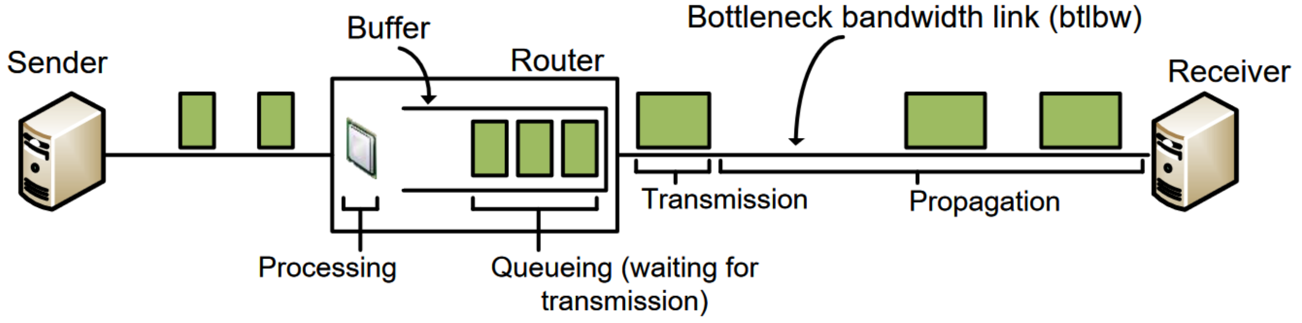
P4BS: Leveraging Passive Measurements from P4 Switches to Dynamically Modify a Router's Buffer Size

Elie Kfoury, **Jose Gomez**, Ali AlSabeh, Jorge Crichigno
College of Engineering and Computing, University of South Carolina
http://ce.sc.edu/cyberinfra/

Intel Headquarters - Santa Clara, CA
April 24-25, 2023

Introduction

- The size of a router's buffer has implications on the network performance
- Large buffer → excessive delays
- Small buffer → packet losses, low link utilization
- *How big should the buffer be*?

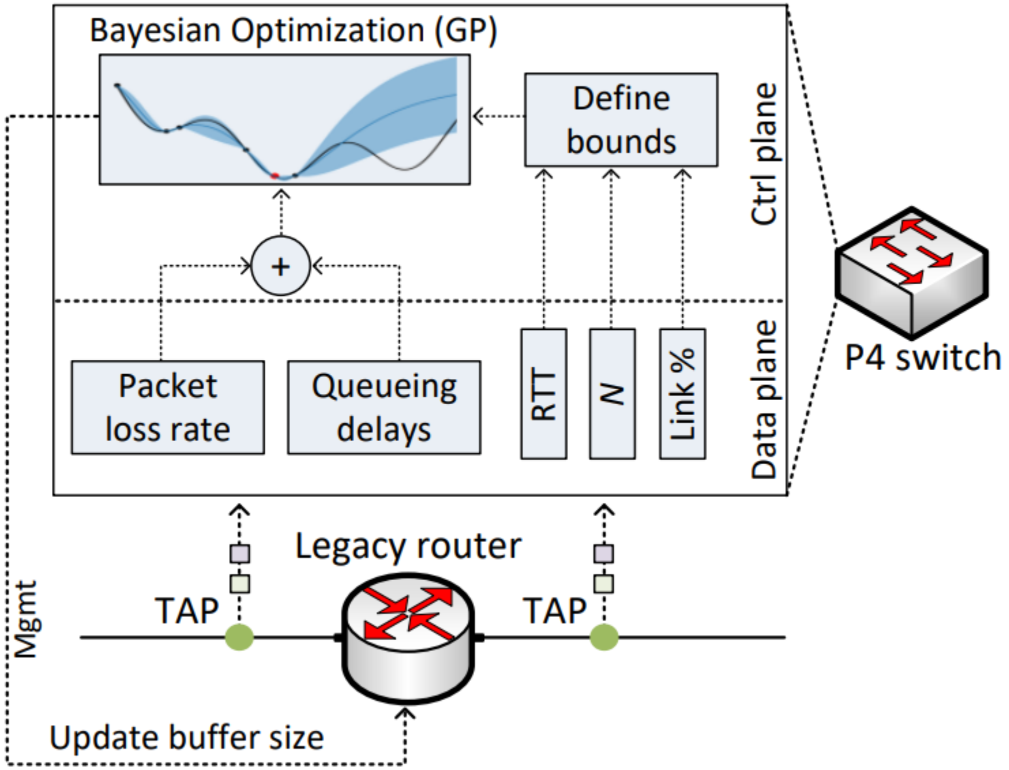


Static Buffer Rules

- Bandwidth-delay product
  - Buffer = $C * RTT$
  - C is the capacity of the port and RTT is the average round-trip time
- Stanford rule:
  - Buffer = $\frac{C * RTT}{\sqrt{N}}$
  - N is the number of long (persistent over time) flows traversing the port
- Operator hardcodes the buffer size based on typical traffic

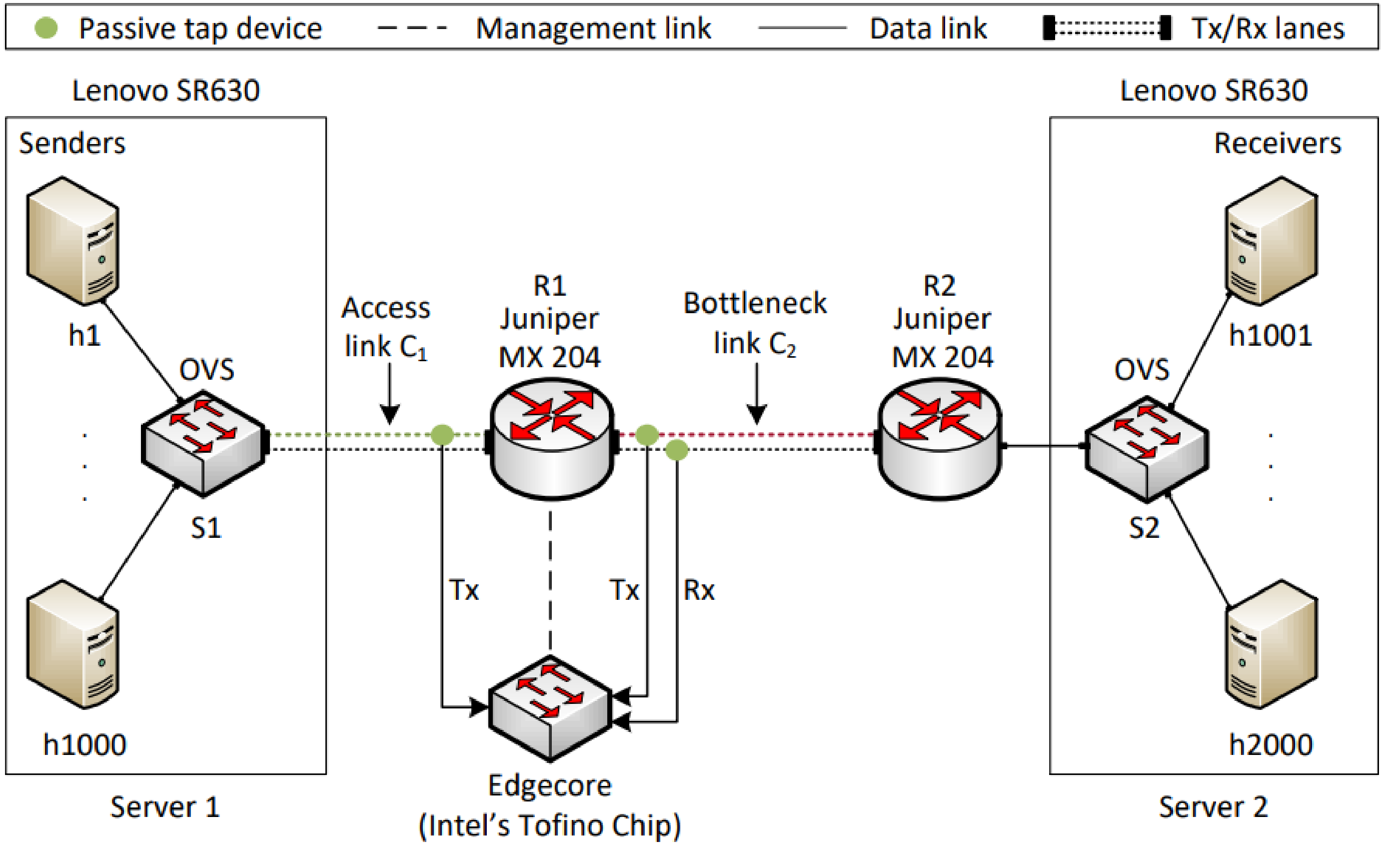
Proposed System

- The buffer size is dynamically modified
- A P4 switch is deployed passively to compute:
  - ➢ Number of long flows
  - ➢ Average RTT
  - ➢ Queueing delays
  - ➢ Packet loss rates
- The control plane sequentially searches for a buffer that minimizes delays and losses
- The searching algorithm is Bayesian Optimization (BO) with Gaussian Processes



Evaluation

- 1000 senders
- P4 switch: Wedge100BF-32X with Intel's Tofino ASIC
- Legacy router: Juniper router MX-204
- Different congestion control algorithms
- Access network:
  - C_1 = 40Gbps, C_2 = 1Gbps
- Core network:
  - C_1 = 10Gbps, C_2 = 2.5Gbps



Results

- Combined metric accounting for packet loss and delay [0, 1] (the lower, the better)
- Top heatmaps: access network
- Bottom heatmaps: core network
- The Mixed scenario combines multiple congestion control algorithms[1]

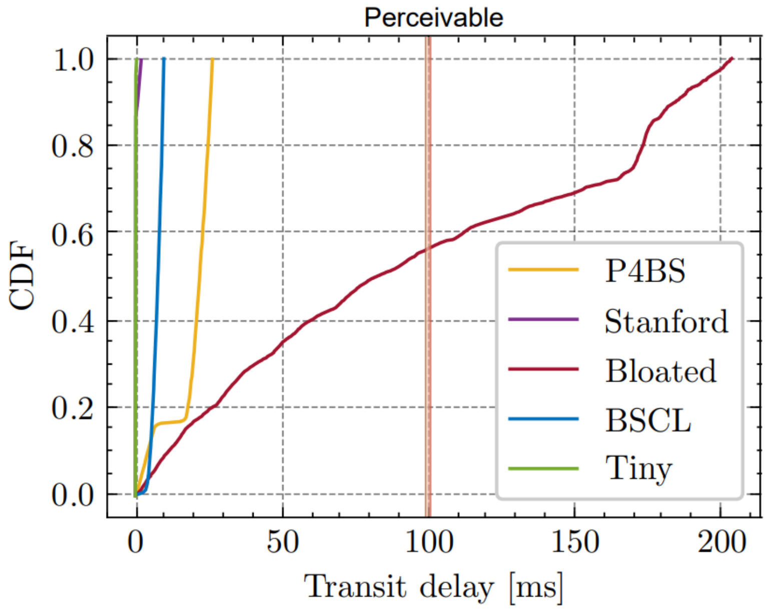
	Tiny				Stanford				BSCL				BDP				Bloated				ADT				P4BS			
N	50	100	250	500	50	100	250	500	50	100	250	500	50	100	250	500	50	100	250	500	50	100	250	500	50	100	250	500
Vegas	0.13	0.24	0.43	0.78	0.08	0.20	0.39	0.79	0.01	0.01	0.29	0.66	0.01	0.03	0.20	0.45	0.20	0.28	0.33	0.42	0.15	0.25	0.46	0.73	0.00	0.01	0.20	0.30
NewReno	0.13	0.23	0.48	0.99	0.09	0.19	0.43	0.98	0.05	0.09	0.33	0.86	0.06	0.09	0.25	0.55	0.47	0.47	0.49	0.56	0.11	0.21	0.47	0.95	0.09	0.08	0.23	0.38
Cubic	0.08	0.13	0.34	0.79	0.03	0.09	0.33	0.78	0.05	0.08	0.25	0.68	0.07	0.09	0.21	0.47	0.48	0.50	0.56	0.65	0.08	0.12	0.46	0.72	0.03	0.07	0.19	0.34
Illinois	0.44	1.00	1.00	1.00	0.14	0.34	1.00	1.00	0.06	0.15	0.85	1.00	0.09	0.14	0.41	1.00	0.46	0.49	0.51	0.53	0.88	0.93	1.00	1.00	0.06	0.16	0.21	0.42
BBRv2	0.23	0.30	0.56	1.00	0.19	0.27	0.62	0.95	0.13	0.18	0.37	0.90	0.16	0.18	0.31	0.75	0.35	0.14	0.26	0.34	0.28	0.38	0.71	1.00	0.07	0.08	0.15	0.33
Mixed	0.23	0.46	1.00	1.00	0.16	0.34	0.93	0.93	0.14	0.22	0.53	0.70	0.12	0.18	0.47	0.71	0.49	0.51	0.54	0.66	0.16	0.41	0.49	0.63	0.05	0.12	0.19	0.30

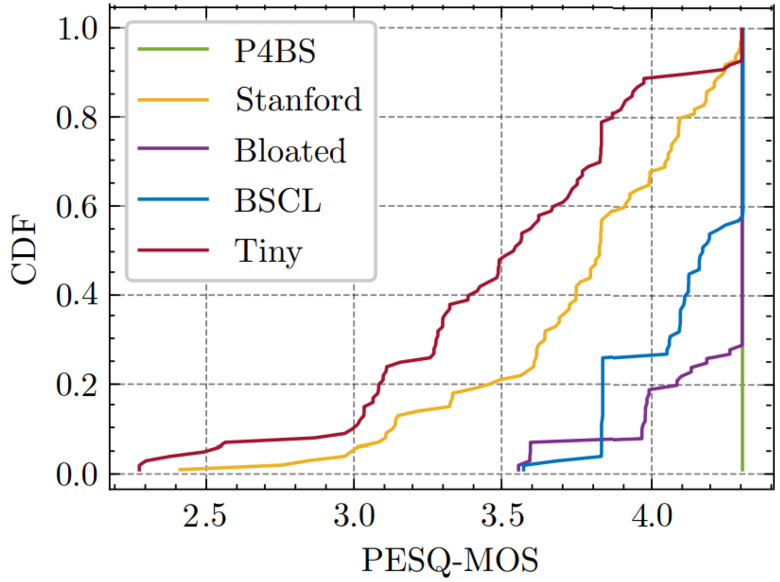
	Tiny				Stanford				BSCL				BDP				Bloated				ADT				P4BS			
N	50	100	250	500	50	100	250	500	50	100	250	500	50	100	250	500	50	100	250	500	50	100	250	500	50	100	250	500
Vegas	0.06	0.10	0.18	0.27	0.02	0.09	0.16	0.27	0.02	0.03	0.07	0.22	0.02	0.03	0.08	0.20	0.17	0.20	0.28	0.47	0.04	0.16	0.29	0.43	0.03	0.07	0.12	0.24
NewReno	0.06	0.11	0.19	0.31	0.05	0.09	0.16	0.32	0.04	0.06	0.13	0.28	0.06	0.07	0.12	0.23	0.46	0.48	0.53	0.60	0.12	0.19	0.31	0.43	0.01	0.10	0.12	0.31
Cubic	0.06	0.08	0.16	0.28	0.04	0.06	0.13	0.28	0.06	0.06	0.13	0.26	0.09	0.08	0.12	0.21	0.50	0.52	0.57	0.68	0.11	0.16	0.29	0.44	0.01	0.02	0.10	0.25
Illinois	0.33	0.52	1.00	1.00	0.10	0.25	1.00	1.00	0.04	0.13	0.80	1.00	0.07	0.11	0.41	1.00	0.48	0.50	0.53	0.61	0.37	0.73	1.00	1.00	0.08	0.11	0.35	0.68
BBRv2	0.18	0.16	0.31	1.00	0.16	0.18	0.31	1.00	0.09	0.16	0.23	1.00	0.10	0.13	0.23	0.85	0.39	0.44	0.40	0.52	0.23	0.32	0.56	1.00	0.11	0.09	0.30	0.51
Mixed	0.11	0.13	0.67	1.00	0.10	0.13	0.63	1.00	0.07	0.10	0.42	0.88	0.08	0.09	0.39	0.69	0.43	0.44	0.45	0.53	0.20	0.30	0.61	1.00	0.04	0.05	0.29	0.59

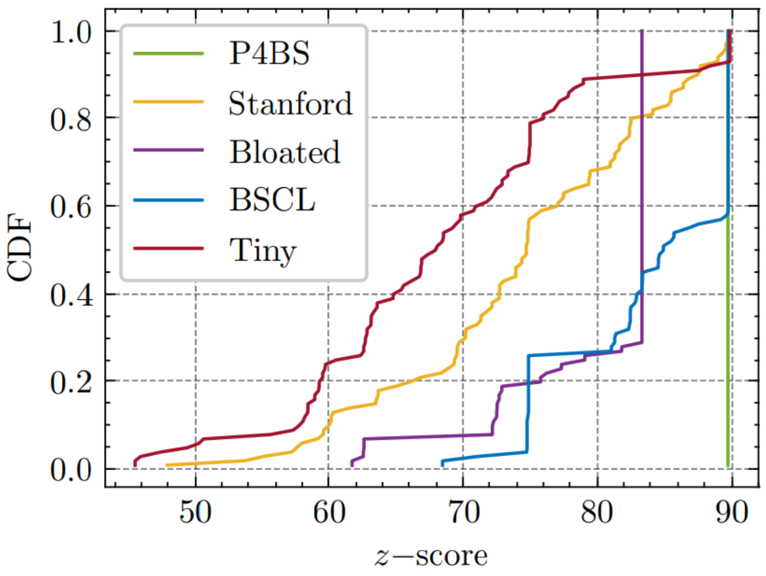
[1] Mishra et al. “The great Internet TCP congestion control census,” ACM on Measurement and Analysis of Computing Systems, 2019

Results

- 100 VoIP calls playing 20 reference speech samples (G.711.a)
- PESQ compares an error-free audio signal to a degraded one (the higher, the better)
- The z-score considers both the delay and the PESQ (the higher, the better)

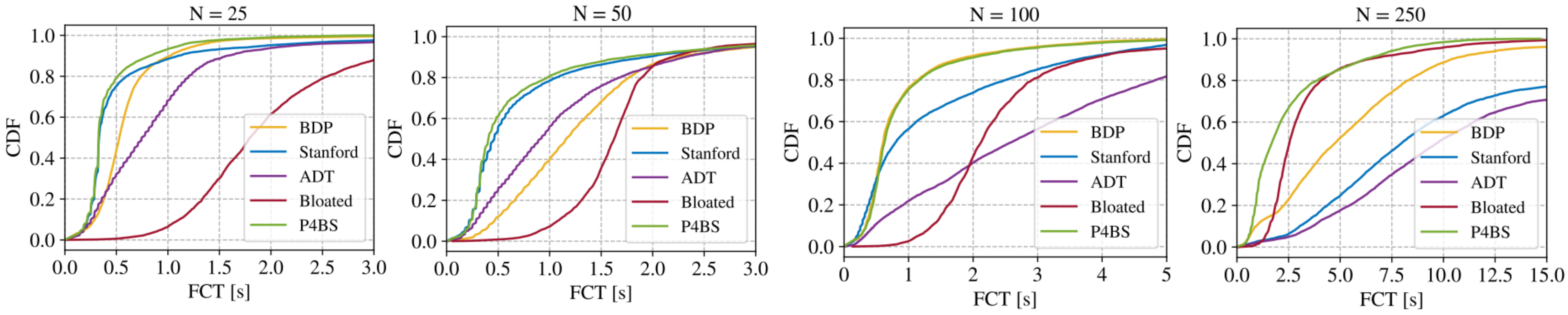






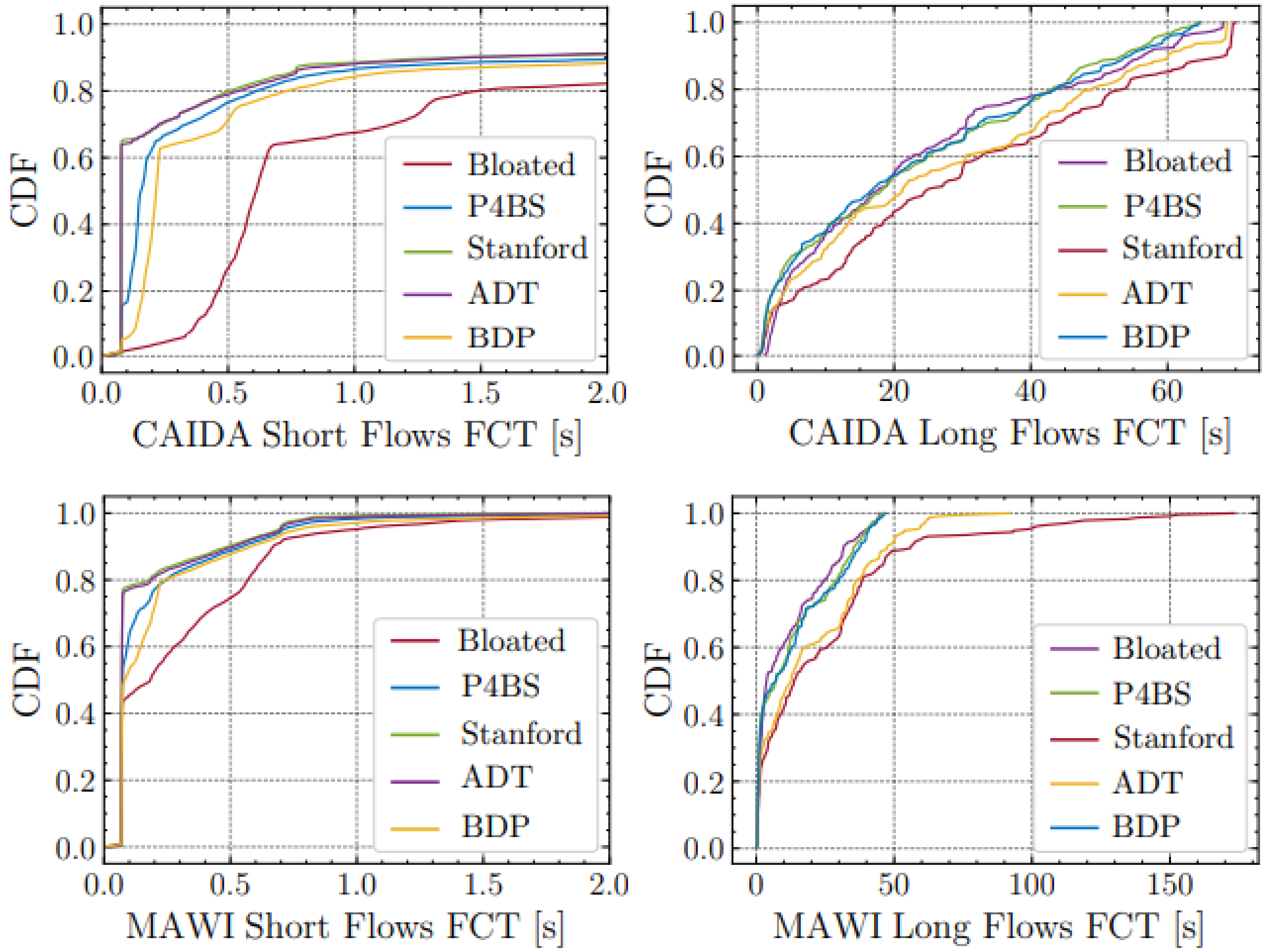
Results

- Web browsing traffic
- Background traffic is generated
  - The sizes of the web pages are in the range [15KB, 2.5MB]



Results

- Real traces
- CAIDA traces from Equinix NYC
- MAWI traces from WIDE
- P4BS found a balance such that:
  - ➢ The FCT of long flows is close to that of the bloated buffer
  - ➢ The FCT of short flows is close to that of the Stanford buffer



This work is supported by NSF awards number 2118311

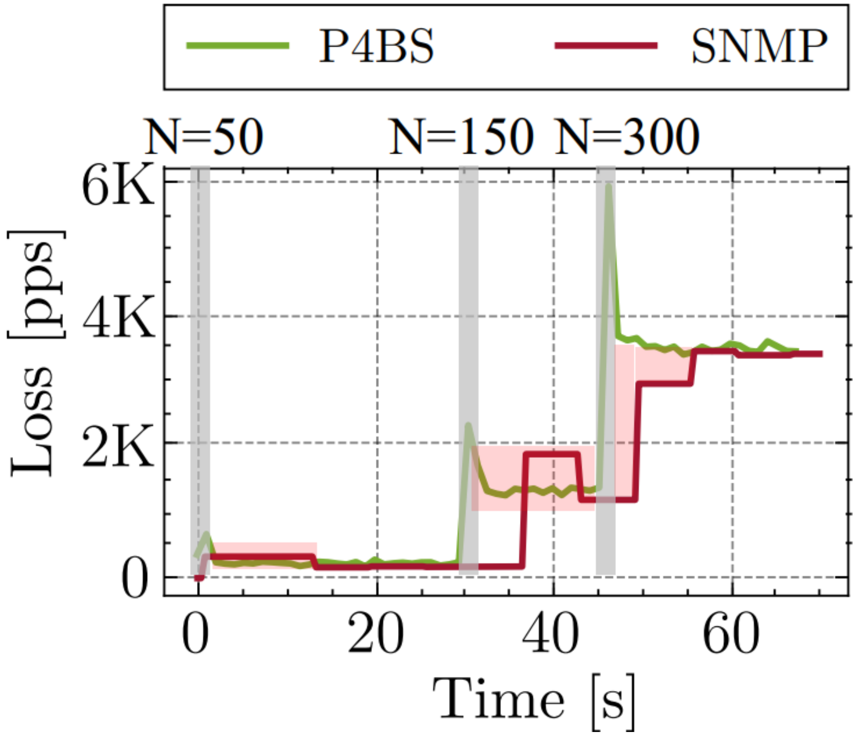
For additional information, please refer to
http://ce.sc.edu/cyberinfra/

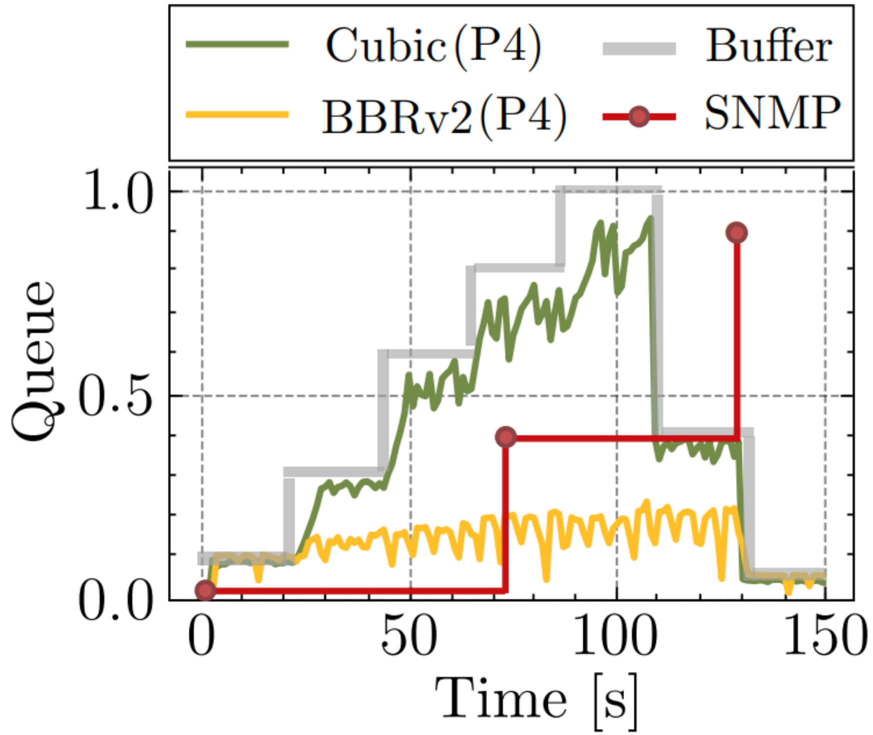
Email: jcrichigno@cec.sc.edu, ekfoury@email.sc.edu, gomezgaj@email.sc.edu

Additional Slides

Metrics Estimation without P4

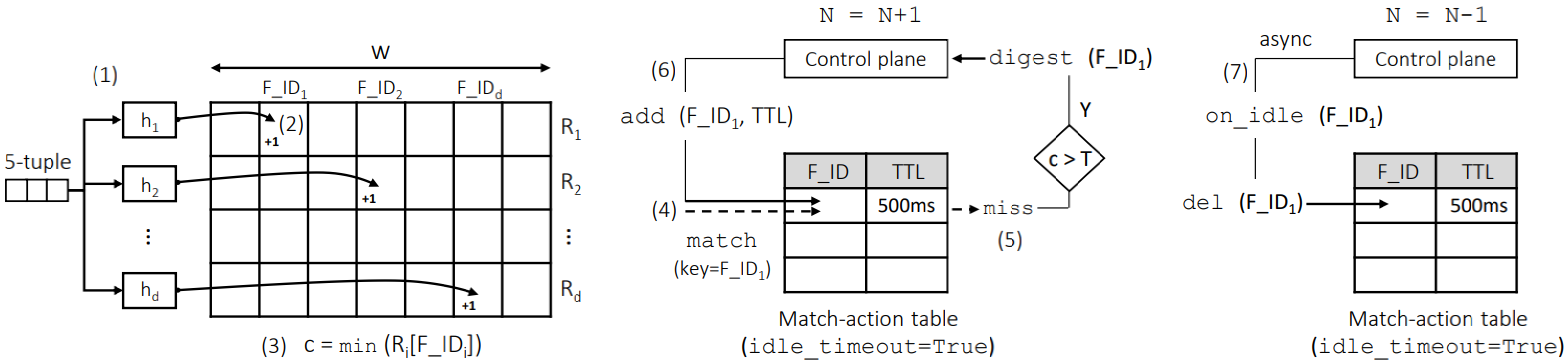
- SNMP produces coarse-grained and stale measurements when estimating the queueing delay and the packet loss rates





Long Flows Counting

- The Count-Min Sketch (CMS) is used to store the counts of the flows
- If the minimum exceeds a predefined threshold, the flows is identified as long flow
- Table timeouts are used to evict flows



Buffer Searching Dynamics

- Left is the acquisition function over time
- Right is the learned function