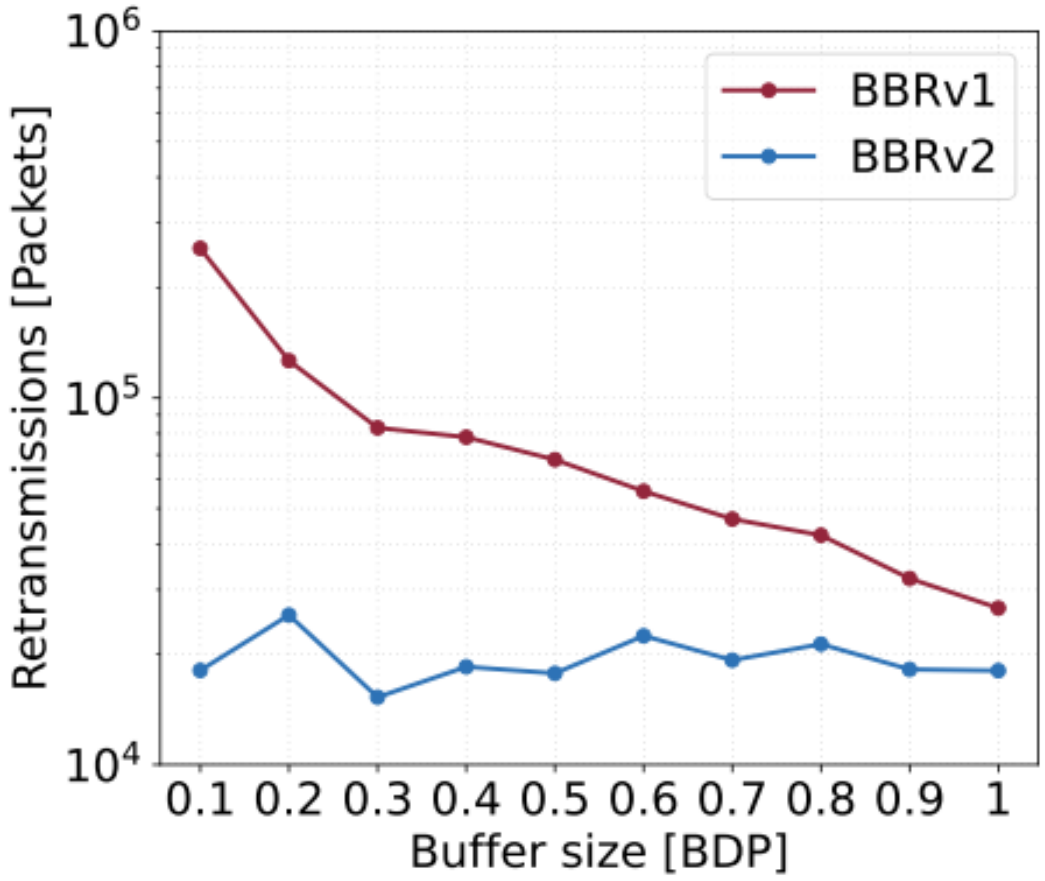




Lessons Learned

- Performance measurements for a single flow, 0.0046% packet loss rate

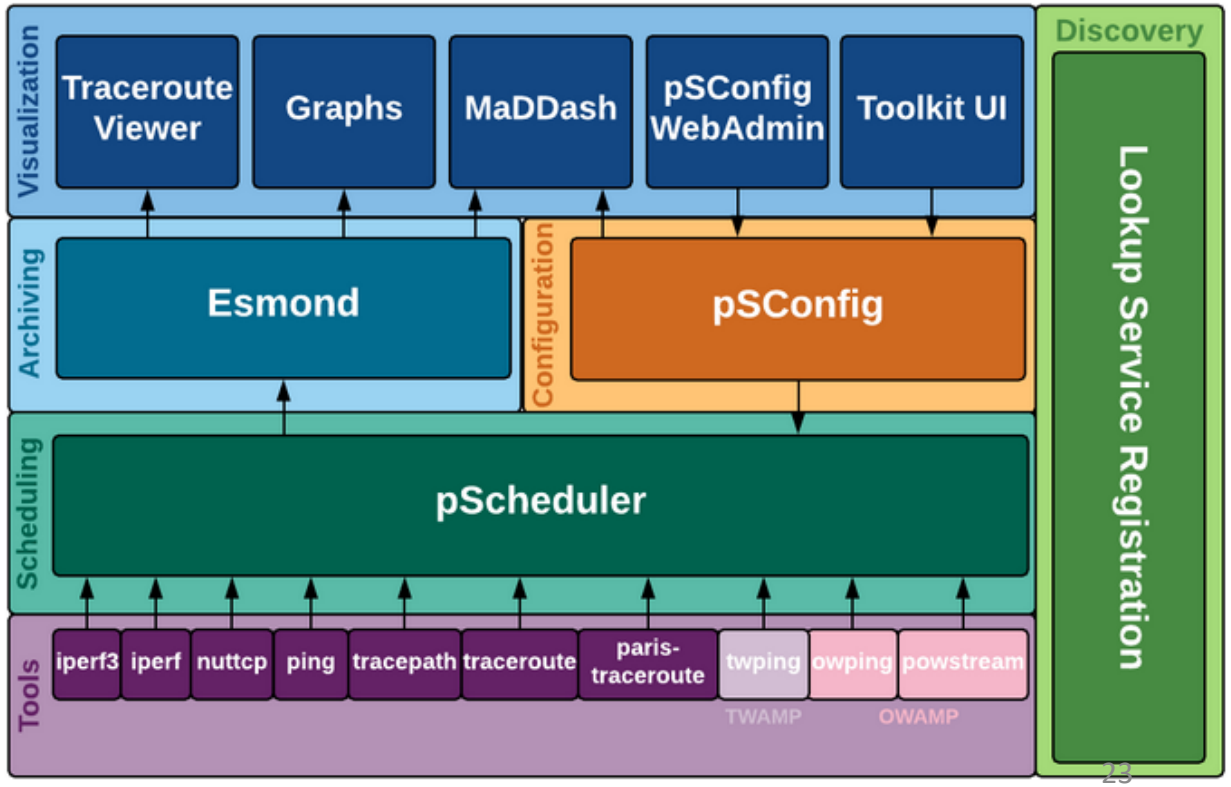




Trends in Network Performance and Measurements

- The networking community has been using the same measurement tools for years
  - Ping: initially released in 1983
  - Traceroute: initially released in 1987
  - SNMP: the original RFC appeared in 1988

perfSONAR layers

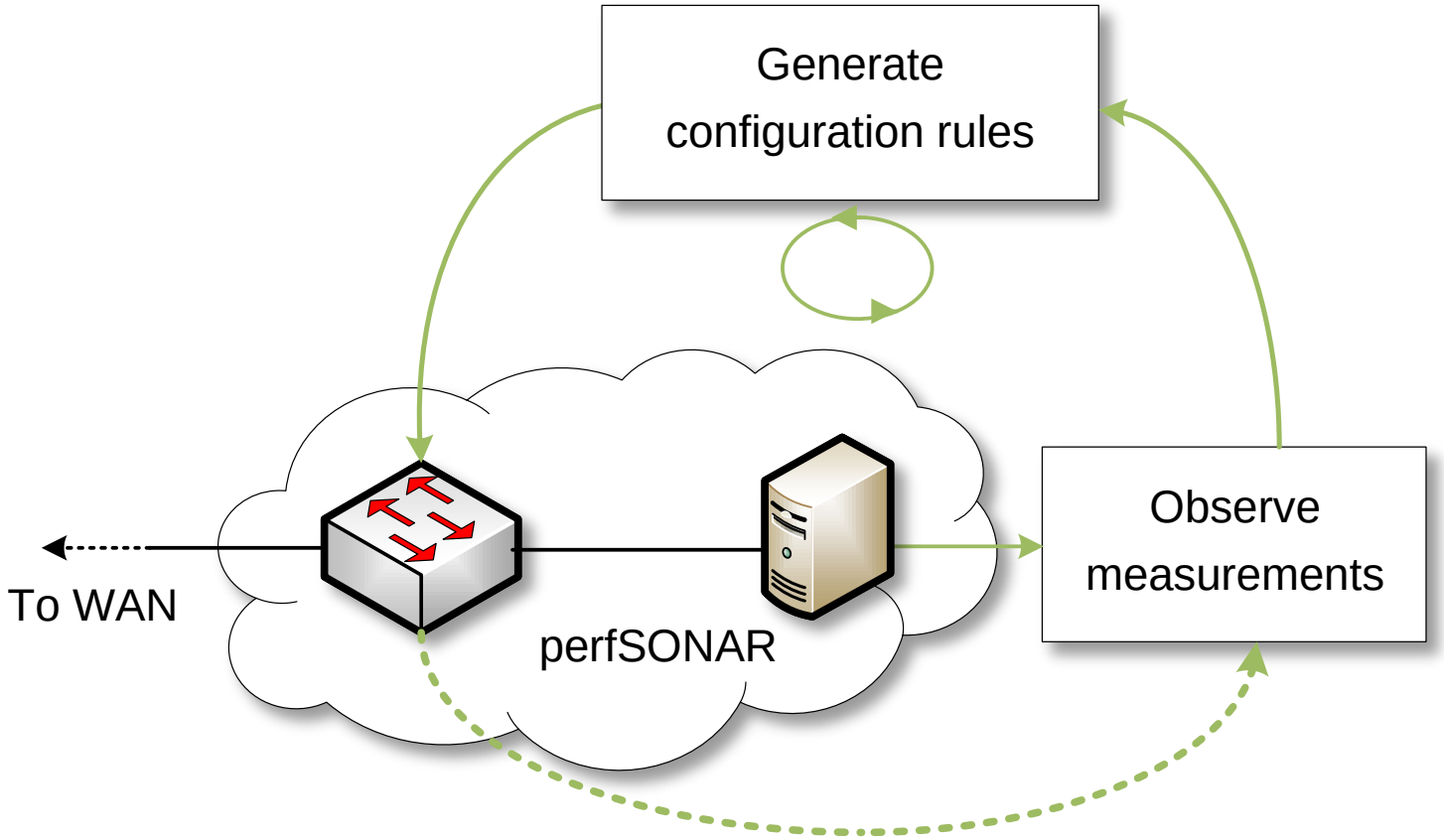


Trends in Network Performance and Measurements

- The networking community has been using the same measurement tools for years
  - Ping: initially released in 1983
  - Traceroute: initially released in 1987
  - SNMP: the original RFC appeared in 1988
- They produce measurements over long time periods (coarse-grained measurements)
- Traceroute packets may follow a different path than the data packets

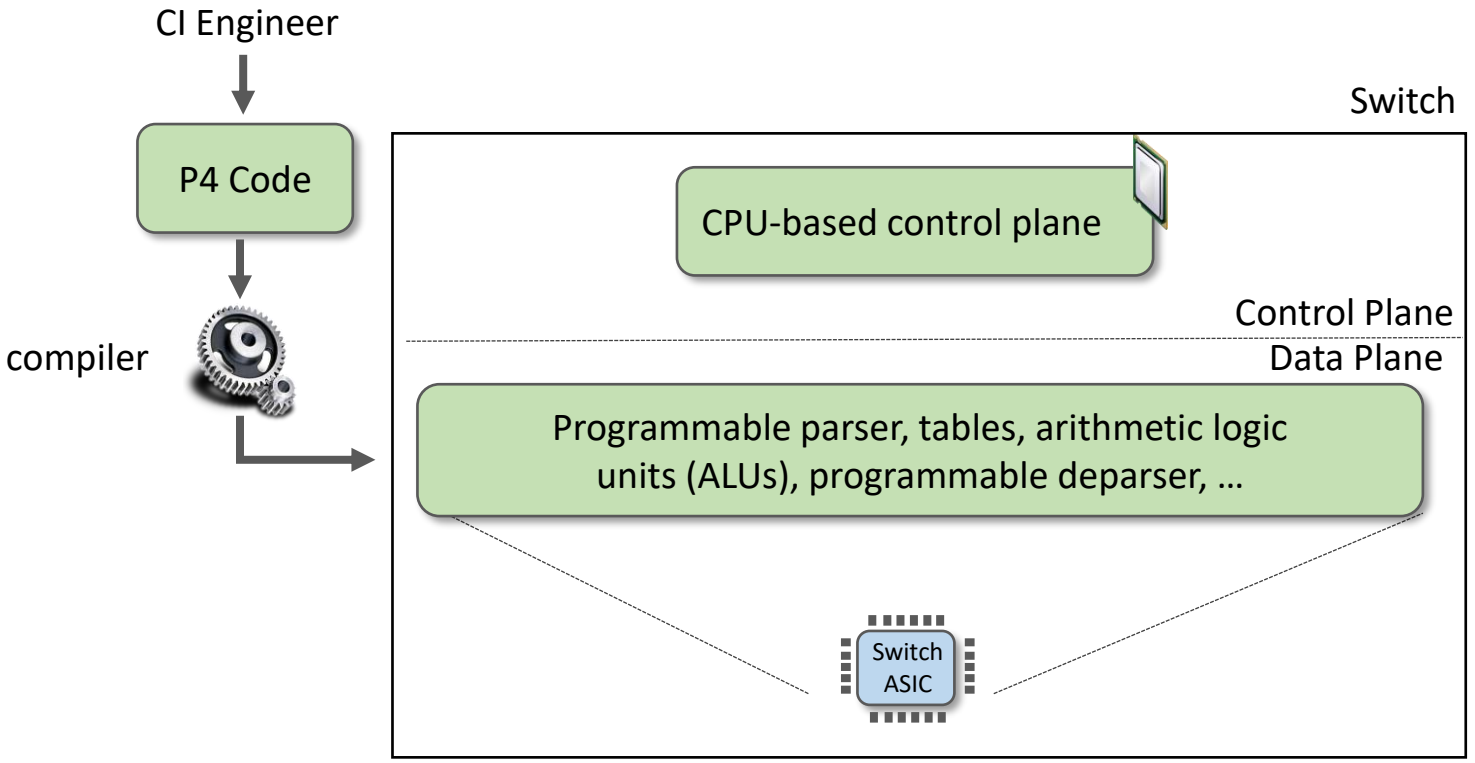
Trends in Network Performance and Measurements

- Can we obtain fine-grained measurements?
- Barrier: the CI engineer is limited by what the equipment vendor implements
  - The switch ASIC is designed with fixed functions (hard-coded) by the chip designer



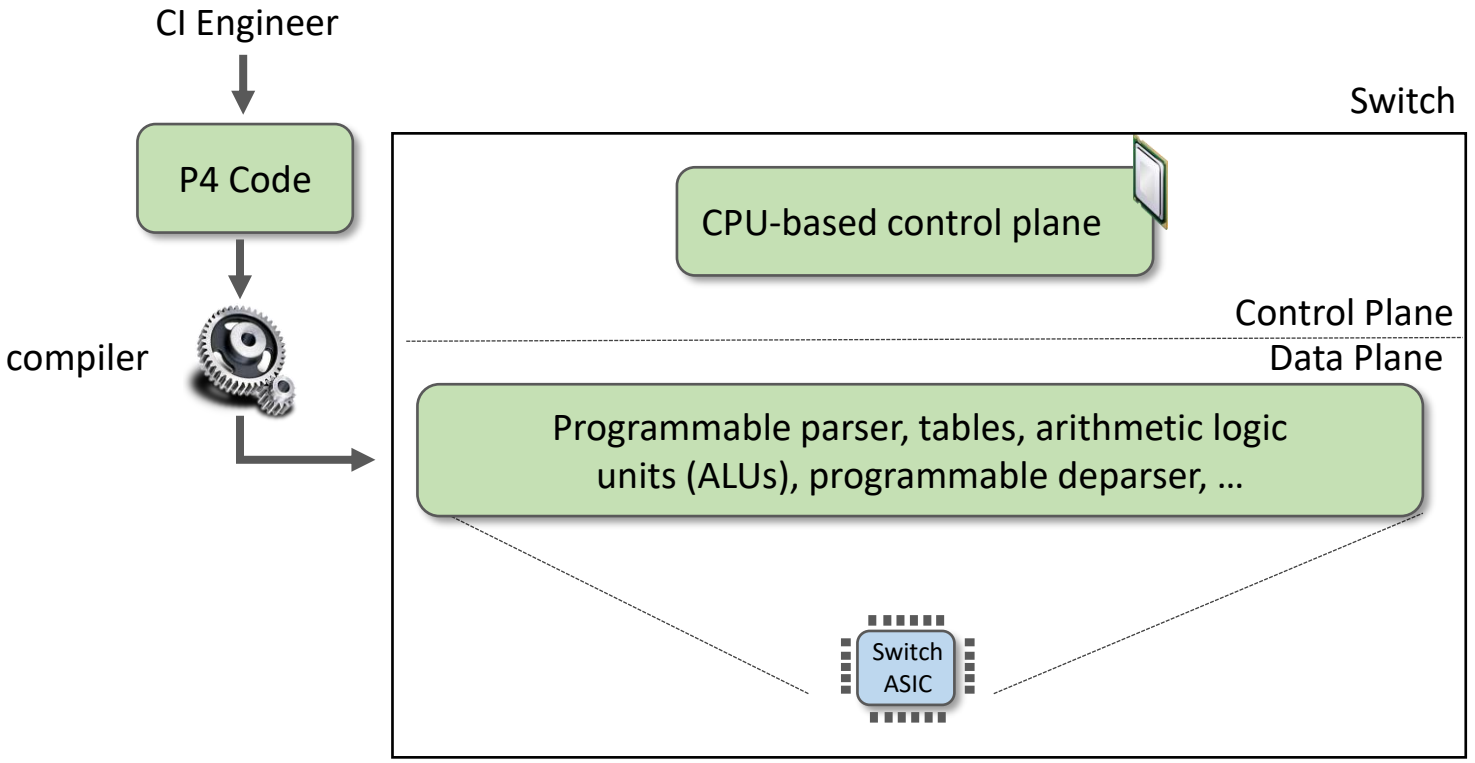
Programmable Switch ASICs

- We now have the technology that permits CI engineers to run customized functions



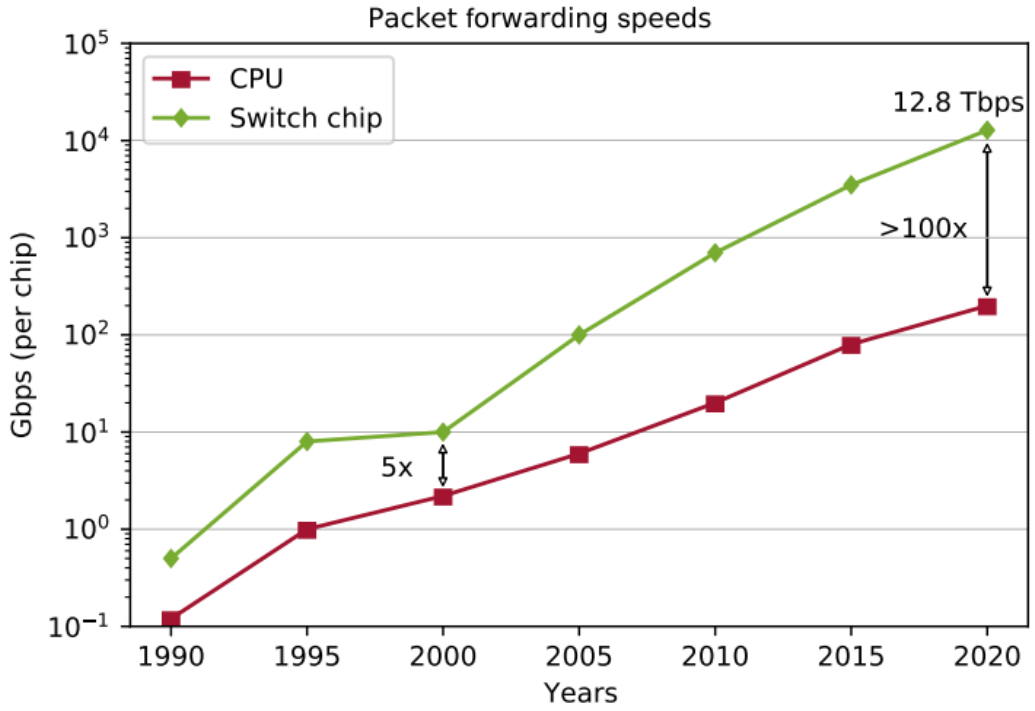
Programmable Switch ASICs

- We now have the technology that permits CI engineers to run customized functions
  - Designed for packet processing operations



Programmable Switch ASICs

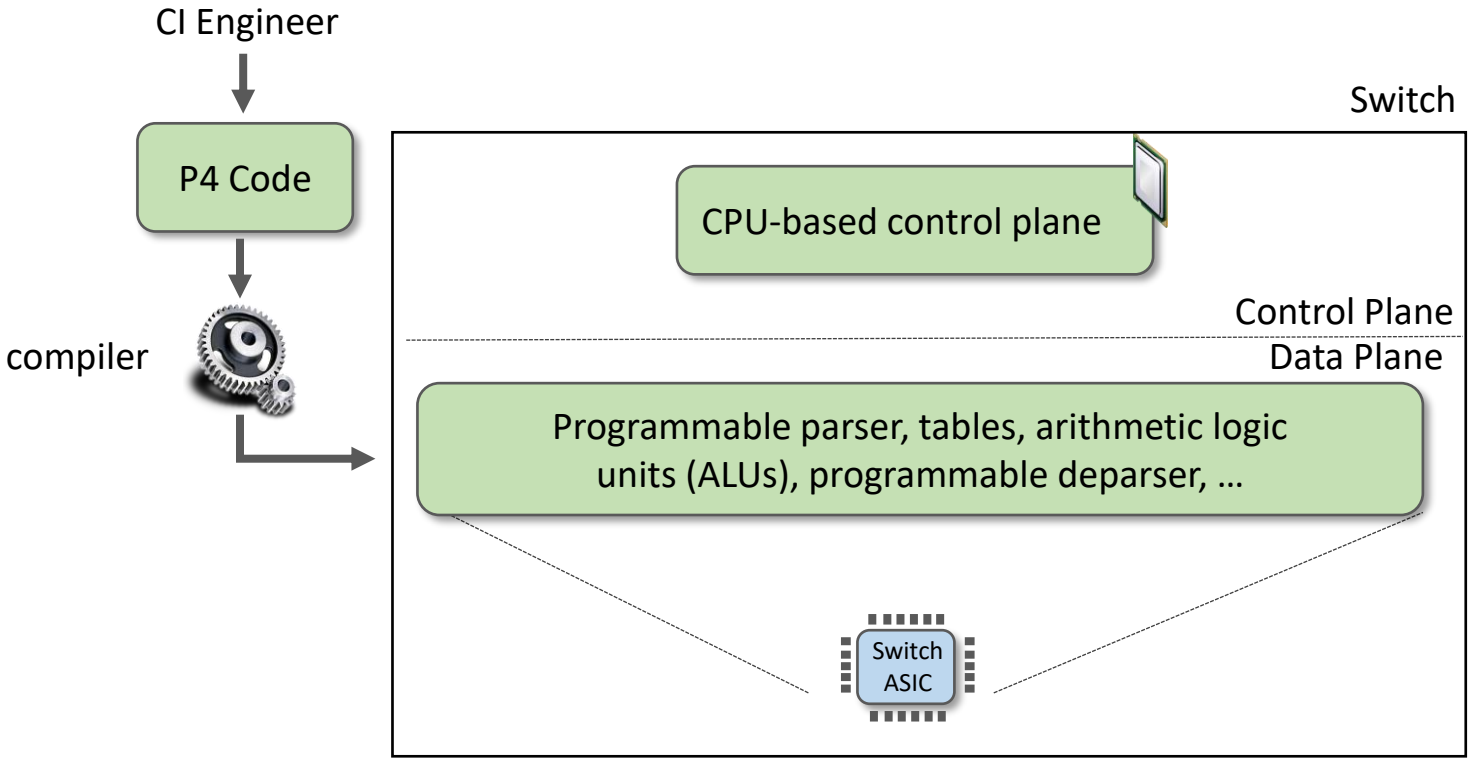
- We now have the technology that permits CI engineers to run customized functions
  - Designed for packet processing operations
  - Much faster than general-purpose CPUs for processing packets



N. McKeown, "Creating an End-to-End Programming Model for Packet Forwarding," Netdev 0x14 Conference 2020, https://www.youtube.com/watch?v=fiBuao6YZl0&t=619s.

Programmable Switch ASICs

- We now have the technology that permits CI engineers to run customized functions
  - Designed for packet processing operations
  - Much faster than general-purpose CPUs for processing packets
  - Limited SRAM memory capacity



Programmable Switch ASICs

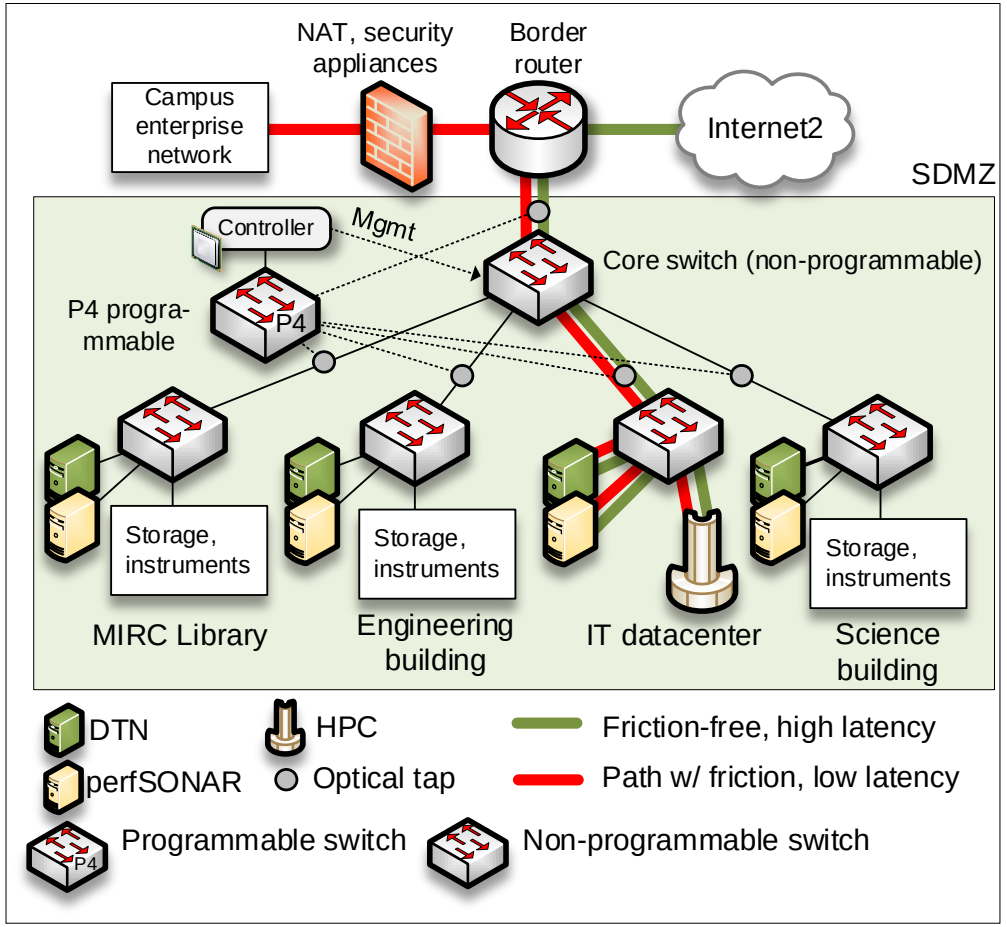
- We can track flows in the data plane
  - Science DMZs do not need to track millions of flows (enough SRAM memory available)
  - NOTE: legacy CPU-based appliances (e.g., firewalls, IDS) cannot process packets fast enough; need large DRAM memory to track millions of flows; eject large flows' control connections prematurely

Programmable Switch ASICs

- Can we exploit the visibility provided by programmable ASICs on non-programmable networks? Solution would:
  - Be less disruptive
  - Foster incremental use of programmable ASICs
  - Not need to deploy complex code at once

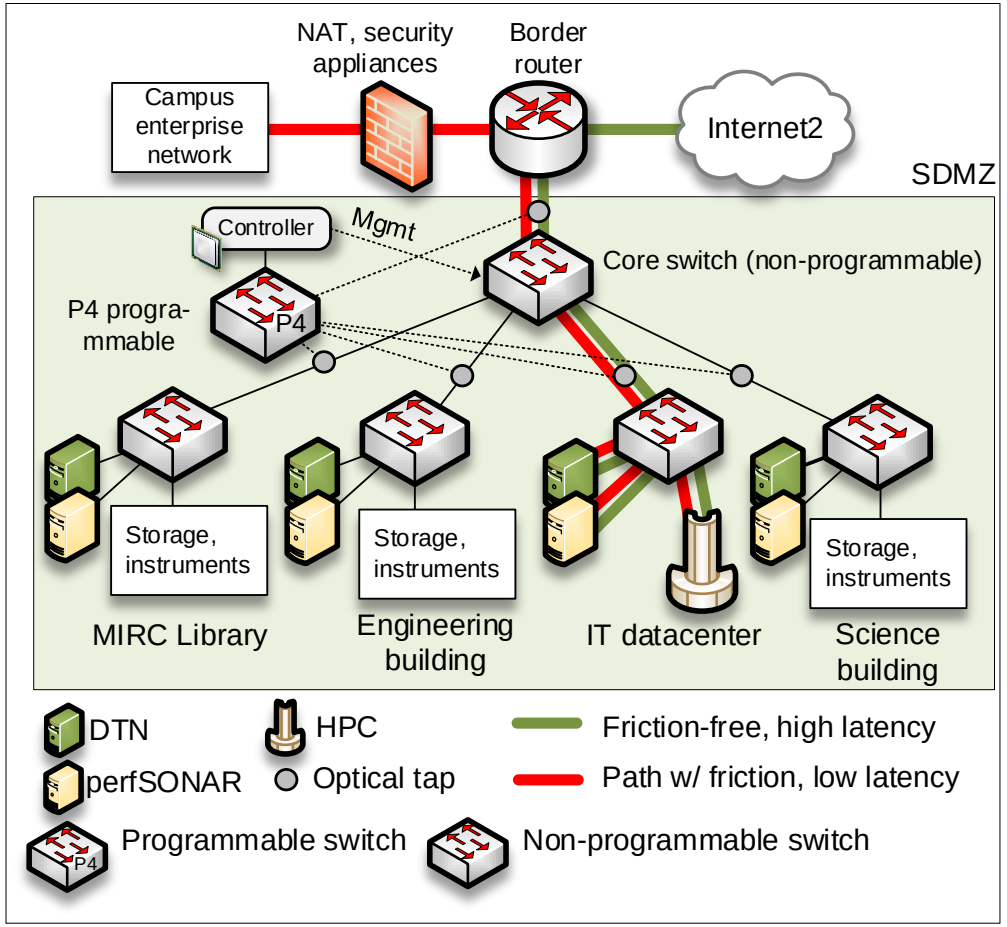
Programmable Switch ASICs

- Traditional Science DMZ with optical passive taps feeding a programmable ASIC

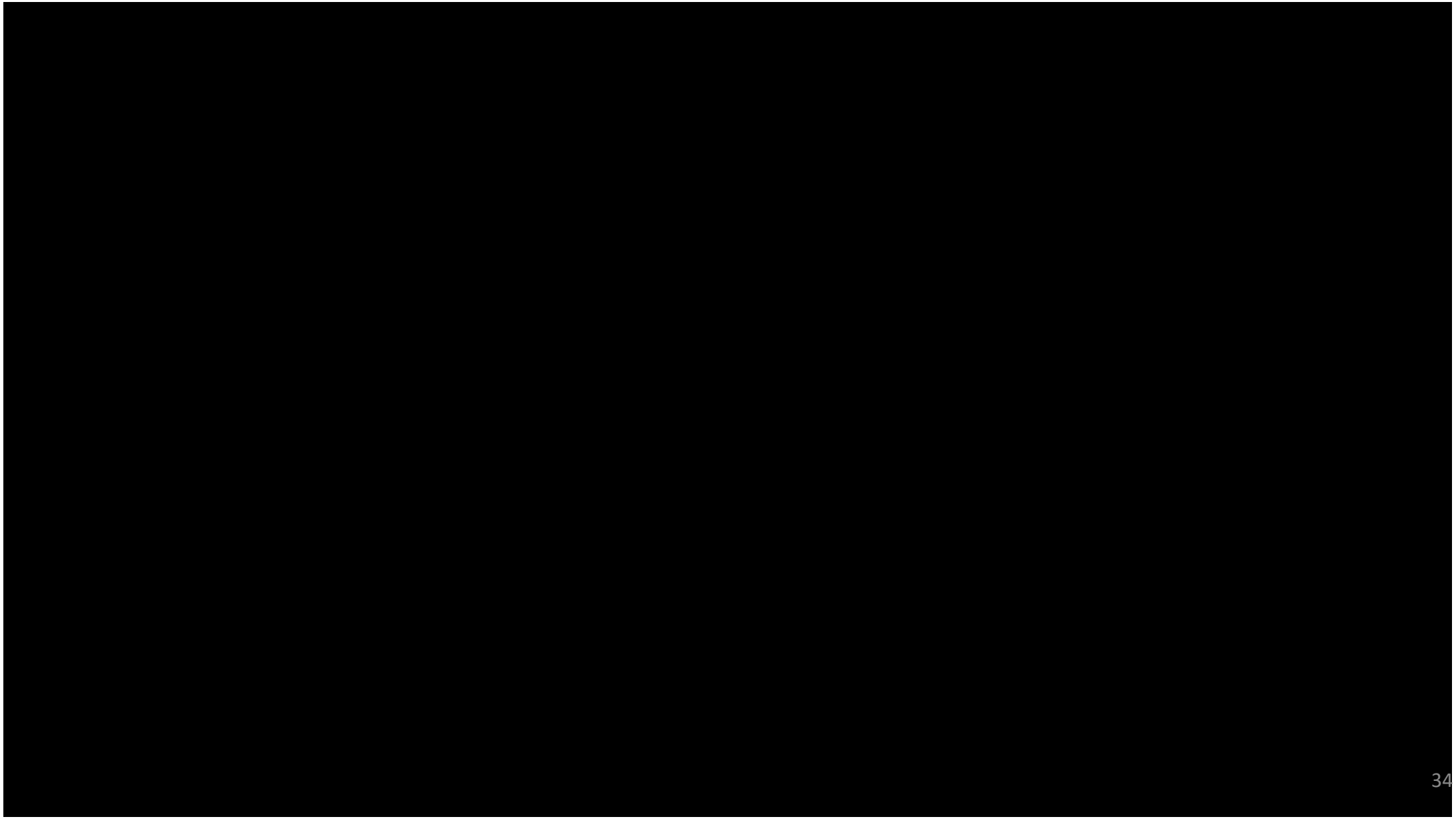


Fine-grained Measurements

- Granular RTT calculation

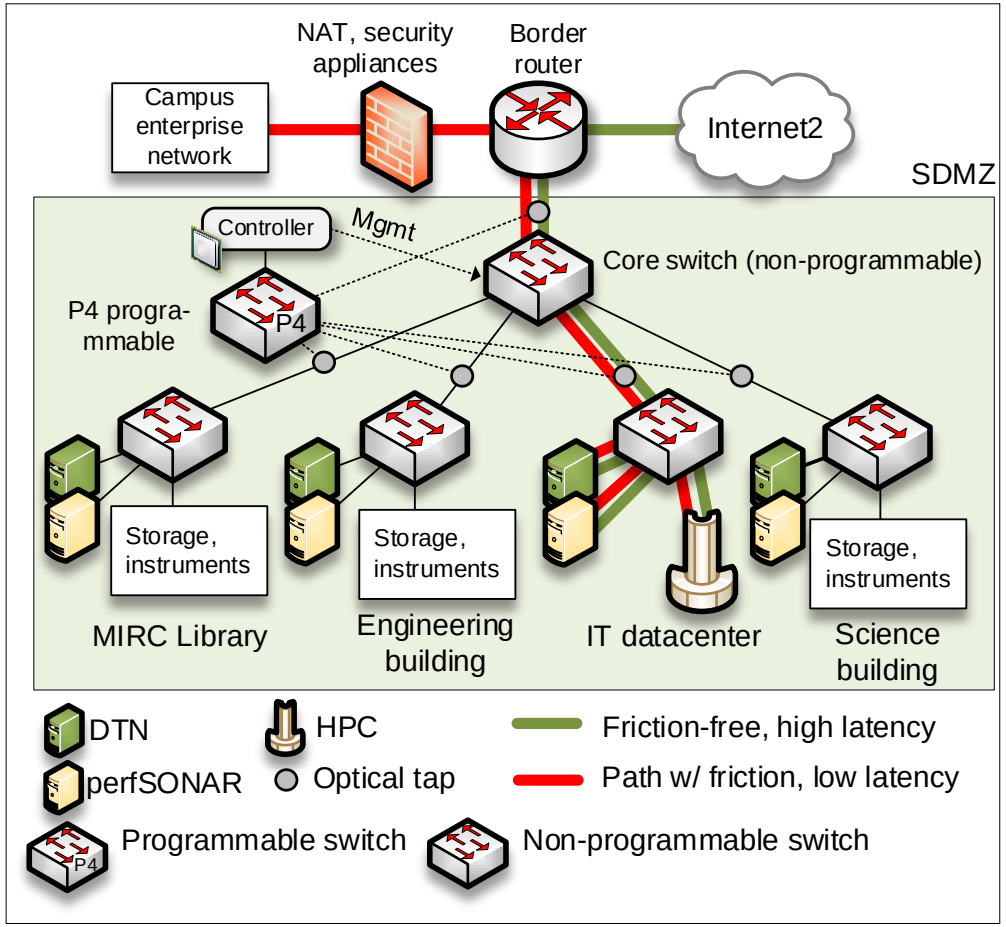


Fine-grained Measurements



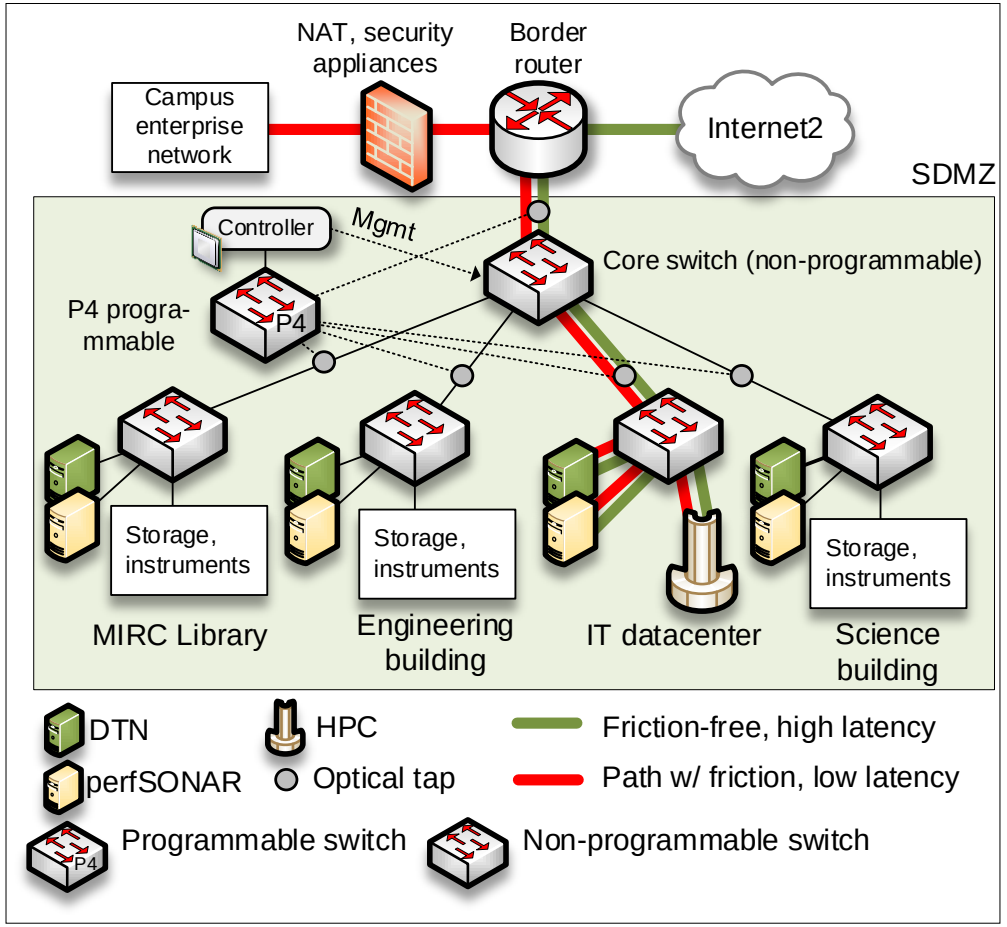
Fine-grained Measurements

- Granular RTT calculation - Applications
  - Calculating the optimal buffer size (a function of the average RTT of all large flows crossing the switch)
  - Detecting bad routing decisions, hijacking, reflected in large RTTs



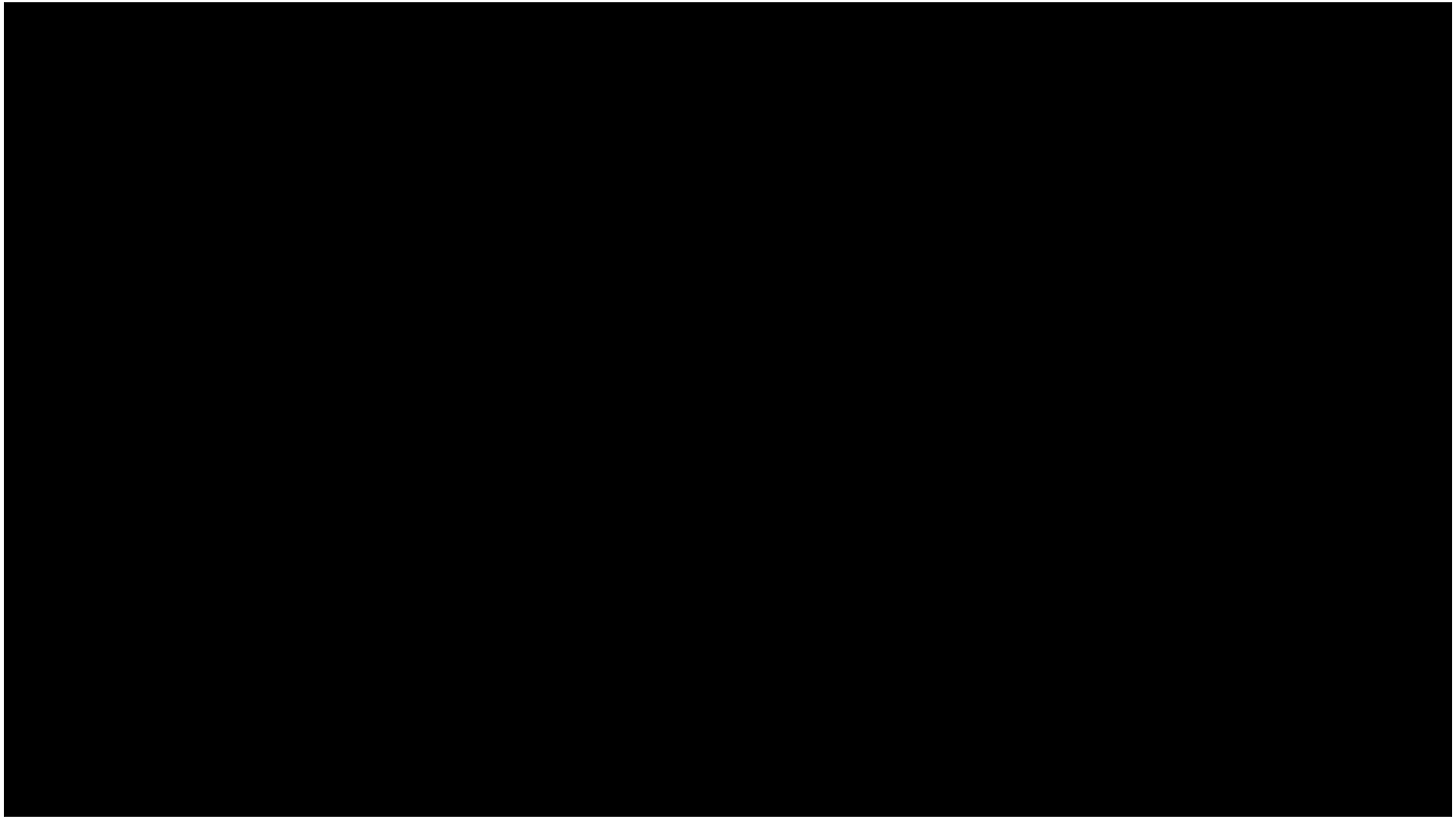
Fine-grained Measurements

- Customized firewall, without adding any additional processing on Science DMZ devices



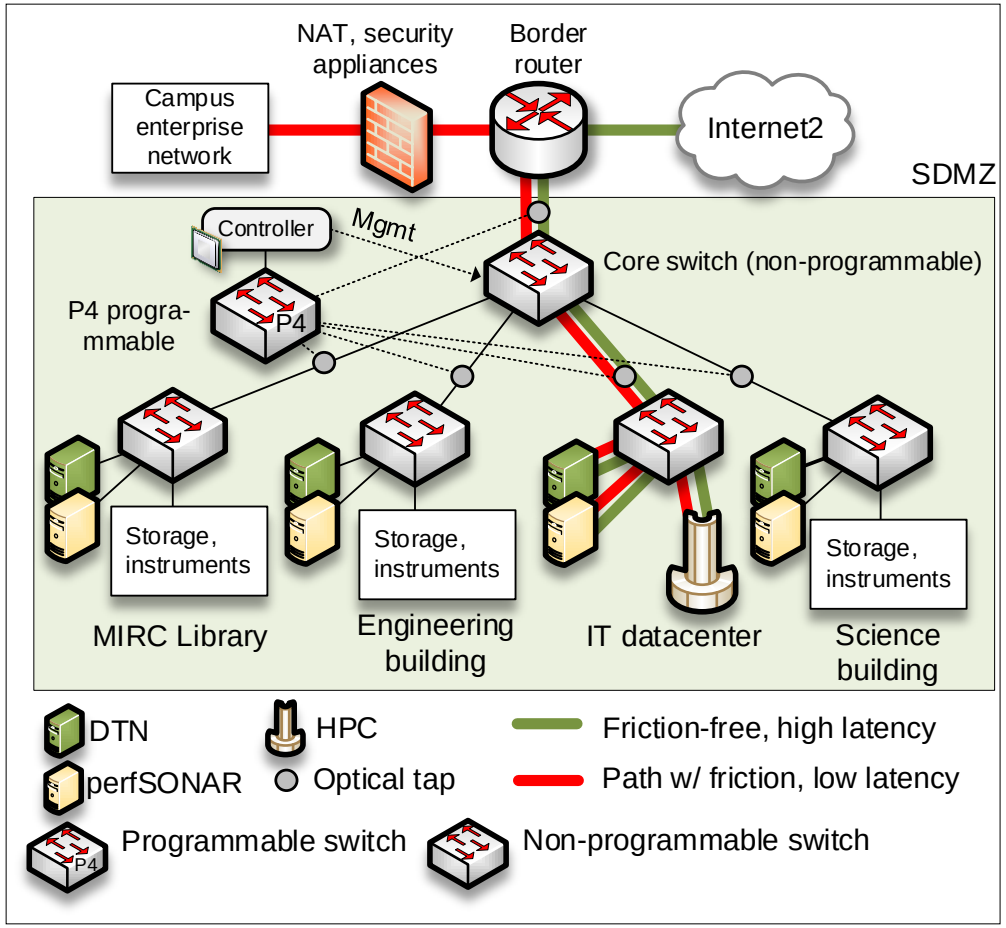
Fine-grained Measurements

- Customized firewall, without adding any additional processing on Science DMZ devices



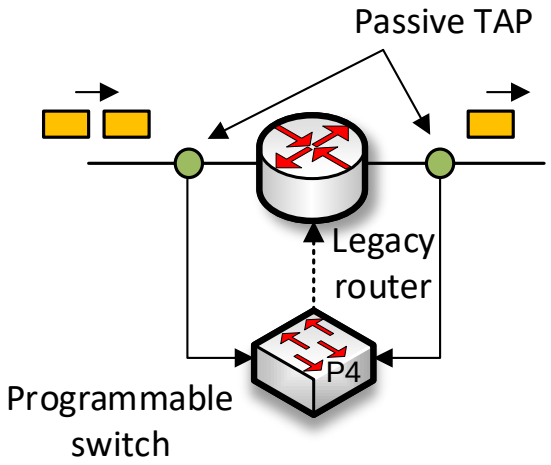
Fine-grained Measurements

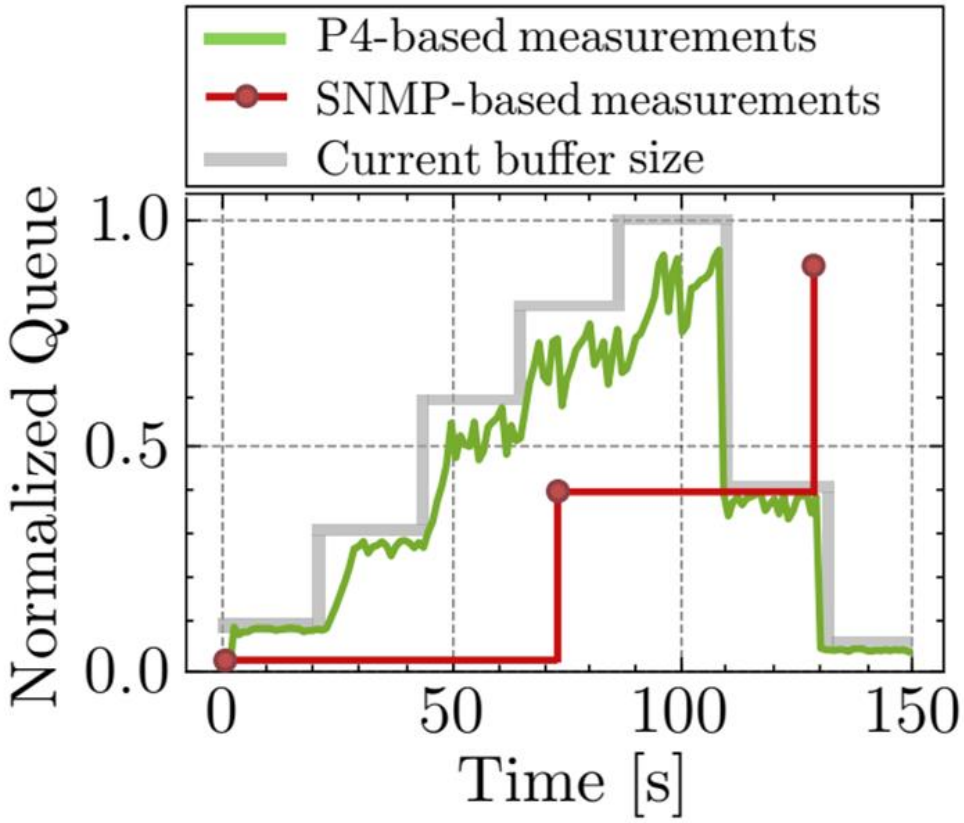
- Fine-grained measurements of buffer occupancy of legacy devices
  - Legacy measurements (e.g., SNMP) provide only coarse-grained measurements
  - E.g., Juniper MX-204 router provides one SNMP sample per minute



Fine-grained Measurements

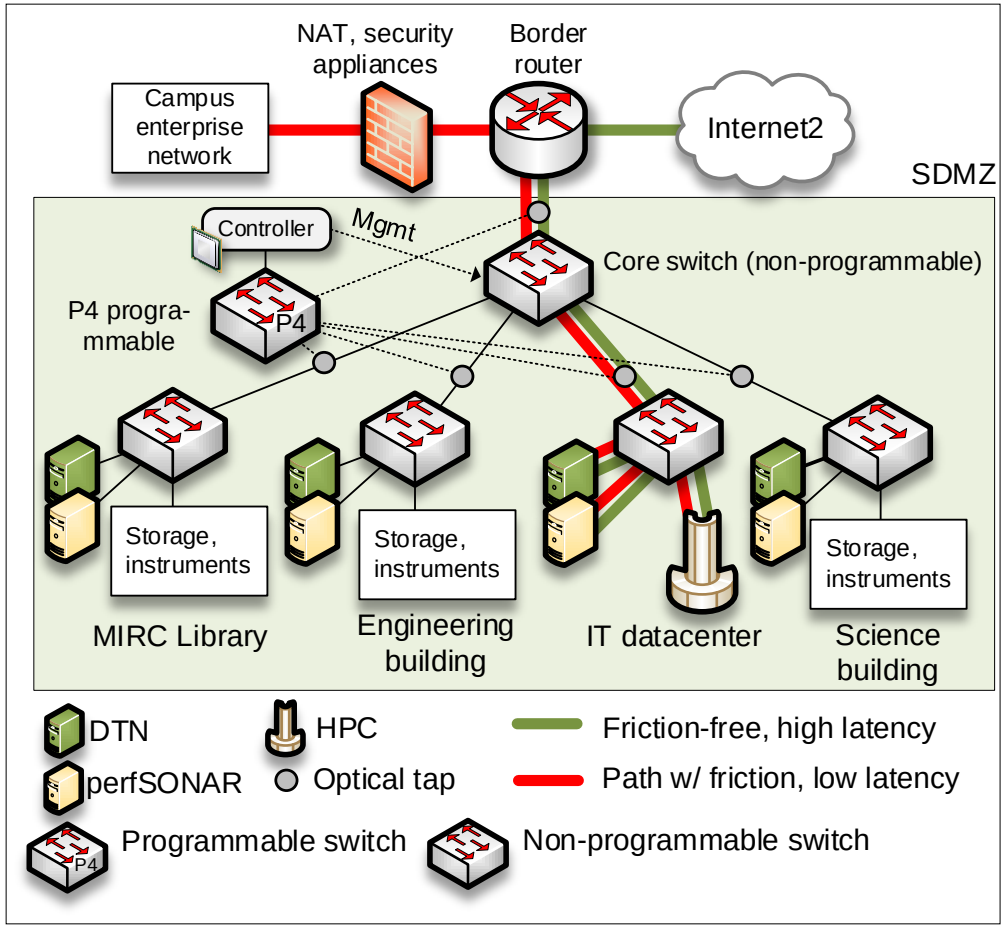
- Fine-grained measurements of buffer occupancy of legacy devices
  - Legacy measurements (e.g., SNMP) provide only coarse-grained measurements
  - E.g., Juniper MX-204 router provides one SNMP sample per minute
  - Programmable switches provide a high precision timer, full visibility of all packets





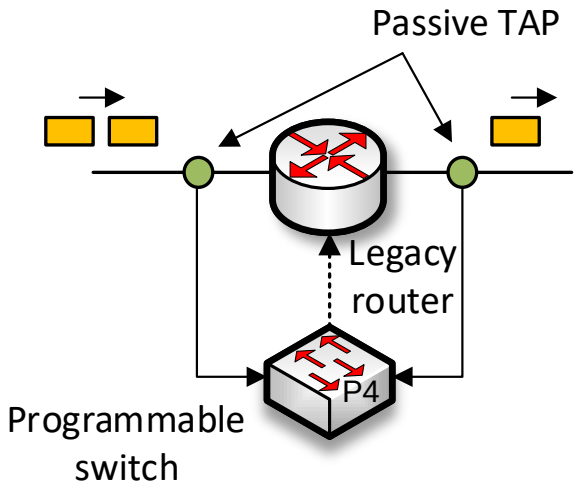
Fine-grained Measurements

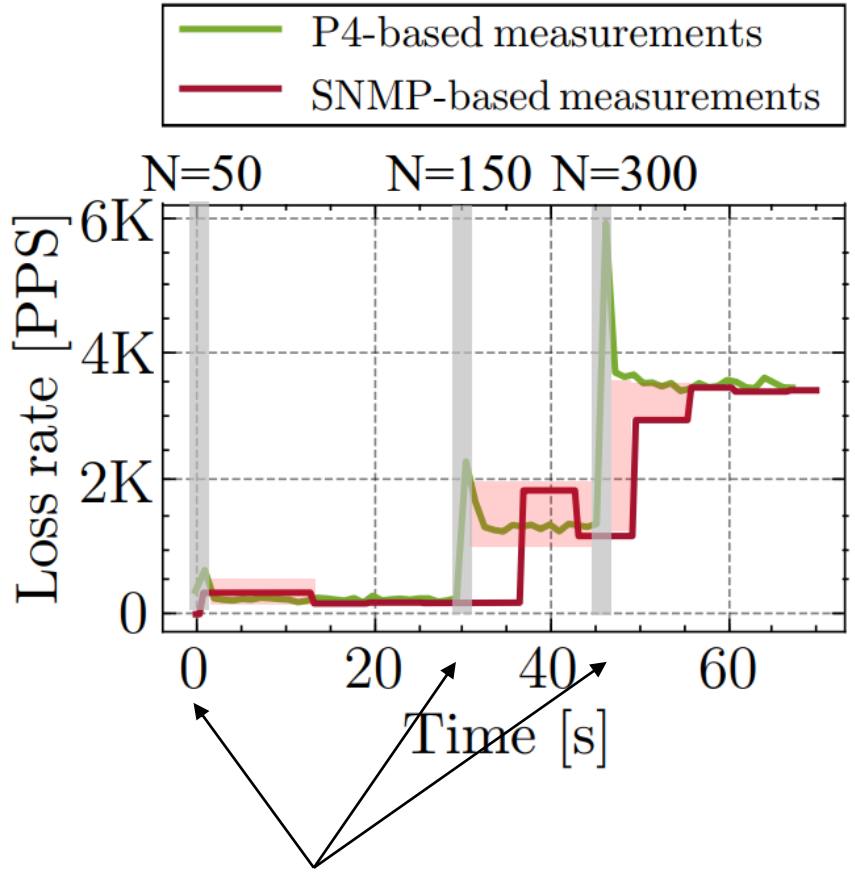
- Fine-grained measurements of packet loss of legacy devices
  - Legacy measurements (e.g., SNMP) provide only coarse-grained measurements (often erroneous)



Fine-grained Measurements

- Fine-grained measurements of packet loss of legacy devices
  - Legacy measurements (e.g., SNMP) provide only coarse-grained measurements (often erroneous)
  - Programmable switches can compute packet loss rates accurately and promptly

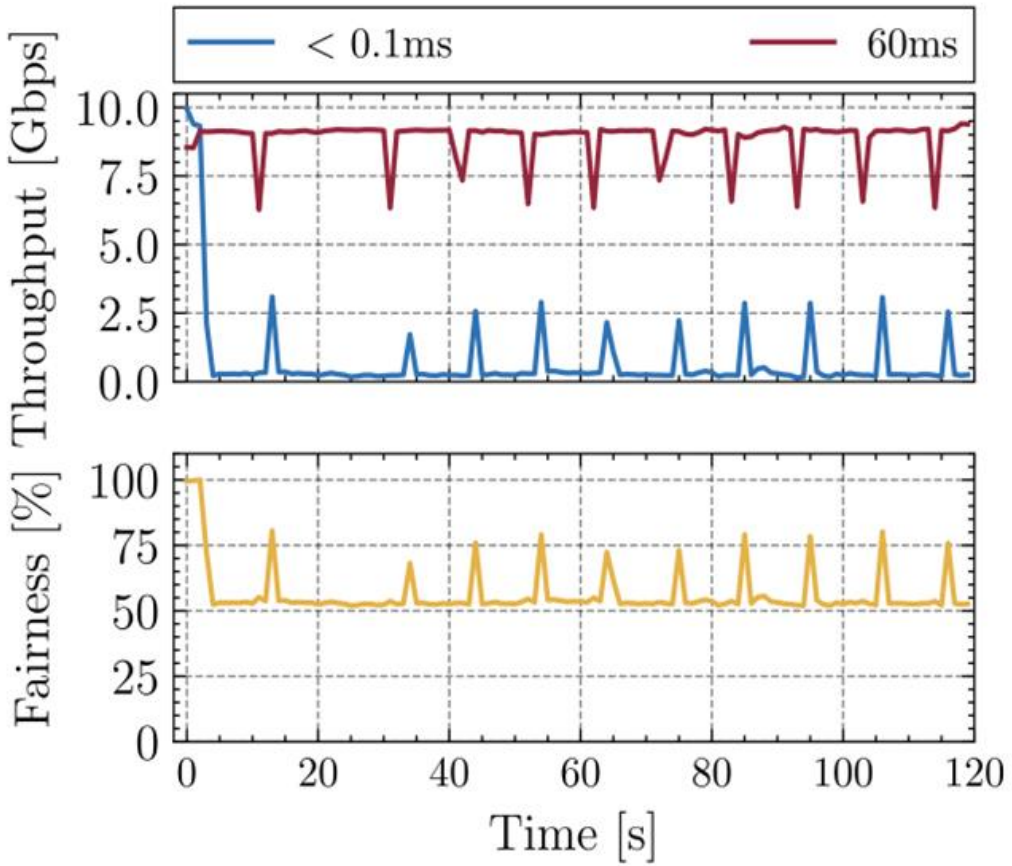




50 / 150 / 300 flows enter the network

Fine-grained Measurements

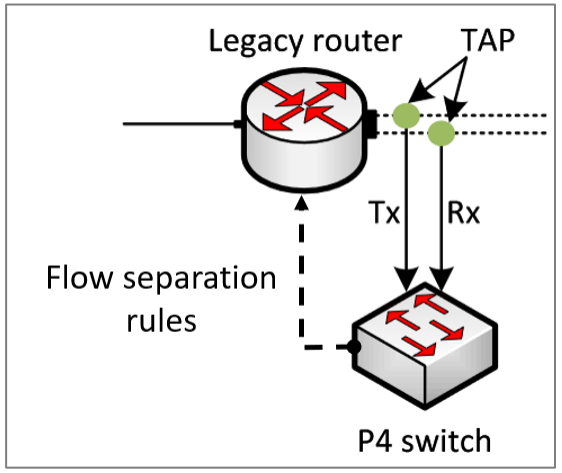
- Traffic separation based on RTT
  - If multiple flows with different RTTs compete, TCP favors one flow over the other

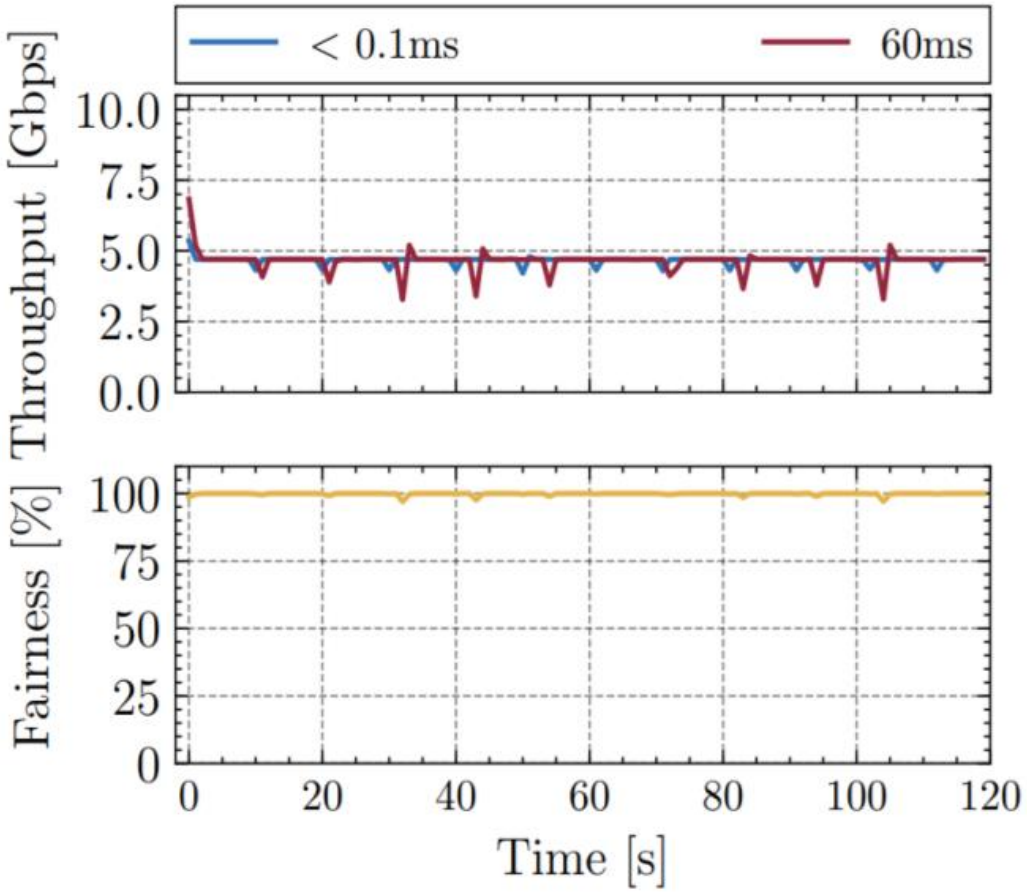


RTT unfairness in a 10 Gbps network. Two BBRv2 flows sharing the network.

Fine-grained Measurements

- Traffic separation based on RTT
  - By accurately measuring RTT in real time, flows can be classified and separated in different queues





RTT unfairness in a 10 Gbps network. Two BBRv2 flows sharing the network.

Programmable ASICs – Opportunities and Challenges

- Fine-grained measurement tools can complement current tools
- P4 code is open
  - Reusing code is simple
- Designing and testing new ideas can be accomplished faster
- There can be more opportunities for collaboration
  - Code can be downloaded from the open-source community, tailored in house

Programmable ASICs – Opportunities and Challenges

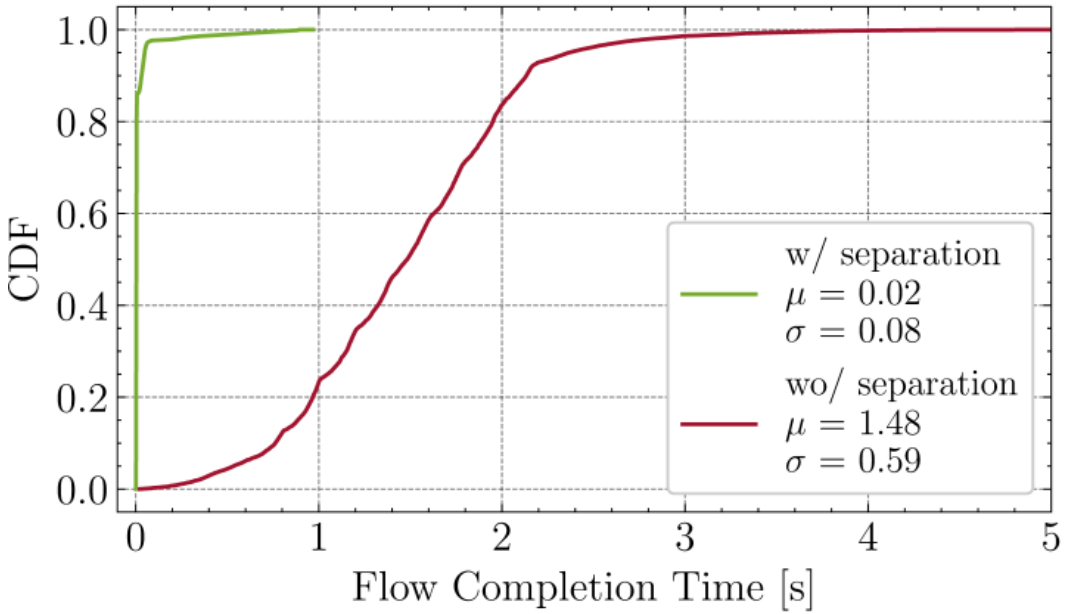
- CI engineers are not yet (fully) familiar with these technologies
- Deploying a fully programmable network is unrealistic
  - Deployment models are still being investigated
- There is no readily available technical support
  - There are few forums, but they focus more on research (e.g., Intel Connectivity Research Program)

Thank you

Fined-grained Measurements

Applications of programmable switches to **enhance the performance** of Science DMZ

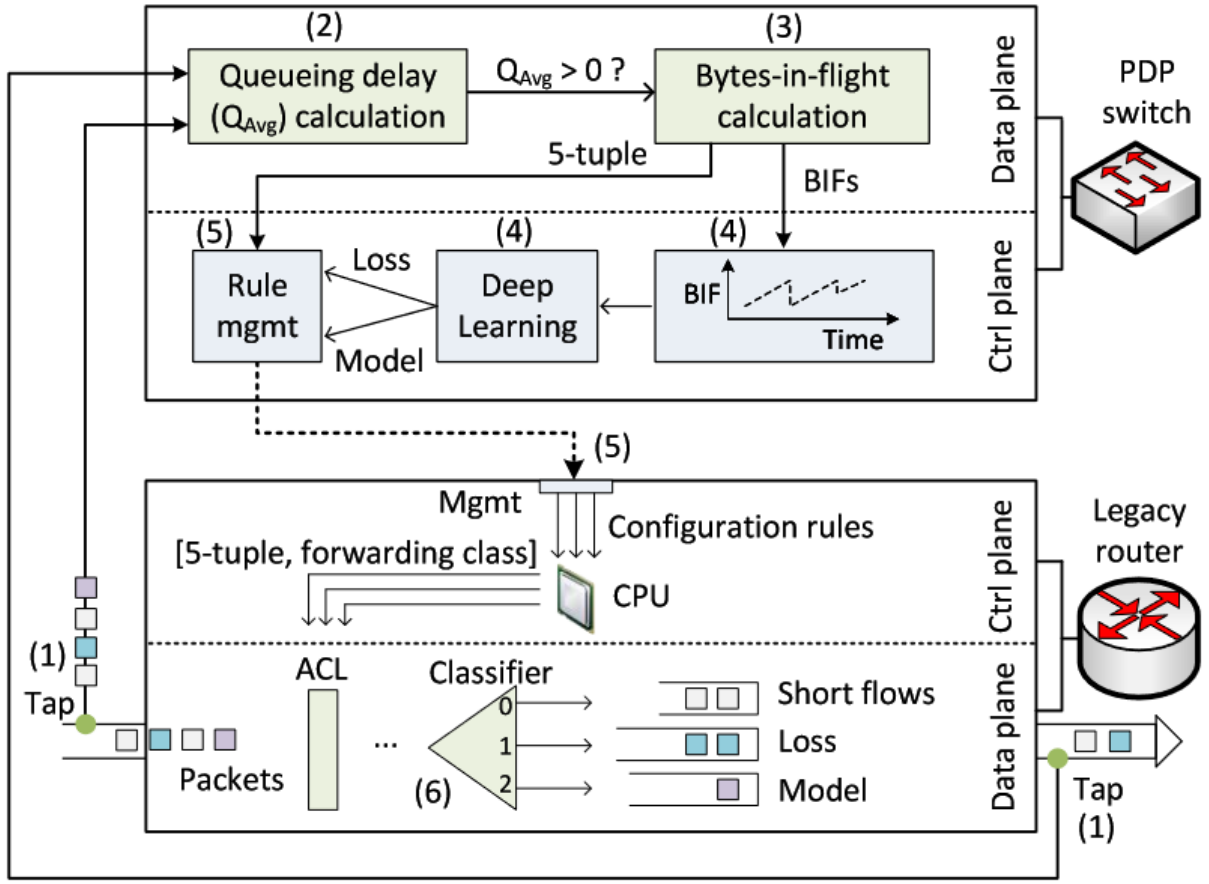
- Traffic separation based on flow size
  - The flow completion time (FCTs) of small flows is significantly impacted when the network is busy
  - A possible solution to prevent the increase of FCTs is to separate small flows from large flows
  - Programmable switches can identify long flows
  - 10,000 small flows whose inter-connection times are generated from an exponential distribution with a mean of one second
  - 10 large flows were started, each with a random starting time over the test duration



Fined-grained Measurements

Applications of programmable switches to **enhance the performance** of Science DMZ

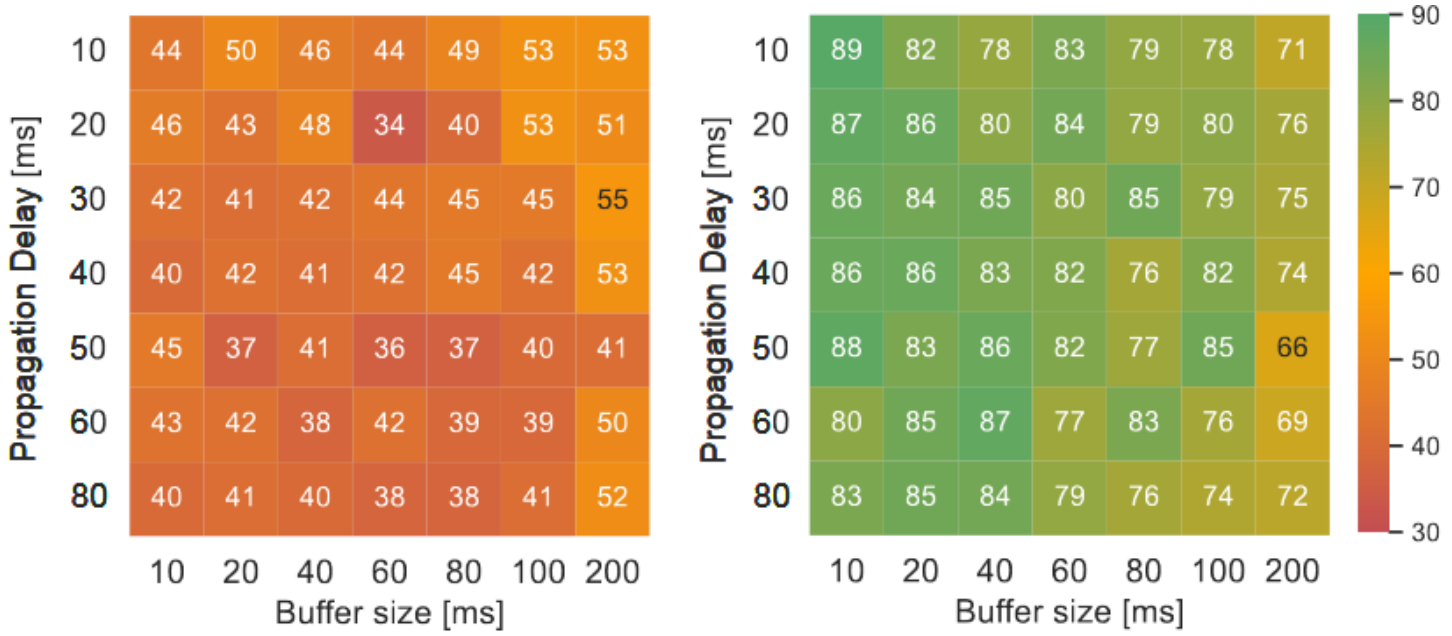
- Traffic separation based on Congestion Control Algorithms (CCA)
  - When flows using different CCAs co-exist on a link, the fairness is significantly impacted
  - One solution is to separate flows into different queues on the router based on their CCAs
  - Programmable switches compute the bytes-in-flight and use Deep Learning algorithms to identify the CCA
  - Flows are then allocated into different queues based on their CCAs



Fined-grained Measurements

Applications of programmable switches to **enhance the performance** of Science DMZ

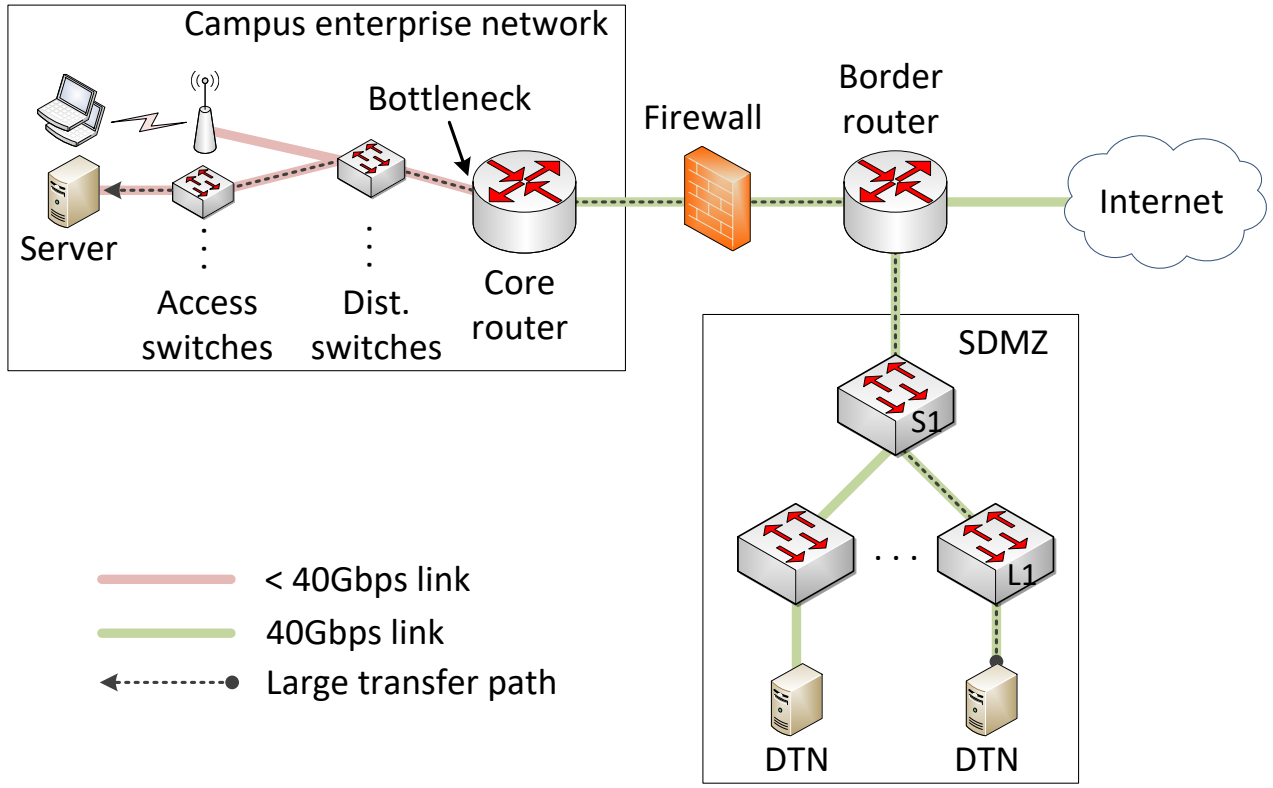
- Traffic separation based on Congestion Control Algorithms (CCA)
  - 10 long flows started at the same time, with alternating CCAs (i.e., Flow1 uses CUBIC, Flow2 uses BBR, Flow3 uses CUBIC, etc.)
  - The figure below shows the fairness index (0 -> unfair, 100 -> fair)
    - Left -> without separation
    - Right -> with separation



Fined-grained Measurements

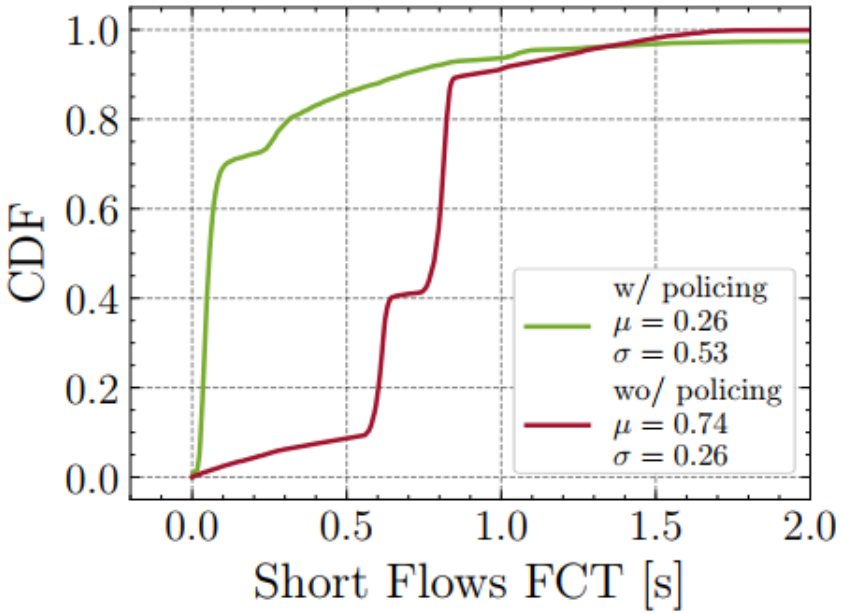
Applications of programmable switches to **enhance the performance** of Science DMZ

- Bottleneck estimation for traffic policing
  - During a large data transfer, a link can get fully utilized and its router's buffer gets filled
  - This increases the latency for short flows sharing the bottleneck link
  - Programmable switches can compute the throughput of flows and identify those that are bottlenecked
  - Such information can be used to force flows to slow their rate, which avoids filling the queues



Fined-grained Measurements

- Bottleneck estimation for traffic policing
  - Flow Completion Time (FCT) of 10,000 short flows whose inter-connection times are generated from an exponential distribution with a mean of one second
  - Long flows which transmitted 75Gbytes:
    - Completed in approximately 85 seconds when policing is not configured
    - Completed in approximately 87 seconds when policing is configured
  - While the FCT of the long flow slightly increased when policing is configured, this increase is acceptable since the flow is not interactive



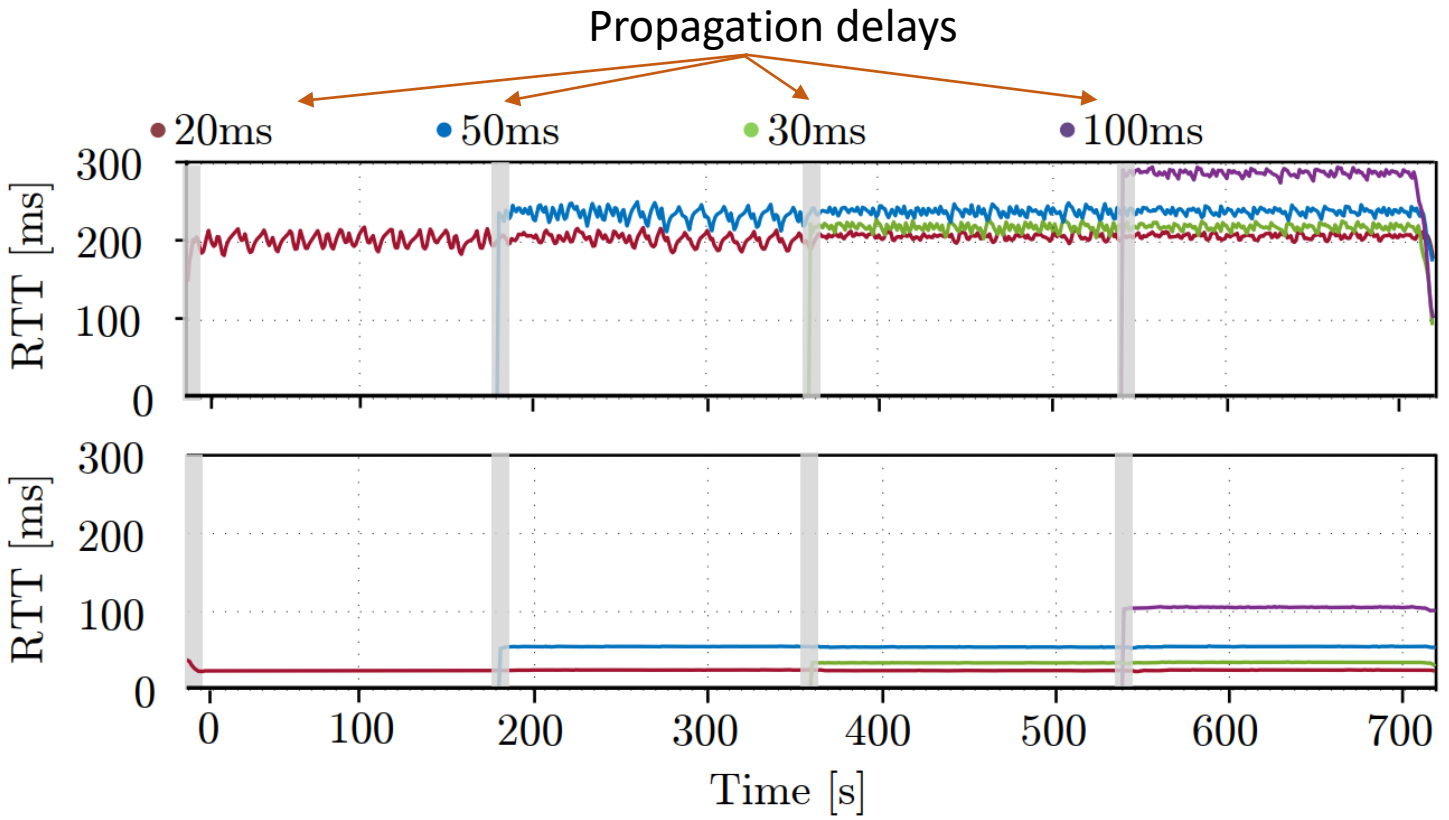
Fined-grained Measurements

Applications of programmable switches to **enhance the performance** of Science DMZ

- Dynamic buffer sizing
  - By leveraging measurements from programmable switches, the buffer size can be dynamically modified

Static buffer

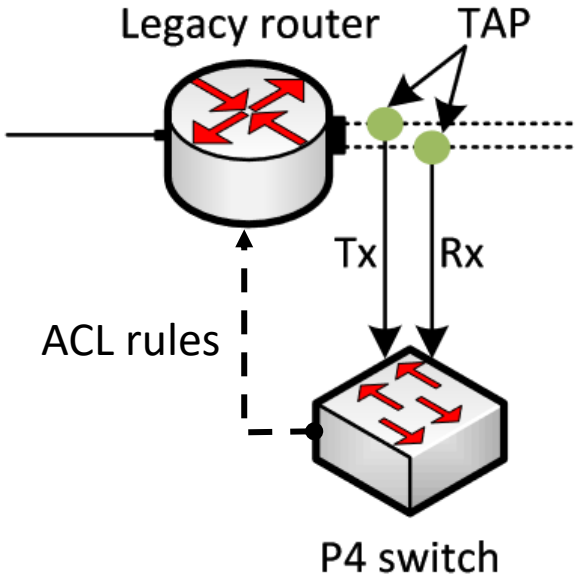
Dynamic buffer



Fined-grained Measurements

Applications of programmable switches to enhance the **security** of Science DMZ

- Detecting DDoS attacks and applying mitigations through ACLs
  - Deploying a firewall or an Intrusion Prevention System (IPS) inline affects the throughput of data transfers in Science DMZ
  - Programmable switches can be deployed passively
  - DDoS detection algorithms are executed in the programmable switches
  - Access Control List (ACL) rules are pushed to the non-programmable router

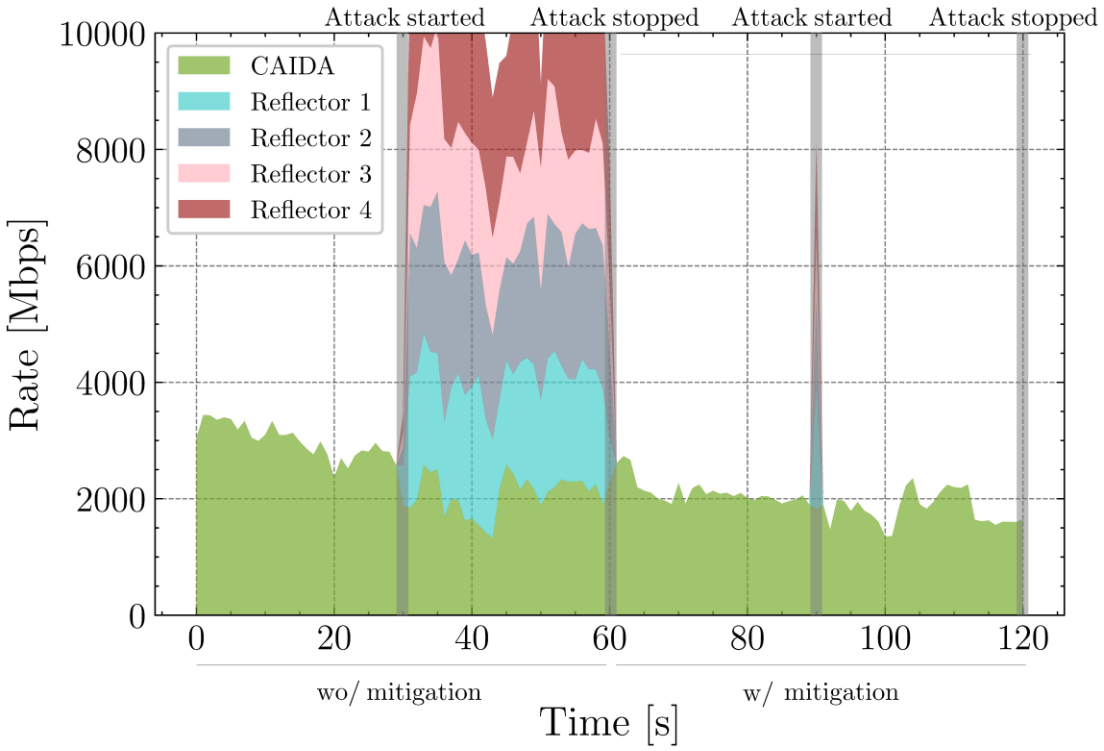


Fined-grained Measurements

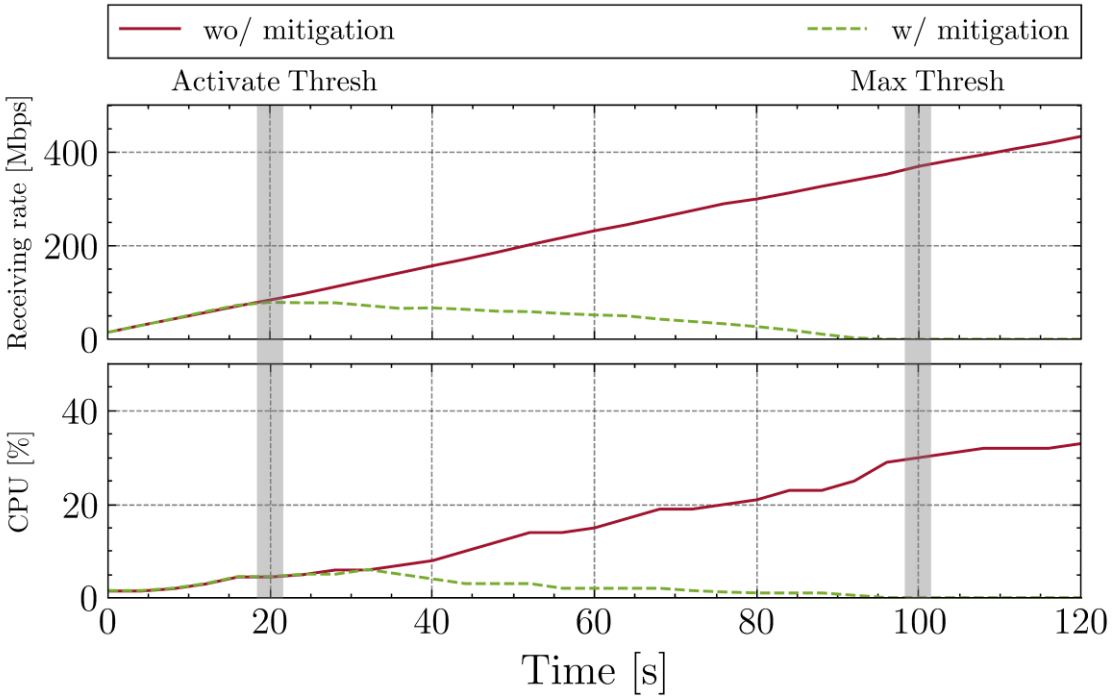
Applications of programmable switches to enhance the **security** of Science DMZ

- Detecting DDoS attacks and applying mitigations through ACLs

DNS amplification attack

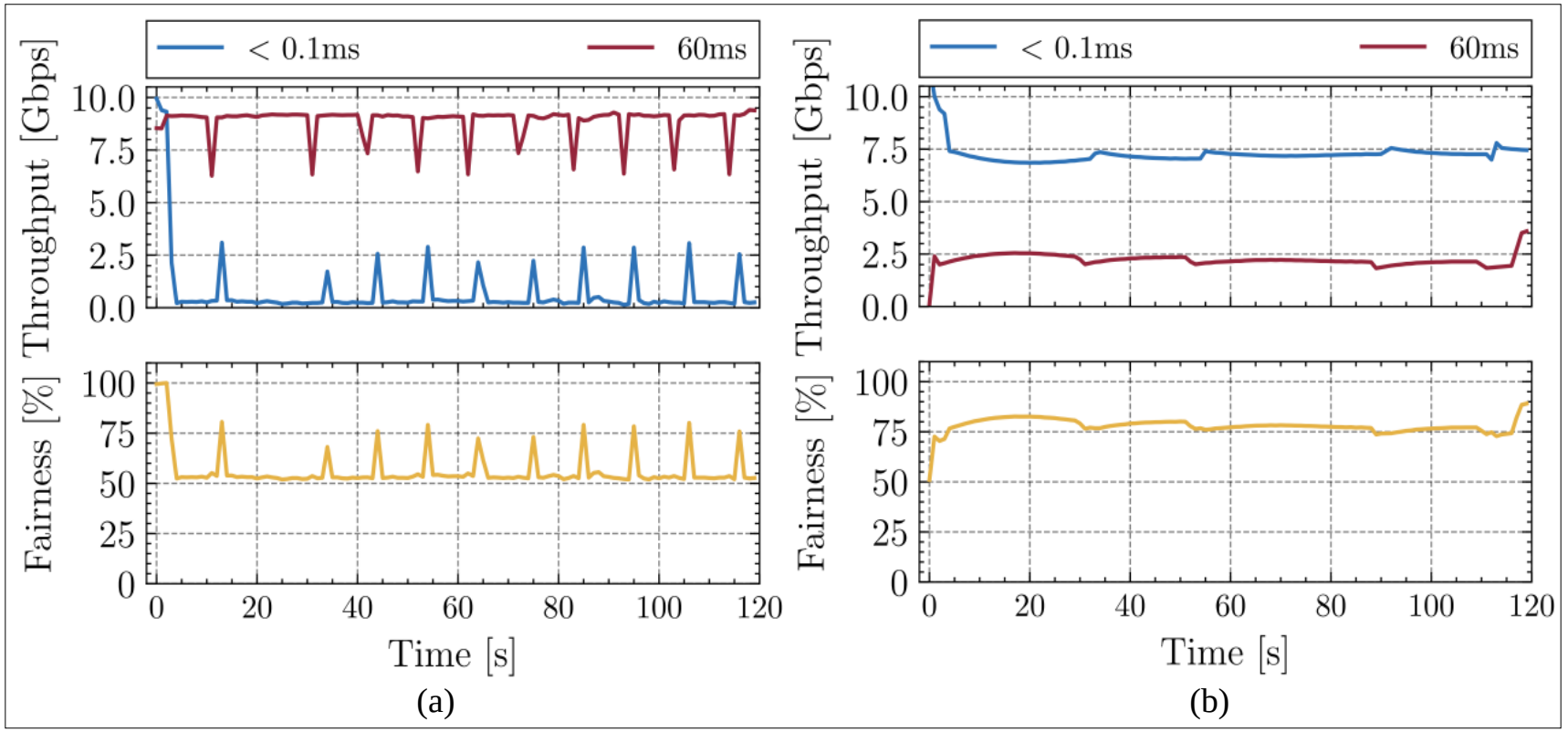


SYN flood attack



Fined-grained Measurements

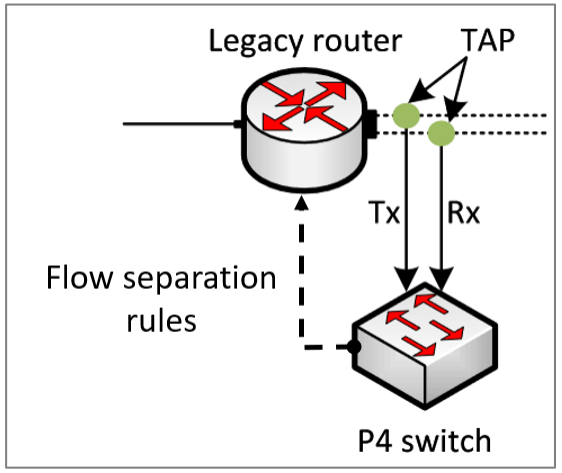
- Traffic separation based on RTT
  - If multiple flows with different RTTs compete, TCP favors one flow over the other

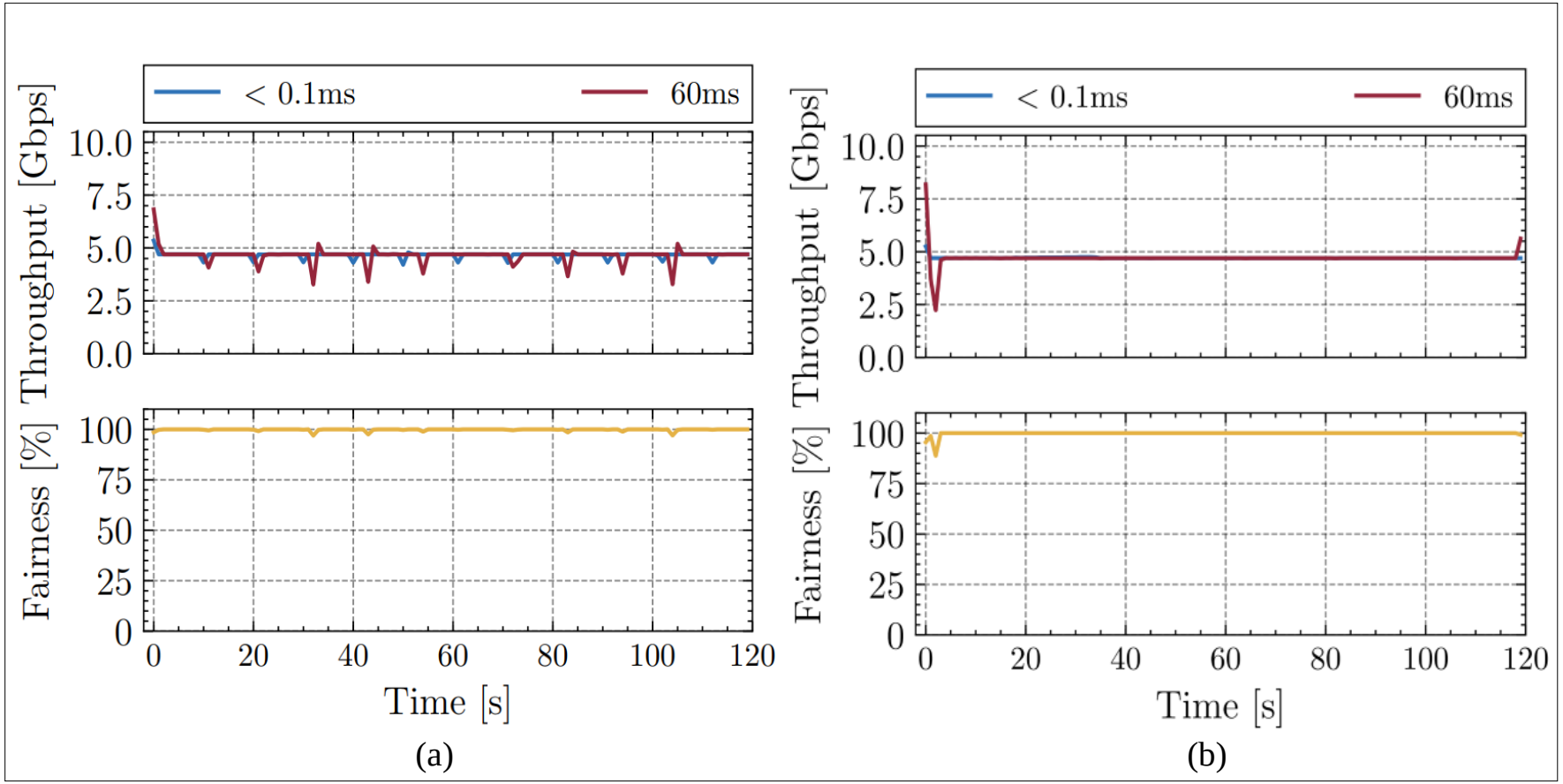


RTT unfairness in a 10 Gbps network. (a) Two BBRv2 flows sharing the network. (b) Two CUBIC flows sharing the network.

Fined-grained Measurements

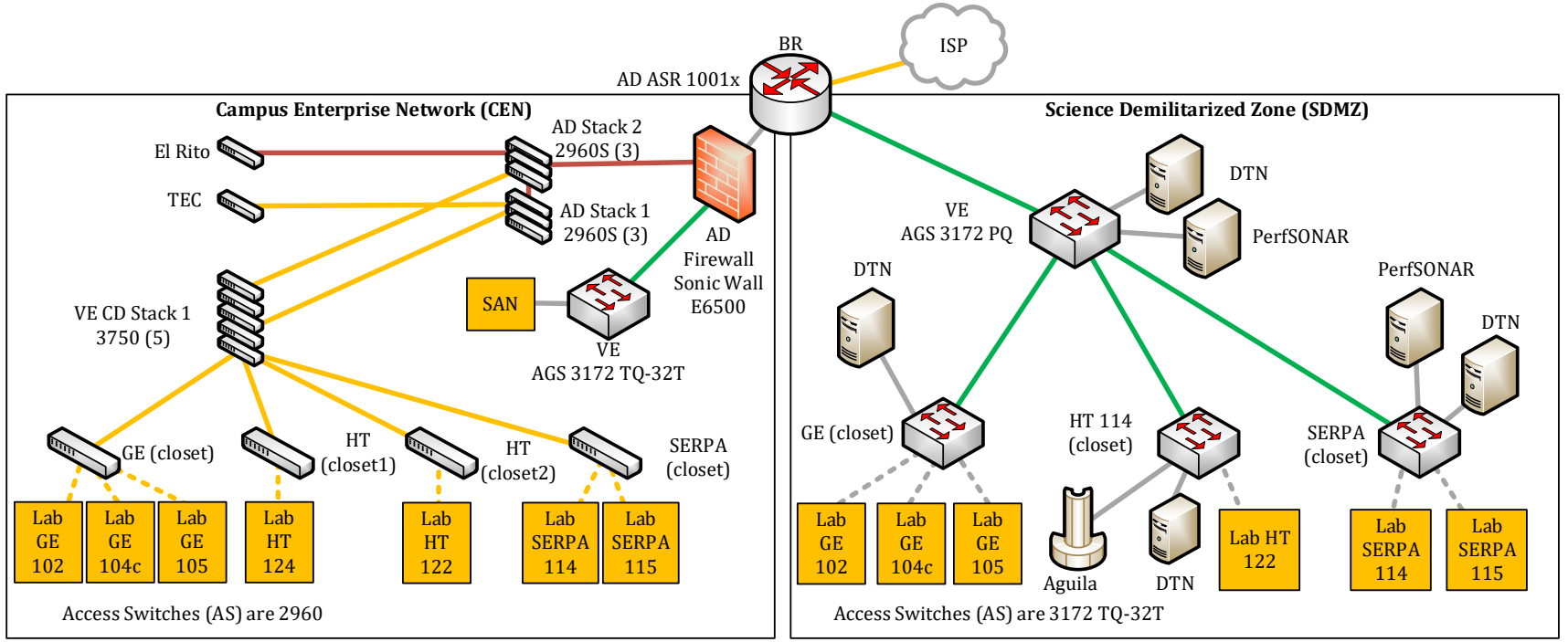
- Traffic separation based on RTT
  - By accurately measuring RTT in real time, flows can be classified and separated in different queues





RTT unfairness in a 10 Gbps network. (a) Two BBRv2 flows sharing the network. (b) Two CUBIC flows sharing the network.

Network Performance and Measurements



Cabling

- 10 Gbps (Single-mode fiber)
- 10 Gbps (10GBase-T)
- 1 Gbps (Multi-mode fiber)
- 1 Gbps (1000Base-T)
- 100 Mbps (100Base-T)
- Horizontal cable (from closet to work area)

Acronyms

AD:	Administration building
AGS:	Aggregation Switch
BR:	Border Router
CD:	Core Distribution
CEN:	Campus Enterprise Network
DTN:	Data Transfer Node
GE:	General Education building
HT:	High Tech building
SDMZ:	Science DMZ
SERPA:	Solar Energy Research Park & Academy building
TEC:	Teaching Education Center
VE:	Vocational Education building

Network Performance and Measurements

- perfSONAR provides information that reflects the state of the network, in a multi-domain basis
- perfSONAR = Measurement Middleware