KNIT 6: A FABRIC Community Workshop

Unconference Sessions

Understanding the Performance of TCP BBRv2 using FABRIC

Jose Gomez, Elie Kfoury, Ali Mazloum, Jorge Crichigno
University of South Carolina
http://ce.sc.edu/cyberinfra

University of South Carolina (USC)

Texas Advanced Computing Center
Austin, TX

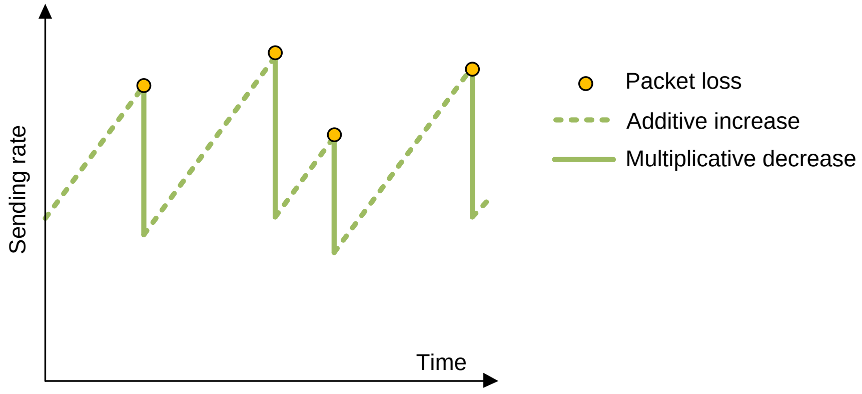
April 25th, 2023

Agenda

- TCP Traditional Congestion Control Algorithms
- BBR
- Motivation
- Results and Evaluations
- Demo
- Limitations
- Lessons Learned

TCP Traditional Congestion Control

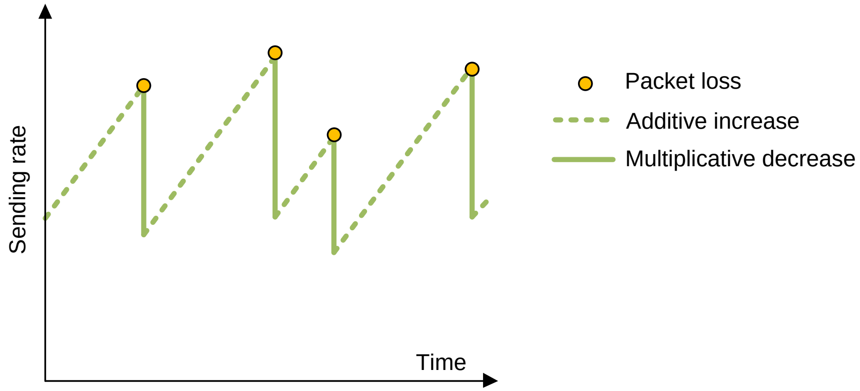
- The principles of window-based CC were described in the 1980s[1]
- Traditional CC algorithms follow the additive-increase multiplicative-decrease (AIMD) form of congestion control

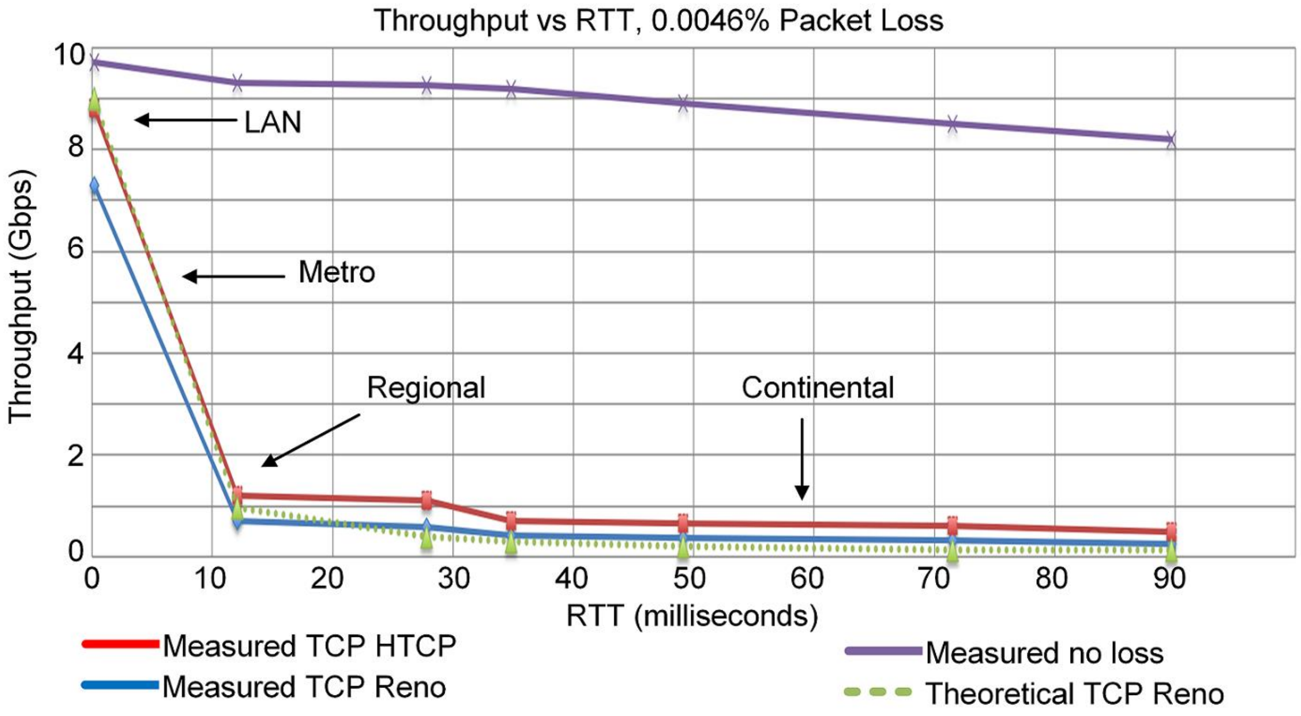


[1] V. Jacobson, M. Karels, Congestion avoidance and control, ACM SIGCOMM Computer Communication Review 18 (4) (1988).

TCP Traditional Congestion Control

- The principles of window-based CC were described in the 1980s[1]
- Traditional CC algorithms follow the additive-increase multiplicative-decrease (AIMD) form of congestion control

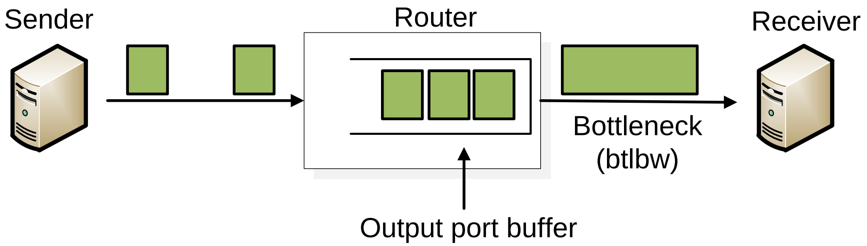


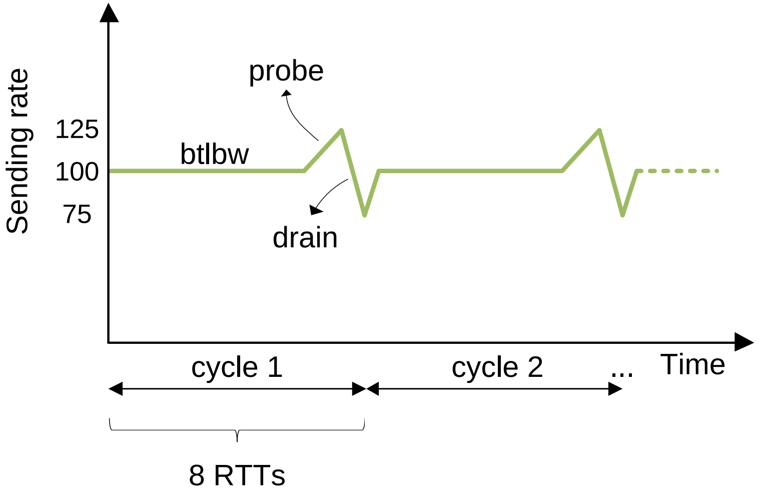


[1] V. Jacobson, M. Karels, Congestion avoidance and control, ACM SIGCOMM Computer Communication Review 18 (4) (1988).

BBR: Model-based CC

- TCP Bottleneck Bandwidth and RTT (BBR) is a rate-based congestion-control algorithm[1]
- BBR represented a disruption to the traditional CC algorithms:
  - is not governed by AIMD control law
  - does not the use packet loss as a signal of congestion
- At any time, a TCP connection has one slowest link bottleneck bandwidth (btlbw)

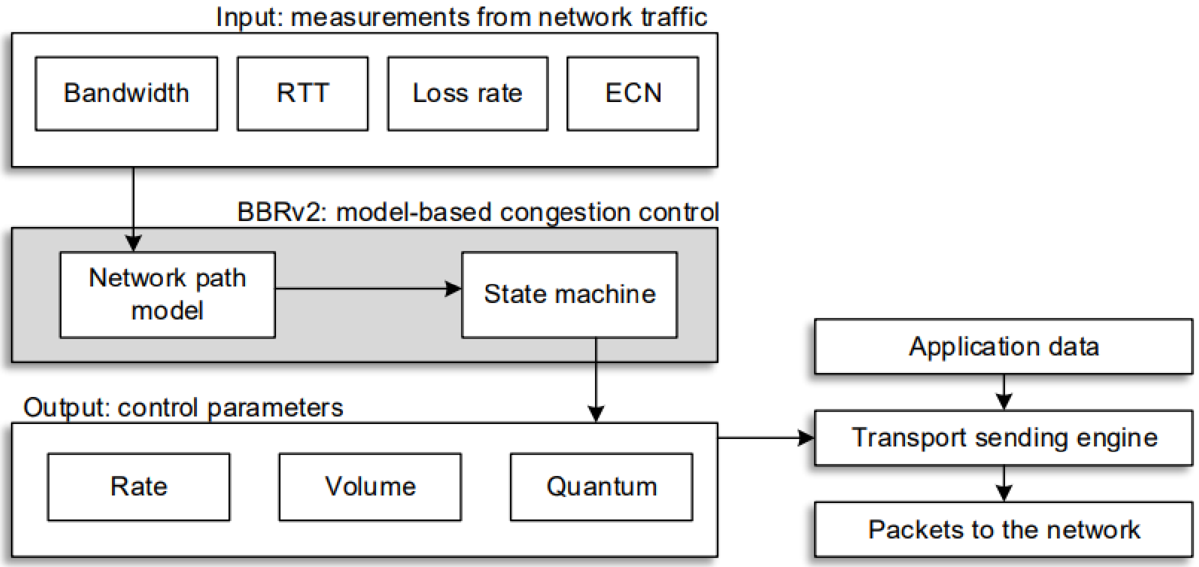




[1] N. Cardwell et al. "BBR v2, A Model-based Congestion Control." IETF 104, March 2019.

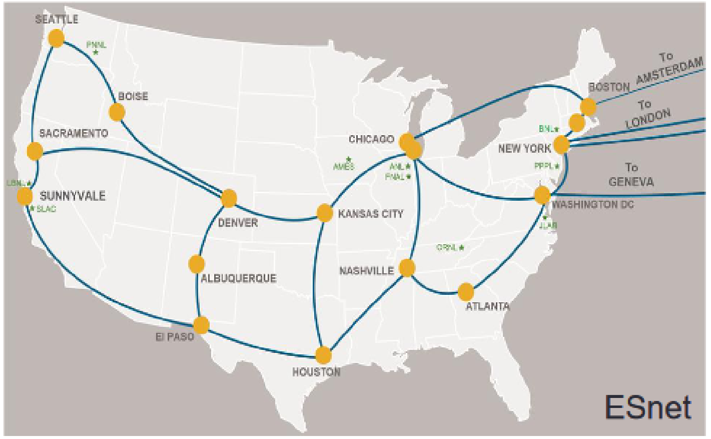
BBRv2

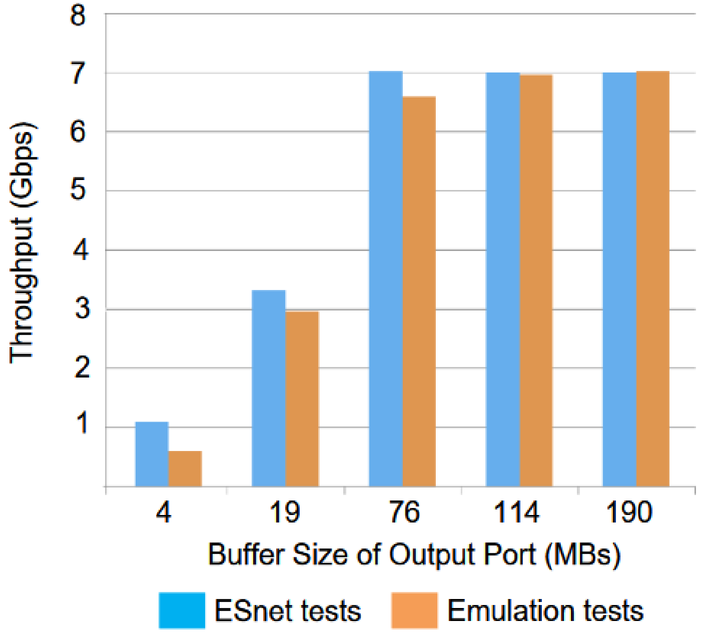
- BBRv2 overcomes the shortcomings of BBRv1
- BBRv2 measures the bandwidth, the RTT, the packet loss rate, and the ECN mark rate
- The measurements are used to estimate the bandwidth-delay product (BDP)
- BBRv2 does not always apply a multiplicative decrease for every round trip where packet loss occurs

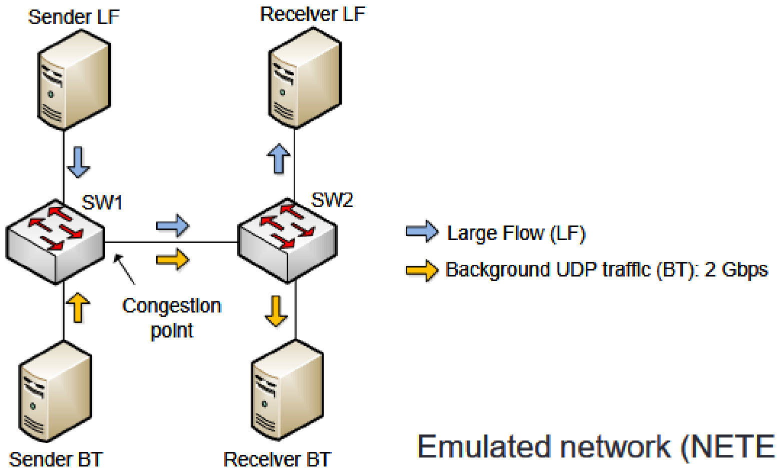


ESnet vs Emulated Network

- ESnet vs emulated network, 10 Gbps links, 70 msec RTT (NETEM emulates delays)





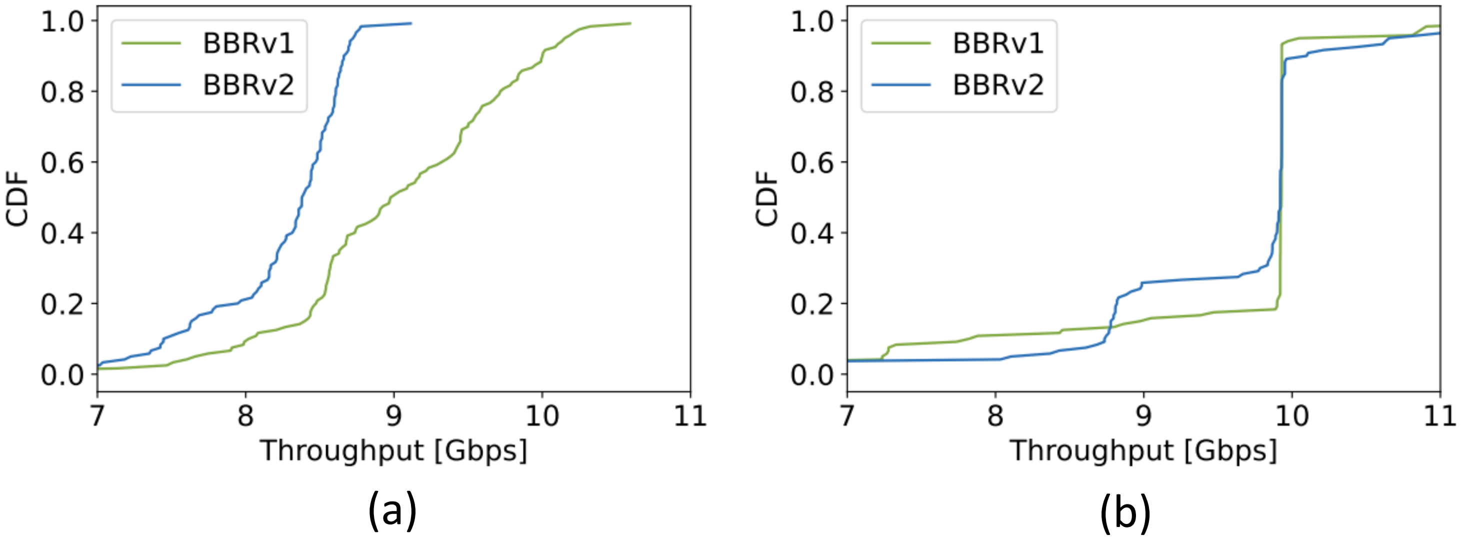


Emulated network (NETEM)

M. Smitasin and B. Tierney, “Evaluating network buffer size requirements,” Proc. Technol. Exchange Workshop, Oct. 2015. Online: https://meetings.internet2.edu/media/medialibrary/2015/10/05/20151005-smitasin-buffersize.pdf

Motivation

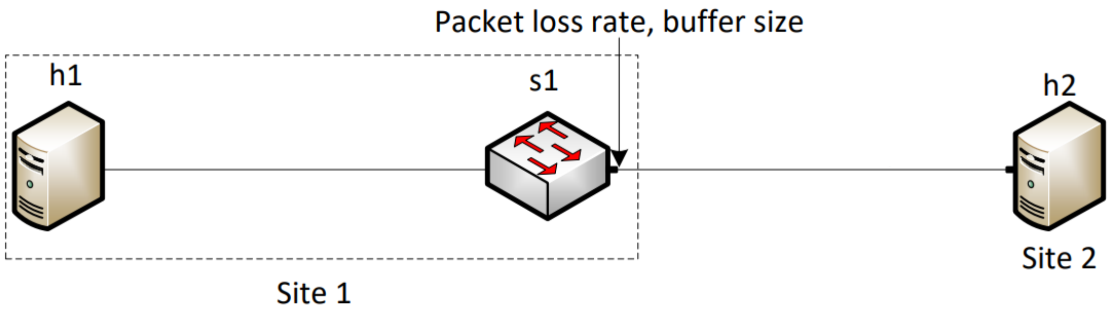
- Understanding the behavior of BBRv2 in a testbed with real propagation delay
- Observing the dynamics of BBRv2 in a Wide Area Network (WAN)
- Analyzing the differences between an emulated environment and a real testbed
- This work leverage the distributed architecture of the FABRIC testbed to reproduce WAN conditions and test the performance of BBRv2



CDF of the bottleneck bandwidth estimation of BBRv1 and BBRv2.
(a) with 45ms emulated delay. (b) with 45ms propagation delay.

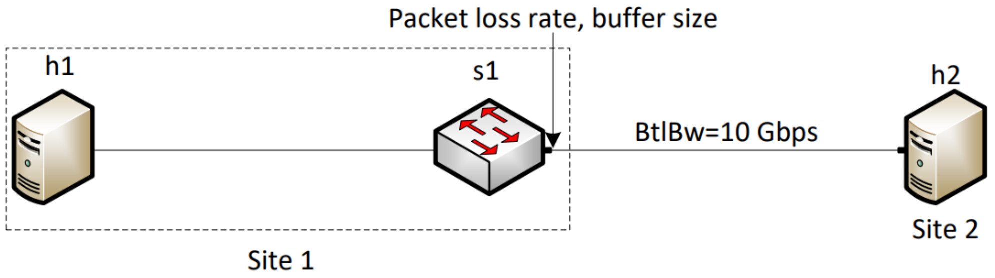
Experimental Setup

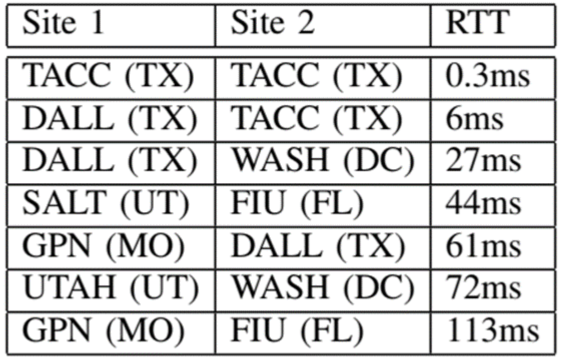
- The experiments used a software switch to limit the rate and emulate packet losses
- The rate is limited using the Token Bucket Filter (TBF) in Linux
- Packet losses are emulated using NETEM
- The sites are selected based on the experiment



Results and Evaluations

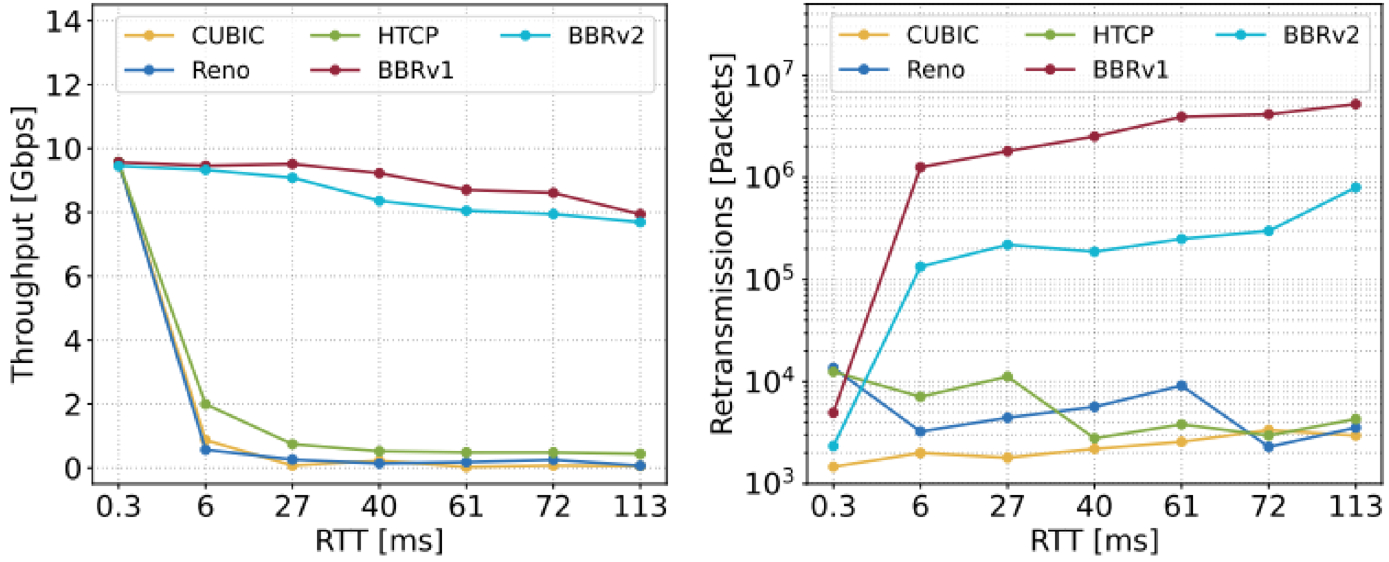
- Experiment 1: Performance in a WAN with packet losses
- The rate is limited to 10Gbps
- The emulated packet loss rate is 0.0046% (i.e., 1/22,000)





Site 1	Site 2	RTT
TACC (TX)	TACC (TX)	0.3ms
DALL (TX)	TACC (TX)	6ms
DALL (TX)	WASH (DC)	27ms
SALT (UT)	FIU (FL)	44ms
GPN (MO)	DALL (TX)	61ms
UTAH (UT)	WASH (DC)	72ms
GPN (MO)	FIU (FL)	113ms

Topology used for the evaluations



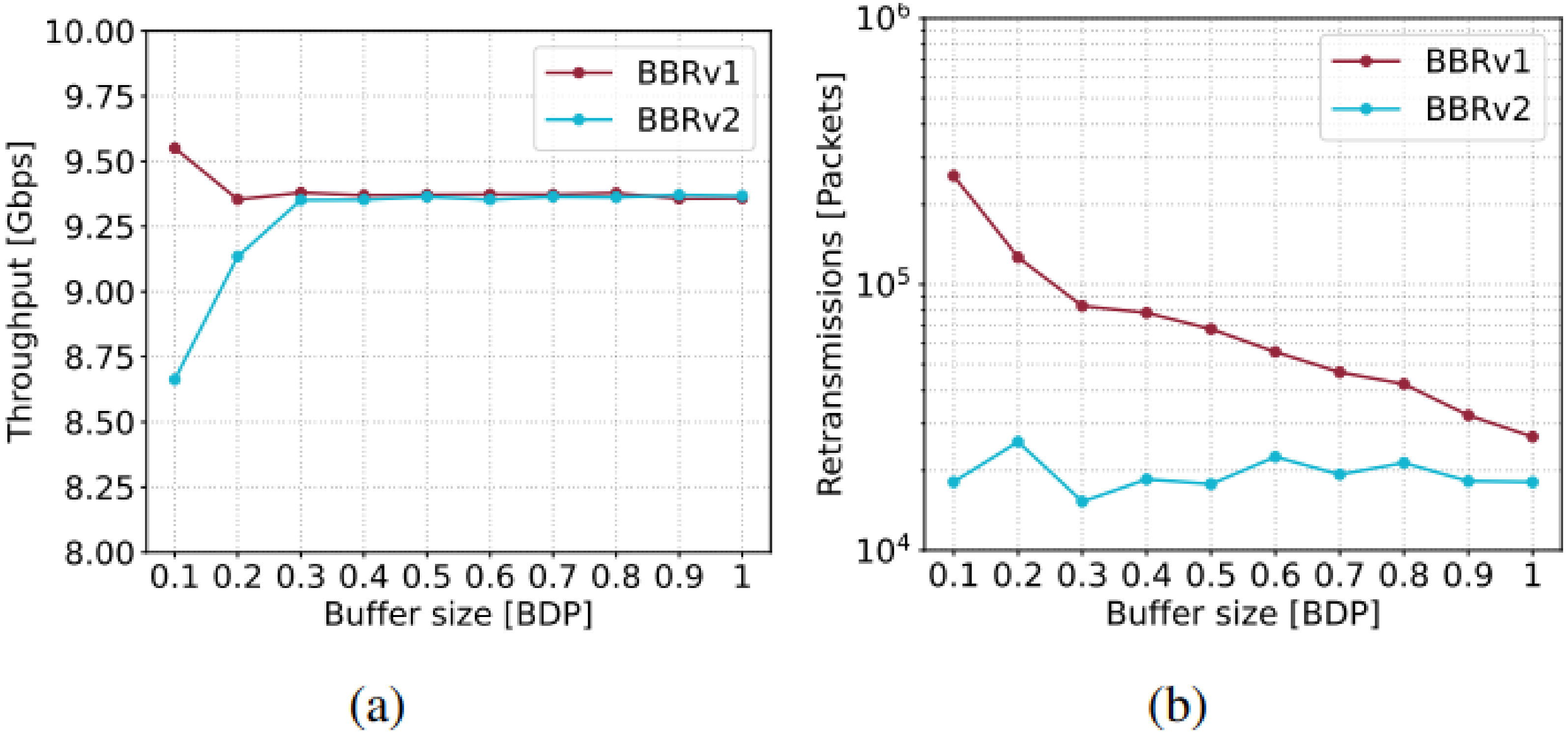
(a)

(b)

Performance of CUBIC, Reno, HTCP, BBRv1, and BBRv2 as a function of the RTT. (a) Throughput. (b) Retransmissions.

Results and Evaluations

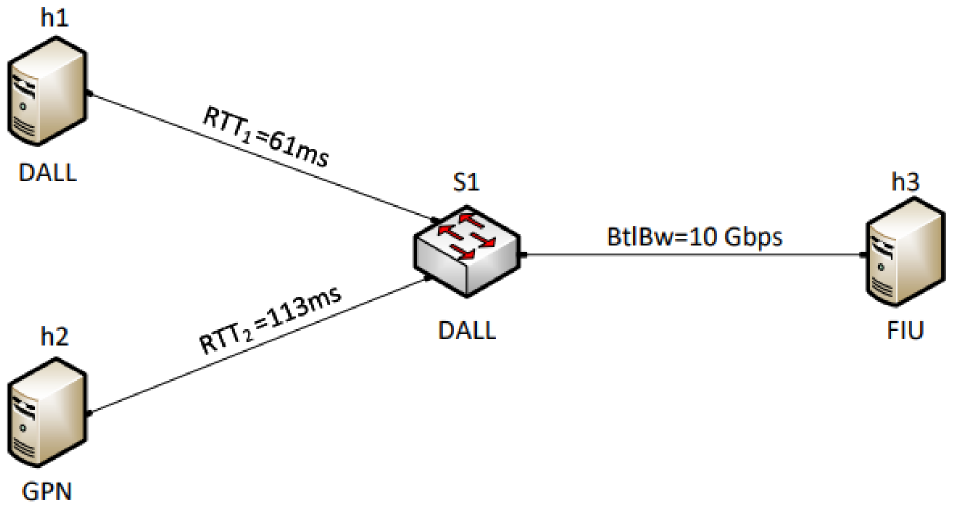
- Experiment 2: Retransmissions as a function of the buffer size
- The RTT between the hosts is 45 milliseconds (SALT, FIU)
- BBRv2 presents lower packet loss rates than BBRv1



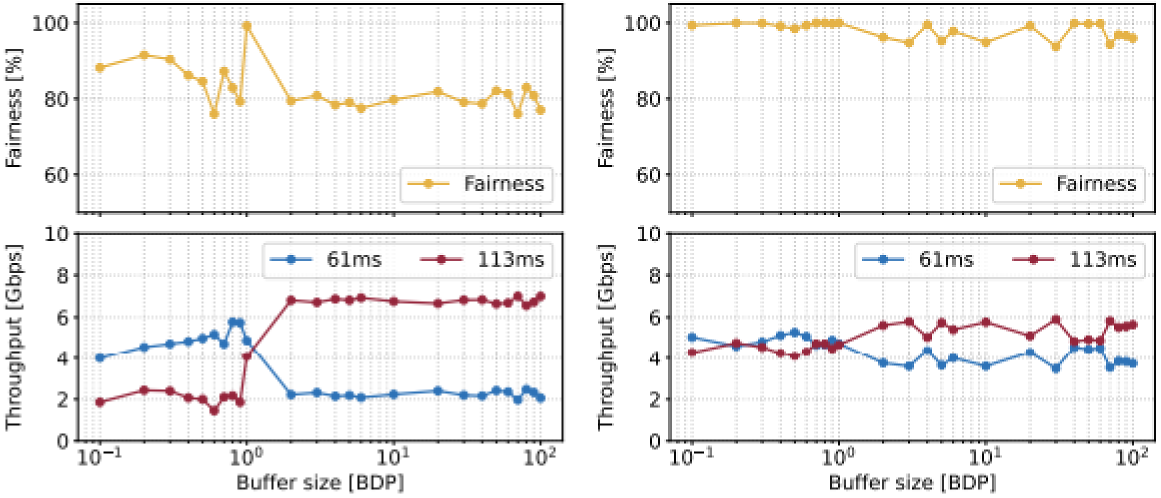
Performance test as a function of the BDP. (a) Throughput. (b) Retransmissions.

Results and Evaluations

- Experiment 3: RTT unfairness
- RTT unfairness occurs when flows with smaller RTTs obtain a higher throughput
- BBRv1 flows present the opposite behavior
- BBRv2 reduces the RTT unfairness of competing flows



Topology used for running the RTT unfairness experiment

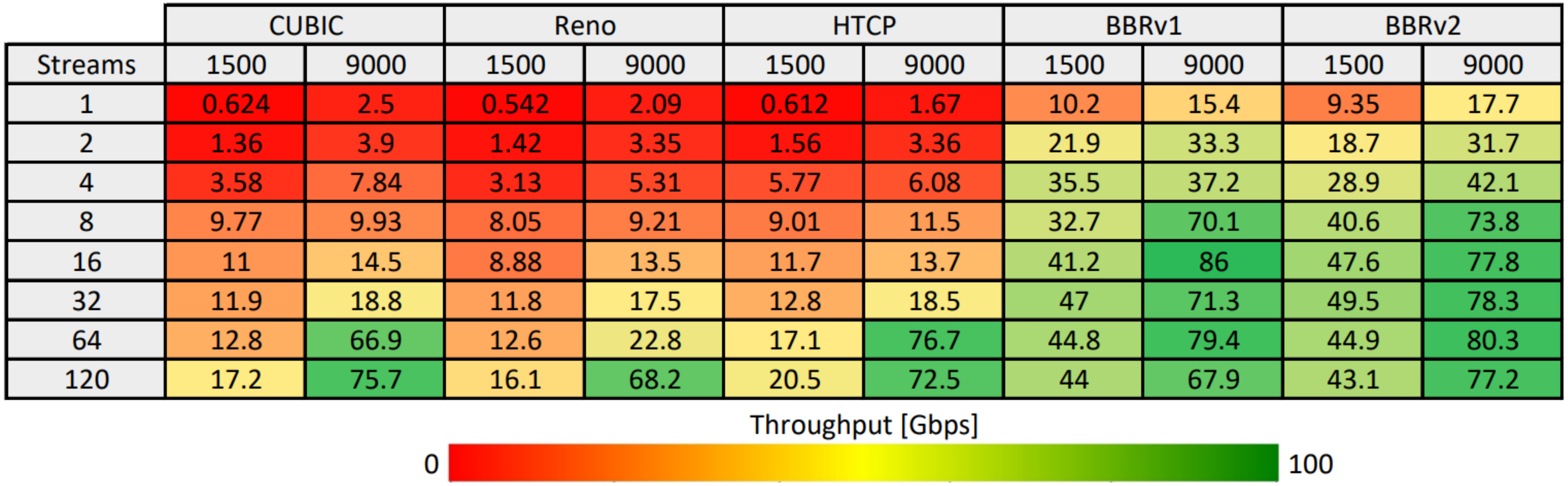


(a) BBRv1 with Tail Drop (b) BBRv2 with Tail Drop

Fairness index and throughput as functions of the buffer size for two competing flows. (a) BBRv1. (b) BBRv2.

Results and Evaluations

- Experiment 4: Parallel streams and different MTUs
- The rate is not limited (i.e., 100Gbps)
- The RTT between the sites is 26 milliseconds (DALL, SALT)
- BBRv1 and BBRv2 achieve throughputs over 70Gbps with eight streams

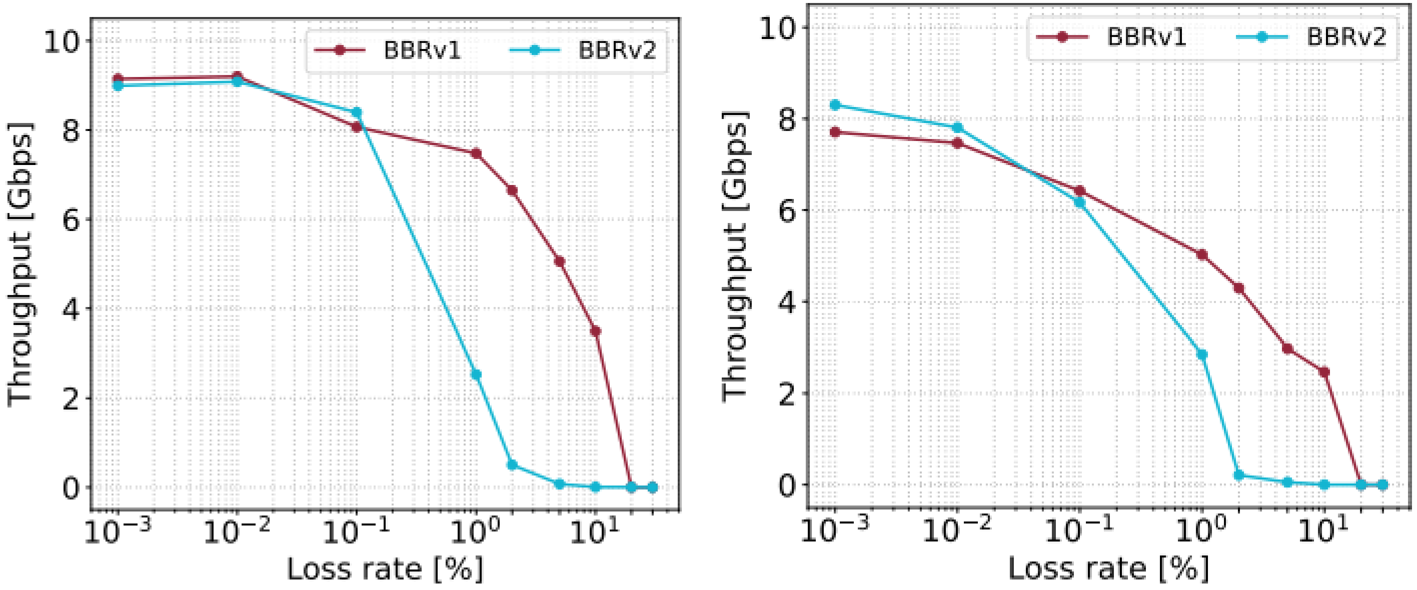


	CUBIC		Reno		HTCP		BBRv1		BBRv2	
Streams	1500	9000	1500	9000	1500	9000	1500	9000	1500	9000
1	0.624	2.5	0.542	2.09	0.612	1.67	10.2	15.4	9.35	17.7
2	1.36	3.9	1.42	3.35	1.56	3.36	21.9	33.3	18.7	31.7
4	3.58	7.84	3.13	5.31	5.77	6.08	35.5	37.2	28.9	42.1
8	9.77	9.93	8.05	9.21	9.01	11.5	32.7	70.1	40.6	73.8
16	11	14.5	8.88	13.5	11.7	13.7	41.2	86	47.6	77.8
32	11.9	18.8	11.8	17.5	12.8	18.5	47	71.3	49.5	78.3
64	12.8	66.9	12.6	22.8	17.1	76.7	44.8	79.4	44.9	80.3
120	17.2	75.7	16.1	68.2	20.5	72.5	44	67.9	43.1	77.2

Average throughput belonging to different CCAs as a function of the number of streams and the MTU.

Results and Evaluations

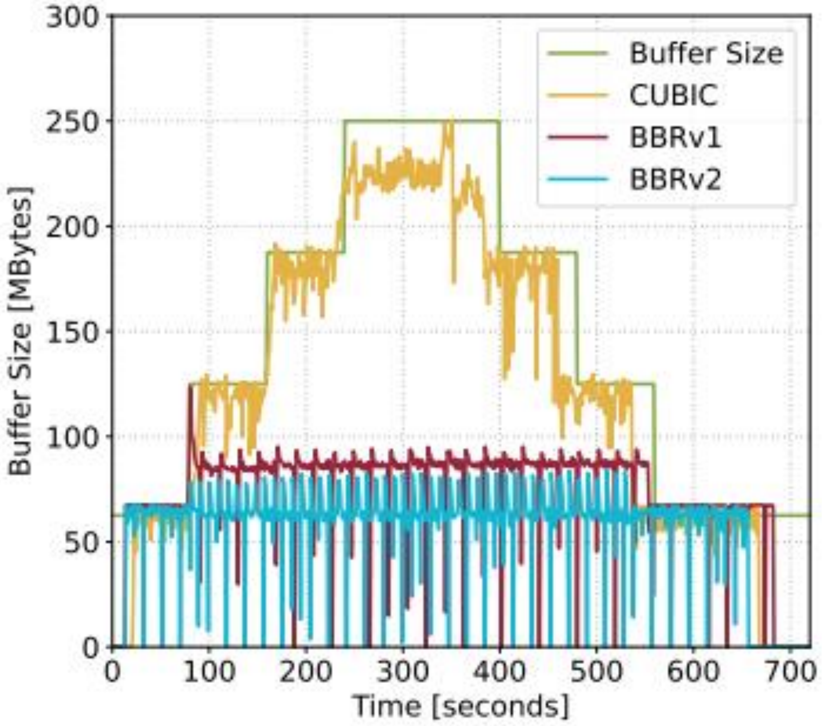
- Experiment 5: Throughput as a function of packet losses
- The performance of BBRv2 is close to that of BBRv1 for loss rates less than 1%
- (a) RTT = 26ms (DALL, SALT)
- (b) RTT = 57ms (UCSD, UMASS)



(a) (b)

Results and Evaluations

- Experiment 6: Queue occupancy
- Link is limited to 10Gbps, the RTT is 50ms
- Bandwidth-delay Product = 10Gbps * 50ms = 62.5MB
- BBRv2 keeps the queue occupancy around BDP, even with bloated buffers



Limitations

- Configuration of the intermediary devices (e.g., routers and switches)
  - Link capacity
  - Router buffer size
  - Queue allocation
- The experiments modified the buffer size of a software switch
- Shared Network Interface Cards (NICs)
- Performance isolation

Lessons Learned

- FABRIC can be used to test protocols and applications under WAN conditions
- The testbed can support a wide variety of experiments
- Its programmable infrastructure allows defining customized network environments
- BBRv2 provides improved fairness compared to BBRv1, particularly when dealing with flows that have different RTTs
- BBRv2 can achieve comparable throughput to BBRv1, while also exhibiting a lower retransmission rate

Future Work

- Leveraging programmable data plane switches for fine-grained measurements
- Replicating the results in FABRIC

UNIVERSITY OF South Carolina

This work is supported by NSF award number 2118311

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

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