OAK RIDGE National Laboratory

Experimental Study of TCP Throughput Profiles and Dynamics Over Dedicated Connections

Nageswara (Nagi) S. V. Rao,

Oak Ridge National Laboratory

ORNL is managed by UT-Battelle LLC for the US Department of Energy

Sponsored by Department of Energy
Advanced Scientific Computing Research office

U.S. DEPARTMENT OF ENERGY | Office of Science

U.S. DEPARTMENT OF ENERGY

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

Outline

1. Introduction
2. Network Testbed and Measurements
   - testbed components
   - measurements collection
3. Throughput Measurements and Profiles
   - throughput measurements
   - profile properties
4. Time Traces
   - throughput variations
   - Poincare maps
5. Throughput Profiles under Losses
6. Conclusions

Data Transfers and Dedicated Connections

Data transfers over wide-area networks:

- integral part of supporting distributed and cloud computing and storage, and supercomputing and science instrument complexes
- support a variety of tasks including moving codes, files and data sets, and also migrating virtual machines and containers.

Dedicated network connections increasingly being provisioned over conventional and cloud network infrastructures to support such data transfers

- provide unimpeded capacity without competing traffic
- offer simpler operational conditions for transport protocols compared to shared connections - do not impose stringent conditions such as fairness and friendliness to other flows

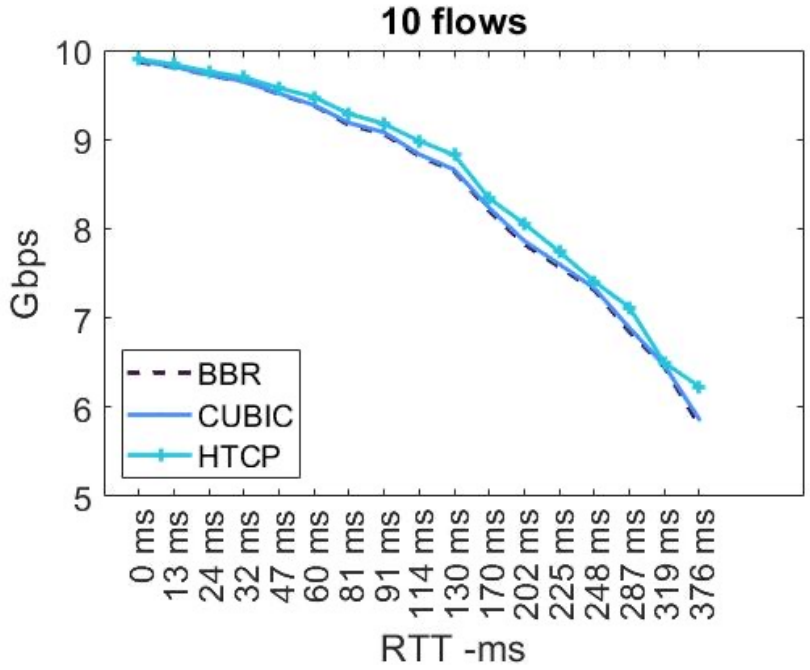
This paper: comprehensive experimental study of eleven different TCP versions

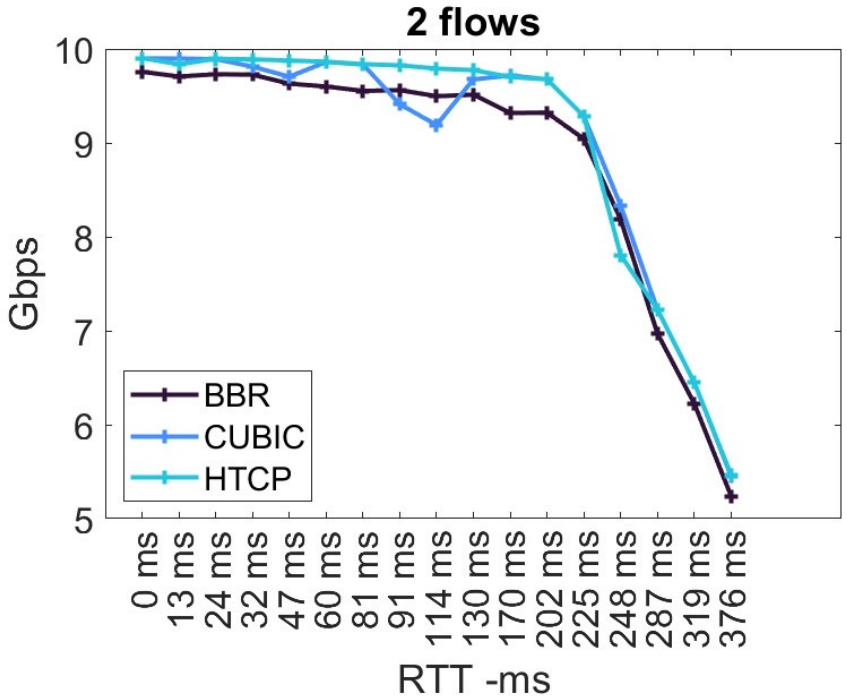
- represent their variety and availability over past decades.
- connections range in length from local to round the earth distances

Overall, our measurements and analytical study shows:

- protocols such as Hamilton TCP (HTCP) and Scalable provide higher throughput, albeit by small amounts in some cases, than Bottleneck Bandwidth and Round-trip propagation time (BBR v1)
- under externally induced losses, BBR provides increasingly higher throughput than CUBIC and HTCP as loss rate is increased

Data Transfers Over Dedicated Connections





Throughput measurements - local to round the earth distances

- 10 Gbps connections with 0-376 ms round trip time (RTT)
- implemented using hardware emulators to closely match TCP throughput and dynamics of corresponding physical connections

Protocols such as Hamilton TCP (HTCP) and Scalable provide higher throughput,

- compared to more recent BBR v1
- Cause: BBR's more complex time dynamics indicated by throughput time traces

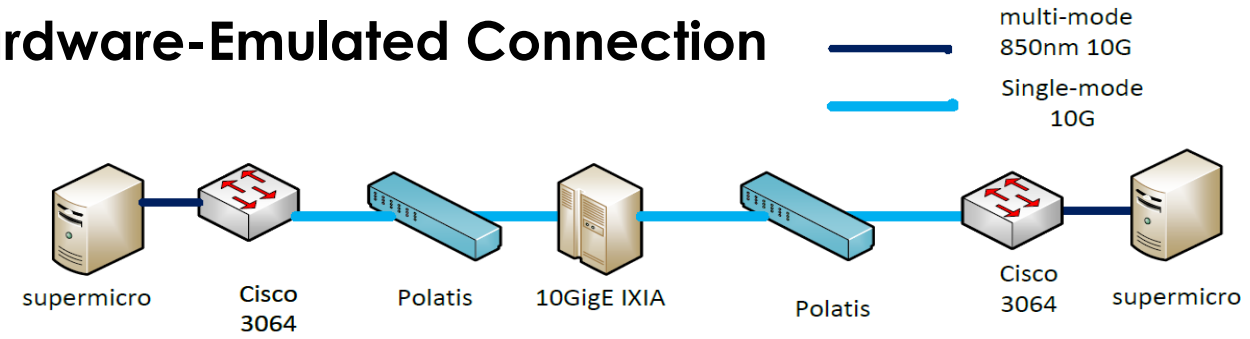
Contributions Summary

- Data transport infrastructure: *Throughput profile* expressed as a function of the round-trip times
  - critical indicator of its level of optimization, particularly, over dedicated connections.
- Our Study: Throughput profiles of eleven TCP versions
  - using measurements collected over dedicated hardware-emulated connections: distances spanning round the earth.
  - Experimental measurements of BBR show overall higher temporal variations:
    - lower throughput profiles compared to five loss-based TCP versions
    - comparison with other TCP versions is mixed.
  - Poincare map regions of throughput time traces:
    - show that richer dynamics of BBR are correlated with its lower throughput profiles
  - We present basic analytical results:
    - higher temporal variations of TCP methods lead to lower throughput profiles - underlying causality
  - Under external losses: BBR achieves sustained high throughput compared to others as the loss rate is increased.
- Overall Results: insights into relationship between time dynamics and throughput profiles
  - Even limited time traces can be indicative of global properties of throughput profiles

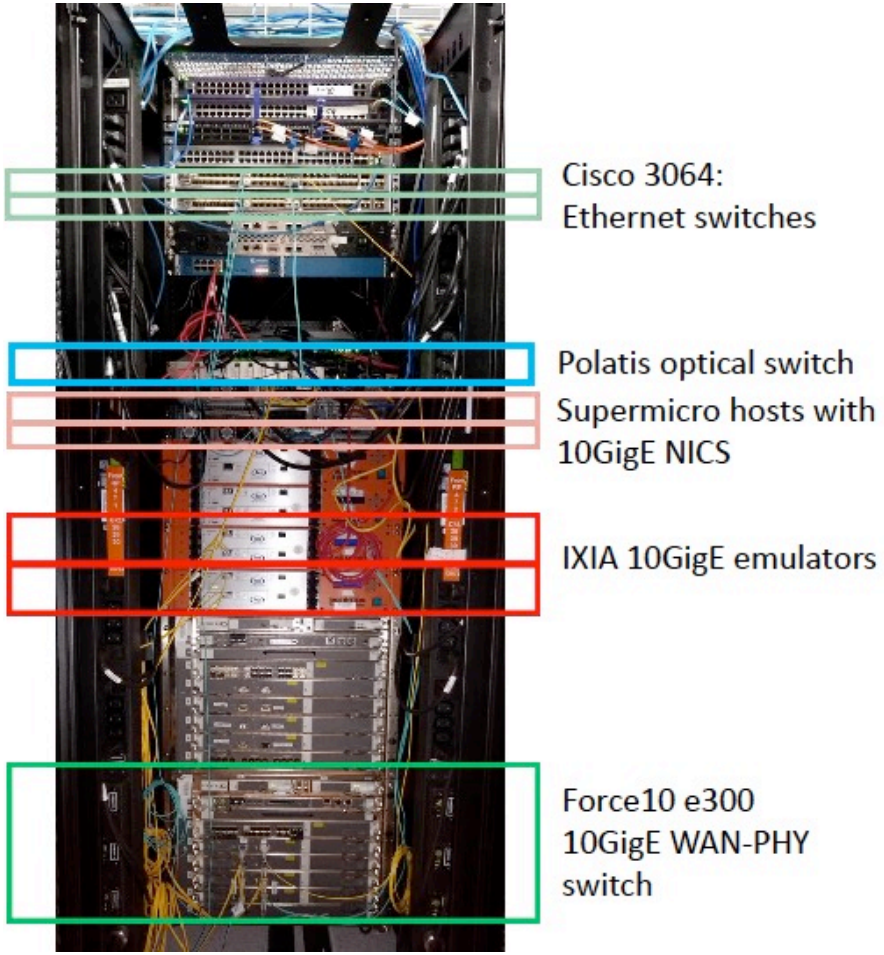
Testbed

- Supermicro Linux servers,
- Cisco 3064 Ethernet switches,
- Polatis all-optical switch
- IXIA 10GigE connection emulators

Hardware-Emulated Connection



- Measurements collected on 32-core Supermicro servers with Redhat Linux RHEL8 kernel used as a DTN
- Two servers identical configurations - connected to two Cisco 3064 Ethernet switches
- connected to 10 GigE IXIA hardware emulator via Polatis switch



TCP dynamics and throughput on host systems and switches closely match equivalent physical connection:

- Ethernet packets are received and processed as per connection RTT and loss rate, and delivered hardware IXIA emulator - process that closely matches a physical connection.
- More accurately reflect physical TCP flows than simulators such as ns-3 and riverbed, and software emulators such as mininet that are subject to host system limitations

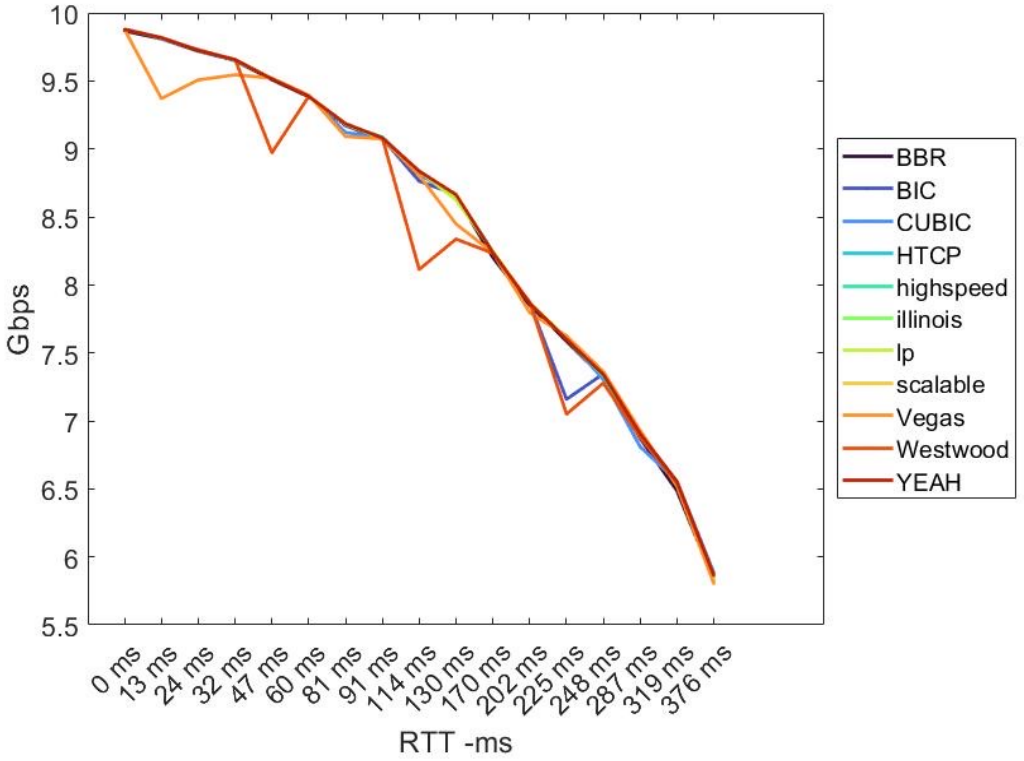
Measurements Collection

Eleven TCP Versions: Memory-to-memory throughput measurements and time traces using the iperf3

- BIC , CUBIC, BBR, HTCP, Highspeed TCP, Illinois, LP, Scalable TCP, Westwood and Yet Another High-performance (YEAH) TCP
- chosen to represent variety and history, and availability as compilable and loadable Linux kernel modules

On our RHEL8 systems:

- CUBIC and BBR are pre-installed
- others are compiled and loaded as kernel modules without rebuilding the kernel (BBRv2 required rebuilding)



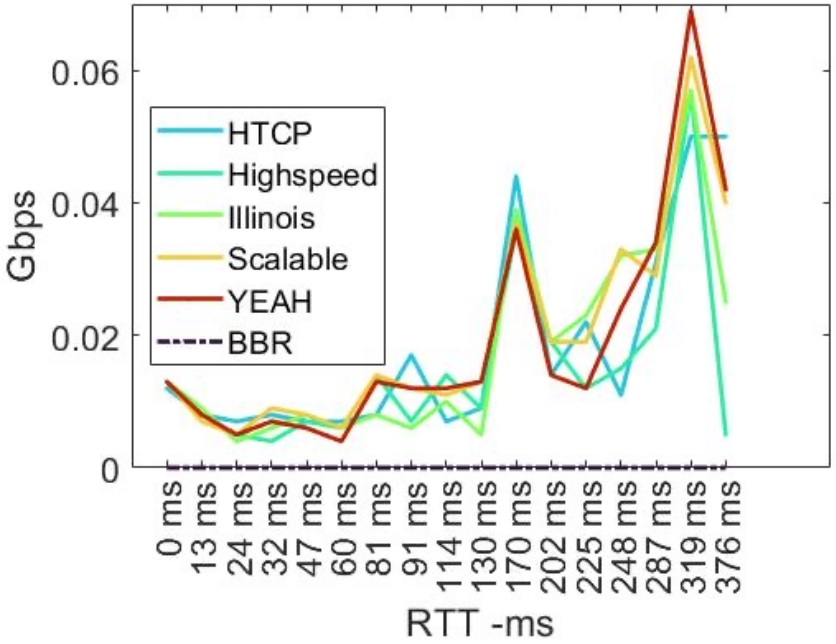
Measurements: Memory-to-memory throughput measurements and time traces using the iperf3:

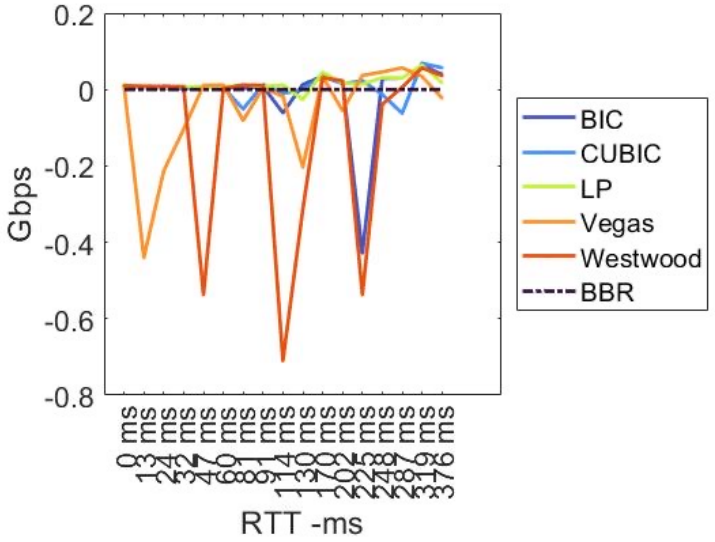
- number of parallel streams is varied from 1 to 10 - repeated 10 times
- automated using bash scripts - curl to configure IXIA;
- each RTT set collection typically takes about 1–2 days for each TCP version
- TCP buffer sizes and BBR parameters set to recommended values for 200 ms RTT and socket buffer parameter for iperf3 is 2 GB

Throughput Measurements

Group1: difference in throughput profiles with respect to BBR are positive for HTCP, Highspeed TCP, Illinois TCP, Scalable TCP, and YEAH:

- their throughput is higher than BBR uniformly for all RTTs,
- measurements observed under different types of dedicated connections are the main motivation for this study. T
- these protocols designed for high performance - adjusting flow rates based on inferring losses in different ways





Group 2: throughput profile differences for BIC, CUBIC, LP, Vegas and Westwood with respect to BBR are mixed

- Vegas and LP are not designed with high throughput as a main goal but their profiles are quite close to others mainly due to dedicated connections

Unlike BBR, Vegas, and LP, Group1 protocols

- maximize throughput based on self-induced losses and inference using different mechanisms
- their dynamics on dedicated connections are self-regulated - indicated by time traces

Profile Properties

$\theta_V^n(\tau, t)$: aggregate throughput of n flows at time t over a connection of RTT τ,
where V is the TCP version.

Throughput profile Θ_V^n integrating throughput over a time window for RTT τ as

$$\Theta_V^n(\tau) = \frac{1}{T_O} \int_0^{T_O} \theta_V^n(\tau, t) dt, \qquad (1)$$

and its generic version is denoted simply by Θ determined by host properties including

- $V = BI, C, BB, H, HS, I, L, S, W, Y$ representing BIC, CUBIC, BBR, HTCP, HighSpeed TCP, Illinois TCP, LP, Scalable TCP, Westwood and YEAH TCP, respectively
- number of parallel streams n, various buffer sizes, and other host parameters

Variations in throughput profile with respect to τ are related to temporal variations of $\theta(\tau, t)$ by special case of Leibniz integral rule applied to Eq. (1)

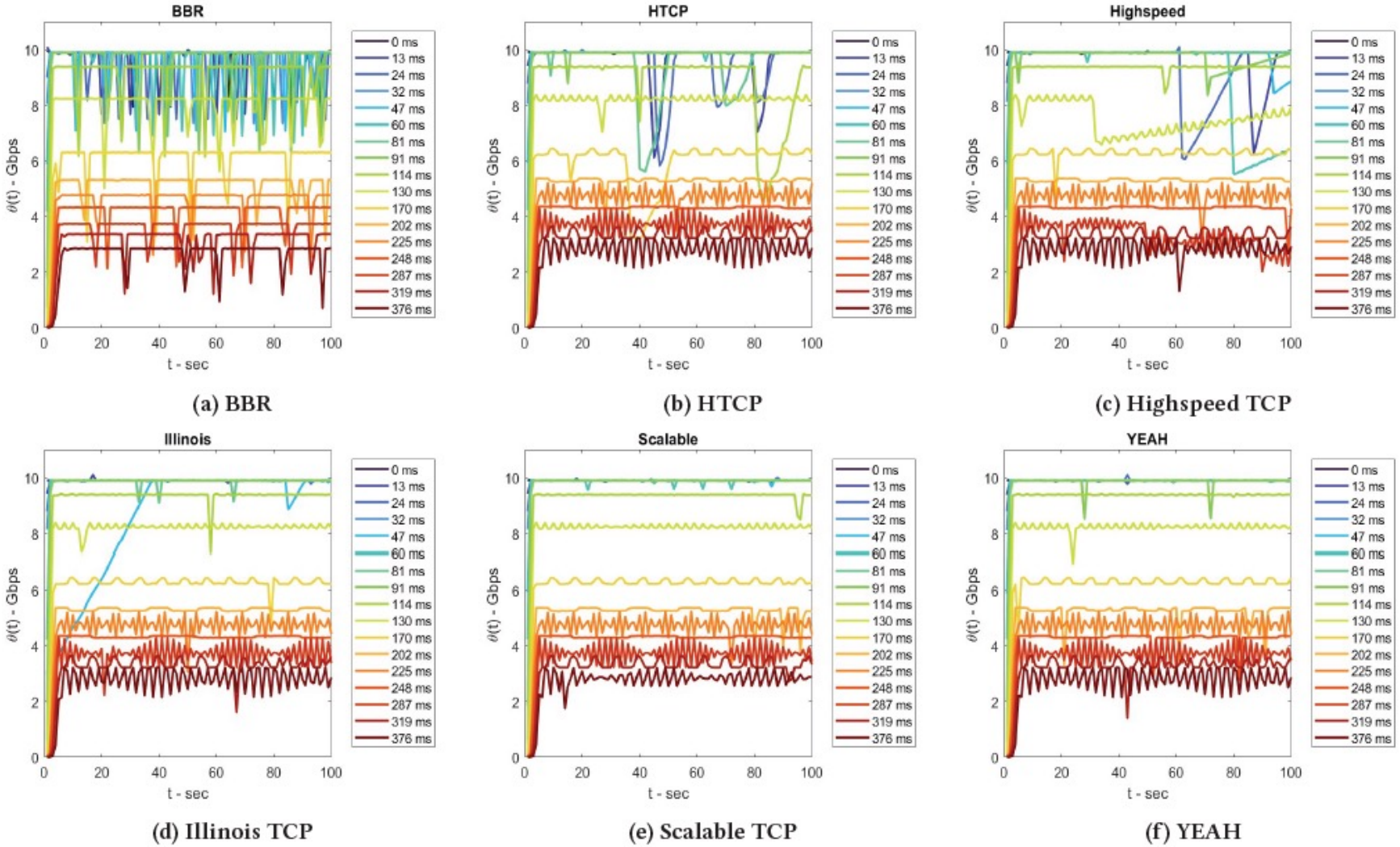
$$\frac{d\Theta}{d\tau} = \frac{1}{T_O} \int_0^{T_O} \frac{\partial \theta(\tau, t)}{\partial \tau} dt, \qquad (2)$$

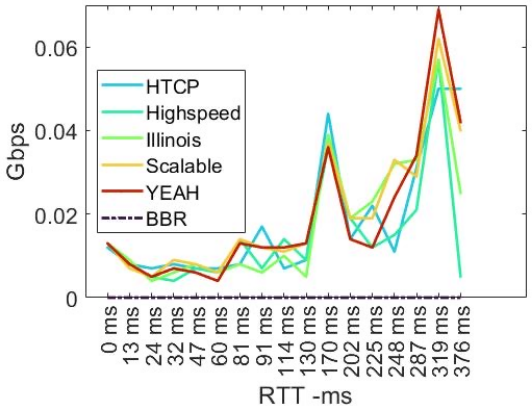
- indicates that lowering of $\theta(\tau, t)$ with respect to τ results in lower through profile
- higher deviations result in lower area $\int_0^{T_O} \theta(\tau, t) dt,$ and hence lower $\Theta(\tau)$
- lowering of throughput profile is consequence of higher variations in time trajectories

We will illustrate these effects using measurements next

Time Traces: Group 1

Group1: difference in throughput profiles with respect to BBR are positive for HTCP, Highspeed TCP, Illinois TCP, Scalable TCP, and YEAH





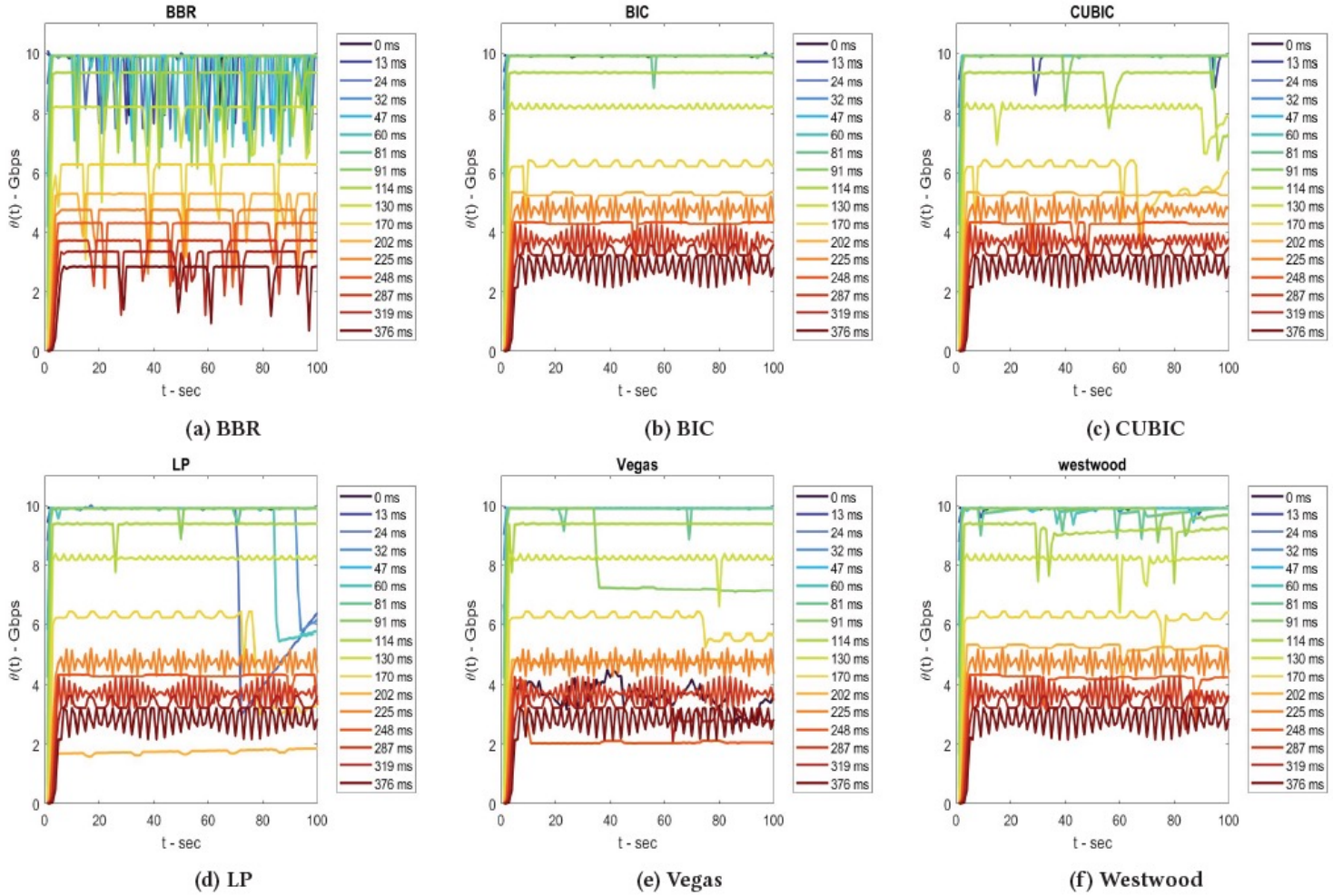
BBR variations are higher at lower RTT

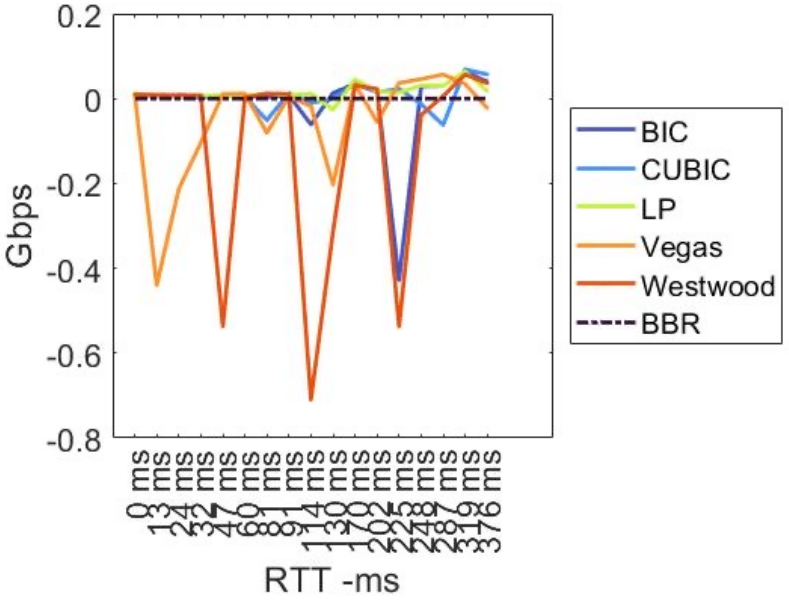
All trajectories have nearly periodic variations

Trajectories of BBR different from others - provide critical insights into its lower throughput

Time Traces: Group 2

- **Group 2**: throughput profile differences for BIC, CUBIC, LP, Vegas and Westwood with respect to BBR are mixed





BBR variations are higher at lower RTT

Some trajectories do not recover well after losses

Time Traces Observations

Collected throughput time traces of eleven TCP versions that correspond to throughput profiles

- In all cases, lowering of θ (τ, .) as τ increased - reflected in lower Θ values as indicated by Eq. (2).
- Lowering of θ visible for RTT larger than 114 ms for all TCP versions- BBR variations are distinct

BBR: variations for lower RTTs are noticeably higher for BBR compared to all others

- Vegas TCP not aggressively loss-based has noticeably smaller variations compared to BBR
- For larger RTT values, approximate periodic trends with two frequencies
  - one slower and other faster - increasing magnitude for all other versions as RTT is increased

BBR: there are nearly periodic trends with two qualitative differences from others:

- extent of variations decreases overall with increasing RTT - opposite of all others
- trajectories have nearly periodic spikes rather than two frequency oscillations in all others

Analytical results:

- higher variations in BBR trajectories result in lower integral in Eq. (1)
- hence lower throughput observed in profiles

Poincare Maps

Poincare map of time trace $\theta(\tau, t)$: value at next time t_{i+1} as a function of current value at time t_i

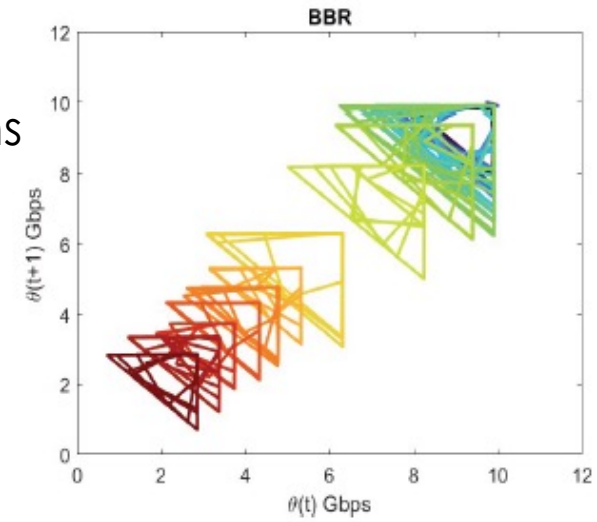
- captures richness of time dynamics as reflected by its geometry - spread and density in our case
- it has been a useful tool to study the dynamics of TCP and UDP based transport methods

TCP traces are collected for 100 seconds at 1 second duration, and the corresponding Poincare maps of BBR, HTCP, Highspeed TCP, Scalable TCP, and YEAH

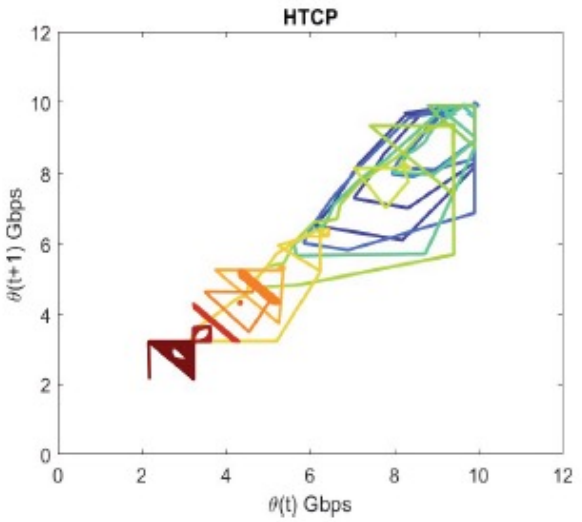
Poincare Maps: BBR, HTCP, Highspeed TCP, Scalable TCP, and YEAH

- BBR larger variations and hence lower throughput profiles
- regions of smaller RTTs are wider

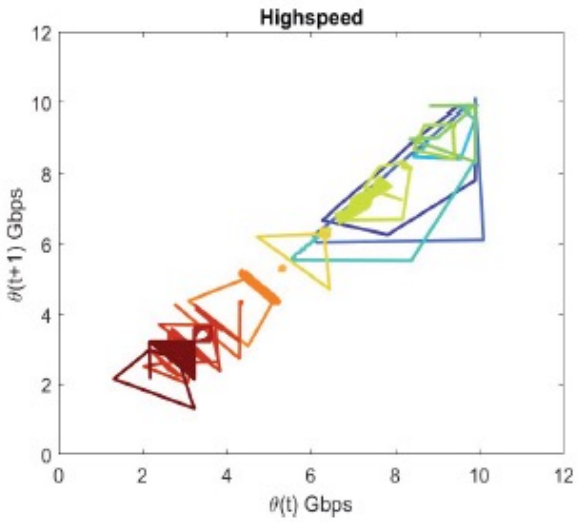
opposite trends in Illinois TCP, Scalable TCP and YEAH



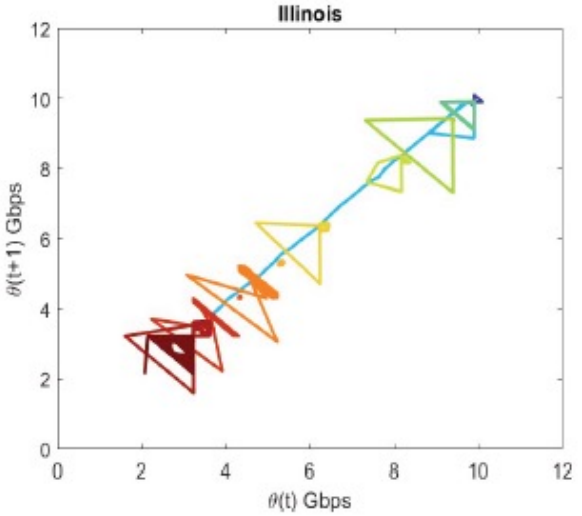
(a) BBR



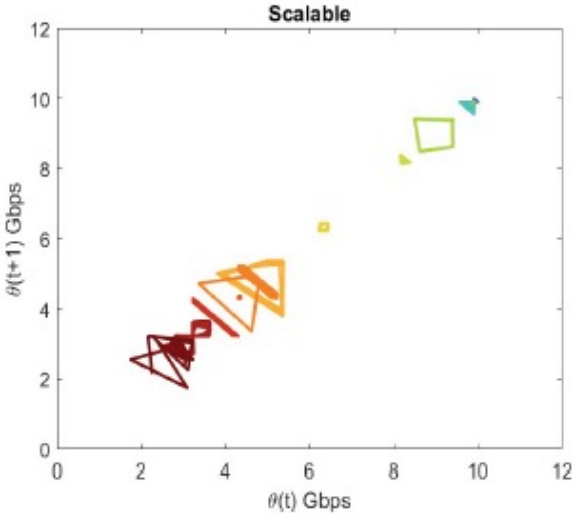
(b) HTCP



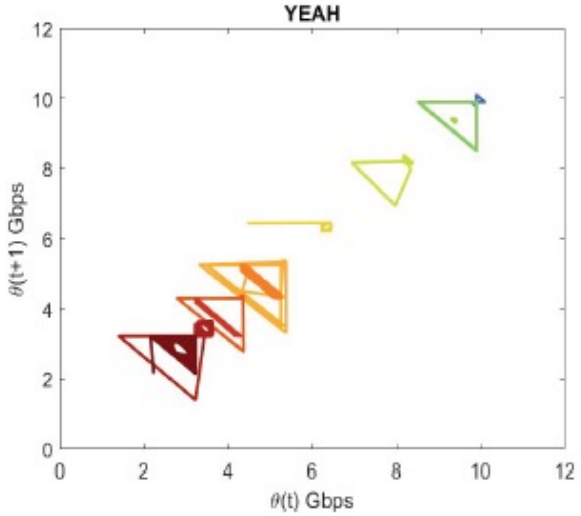
(c) Hispeed TCP



(d) Illinois TCP



(e) Scalable TCP



(f) YEAH

Poincare Maps

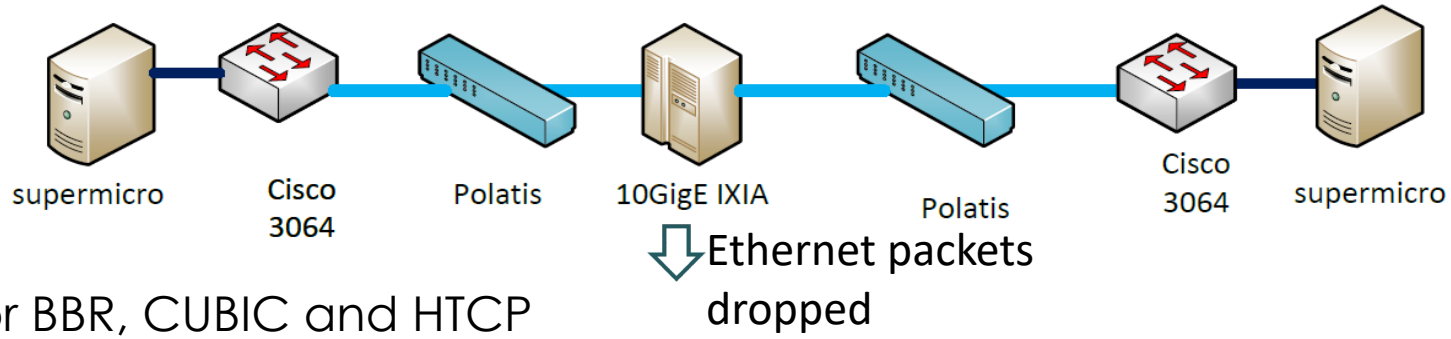
Poincare regions for a fixed RTT τ mostly consists of near triangles indicating nearly cyclic behavior of the trajectories observed in all cases

- simpler than more complex regions of chaotic systems

Poincare map of BBR is different from others - provides critical insights into its lower throughput:

- Poincare regions associated with different RTT values are wider for BBR compared to the corresponding regions of others
  - larger variations and hence lower throughput profiles. Second, the regions corresponding to smaller RTTs are wider for BBR and they shrink as RTT is increased.
- Overall, opposite trend in others, particularly, Illinois TCP, Scalable TCP and YEAH

Throughput Profiles: External Losses



Collected throughput measurements for BBR, CUBIC and HTCP

- losses using IXIA emulator - drops Ethernet packets periodically or under uniform distribution
- losses are external to those created by TCP loss-based protocols
  - factors such as IP, Ethernet and physical losses, and also competing traffic flows over shared networks
- limited in capturing underlying processes and structure of losses on shared networks of very varied origins

Loss-based TCP versions have different effects compared to self-induced losses

- used to adjust window sizes or flow rates during the congestion avoidance phases;
  - self-induced losses are controlled by protocol dynamics
  - external loss rate may be highly varying

BBR dynamics: effected in a fundamentally different way and result in:

- profiles remain concave as Ethernet packets dropped at an increasing rate up to 1 in 500
- CUBIC and BBR start off as concave at zero losses, begin transition to convex and become entirely convex as the loss rate is increased to this value
- measurements on next slide

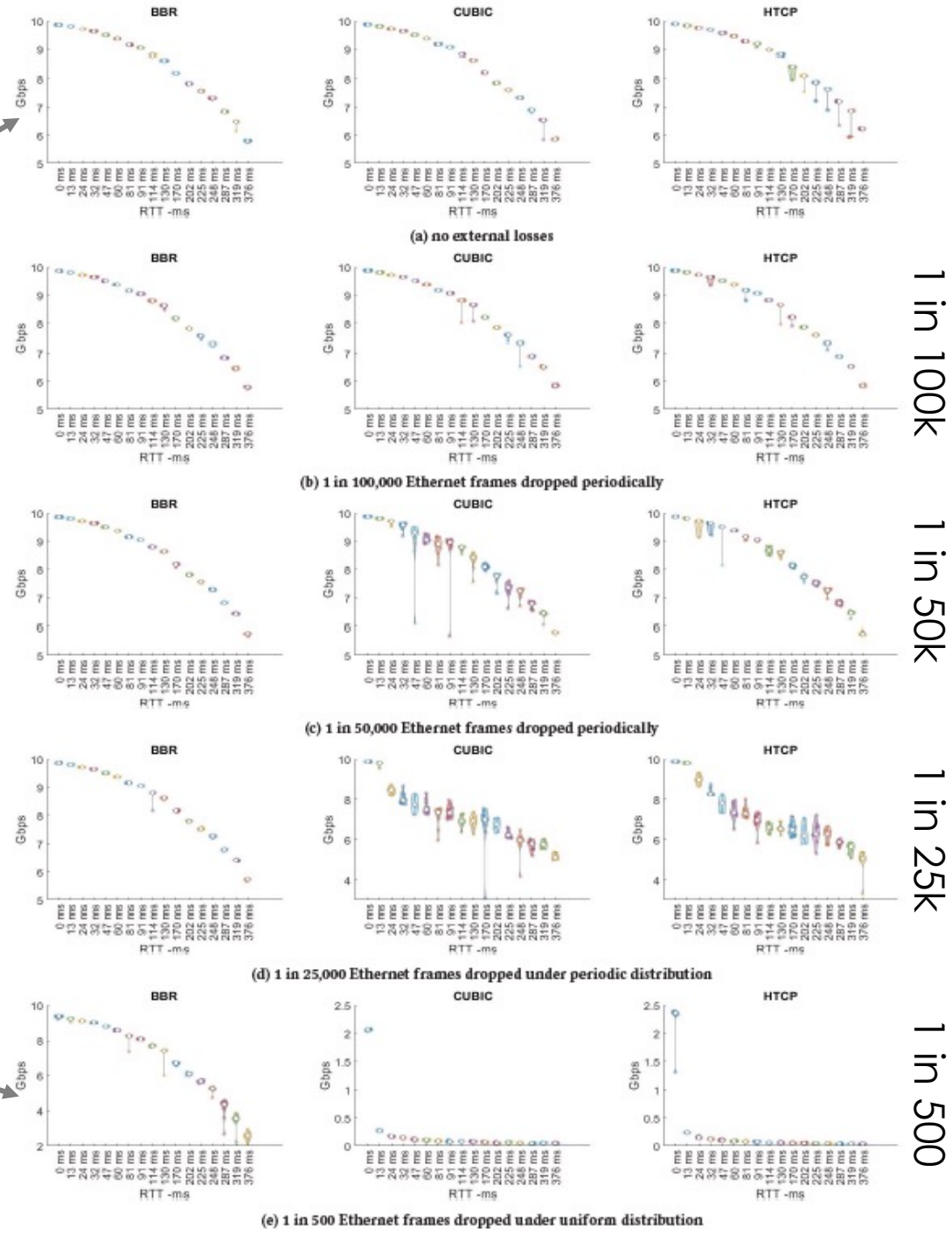
Throughput Profiles Under Losses

BBR throughput profile remains concave:
- higher than CUBIC and HTCP - they become increasingly convex with increased loss rate
- BBR provides higher and sustained throughput

BBR, CUBIC and HTCP: remain concave when
- periodic loss rate is 1 in 100,000
- retain overall shape at loss rate 1 in 50,000
- CUBIC shows more variability

CUBIC and HTCP: begin transition to convex with loss rate doubled to 1 in 25,000 under periodic losses
- transitions of CUBIC and HTCP different for periodic and uniform random losses

Drastic difference at loss rate of 1 in 500:
- CUBIC and HTCP significantly lower and convex
- BBR's profile remains concave throughput
- 9.3 Gbps for shorter- 2 Gbps for longest connection



Conclusions

Summary:

Extensive throughput measurements and time traces:

- dedicated connections for eleven TCP versions available for typical Linux servers
- time dynamics of these protocols provide critical insights into throughput profiles
- distinctly different time dynamics of model-based BBR compared to loss-based TCP versions

Practically useful information:

- under low external losses: higher temporal variations of BBR at smaller RTTs
  - result in lower throughput profiles compared to others such as Hamilton TCP
- Poincare map regions generated from the time traces indicate:
  - nearly periodic dynamics but with observable differences between BBR and other versions
  - Higher temporal variations lead to lower throughput profiles

Future Work:

- Detailed analysis of time trajectories and Poincare maps under various types of external losses
  - insights on conditions under which BBR profiles will be superior to others.
- Comparison of emulation-based results with physical network measurements
  - Google cloud networks and UltraScience Net
- UDP based protocols for dedicated connections, e.g., QUIC
  - UDT exhibits complex dynamics which can be related to its throughput profiles.
- Analytical TCP models for dedicated connections
  - extend current models and analyses based on Poincare maps and throughput profiles

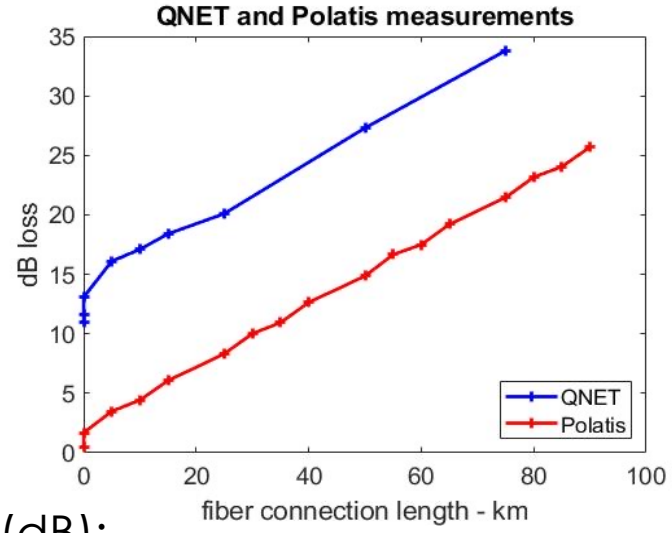
Thank you

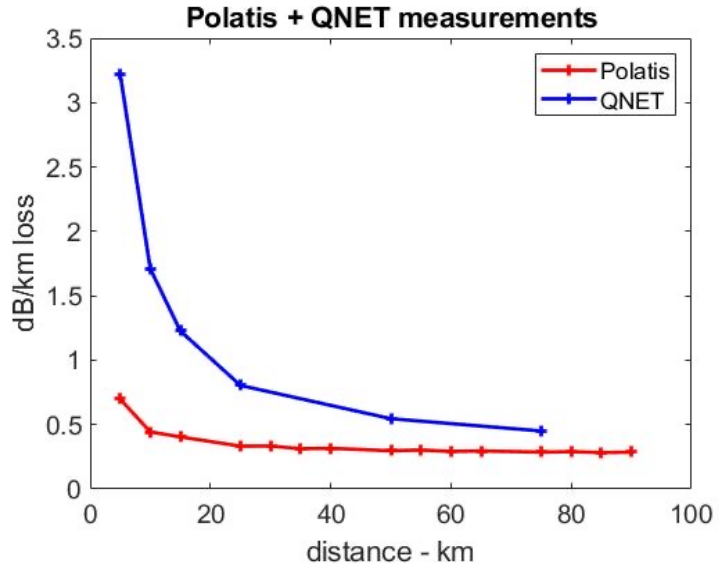
19

Light-level Measurements

For conventional and quantum connection
- light levels (dBm) measured on all-optical switch - Polatis measurements.

For quantum connections,
- additional light level measurements at source and detectors in node Alice - QNET measurements





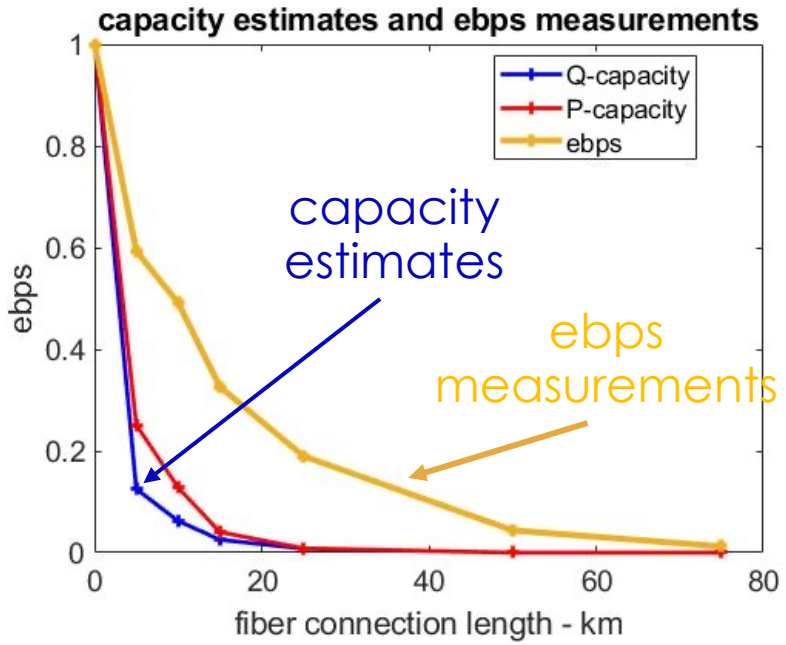
Connection loss (dB):
- subtract destination from source levels
- function of connection length in km - nearly linear
- constant additional 15 -20 dB loss for quantum connections
  - additional fiber connections to Alice and Dave - direct and via Bob and at source and detectors.

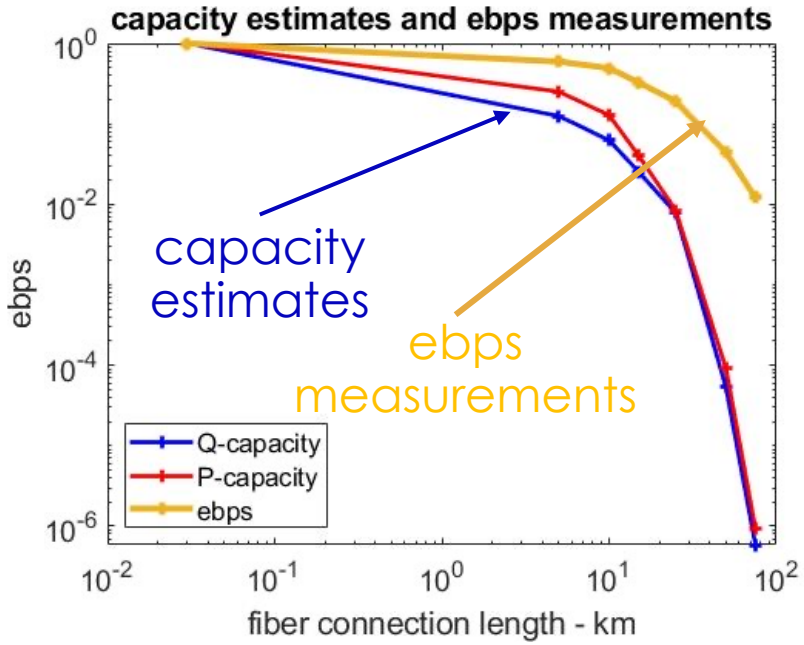
Loss rate per distance estimate - divide connection losses by length,
- decreasing trend with connection length
- higher values at shorter connections
  - higher fraction of losses due to
  - fiber patches at nodes, cross-connects optical switch, and at source and detectors

Comparison: Models and Physical Connections

Both ebps measurements and corresponding capacity estimates

- while both decrease rapidly, **ebps measurements are higher than capacity estimates**





Postulation: degree of misalignment between

- assumptions used for the capacity estimation and
- QNET conditions under which the light intensity measurements are collected

Further refinements needed to correlate measurements with theoretical estimates

Comparison: somewhat similar to Shannon limit estimation (conventional optical connections) several refinements needed to correlate measurements with theoretical estimates