

Enhancing perfSONAR Measurement Capabilities using P4 Programmable Data Planes

Ali Mazloum, Jose Gomez, Elie F. Kfoury, **Jorge Crichigno**
College of Engineering and Computing, University of South Carolina

<https://research.cec.sc.edu/cyberinfra/>

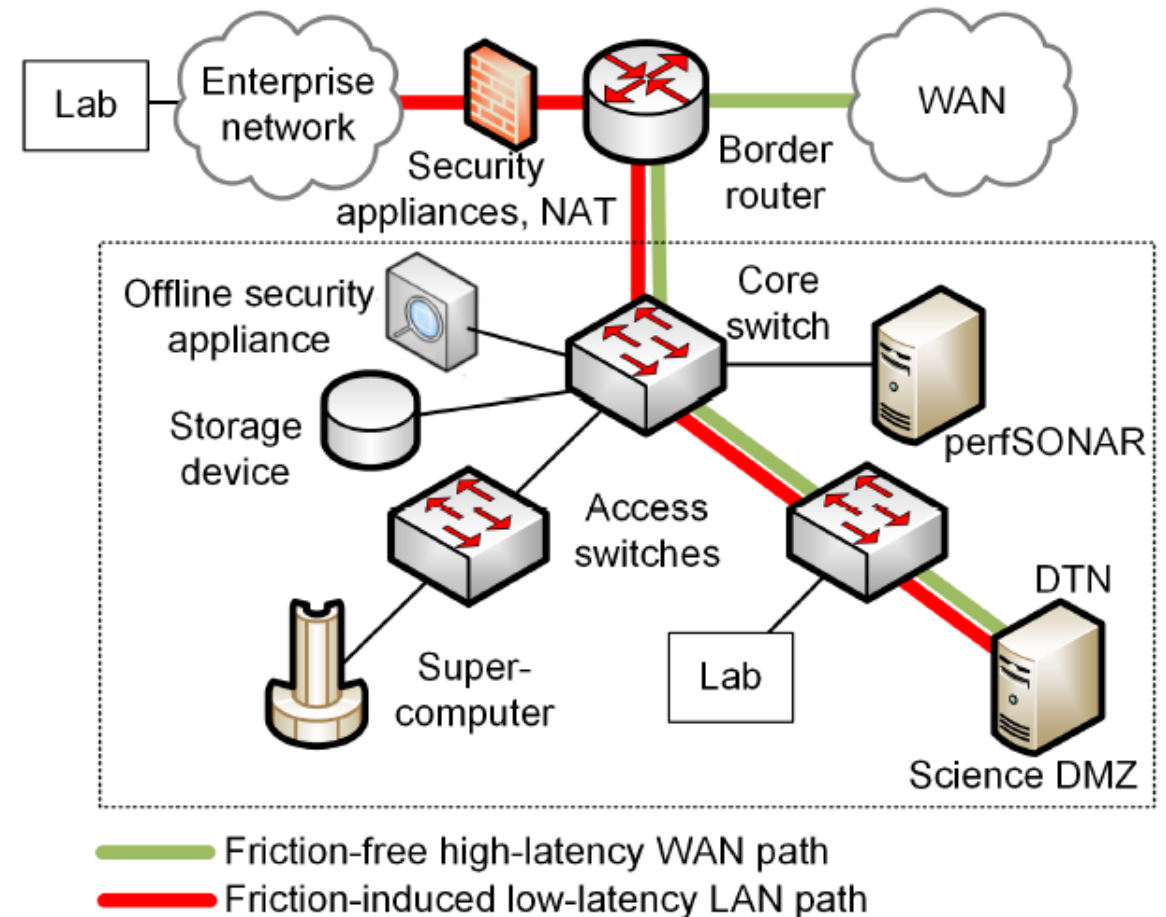
Innovating the Network for Data-Intensive Science (INDIS) Workshop
November 12, 2023
Denver, CO

Agenda

- Motivation
- Background information
 - Science demilitarized zone (DMZ)
 - perfSONAR
 - P4 programmable data planes
- Proposed system
- Experimental setup
- Results
- Conclusion

Science DMZ

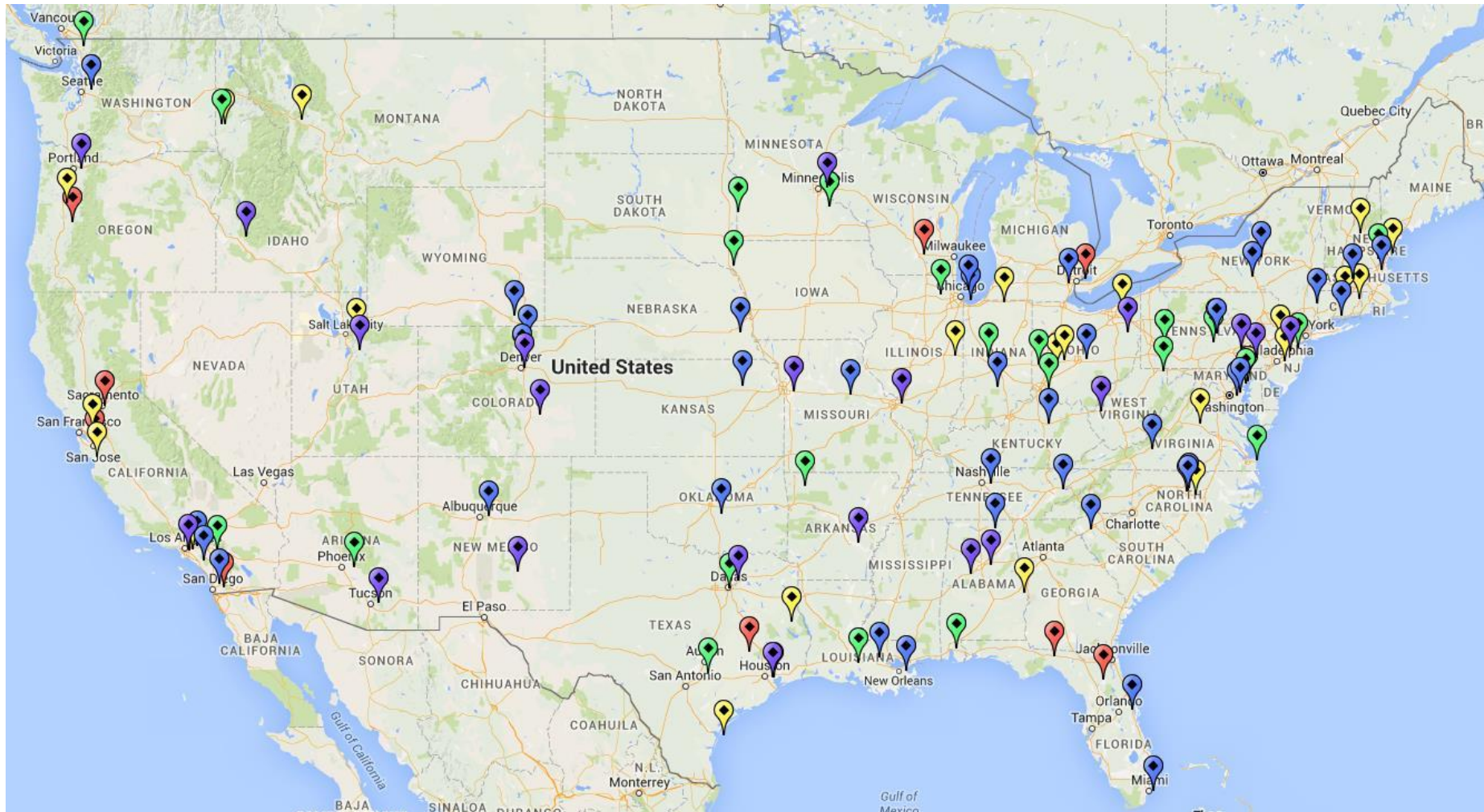
- The Science DMZ is a network designed for big science data¹
- Main elements:
 - High throughput, friction-free WAN paths
 - Security tailored for high speeds
 - Data Transfer Nodes (DTNs)
 - **End-to-end monitoring / perfSONAR**



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

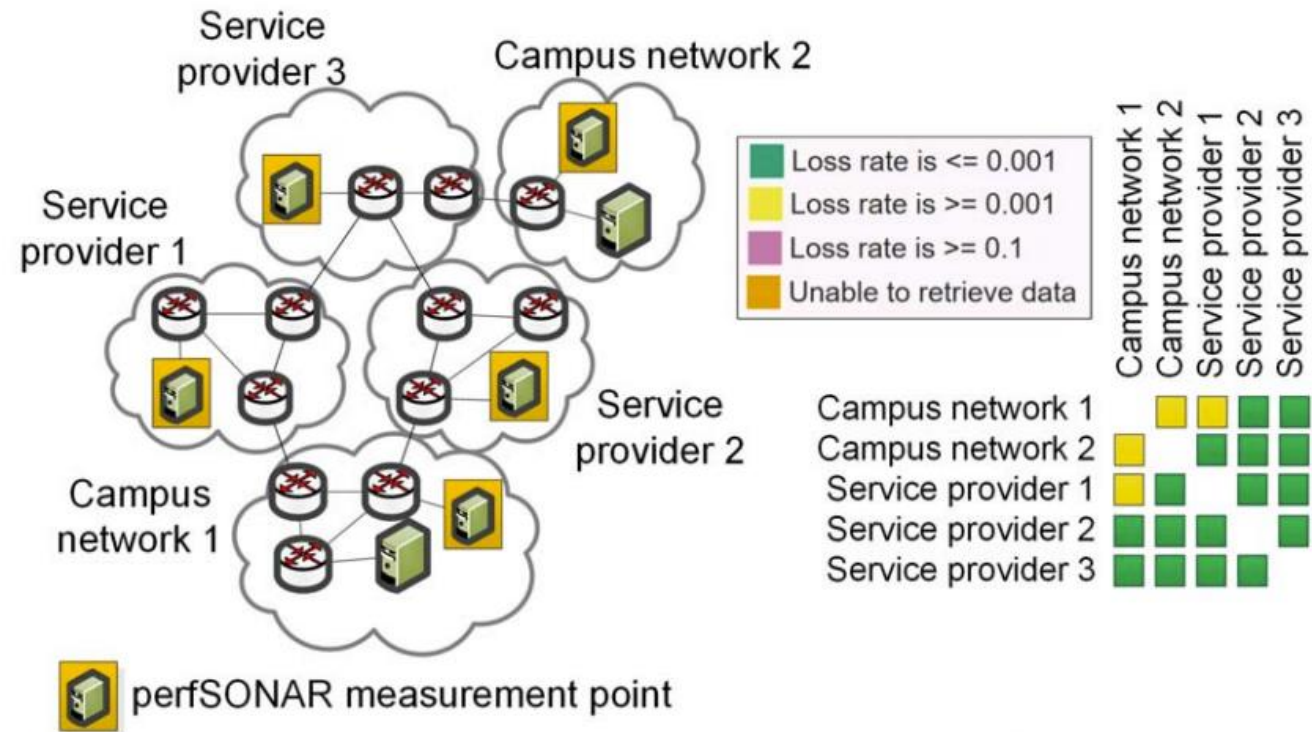
Science DMZ

- Science DMZ deployments, U.S.



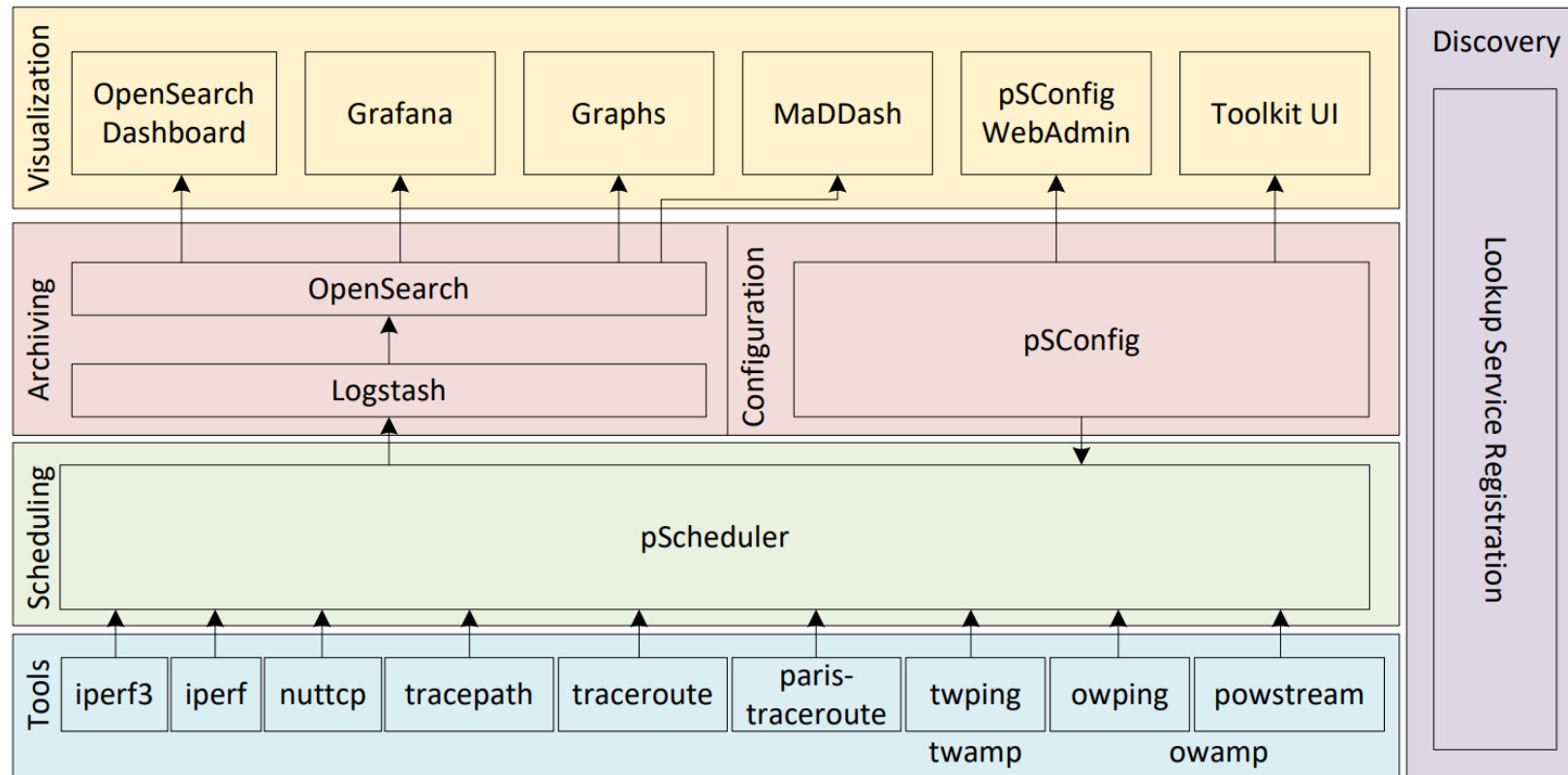
perfSONAR

- perfSONAR is a measurement tool that provides federated coverage of paths
- It helps troubleshoot performance issues (e.g., finding soft failures)
- perfSONAR is a key component of the Science DMZ



perfSONAR

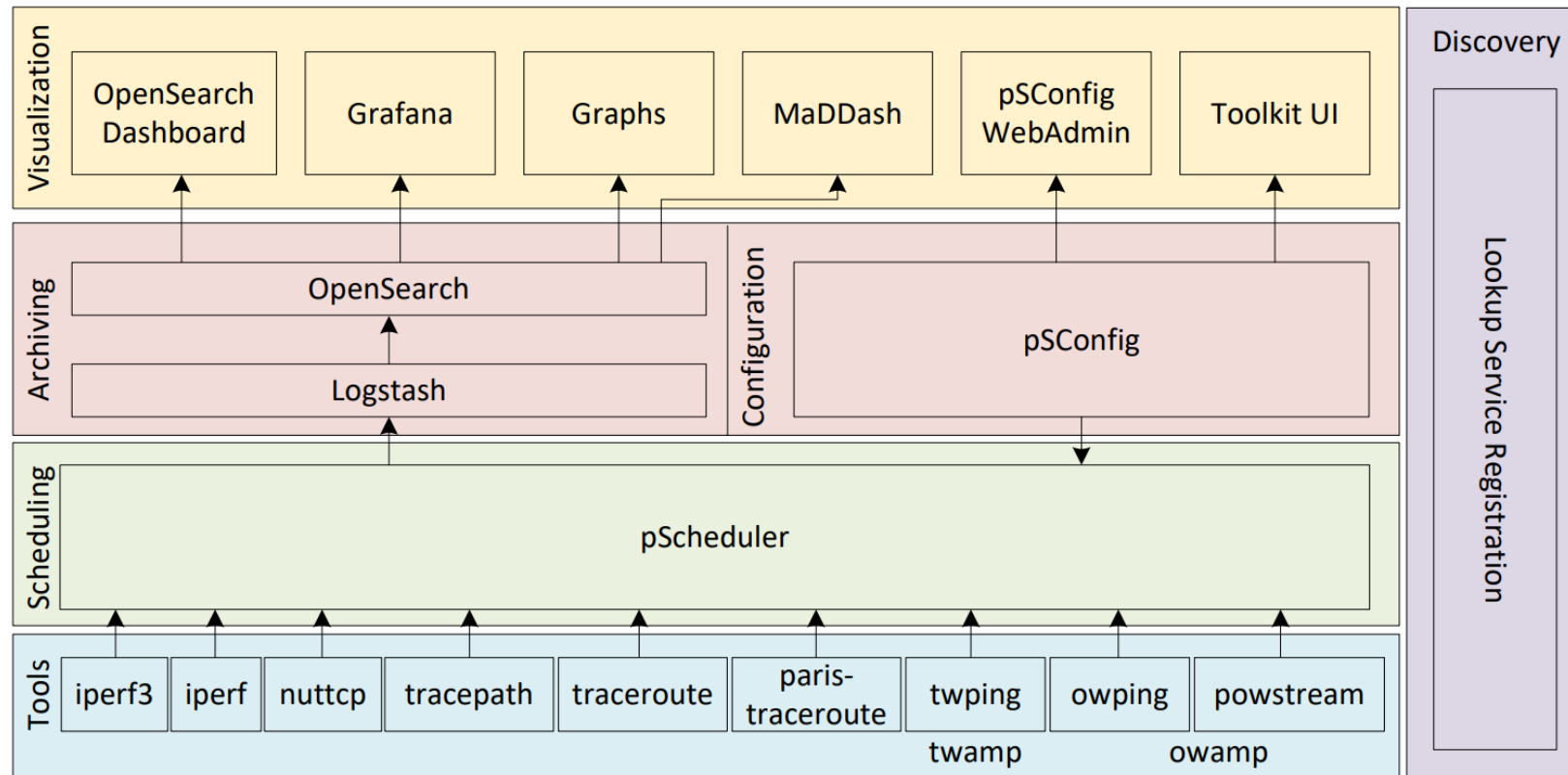
- perfSONAR is a measurement tool that provides federated coverage of paths
- It helps troubleshoot performance issues (e.g., finding soft failures)
- perfSONAR is a key component of the Science DMZ



perfSONAR architecture

perfSONAR

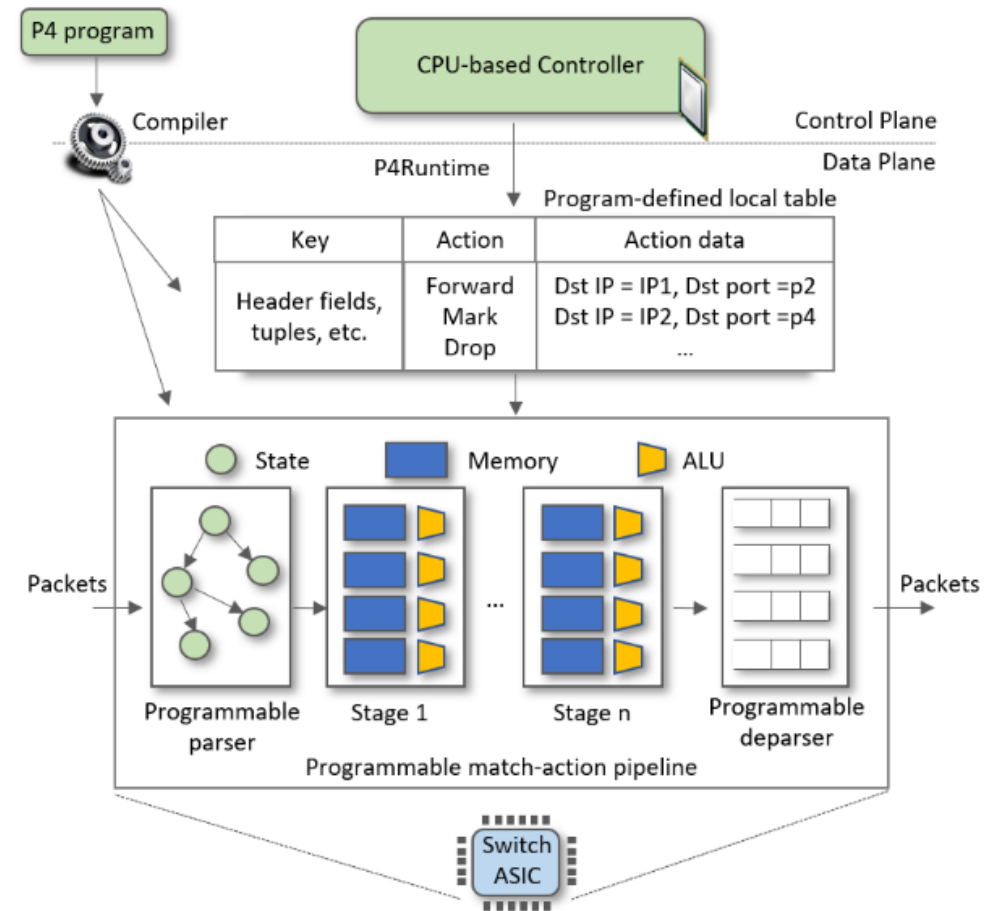
- perfSONAR
 - relies on tools that provide coarse-grained measurements
 - depends on active measurements
 - provides APIs that enable a programmer to extend its functionality



perfSONAR architecture

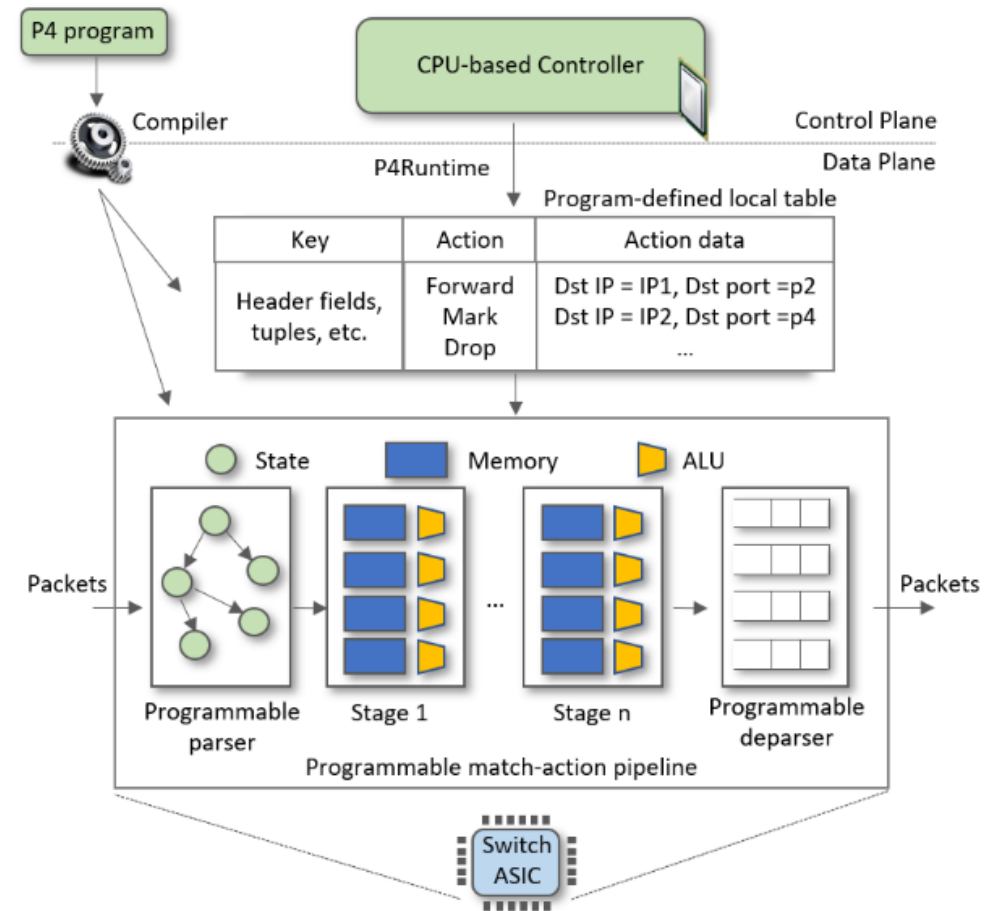
P4 Programmable Data Planes

- A P4¹ Programmable Data Planes (PDP) is a domain-specific processor for networking



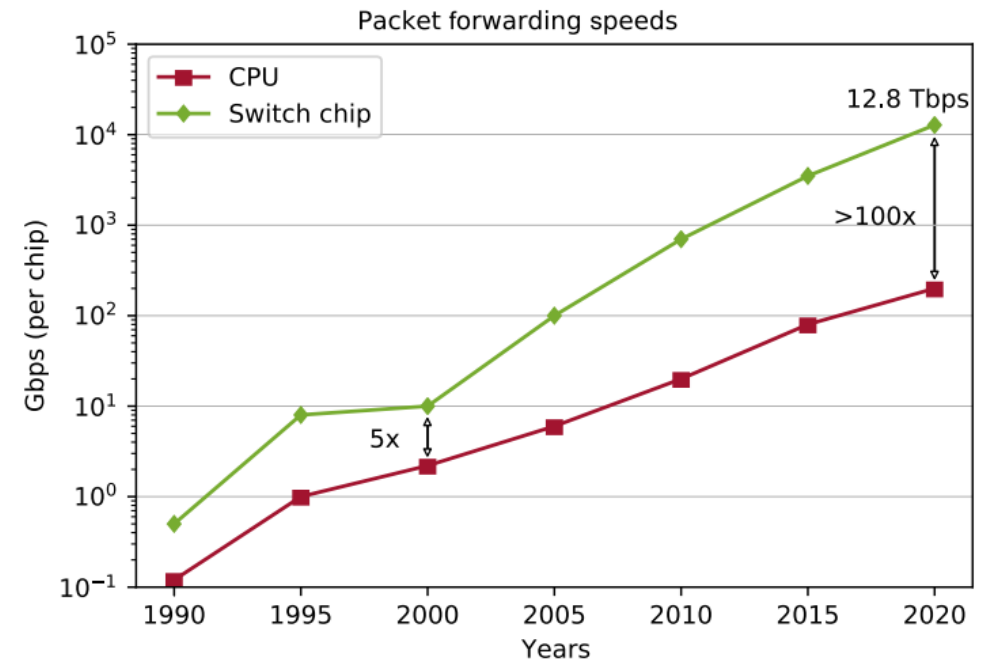
P4 Programmable Data Planes

- A P4¹ Programmable Data Planes (PDP) is a domain-specific processor for networking
- It enables the programmer to
 - define and parse new protocols
 - measure events with high precision (nanosecond resolution)
 - run custom applications at line rate



P4 Programmable Data Planes

- A P4¹ Programmable Data Planes (PDP) is a domain-specific processor for networking
- It enables the programmer to
 - define and parse new protocols
 - measure events with high precision (nanosecond resolution)
 - run custom applications at line rate

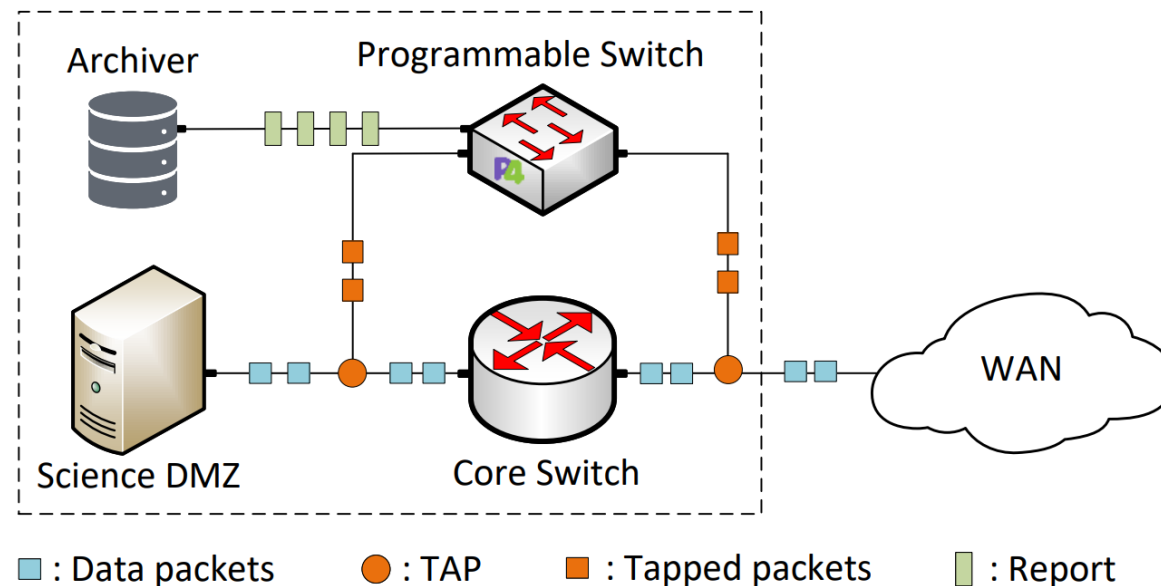


Evolution of the packet forwarding speeds¹

1. Reproduced from N. McKeown. Creating an End-to-End Programming Model for Packet Forwarding. Available: <https://www.youtube.com/watch?v=fiBuao6YZI0&t=634s>

Proposed System - Overview

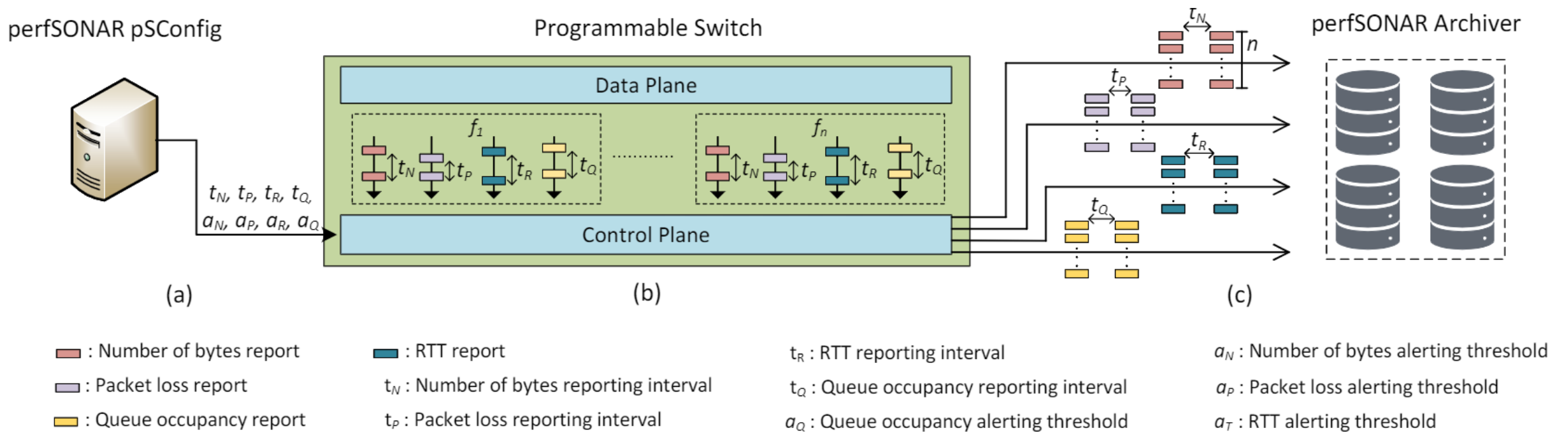
- The scheme uses optical passive taps¹ to mirror traffic which is then forward to a PDP
- The PDP
 - continuously generates fine-grained measurements at line rate (e.g., RTT, loss rate, throughput)
 - periodically reports to the archiver (perfSONAR) on a per-flow basis
 - introduces new measures (e.g., queue occupancy, packet interarrival time)
 - detects microbursts



1. Optical taps operate at the physical layer by splitting the light traveling in the fiber

Proposed System - Components

- The system has four components
 - Configuration unit
 - Data collection unit
 - Data extraction unit
 - Data storage unit



Interaction between the different components of the proposed system

Proposed System - Measurements

- Currently, the system computes the following per-flow statistics
 - Packet loss rate
 - RTT
 - Throughput
 - Queueing occupancy and queueing delay
 - Packet interarrival time

Proposed System - Measurements

- Currently, the system computes the following per-flow statistics
 - Packet loss rate
 - RTT
 - Throughput
 - Queueing occupancy and queueing delay
 - Packet interarrival time
- Based on the above statistics, other computations are executed in the control plane
 - Jain's fairness index¹
 - Link utilization

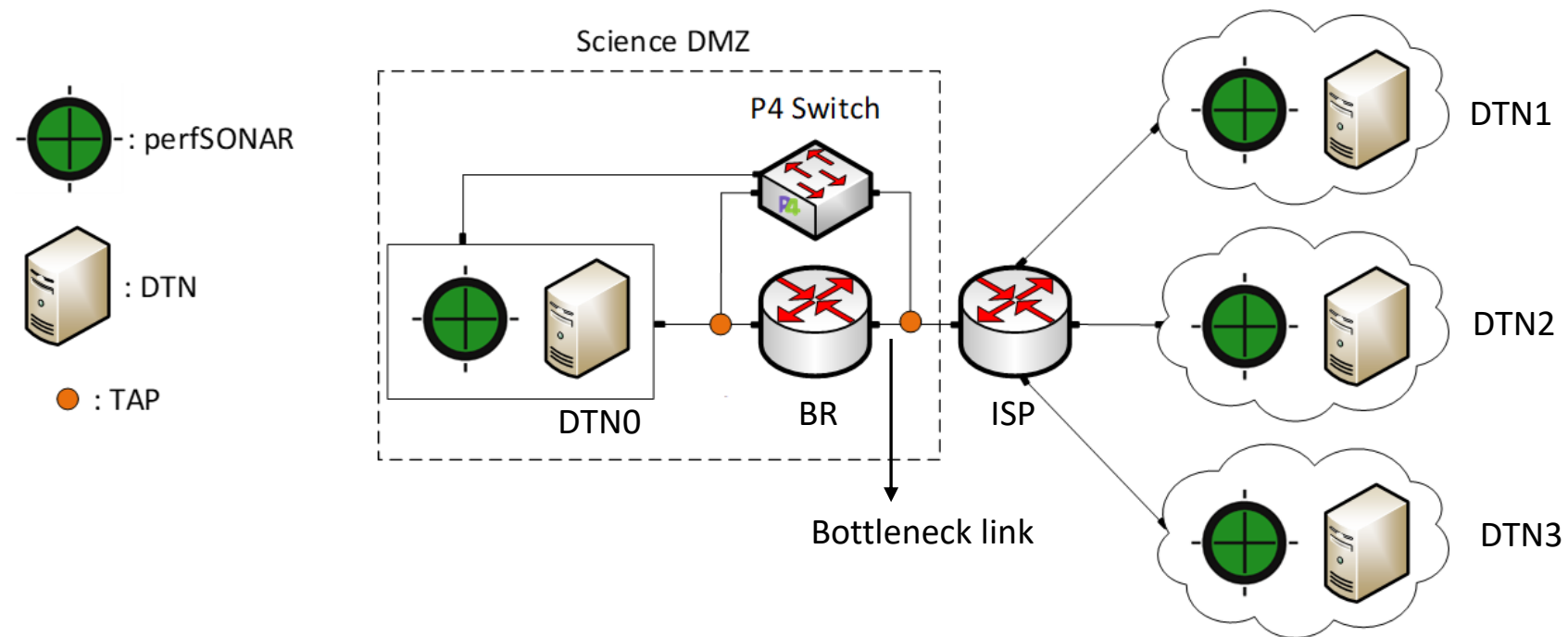
1. R. Jain, A. Duresi, and G. Babic, "Throughput fairness index: An explanation," in ATM Forum contribution, vol. 99, 1999.

Proposed System - Features

	P4-perfSONAR	perfSONAR	Comments
Measurement type	Active and passive measurements	Active measurement	Passive measurements do not induce overhead
Measurement source	Real traffic	Injected traffic	More accurate measurements are collected with real traffic
Granularity	Per-flow and per-packet granularity	Limited	P4-perfSONAR produces accurate, high-resolution measurements
Visibility	Real-time visibility over all data transfers	Limited by active tests	P4-perfSONAR provides high visibility
Microburst detection	Supported	Not supported	P4-perfSONAR monitors the router's buffer
End-point failures	Supported	Not supported	P4-perfSONAR can track every flow in the network and identify anomalies on individual flows

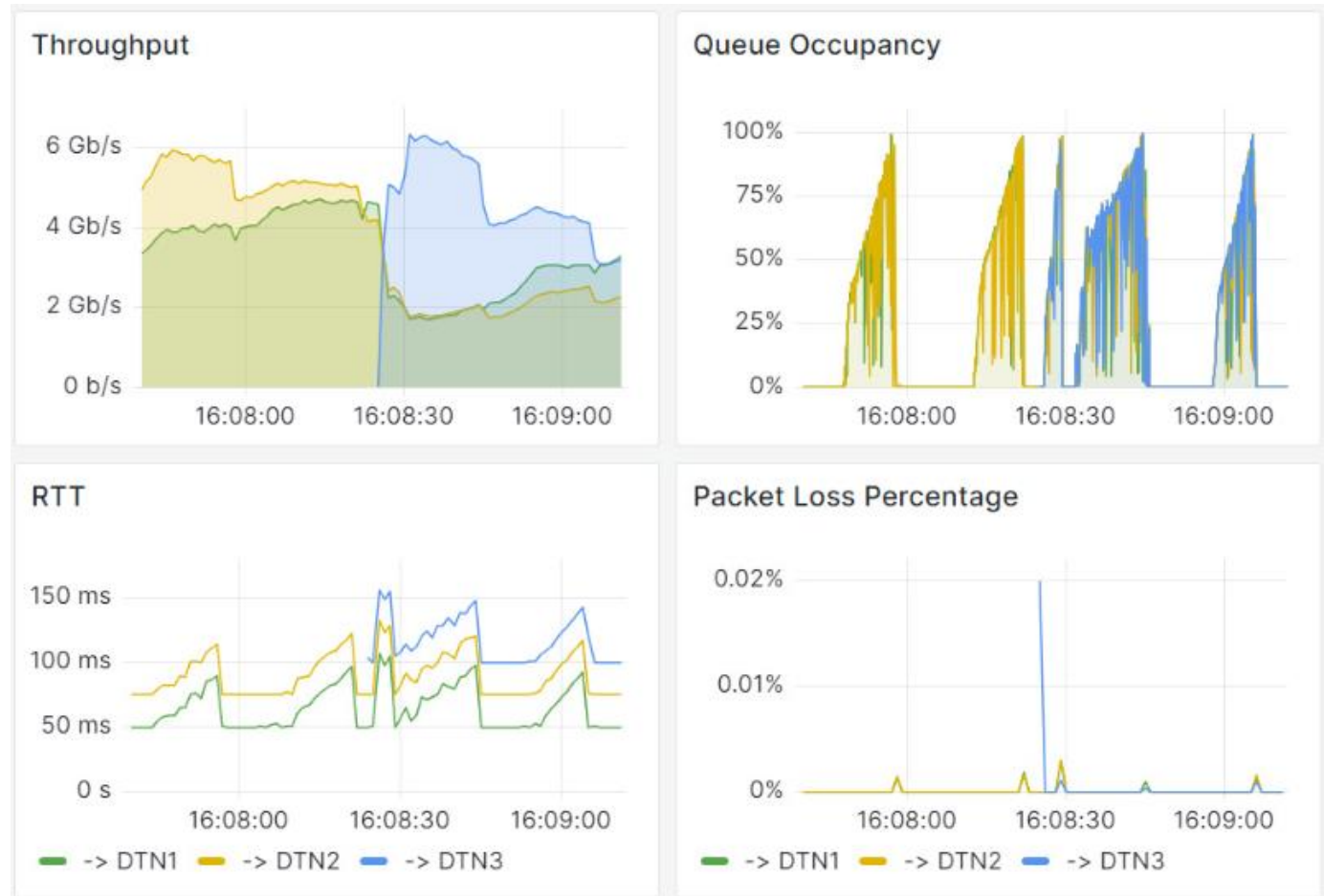
Experimental Setup

- The topology consists of a Science DMZ connected to a WAN (emulated w/ NETEM)
- The BR and ISP routers are Juniper MX 204
- The optical TAPs copy the traffic at the ingress and egress interfaces of the BR, and forward the copy to a P4 switch
- The capacity of the bottleneck link is 10 Gbps



Results – Per-flow Monitoring

- At t=0, there are two flows: between DTN0 and DTN1, and between DTN0 and DTN2
- At t=16:08:25, another flow is introduced, between DTN0 and DTN3
- The propagation delays are:
 - DTN0-DTN1: 50ms
 - DTN0-DTN2: 75ms
 - DTN0-DTN3: 100ms



Results – Fairness and Link Utilization

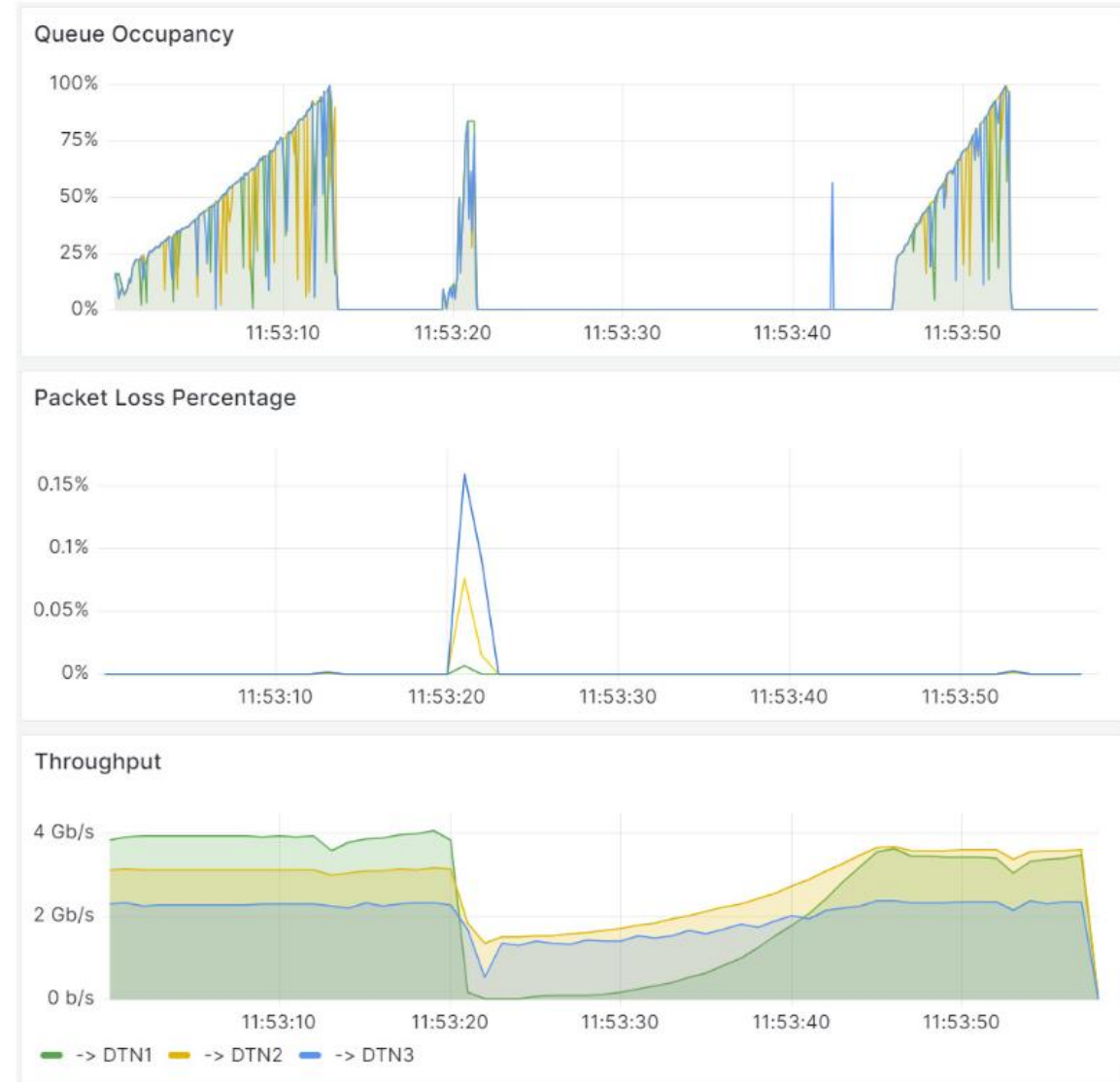
- Link utilization is computed as the aggregated throughput over the link capacity, in percentage (this measure is for the bottleneck link)
- Fairness is given by the Jain's fairness index¹
 - A totally fair system has an index of 1 and a totally unfair system has an index of 0



1. R. Jain, A. Duresi, and G. Babic, "Throughput fairness index: An explanation," in ATM Forum contribution, vol. 99, 1999.

Use-Case 1: Small Buffer

- In this use case, the BR router is configured with a small (BDP/4)¹
- The propagation delays between DTN0 and the remaining DTNs are 50ms (DTN1), 100ms (DTN2), and 150ms (DTN3)
- A traffic burst is generated at t=11:53:19



1. BDP stands for bandwidth-delay product. It is computed the average propagation delay, multiplied by the link capacity of the bottleneck link

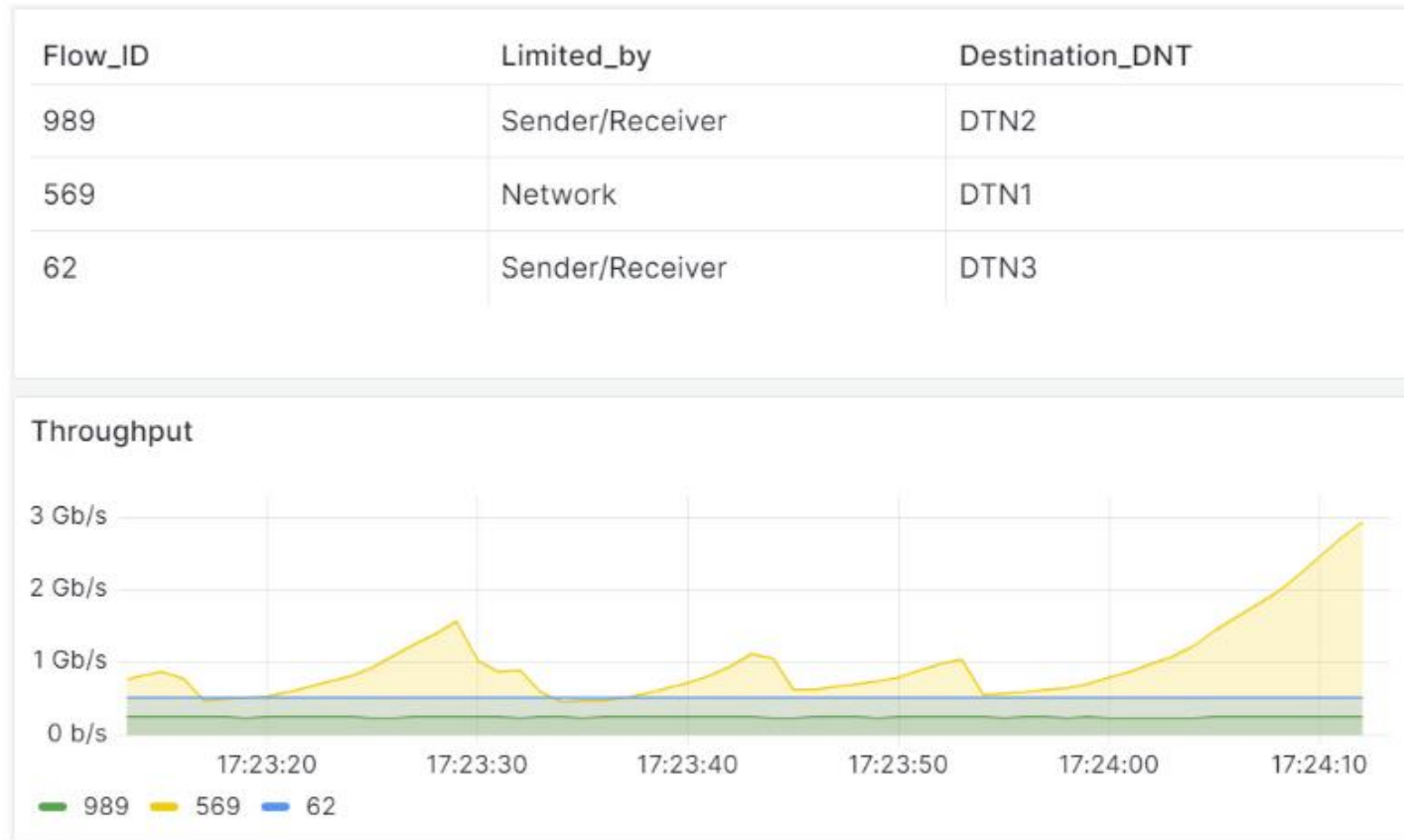
Use-Case 2: Network- or Host-limited Throughput

- In this use case, three tests are performed. The results visualized in perfSONAR help troubleshoot the problems
 - Test 1: the throughput is limited by the small TCP buffer configured at the receiver (DTN0->DTN2)
 - Test 2: the throughput is limited by the network, which excessively drops packets (0.01%) (DTN0->DTN1)
 - Test 3: the throughput is limited by low maximum sending rate at the sender (500 Mbps) (DTN0->DTN3)

Test 1

Test 2

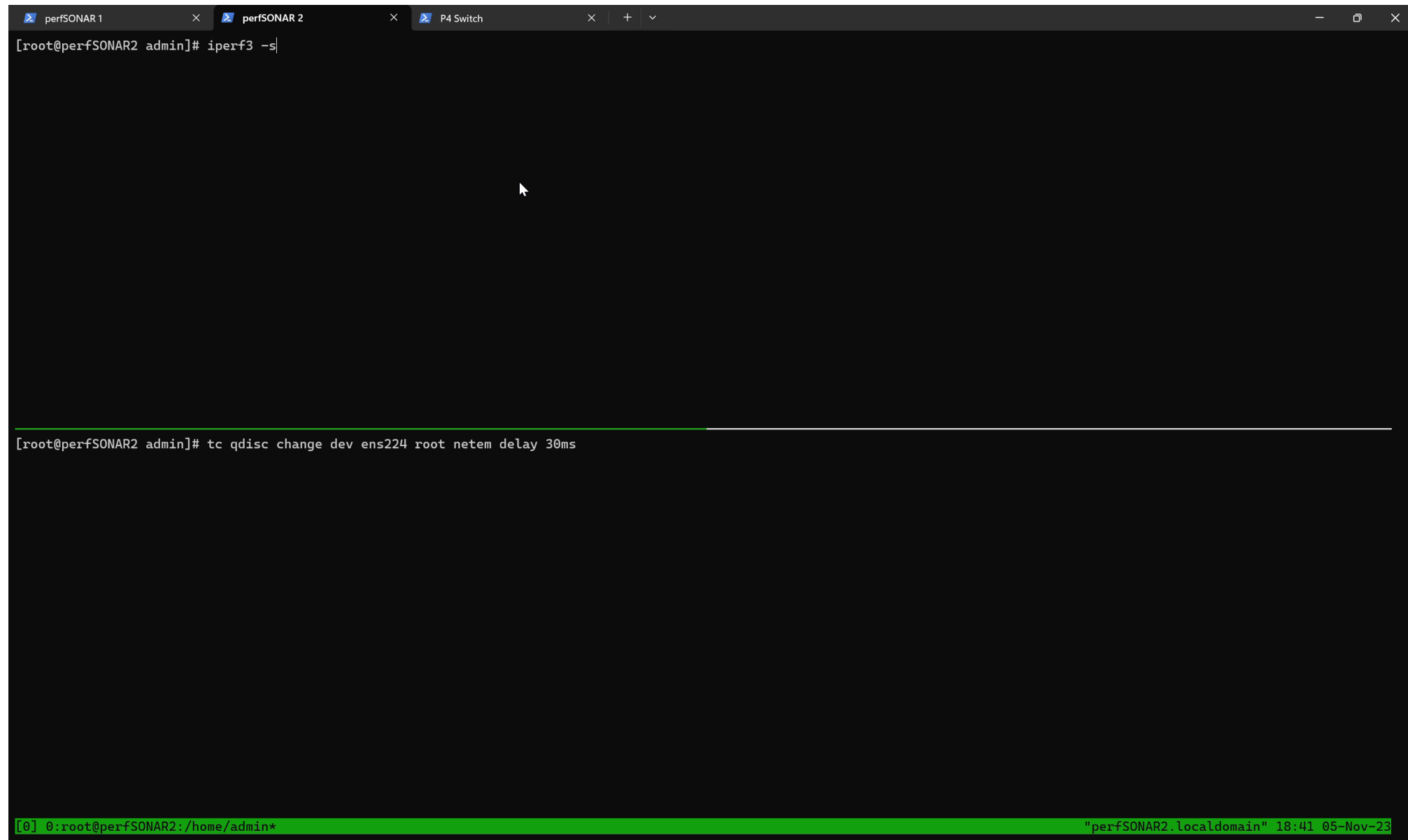
Test 3



Use-Case 3: Adaptive Reporting Rate by the Data Plane

- Instead of enforcing the P4 switch to report measurements periodically to the archiver (perfSONAR), the P4 switch adaptively sends reports
 - Reporting updates are not needed if measurements do not change significantly
- The system uses the Linear Prediction Coefficient¹

Use-Case 3: Adaptive Reporting Rate by the Data Plane



A terminal window with three tabs: 'perfSONAR 1', 'perfSONAR 2', and 'P4 Switch'. The active tab is 'perfSONAR 2'. The prompt is '[root@perfSONAR2 admin]#'. The first command entered is 'iperf3 -s'. The second command, entered on a new line, is 'tc qdisc change dev ens224 root netem delay 30ms'. The terminal background is black with white text. A green status bar at the bottom shows the shell prompt '0:root@perfSONAR2:/home/admin*' and the system time '18:41 05-Nov-23'.

```
[root@perfSONAR2 admin]# iperf3 -s  
  
[root@perfSONAR2 admin]# tc qdisc change dev ens224 root netem delay 30ms
```

[0] 0:root@perfSONAR2:/home/admin* "perfSONAR2.localdomain" 18:41 05-Nov-23

Conclusion

- This presentation described an extension of perfSONAR with P4 PDP switches
- The P4 PDP switches provide per-packet visibility, fine-grained measurements, and line-rate computation
- The scheme augments perfSONAR by tracking flows individually, providing high-resolution measurements, and operating over passive traffic
- The system enables the programmer to compute new statistics such as queueing delay, fairness, link utilization, etc.
- By using an offline PDP switch operating over a copy of the traffic, the system fosters an incremental use of this technology (no need to deploy complex P4 code at once)
- Future work may include deploying the scheme on production networks (ISPs, campus networks)



UNIVERSITY OF
South Carolina



This work is supported by NSF award number 2118311

For additional information, please refer to
<https://research.cec.sc.edu/cyberinfra>

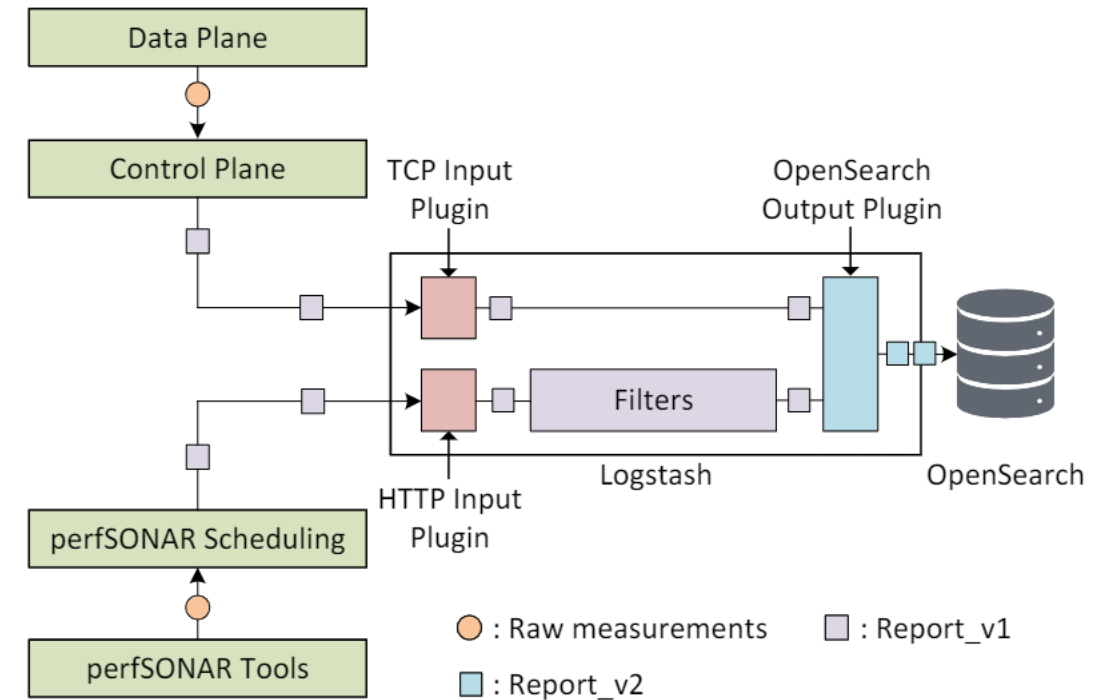
Email: amazloum@email.sc.edu, jcrichigno@cec.sc.edu

Proposed System - Features

- The system is seamlessly integrated to perfSONAR

- `psconfig config-P4 --metric throughput --samples_per_second 1`
- `psconfig config-P4 --metric RTT --samples_per_second 2`
- `psconfig config-P4 --metric queue_occupancy --alert --threshold 30 --samples_per_second 10`

Configuration example



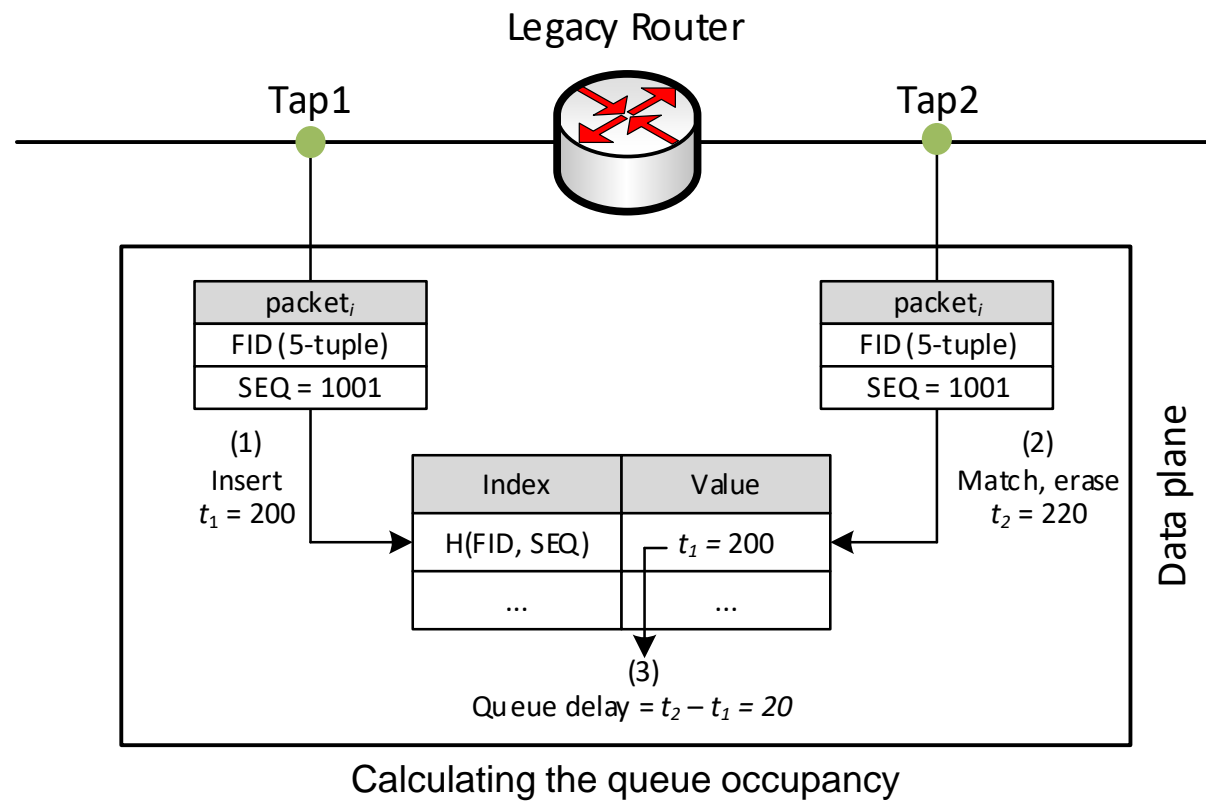
Connecting the proposed system with perfSONAR's archiver

Throughput Monitoring

- Packets are grouped into flows using the 5-tuple
- The total number of bytes of each flow is updated on a per-packet basis
- The control plane extracts the number of bytes and calculates the throughput

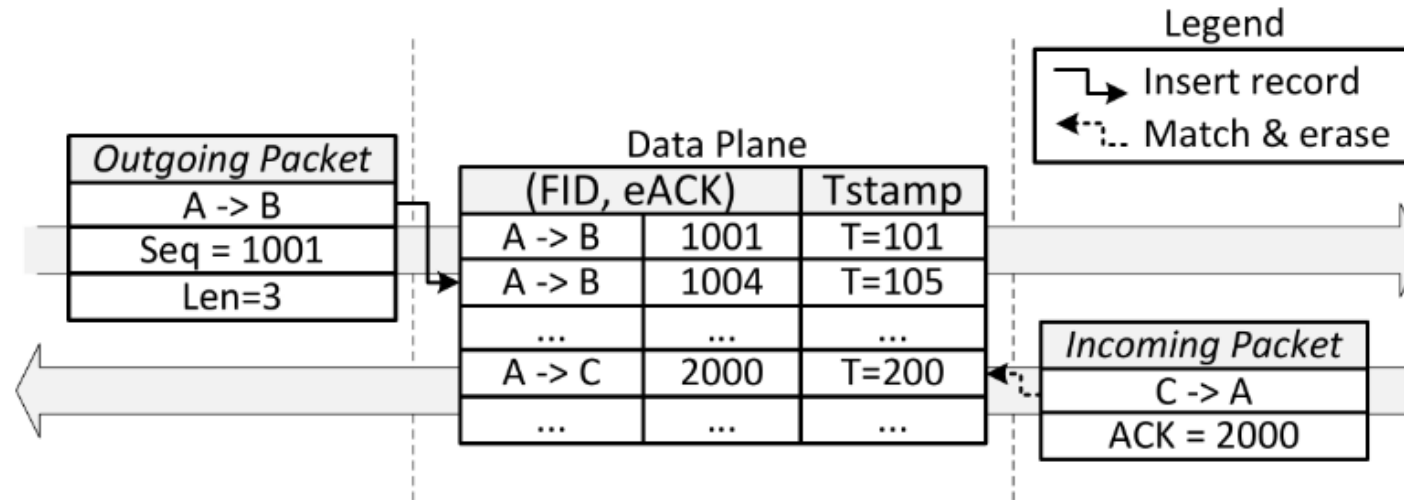
Queue Occupancy Monitoring

- The queueing occupancy is calculated by leveraging the precise timestamp of the hardware switch (nanosecond resolution)



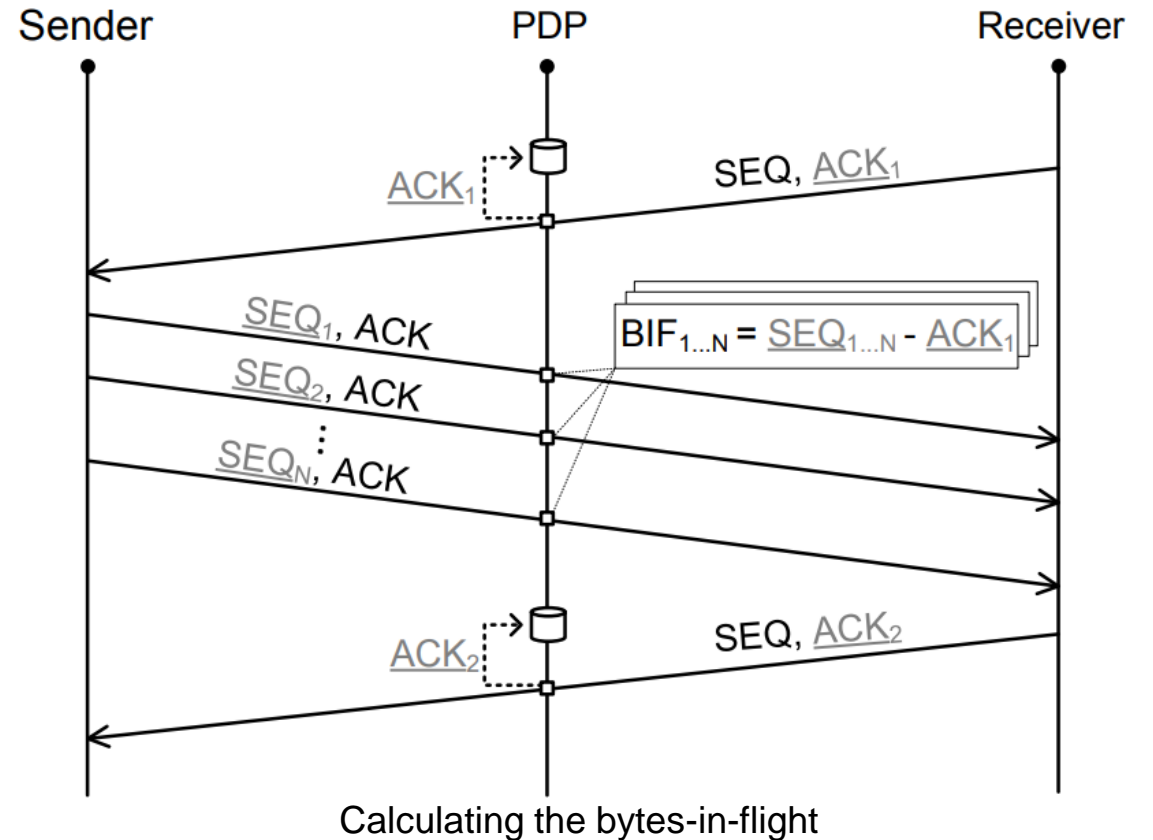
RTT Calculation

- RTT is calculated for flows by monitoring the time difference between receiving a packet and its acknowledgment
- Packet loss is calculated by counting retransmissions



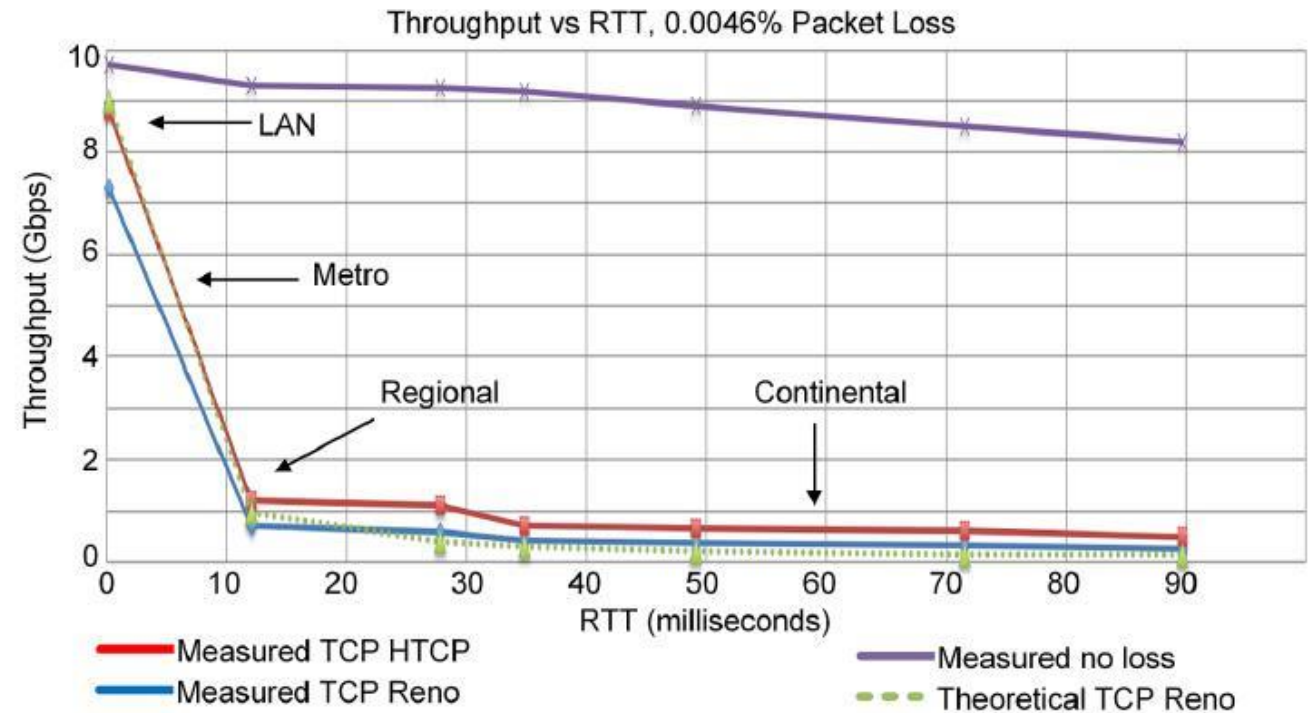
Detecting Flows not Constrained by the Network

- Bytes-in-flight (BIF) is the amount of data sent but not yet acknowledged
- The bytes in flight of connections limited by the network will continue increasing until a packet loss occurs
- If packet loss does not constrain the increase in the bytes in flight, then the problem is at the sender/receiver



Science DMZ

- The Science DMZ is a network designed for big science data¹
- Main elements:
 - High throughput, friction-free WAN paths
 - Security tailored for high speeds
 - Data Transfer Nodes (DTNs)
 - **End-to-end monitoring / perfSONAR**



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

Proposed System - Overview

- The system visualizes network telemetry produced by P4 PDP switches
- The reasons for extending perfSONAR rather than creating a new application include:
 - perfSONAR has been deployed worldwide (campus networks, service providers)
 - perfSONAR already provides APIs to add functionality to it
 - the proposed extension complements the current functionalities of perfSONAR