ESnet
ENERGY SCIENCES NETWORK

ESnet6 High Touch Services

Real-Time Precision Telemetry and ML-based TCP Classification

Richard Cziva, Ph.D.
Energy Sciences Network
Lawrence Berkeley National Laboratory

Jan 15, 2020
SIG-NGN

Overview

- ESnet introduction
- High Touch Services - Why are we doing this?
- ESnet6 Network Precision Telemetry
- Use-case 1: Real-time TCP Rate Monitoring
- Use-case 2: ML-based TCP Congestion Control Classification

Acknowledgements to the High Touch Team:

Chin Guok <chin@es.net>

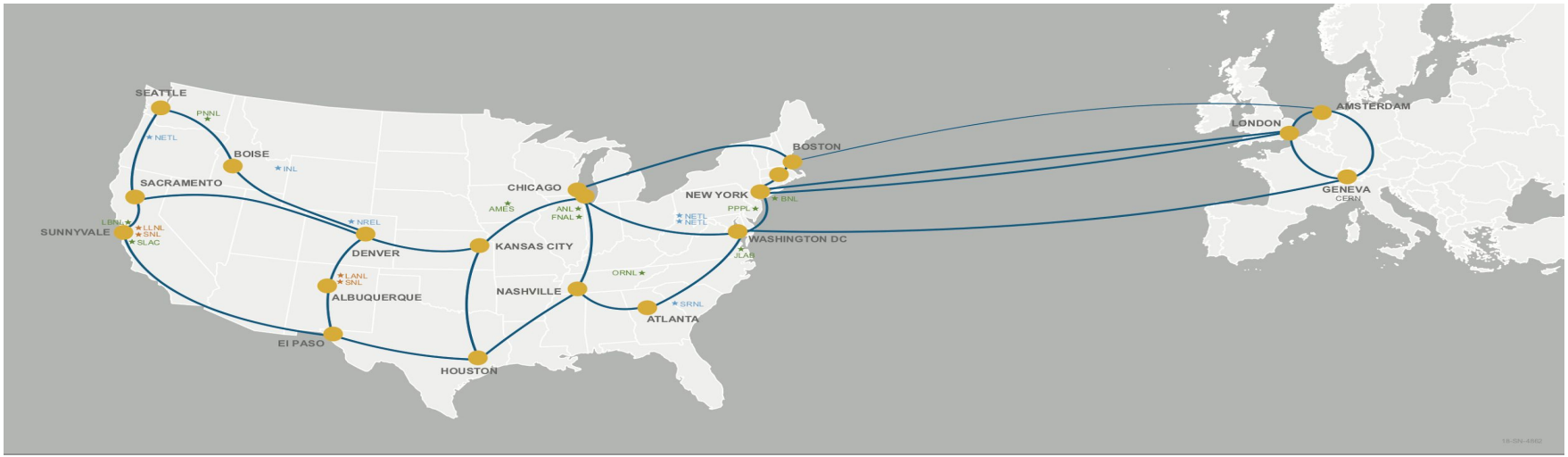
Yatish Kumar <yak@es.net>

Bruce Mah <bmah@es.net>

Kyle Simpson <kasimpson@lbl.gov>

ESnet

ESnet: DOE's high-performance network (HPN) user facility optimized for enabling big-data science



ESnet provides connectivity to
all of the DOE labs, experiment sites, & supercomputers

Motivation of ESnet6 High Touch Services

- Enhance the user experience
  - Real-time profiling of data transfer performance to proactively address issues*.
  - Predict network component failures to avoid unscheduled maintenance.
  - Predict usage patterns to determine least disruptive window to schedule preventive maintenance*.
- Increase network efficiently
  - Identify elephant flows to steer traffic over uncongested paths.*
  - Predict usage patterns to dynamically traffic engineer the network to eliminate hotspots*.
- Improve the security of the network
  - Detect usage abnormalities for further investigation.

**NB: Areas that ESnet is actively exploring*

ESnet

Why do we need High Touch?

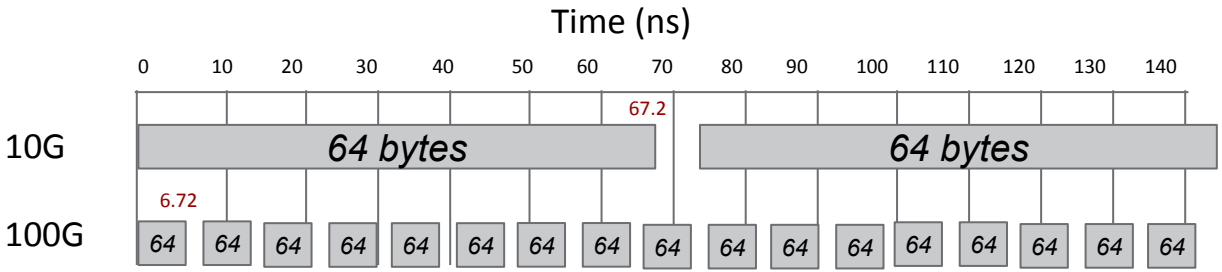
- Network monitoring today:
  - SNMP counters: per-interface - aggregate
  - Flow-based: Netflow, sFlow - approximate, sampled

- Enablers of per-packet telemetry:
  - Software-Defined Networking - more control over forwarding elements
  - Programmable network hardware with accurate timestamps (P4)
  - High-speed packet processing libraries (XDP, DPDK…)

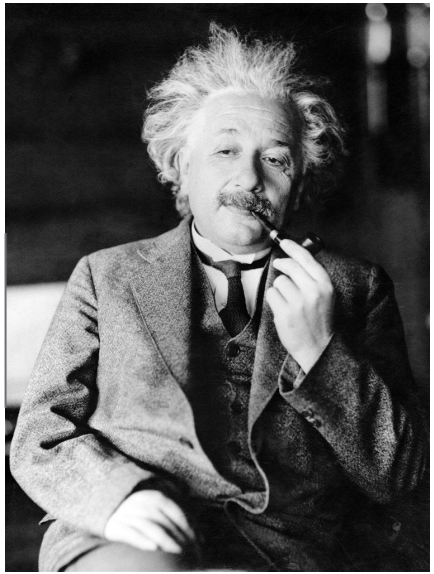
ESnet

The need for higher precision timing

- Increasing speed of the network links
  10G->40G->100G->400G->800G->...



At 100 Gbps, there can be as little as 6.7 nanoseconds between packets that need to be analyzed.



"The only reason for time is so that everything doesn't happen at once."

- Goal: going from **microsecond precision** to **nanosecond precision**

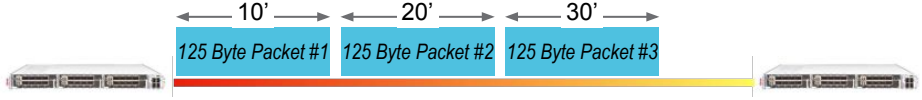
ESnet

ESnet6 High-Touch Precision Network Telemetry

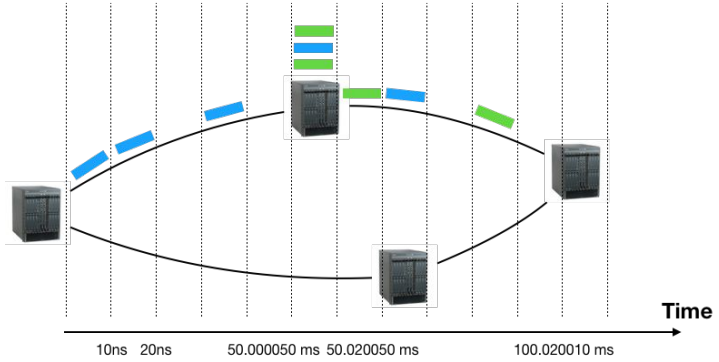
1. A 100 Mhz Clock has a period of 10 ns:
   Any contemporary FPGA / ASIC can operate counters at this rate.
   A 100 Gbit serial bit-stream transmits 1,000 bits in 10 ns.
   A small 125 byte packet contains 1,000 bits.

With 10 ns resolution timestamps, we can distinguish individual packets.

2. The speed of light is 1 ns / foot (in a vacuum).



With 10 ns resolution timestamps, we can locate a packet within 10 feet. The 3,000 mile distance from Berkeley to New York represents 15 million feet, or 1.5 million packets on the optical cable.

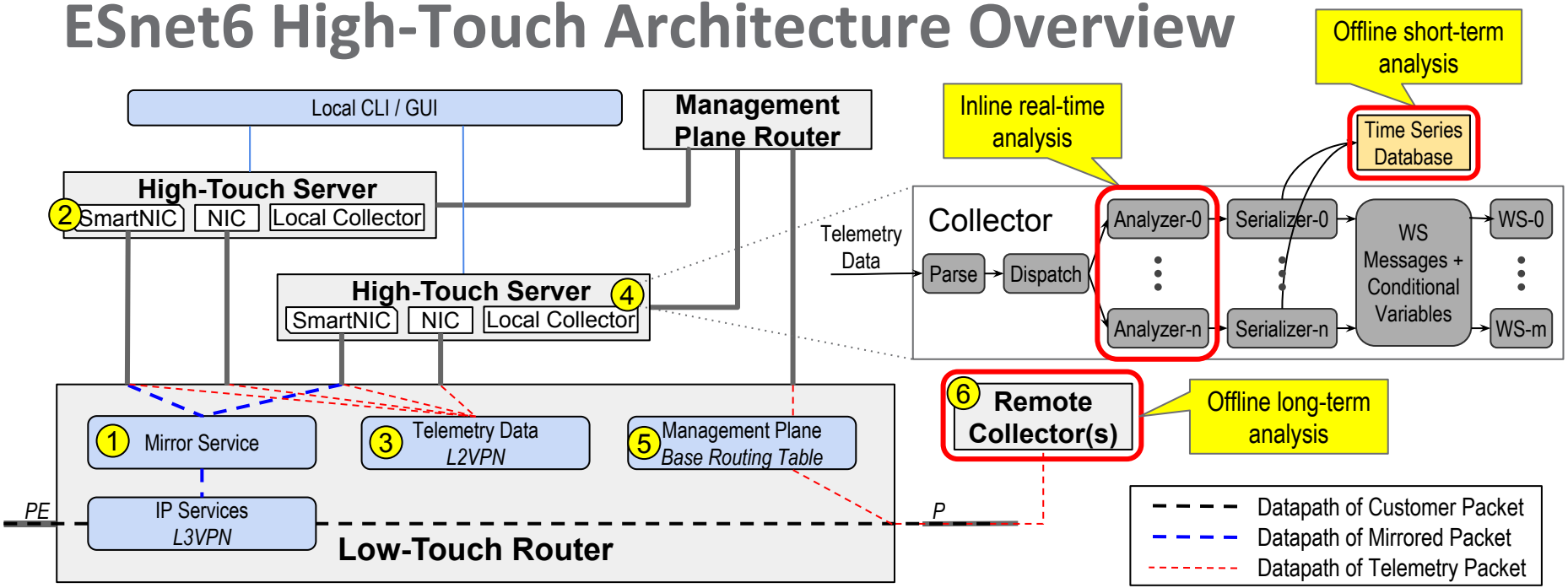


We can get detailed insights into how the network is behaving.

We can profile how flows are performing in our network and take proactive action.

We can use the detailed flow information for traffic engineering, capacity planning, or anomaly detection (e.g., AL/ML applications)

ESnet6 High-Touch Architecture Overview



1. Mirror Service - Allows selective flows in the dataplane to be duplicated and sent to the SmartNIC for processing.
2. SmartNIC - Appends meta-data and repackages packet for transmission to Collector code.
3. Telemetry Data L2VPN - Provides option to connect SmartNIC and Collector and bypass PCIe bus if needed.
4. Collector - Performs (limited) in-line real-time analysis as well as inserts telemetry data into database for offline local (short-term 1-2 hr) analysis.
5. Management Plane Base Routing Table - Provides connectivity to remote collector where aggregated telemetry data is sent for offline global analysis.
6. Remote Collector - Stores aggregated telemetry data for long-term global analysis.

“High-Touch” vs “Low-Touch” Hardware

“High-Touch”
Programmable data-plane

Pros:
- Flexible to customize for specialized use cases

Cons:
- Complexity of designing / implementing specialized use cases
- Higher cost

BarefExchip NP-4 (NPU)


Barefoot Tofino (P4)


Xilinx Kintex UltraScale (FPGA)+

“Low-Touch” (and “No-Touch”)
Application-Specific Integrated Circuits (ASIC) based data-plane

Pros:
- Optimized for specific tasks
- Lower cost

Cons:
- Inflexible

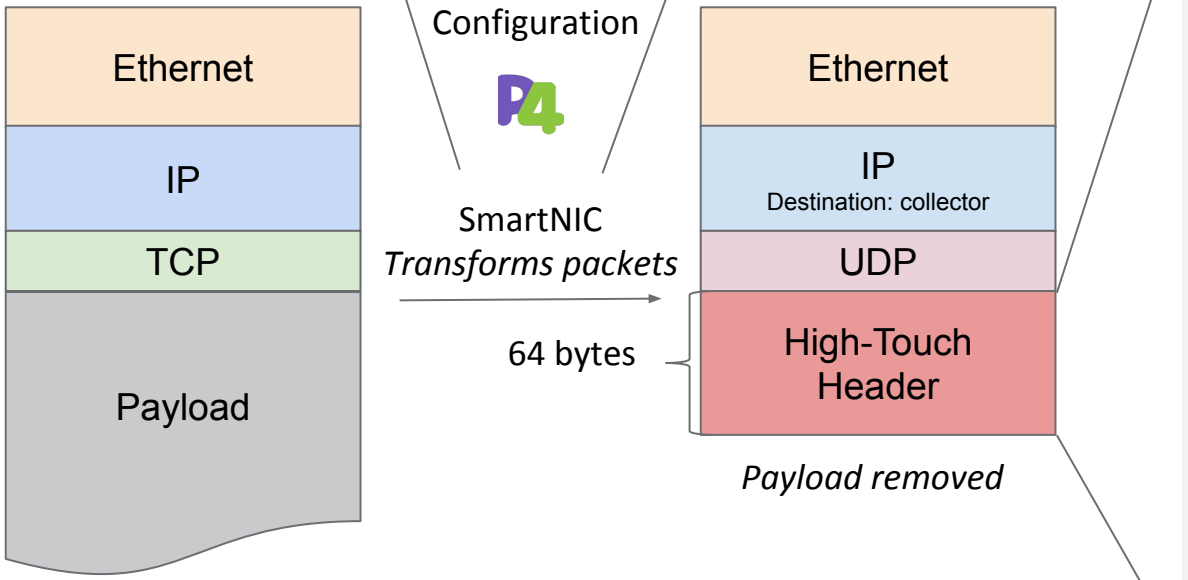
Broadcom Jericho

Juniper Trio

Nokia FP4

ESnet

Telemetry Producers - SmartNIC




```
type HighTouchLayer struct {
        Version                     string
        SensorID                    uint8
        IngressTimestampSeconds     uint32
        IngressTimestampNanoSecondsuint32
        // IP data or original packet
        IpVersion        uint8
        IpDiffserv       uint8
        IpTotalLen       uint16
        IpFlags          uint8
        IpTtl            uint8
        IpSrcAddr        net.IP
        IpDstAddr        net.IP
        // TCP data of original packet
        TcpSrcPort   uint16
        TcpDstPort   uint16
        TcpSeqNo     uint32
        TcpAckNo     uint32
        TcpEcn       uint8
        TcpWindow    uint16
        TcpUrgentPtr uint16
        // Aggregate counters
        FlowPktCount  uint32
        FlowByteCount uint32
}
```

Copy of original packet of a TCP flow

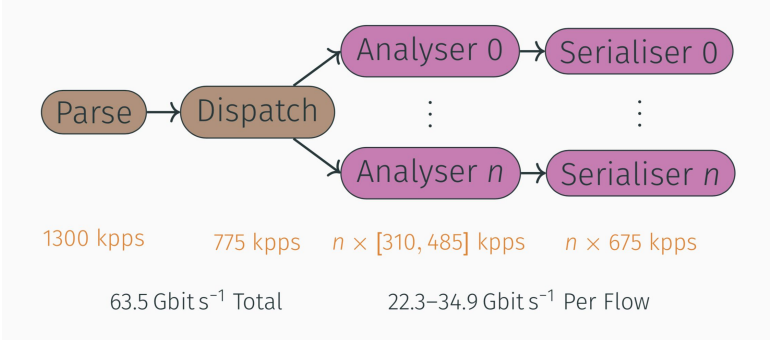
HighTouch Telemetry Packet

HighTouch Telemetry Packet Format v1

High-Touch Analysis and Algorithms Prototype

Collector: Our highly-parallelized software that ingests our telemetry packets and computes:

- Rate (point rate vs. sliding window)
- Retransmission and loss detection
- Initial SRTT estimation
- Online half-SRTT estimation
- Bytes-in-flight
- Congestion window estimation



**NB: Performance depends on the selected algorithms we run per flow - worst case 22 Gbit/s (all analysis computer for each flow) using a mid-range machine up to 35 Gbit/s (only rates are computer for each flow).*

Collector will run on powerful x86 machines [780 GB RAM, 32 cores Intel Gold 6242, 25 TB local NVME storage] being deployed at over 30 locations of our network.

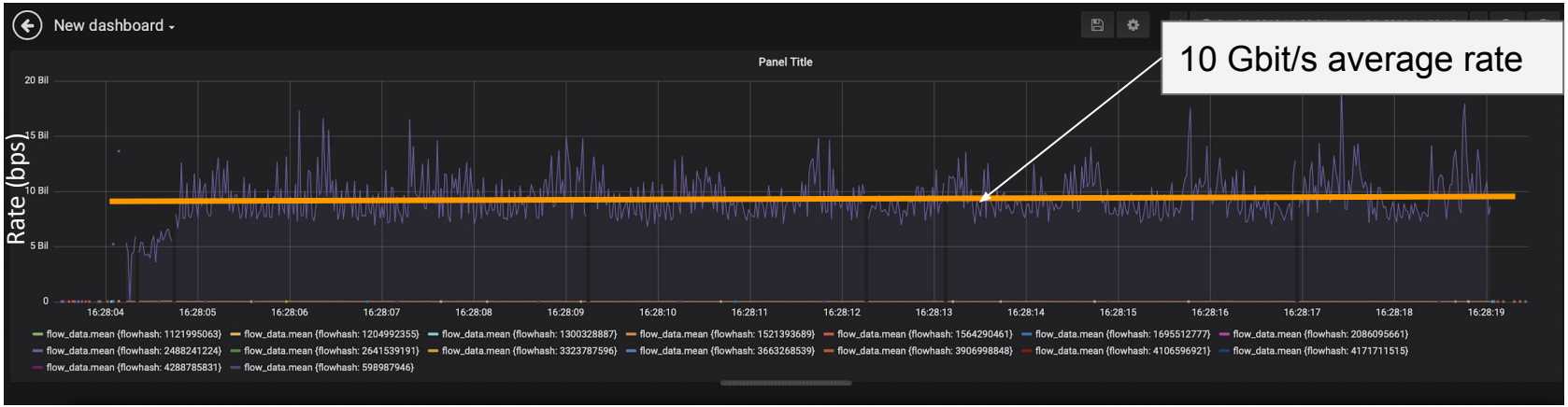
ESnet

Real-time data analysis use-case: TCP Rate Monitoring

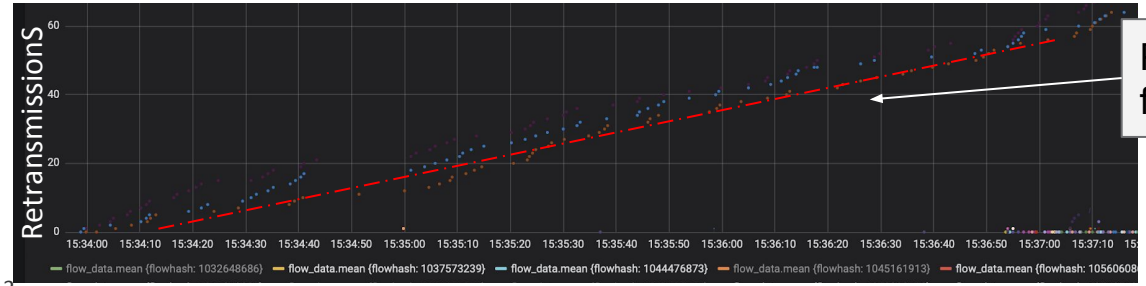
ESnet

Visualizing Real-Time Telemetry Data

- We can plot metrics for every packet in a flow using InfluxDB / Grafana



A sample PerfSonar 10Gbit/s test measured by High Touch Rate Monitor

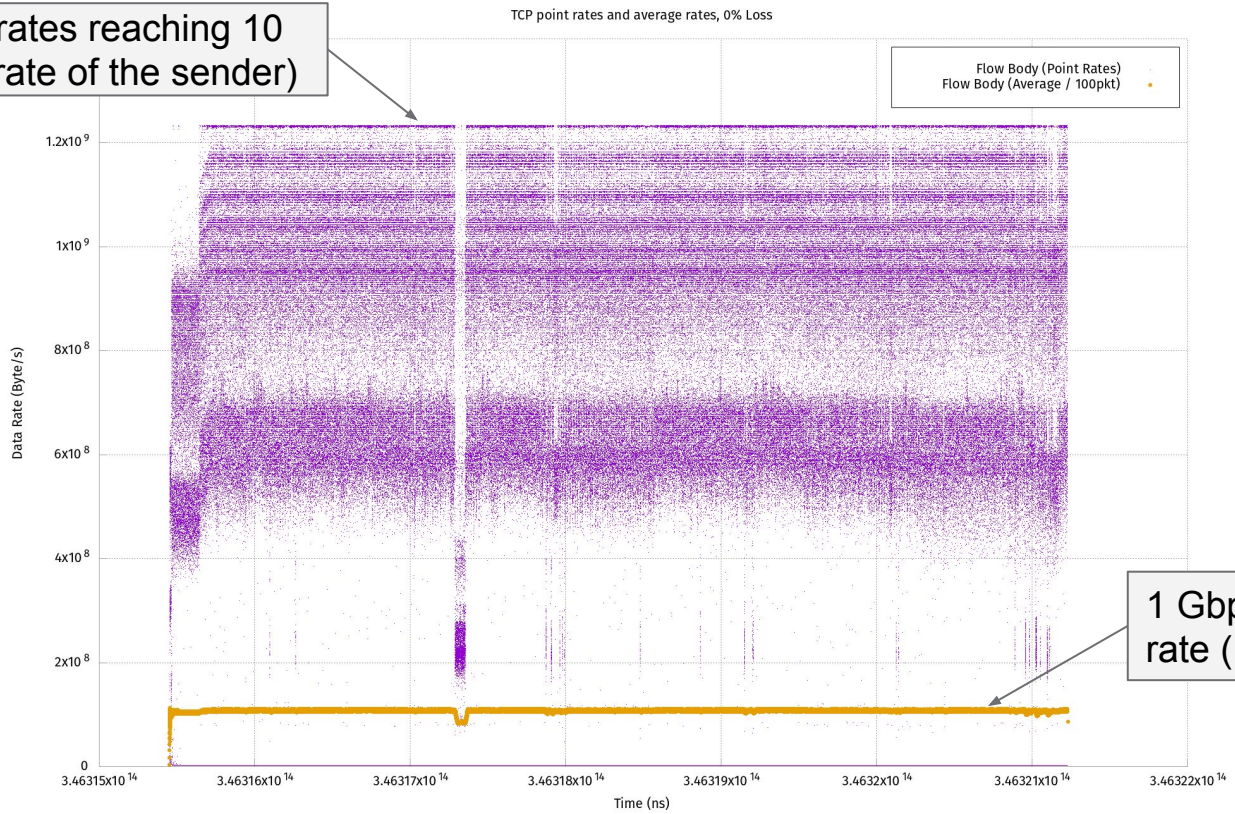


Packet retransmissions for flows **indicating an issue**

ESnet

1 Gbps iPerf flow - 600,000 packets

Per-packet rates reaching 10 Gbit/s (line rate of the sender)

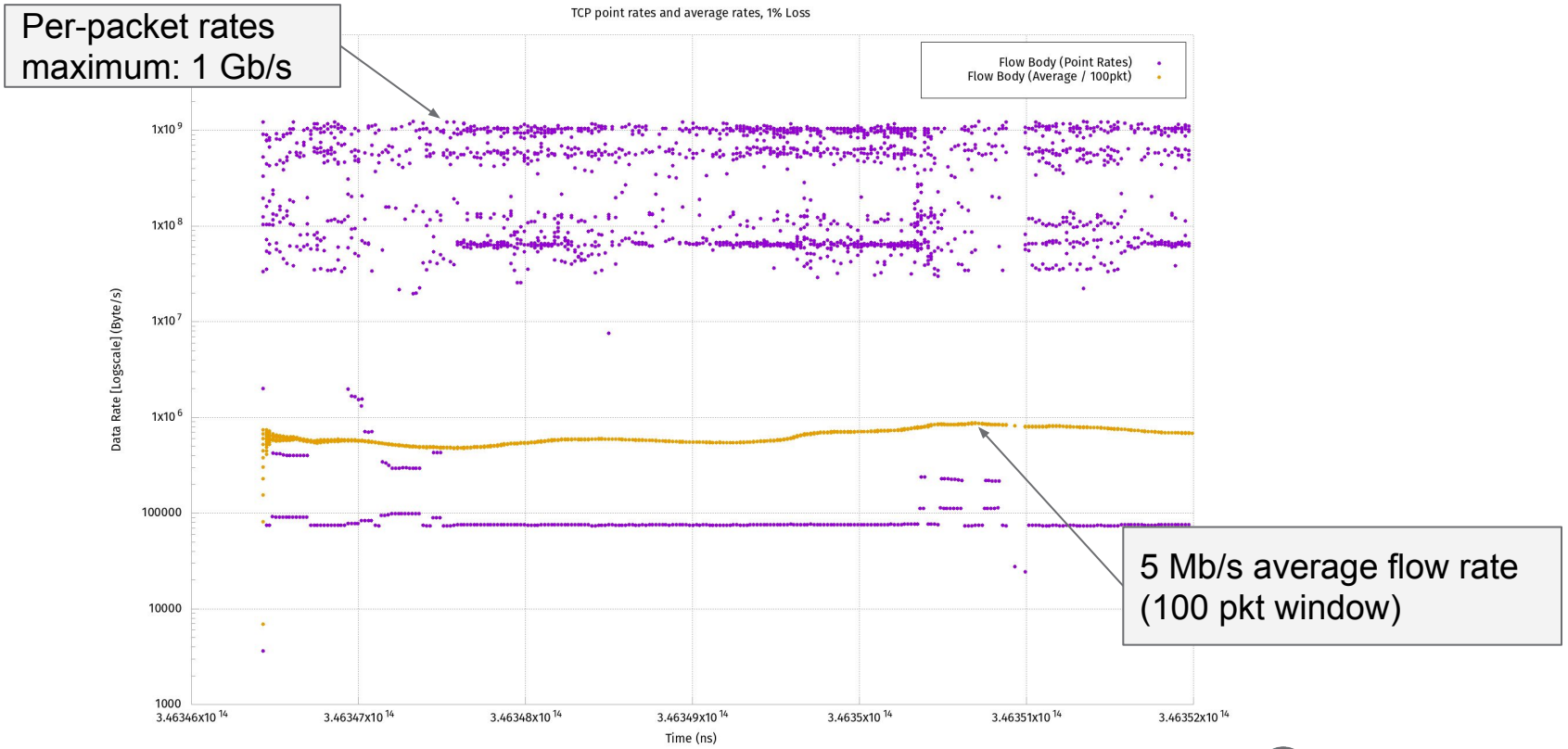


1 Gbp/s average flow rate (100 pkt window)

Note: Average rate is calculated using a time-weighted average of per-packet rates.

ESnet

1 Gbps iPerf flow - 1% packet drop

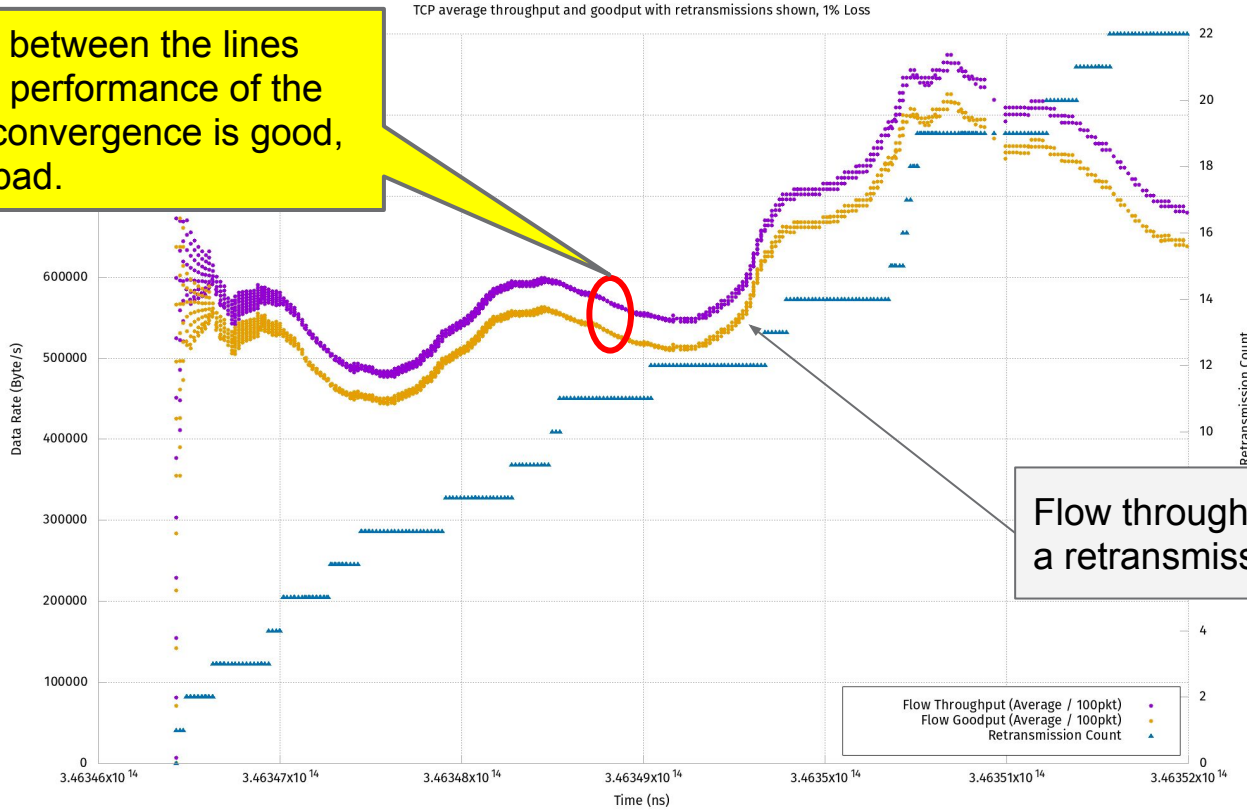


Note: only 23 packets were dropped all together, taking bandwidth down to 5 Mb/s from 1 Gb/s.

ESnet

1 Gbps iPerf flow - 1% packet drop - cont'd

The difference between the lines represents the performance of the data transfer, convergence is good, divergence is bad.



Flow throughput drops when a retransmission happens

ESnet

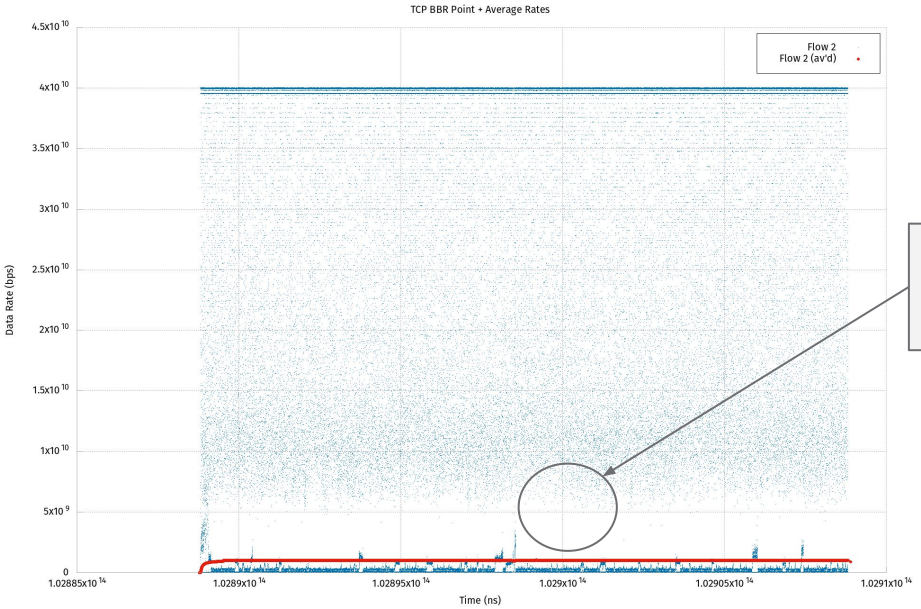
Offline data analysis use-case:

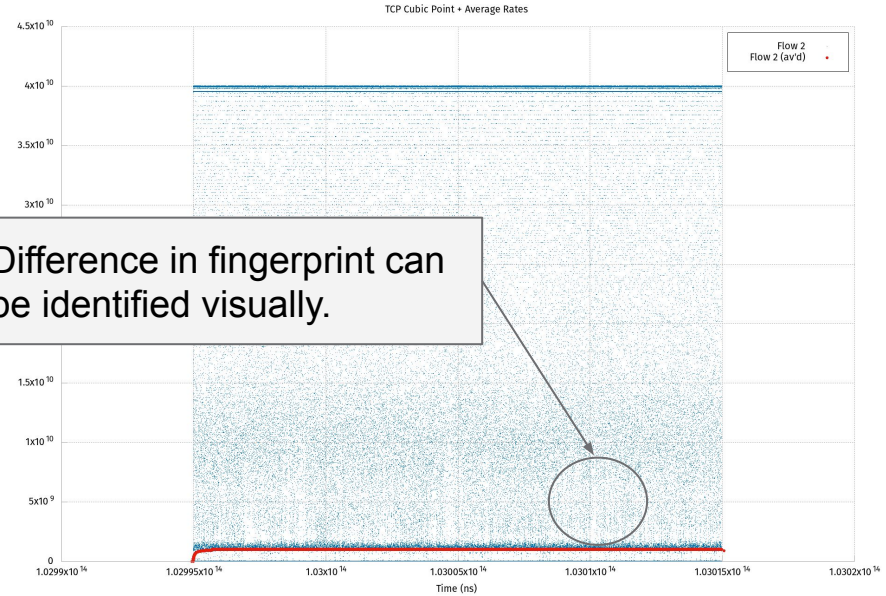
- TCP flows can take more of their fair-share…
- Finding misconfigured flows will allow us:
  - Tuning our DTNs
  - Notifying our sites automatically (periodic reports)

Can we infer TCP congestion control properties?

ESnet

BBR vs Cubic - Point rates





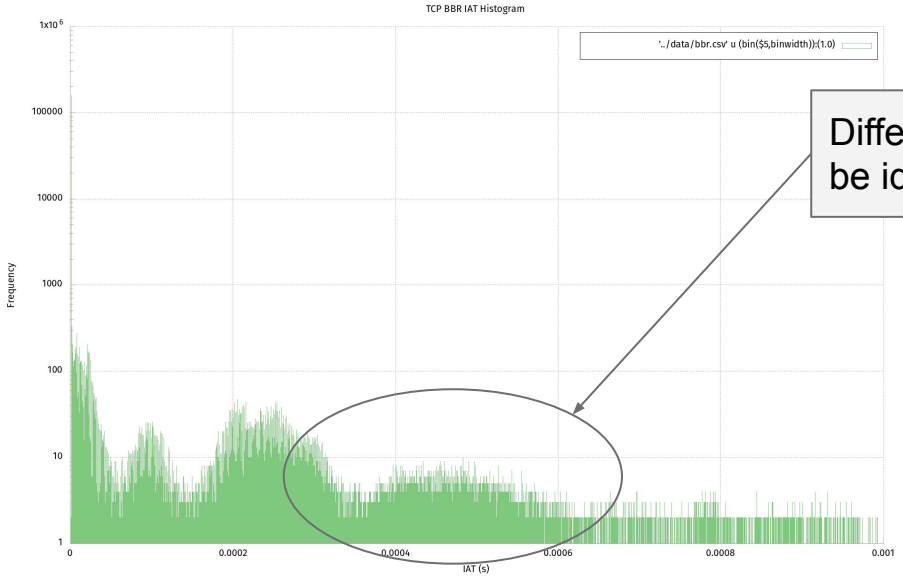
Difference in fingerprint can be identified visually.

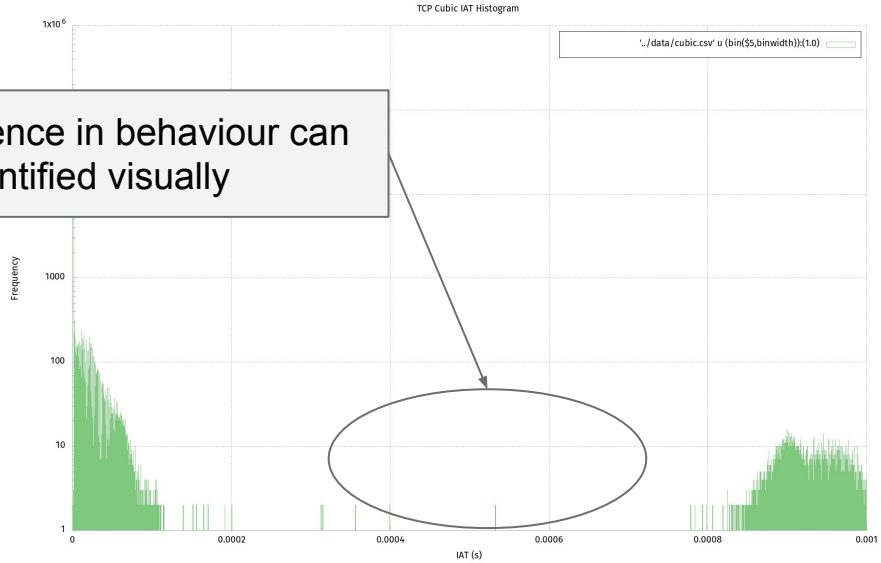
TCP BBR

TCP Cubic

2 millions of data points shown (around 600.000 points a second generated)

BBR vs Cubic - Inter-Arrival Time histogram





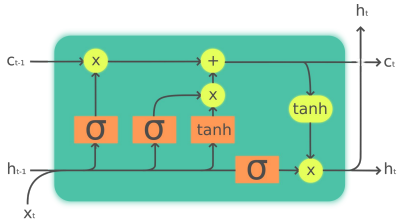
Difference in behaviour can be identified visually

TCP BBR (delay-based)

TCP Cubic (loss-based)

BBR: inter-packet timing is more widespread than other congestion control algorithms.

Classification Specifics



An LSTM network

Input:

- 5--20 observations from the collector (per flow)
- 18 features, all collector outputs
- Flows from iPerf (TCP BBR/Cubic/Reno/Vegas), 0.1--1.0 Gbps

LSTM properties using the Keras Python library

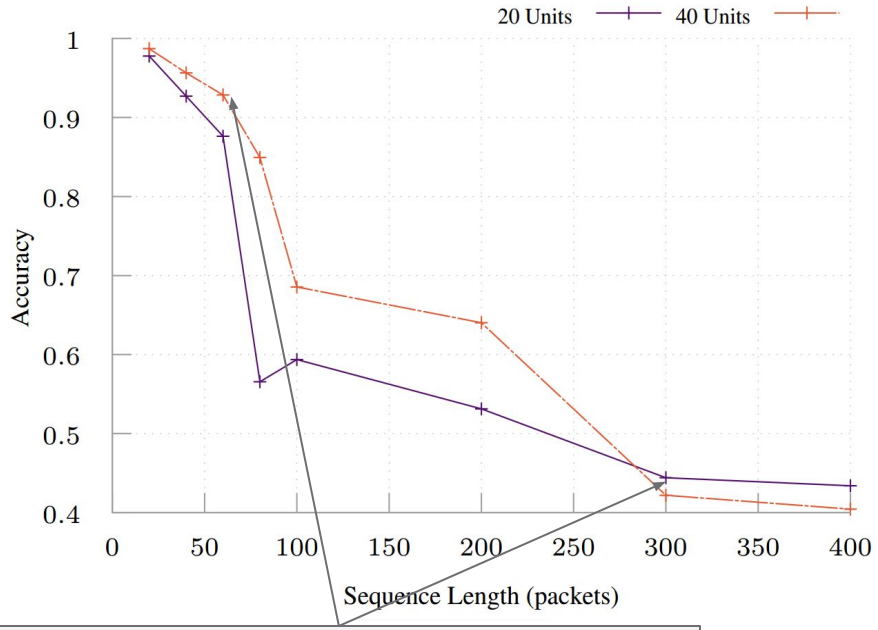
- Automatic NN feature extraction: 20 / 40 units tested.
  - Units => dimensionality of input and forget gates, output, inner state
- Softmax activation function - map output likelihoods to classes
- Adam (adaptive moment estimation) model optimization
- Dropout 0.1 - reducing overfitting

ESnet

Performance -- Sequence/Unit Count

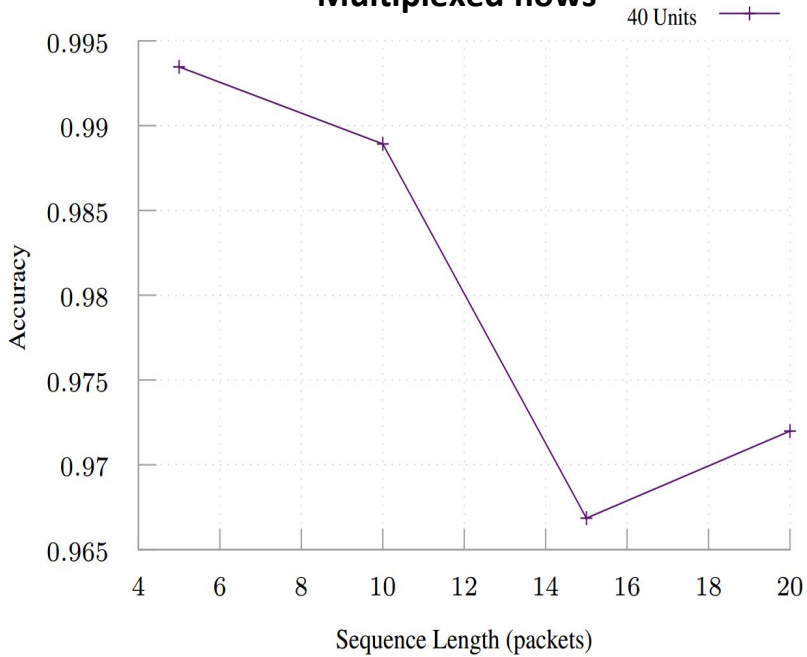
Classifying mixed flows is slightly harder (drops 98.7->97.2%)

Solo flows



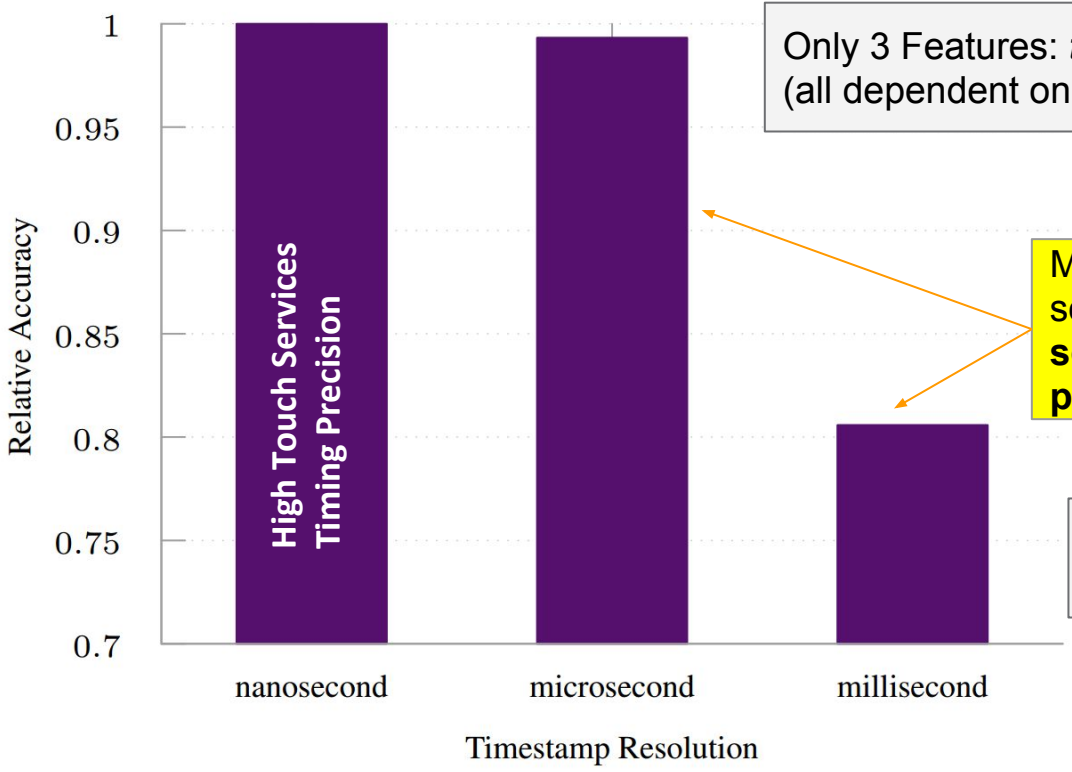
More units mean a more powerful model, but need more training data. (Peak 98.7%)

Multiplexed flows



Still effective down to 5 packet sequences due to extra training data.

Effect of timestamp precision (normalized)



Only 3 Features: *time, dt, point rate*
(all dependent on time accuracy)

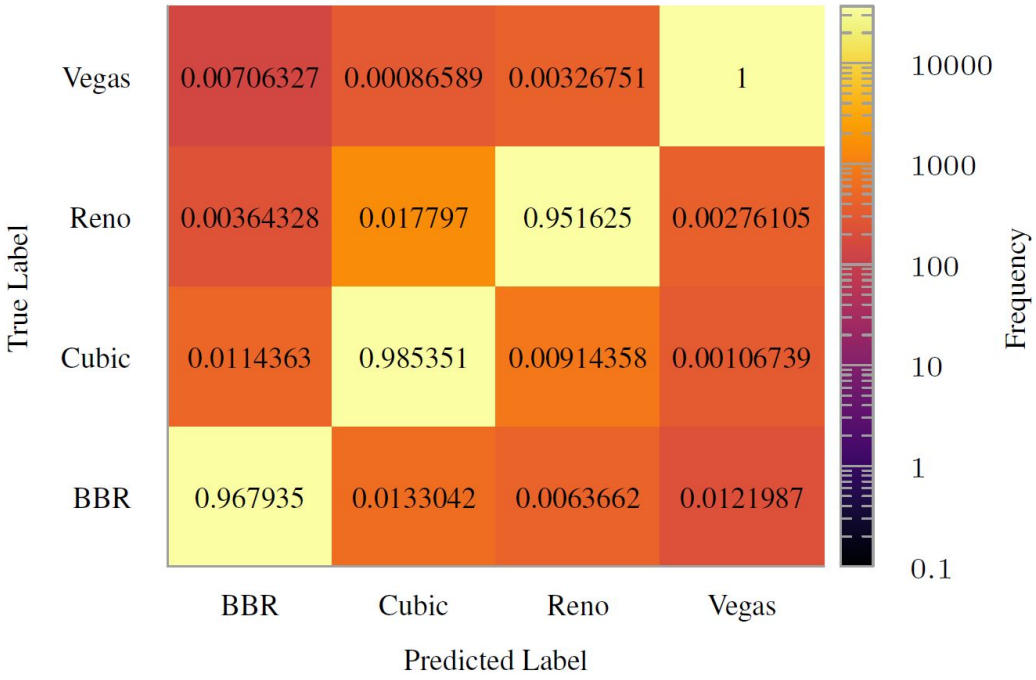
Most packets have dts on the order of *µs*, so sharpest decrease here. **Traditional software tools provide timestamp precision between these resolutions.**

NOTE: Loss of other features significant. (98.7->34.2%)

ESnet

Classification -- confusion matrix

- Visualising the likelihood of how samples might be mislabelled.
- Sequences rarely misclassified...
- BUT, Reno and Cubic most likely to be confused.



Results are under peer-review at IEEE ICC.

ESnet

High Touch Services - other possibilities

- Programmable streaming telemetry
  - Select flows, precision, destination collectors
- Enhanced security services
  - Programmable traffic selection, packet truncate
  - Forward selected packets to IDS systems
  - Alert operators when specific header signatures are seen
- In-band network telemetry (INT)
  - Hop-by-hop packet tracing

ESnet

Summary

- ESnet is developing High Touch Services:
  - A platform for *precision* network telemetry
  - Uses *programmable* data-plane for telemetry producers
  - Provides *nanosecond-accurate* timing for each packet
  - A powerful, scalable telemetry collector allows:
    - Online data analysis, storing data in time-series DB
    - Data access via APIs for offline analysis
- We presented two use-cases:
  - Real-time visualization: finding retransmissions, poor throughput
  - Offline analysis example: TCP congestion control identification

ESnet

Questions…

richard@es.net

ESnet