perfSONAR: Enhancing Data Collection through Adaptive Sampling

Sergio Elizalde, Ali Mazloum
College of Engineering and Computing, University of South Carolina
https://research.cec.sc.edu/cyberinfra/

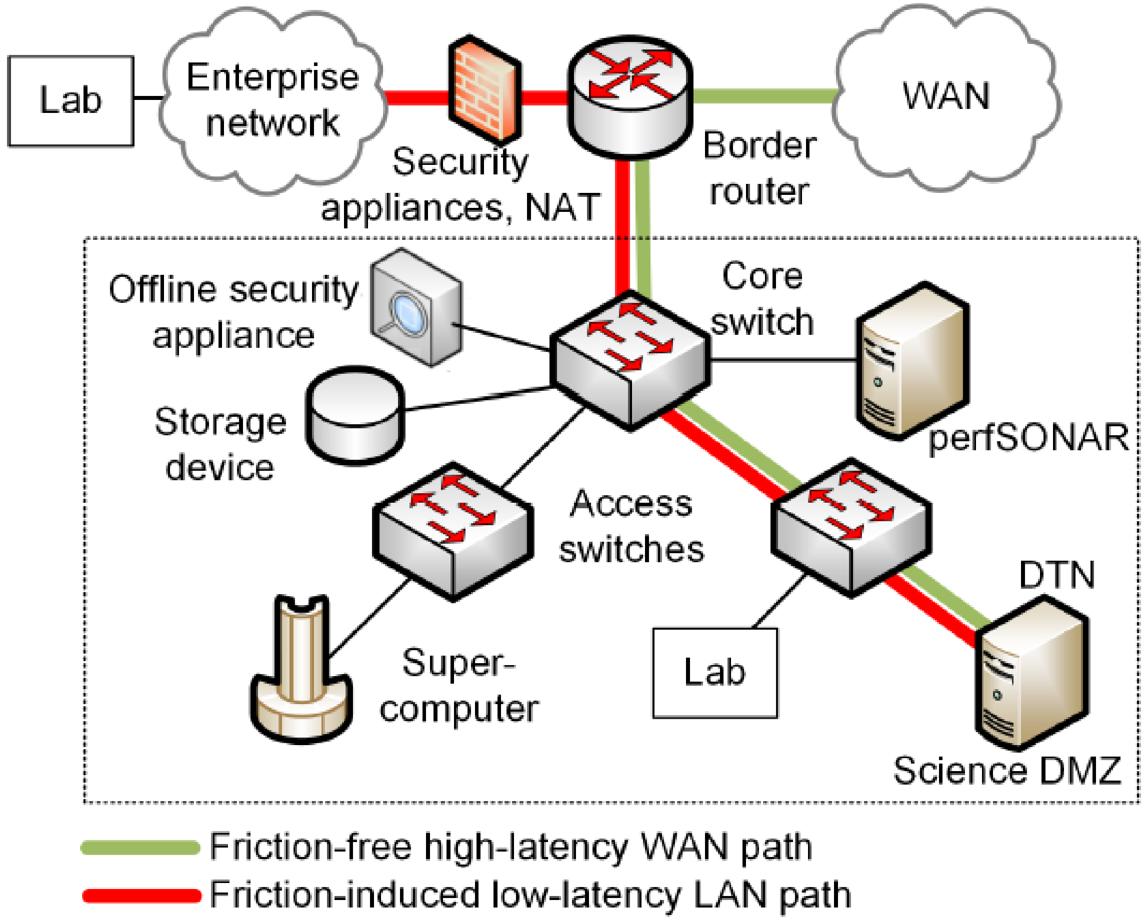
GMI-AIMS-5
February, 2025
San Diego, California

Agenda

- Background information on Science DMZ, perfSONAR, P4, Linear Prediction
- Proposed system
- Experimental setup
- Results
- Demo
- Conclusion

Science DMZ

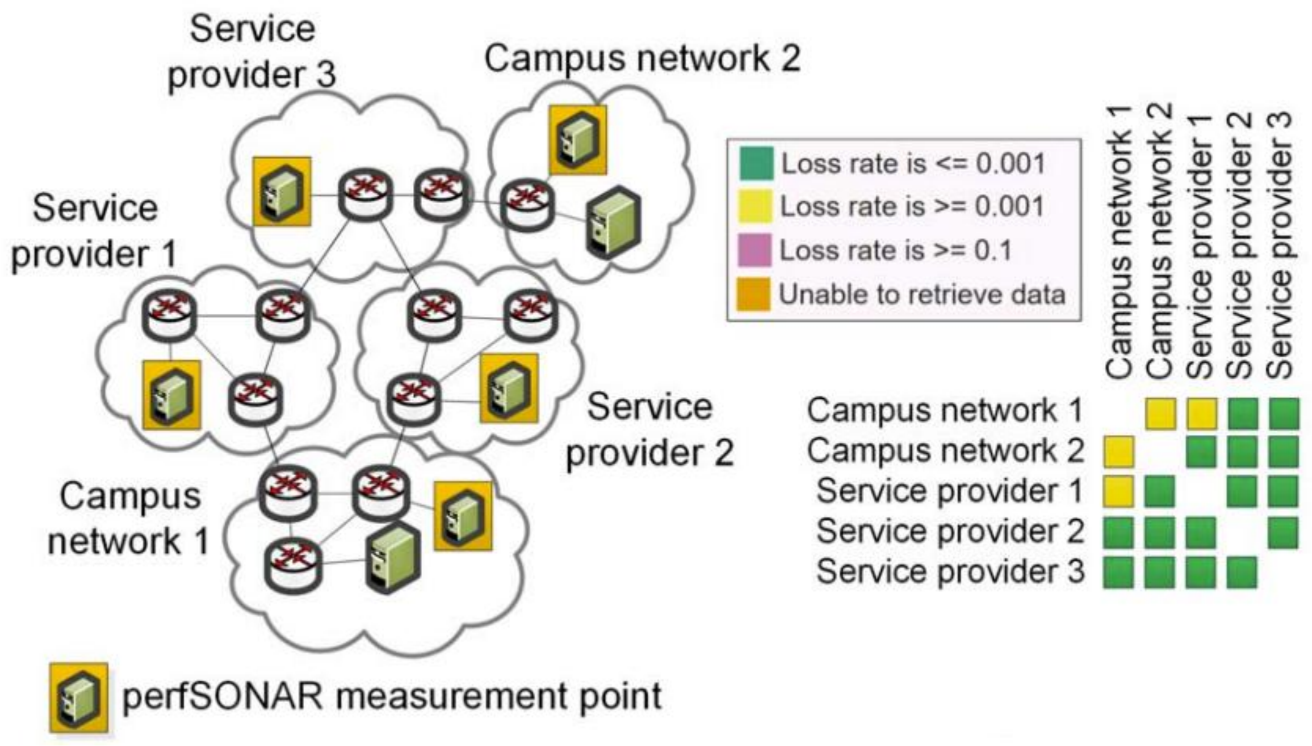
- The Science DMZ is a network designed for big science data transfers on WANs[1]
- Main elements:
  - High throughput, friction-free WAN paths
  - Security tailored for high speeds
  - Data Transfer Nodes (DTNs)
  - **End-to-end monitoring / perfSONAR**



[1]E. 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.

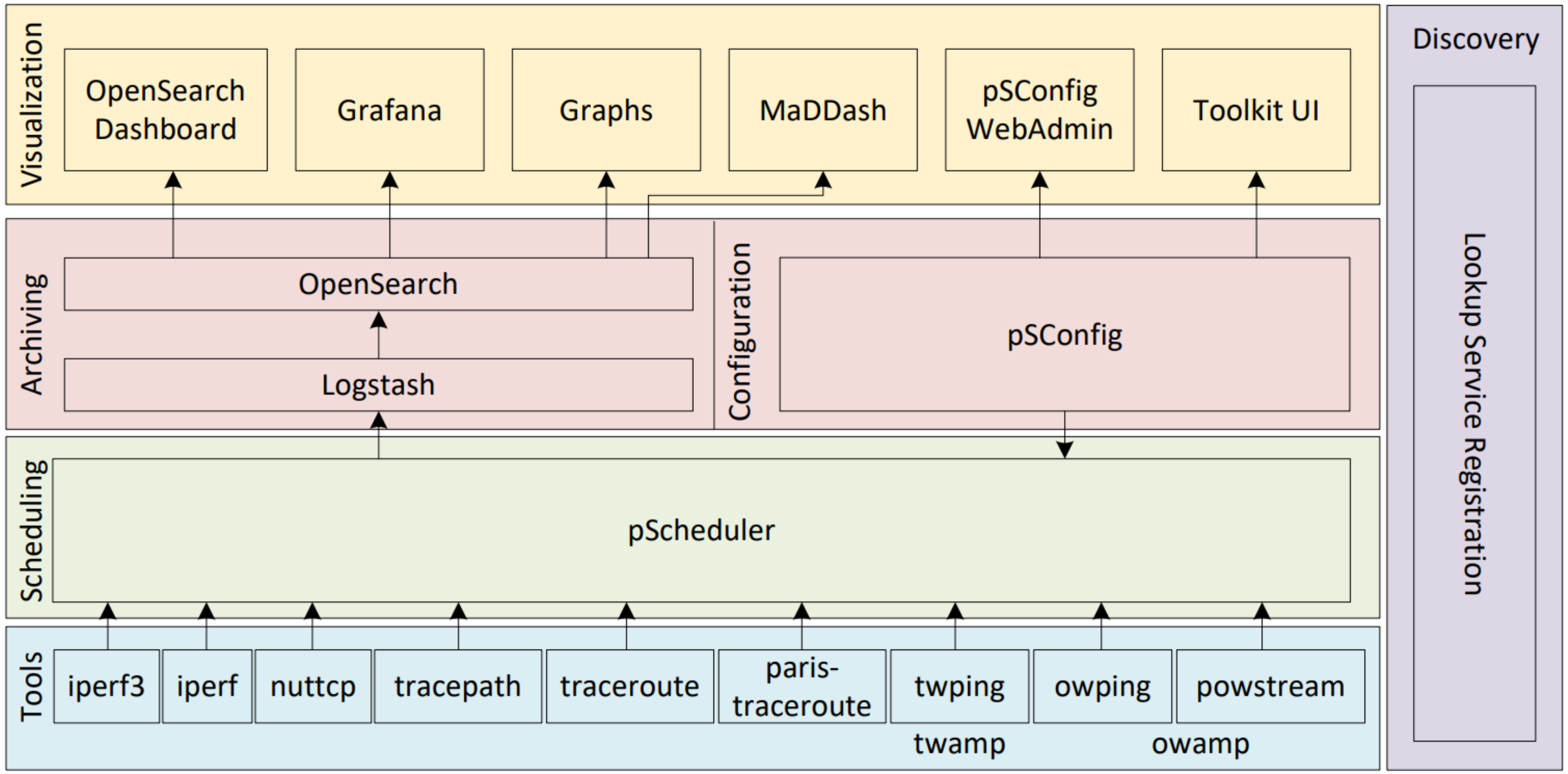
perfSONAR

- perfSONAR is a measurement tool for end-to-end paths
- It helps troubleshoot performance issues (e.g., finding soft failures)
- perfSONAR is a key component of the Science DMZ



perfSONAR

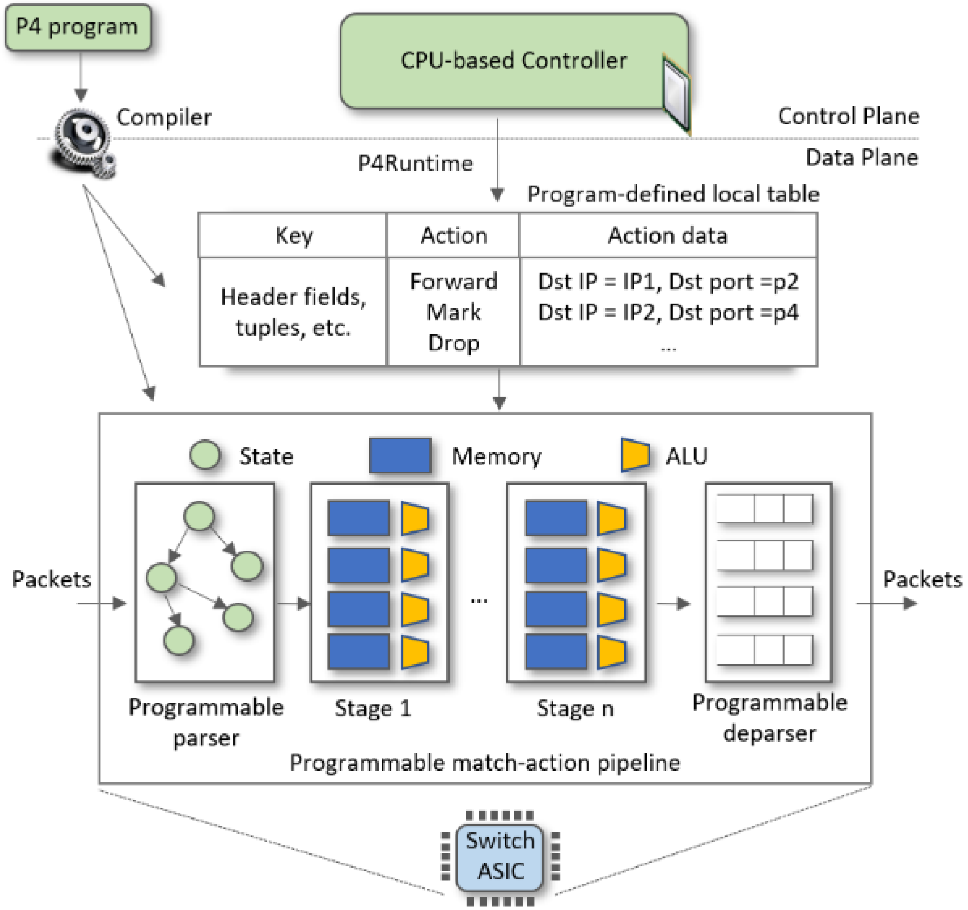
- perfSONAR
  - relies on tools that provide coarse-grained measurements
  - depends on active measurements
  - provides APIs that enable a programmer to extend its functionality



perfSONAR architecture

P4 Programmable Data Planes

- A P4[1] 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



1. P4 stands for stands for Programming Protocol-independent Packet Processors

Linear Prediction Method

- The Linear Prediction method (LP) enables a system to predict the future value of a variable based on the observed values
- A future value is calculated by using the last observed sample and the average rate of change of the last *N-1* samples[1]:

$$x_{p_{N+1}} = x_N + \frac{\Delta T_{current}}{N-1} \sum_{i=1}^{N-1} \frac{x_{i+1} - x_i}{t_{i+1} - t_i}$$

Where:

- $x_{p_{N+1}}$ is the predicted value for the next sample
- x_N is the most recent (actual) sample
- N is the number of (actual) samples considered so far
- x_i *and* t_i represent the value and arrival time of the i*th* sample, respectively
- $\Delta T_{current}$ is the time gap between the most recent sample and the next sample

[1]E.A. Hernandez, M. C. Chidester, and A. D. George, "Adaptive sampling for network management," Journal of Network and Systems Management, vol. 9, pp. 409–434, 2001

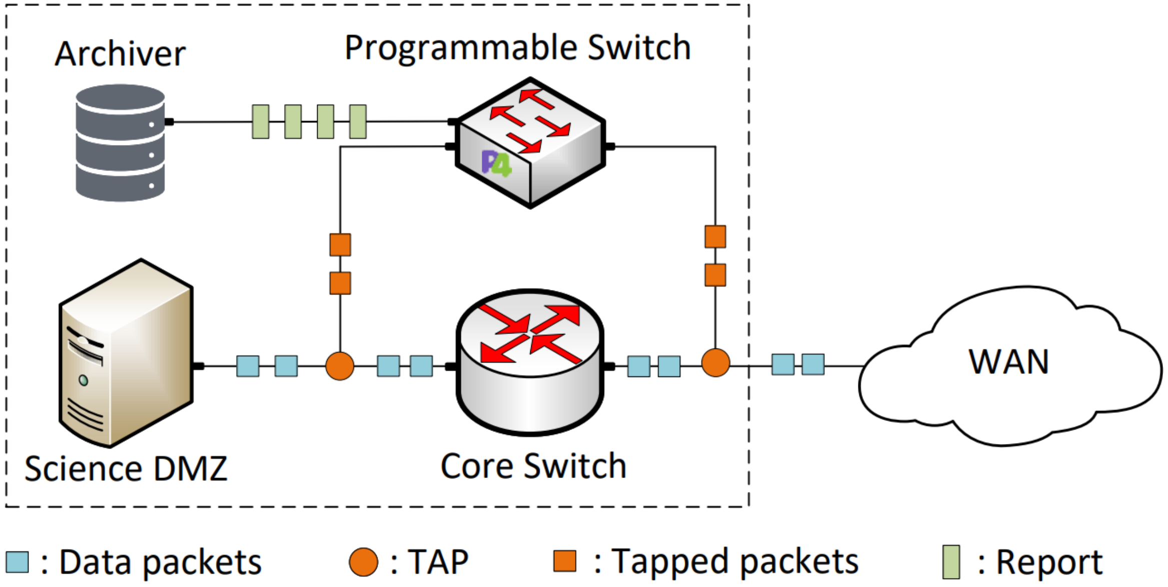
Goal of the Proposed Application

- Extend perfSONAR with P4 switches to
  - enable per-packet visibility
  - track flows individually
  - compatible with current Science DMZ / non-programmable devices
  - produce high-resolution measurements while **minimizing the reporting rate**

	perfSONAR	P4-perfSONAR	Comments
Measurement type	Active measurement	Active and passive measurements	Passive measurements do not introduce overhead
Measurement source	Injected traffic	Real traffic	More accurate measurements are collected with real traffic
Granularity	Limited	Per-flow and per-packet granularity	P4-perfSONAR produces accurate, high-resolution measurements
Visibility	Limited by active tests	Real-time visibility over all data transfers	P4-perfSONAR provides high visibility
Reporting rate	Fixed	Adaptive	Reporting rate adapts according to the input signal

Proposed System - Overview

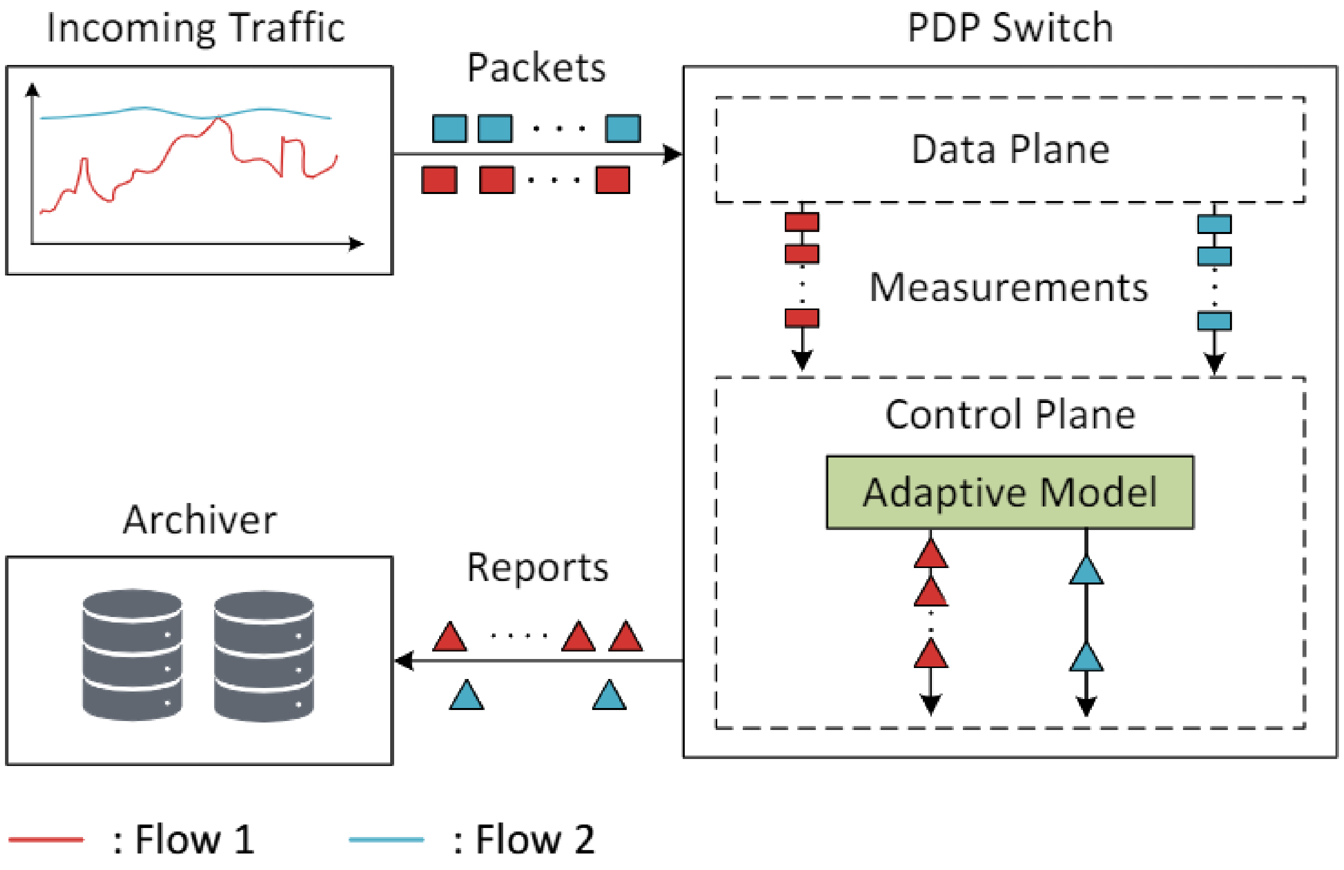
- The scheme uses optical passive taps[1] to mirror traffic which is then forward to a PDP
- The PDP
  - ➢ continuously generates fine-grained measurements at line rate (e.g., RTT, loss rate, throughput)
  - ➢ selectively reports to the archiver (perfSONAR) on a per-flow basis
  - ➢ introduces new measures (e.g., queue occupancy, packet interarrival time)



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

Proposed System - Overview

- An adaptive model runs on the control plane and selectively reports the measurements to the archiver
- The higher the variation in the observed samples, the higher the reporting rate



Proposed System – Adaptive Model

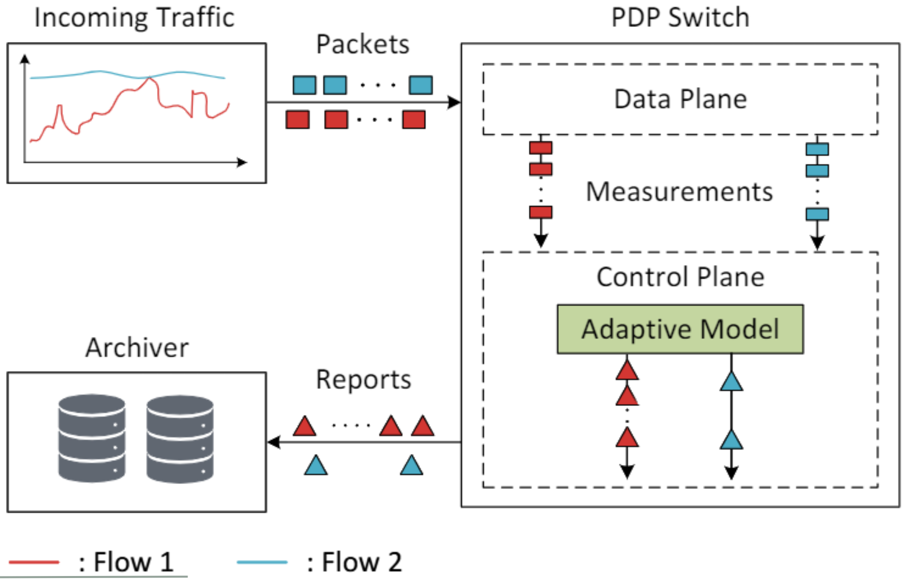
- The control plane is configured to extract measurements at a constant rate
- By considering a constant extraction rate, the prediction can be simplified as:

$$xp_{N+1} = x_N + \frac{x_N - x_1}{N - 1}$$

- After extracting the *N+1* sample, the error in prediction is calculated as follows[1]:

$$m = \frac{xp_{N+1}}{x_{N+1}}$$

- A prediction is considered accurate if $1 - \sigma < m < 1 + \sigma$ such that $0 < \sigma < 1$



[1]E.A. Hernandez, M. C. Chidester, and A. D. George, “Adaptive sampling for network management,” Journal of Network and Systems Management, vol. 9, pp. 409–434, 2001

Proposed System - Measurements

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

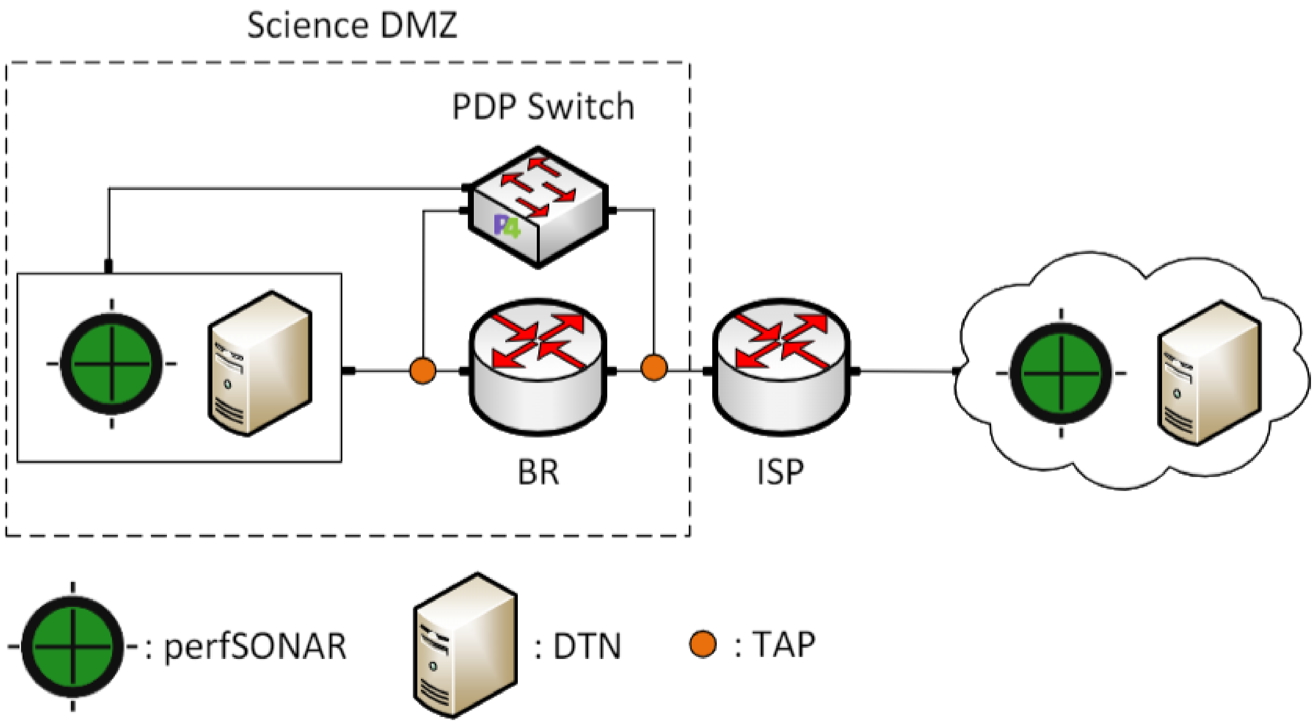
Proposed System - Measurements

- 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 index[1]
  - Link utilization

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

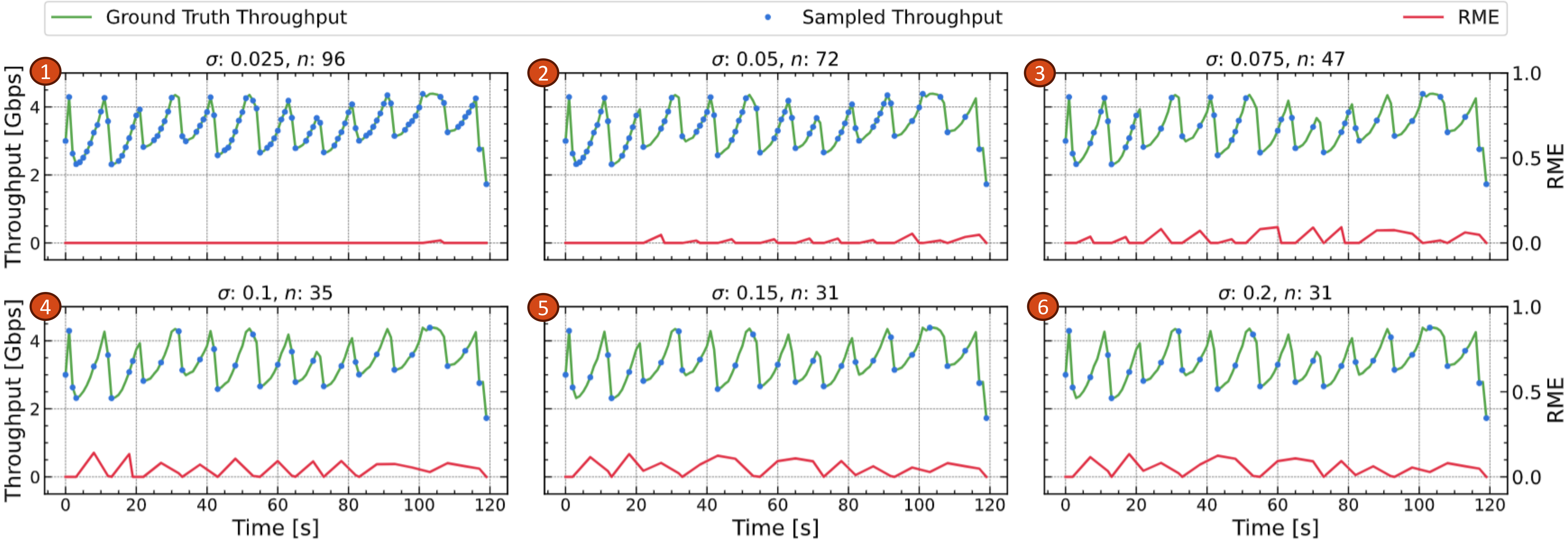
Experimental Setup

- 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 PDP switch



Results – Evaluating the Impact of σ

- This experiment aims to evaluate the impact of σ on the RME and the number of reported samples
- The larger the value of σ, the larger the tolerance to error
- A two-minute test is performed between the DTNs, and six values of σ are used



Demo 1
Throughput Measurements

Demo 2
Throughput and RTT Measurements

perfSONAR 1 | perfSONAR 2 | P4 Switch

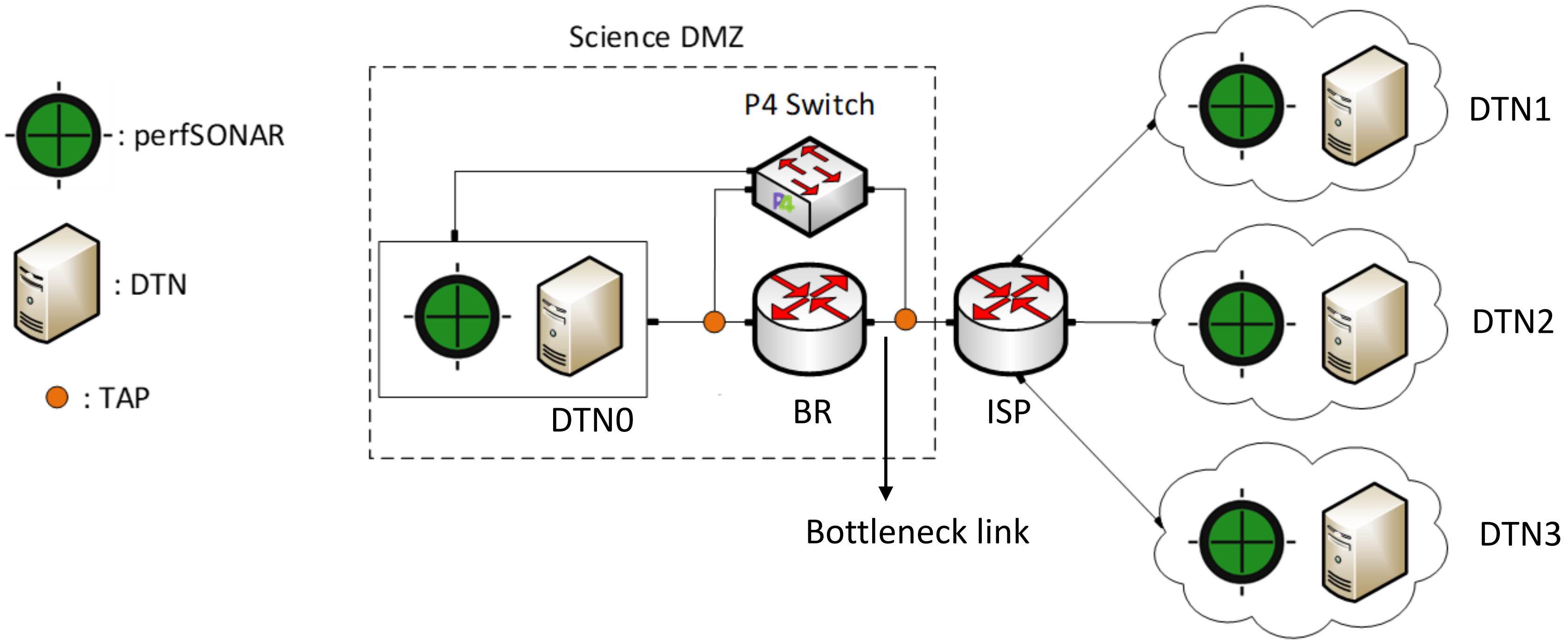
```
[root@perfSONAR2 admin]# iperf3 -s
```

```
[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
```

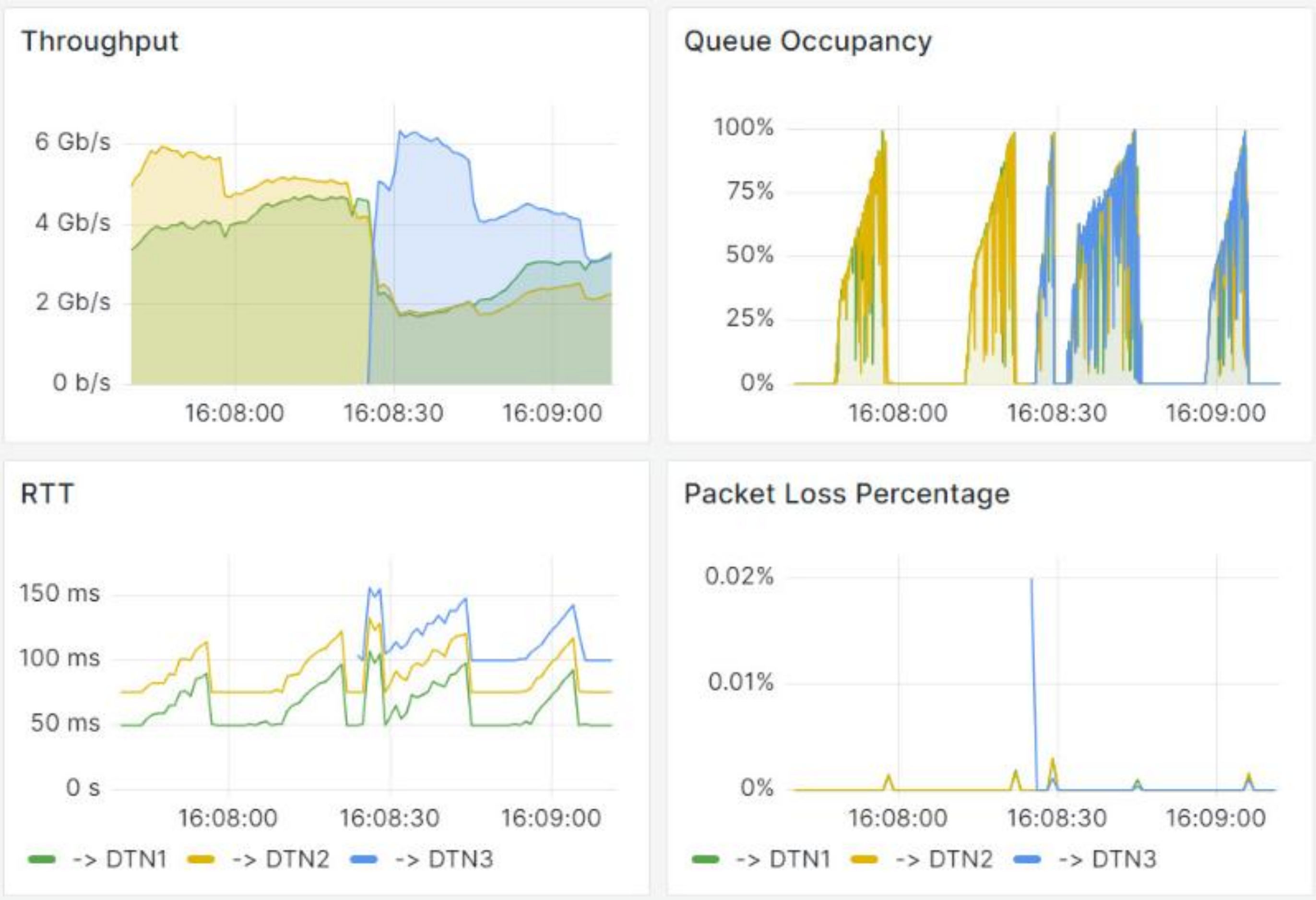
Experimental Setup 2

- 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



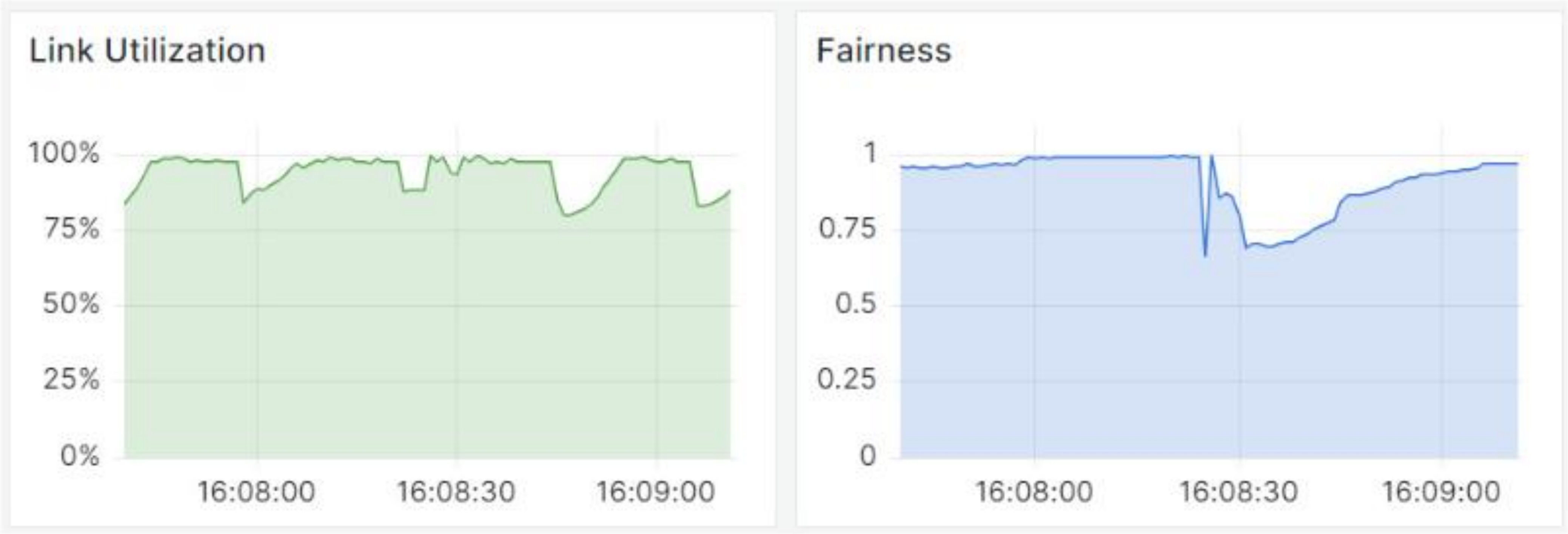
Results – Per-flow Monitoring

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



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

Conclusion

- This presentation described an extension of perfSONAR with P4 switches
- The P4 switches provide per-packet visibility and line-rate computation
- The scheme augments perfSONAR by tracking flows individually, providing high-resolution measurements, and operating over passive traffic
- The system uses an adaptive model that selectively reports measurements

	perfSONAR	P4-perfSONAR	Comments
Measurement type	Active measurement	Active and passive measurements	Passive measurements do not introduce overhead
Measurement source	Injected traffic	Real traffic	More accurate measurements are collected with real traffic
Granularity	Limited	Per-flow and per-packet granularity	P4-perfSONAR produces accurate, high-resolution measurements
Visibility	Limited by active tests	Real-time visibility over all data transfers	P4-perfSONAR provides high visibility
Reporting rate	Fixed	Adaptive	Reporting rate adapts according to the input signal

This work was supported by NSF award number 2118311

Article: https://tinyurl.com/273dfakn

Ali Mazloum, Ali AlSabeh, Elie F. Kfoury, Jorge Crichigno
College of Engineering and Computing, University of South Carolina
https://research.cec.sc.edu/cyberinfra/