BBR v2: A Model-based Congestion Control

IETF 105 Update

Neal Cardwell, Yuchung Cheng,

Soheil Hassas Yeganeh, Priyaranjan Jha, Yousuk Seung,

Ian Swett, Victor Vasiliev, Bin Wu, Matt Mathis

Van Jacobson

https://groups.google.com/d/forum/bbr-dev

IETF 105: Montreal, July 2019

Google

Outline

- BBR v2 open source "alpha/preview" release
  - Status of the BBR v2 code
  - Lab test results
  - Deployment status
- Conclusion

BBR v2 open source alpha/preview release

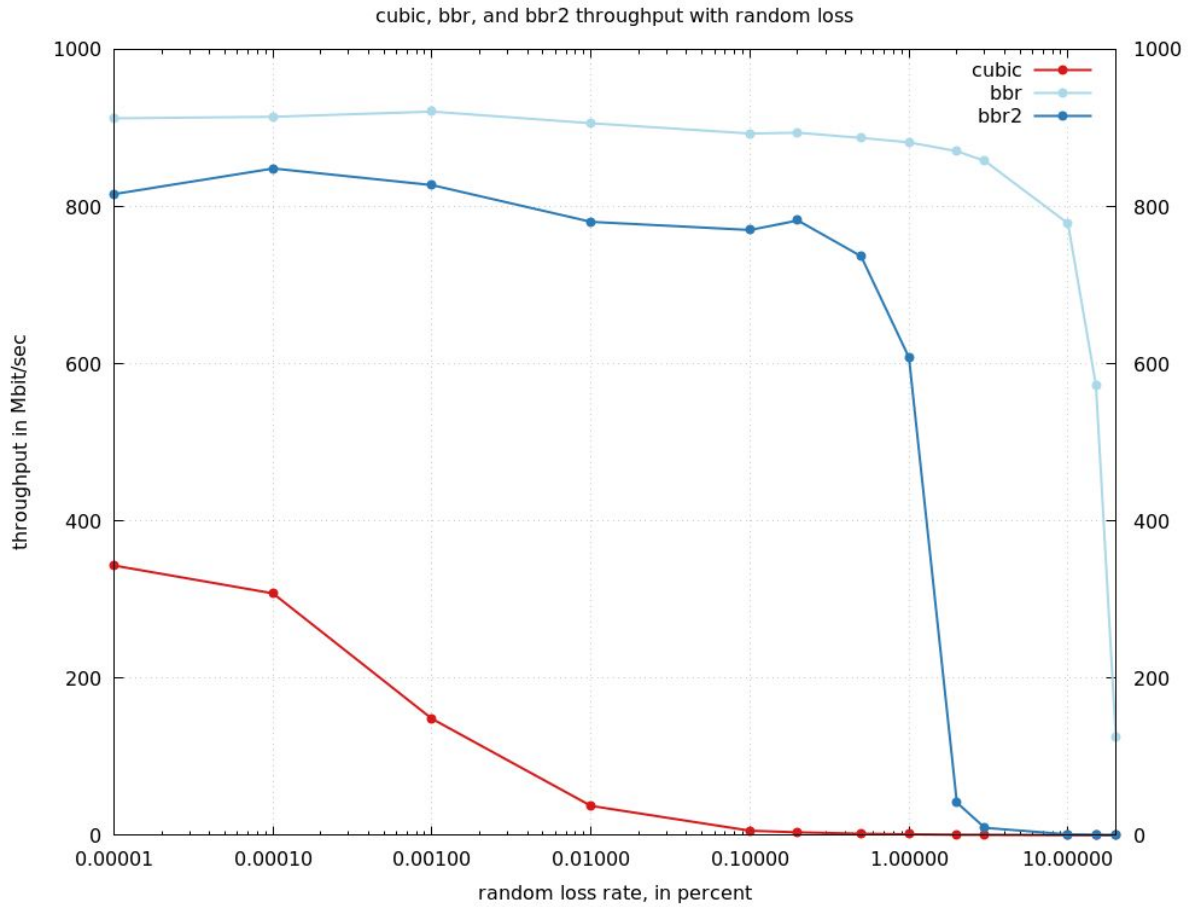
- Goal of this release: enable research collaboration and wider real-world testing
- We encourage researchers to dive in and help evaluate/improve BBR
  - We welcome patches with good solutions to issues
- BBR v2.0.alpha.1 (preview) code available as open source 2019-07-22 (IETF 105):
  - Linux TCP (dual GPLv2/BSD): **github.com/google/bbr/blob/v2alpha/README.md**
  - Chromium QUIC (BSD): on chromium.org in bbr2_sender.{ cc, h }
- TCP BBR v2 release includes test scripts used to generate graphs for these slides
  - These tests use network emulation via netem
- BBR v2 algorithm was described at IETF 104 [slides | video]

BBR v2: what's new?

- Properties maintained between BBR v1 and BBR v2:
  - High throughput with a targeted level of random packet loss
  - Bounded queuing delay, despite bloated buffers
- Improvements from BBR v1 to BBR v2 (as discussed at IETF 104 [slides | video]):
  - Improved coexistence when sharing bottleneck with Reno/CUBIC
  - Much lower loss rates for cases where bottleneck queue < 1.5*BDP
  - High throughput for paths with high degrees of aggregation (e.g. wifi)
  - Using DCTCP/L4S-style ECN signals
  - Vastly reduced the throughput reduction in PROBE_RTT
- Following are a few tests, to illustrate the core properties maintained and improved...
  - Metrics we're evaluating in these:
    - throughput, queuing latency, retransmit rate, fairness

BBR v2.0.alpha.1 lab test results

High throughput with target of 1% random loss



Bulk throughput

1 cubic, bbr, or bbr2

bw = 1Gbit/sec, min_rtt = 100ms

buf = 1*BDP

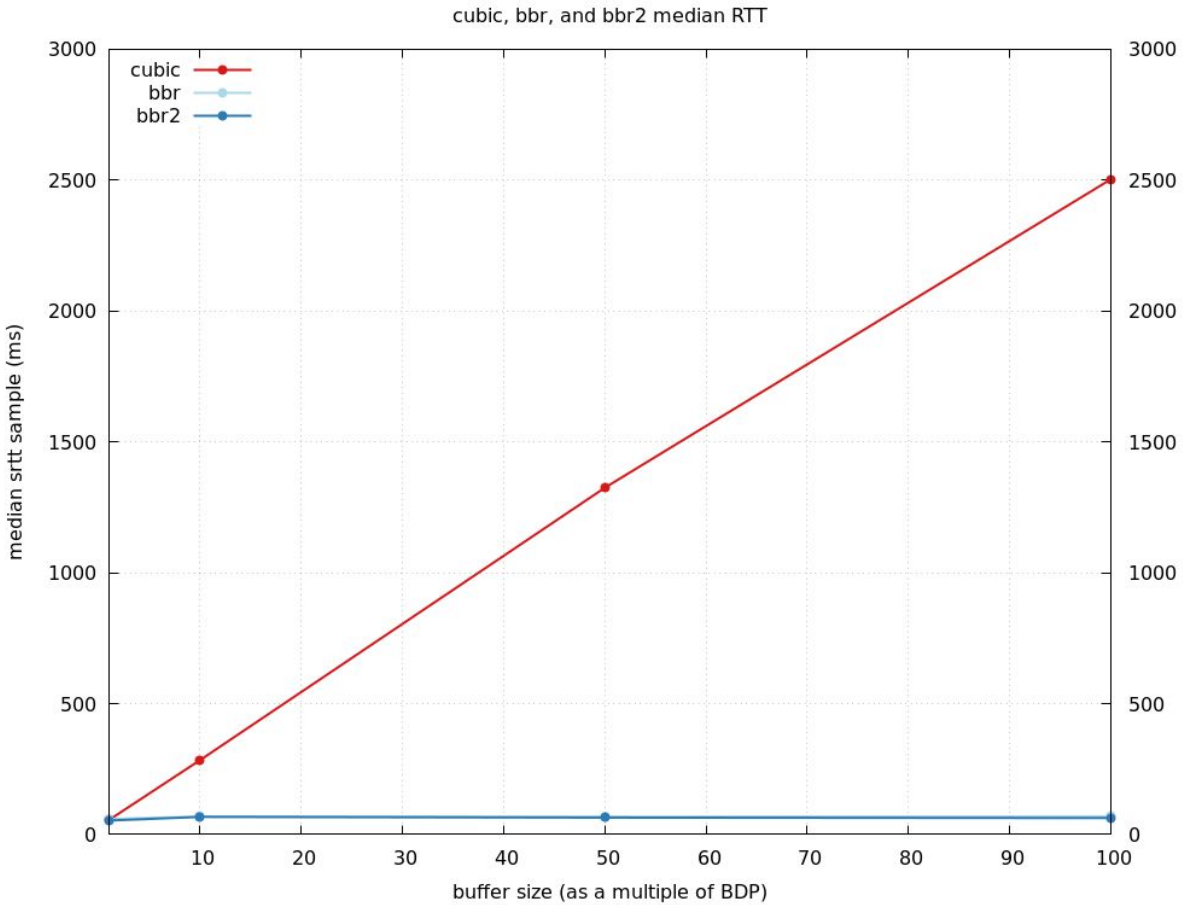
2 min. netperf TCP stream

loss={10^-5 , ..., 10^1, 15, 20} %

(Knee for bbr2 is bounded by explicit loss_thresh=2% design parameter.)

Y axis: p50 throughput of 10 trials

Low queue delay, despite bloated buffers



Latency from bulk flows

2 cubic or 2 bbr2

1st flow at t=0, 2nd at t=2s

bw = 50Mbit/sec, min_rtt = 30ms

2 min. netperf TCP stream

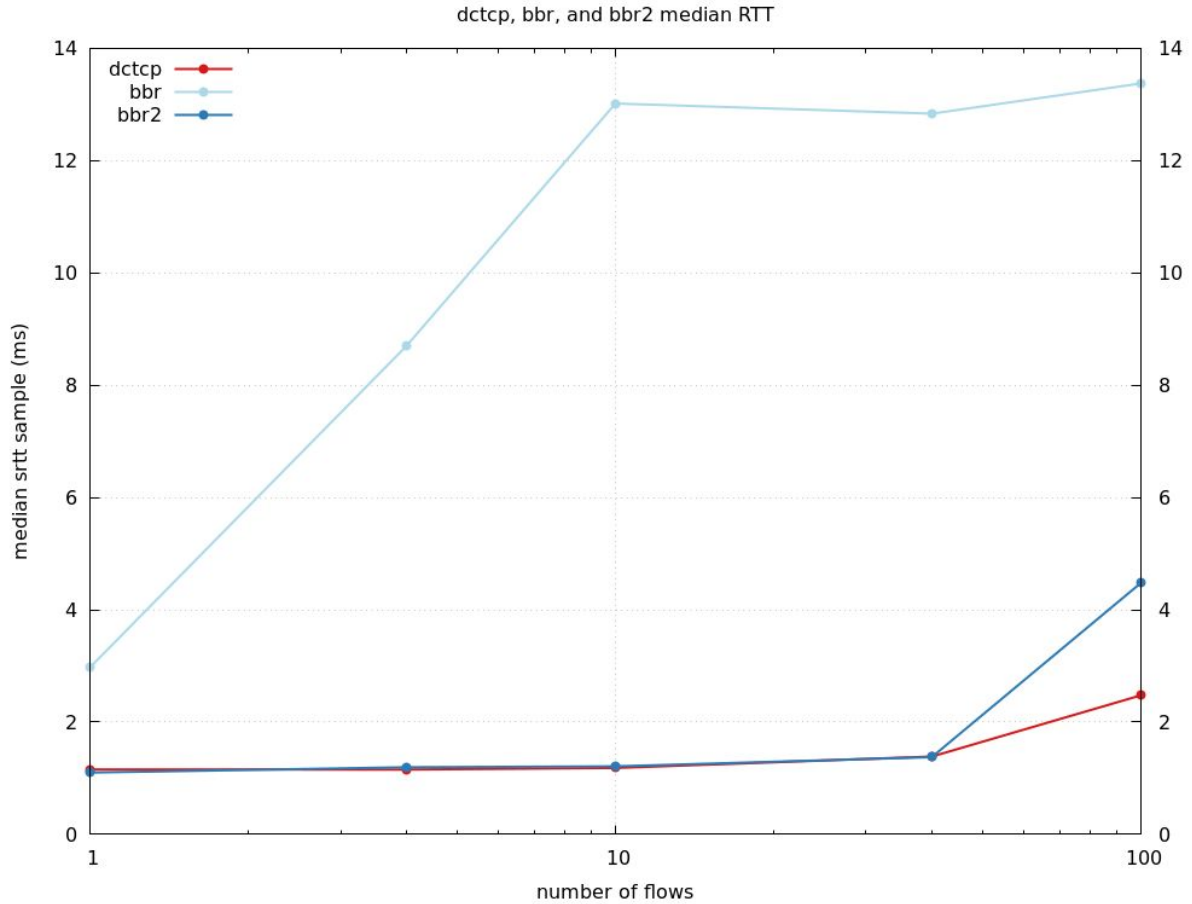
buf={1, 10, 100}xBDP

Y axis: p50 srtt sampled

(bbr and bbr2 overlap, at 53-69ms)

ECN is disabled

Low latency using DCTCP/L4S-style ECN signals (1/2)



Latency from bulk flows w/ ECN

N dctcp, bbr, or bbr2

num_flows = {1, 4, 10, 40, 100}

bw = 1Gbit/sec, min_rtt = 1ms

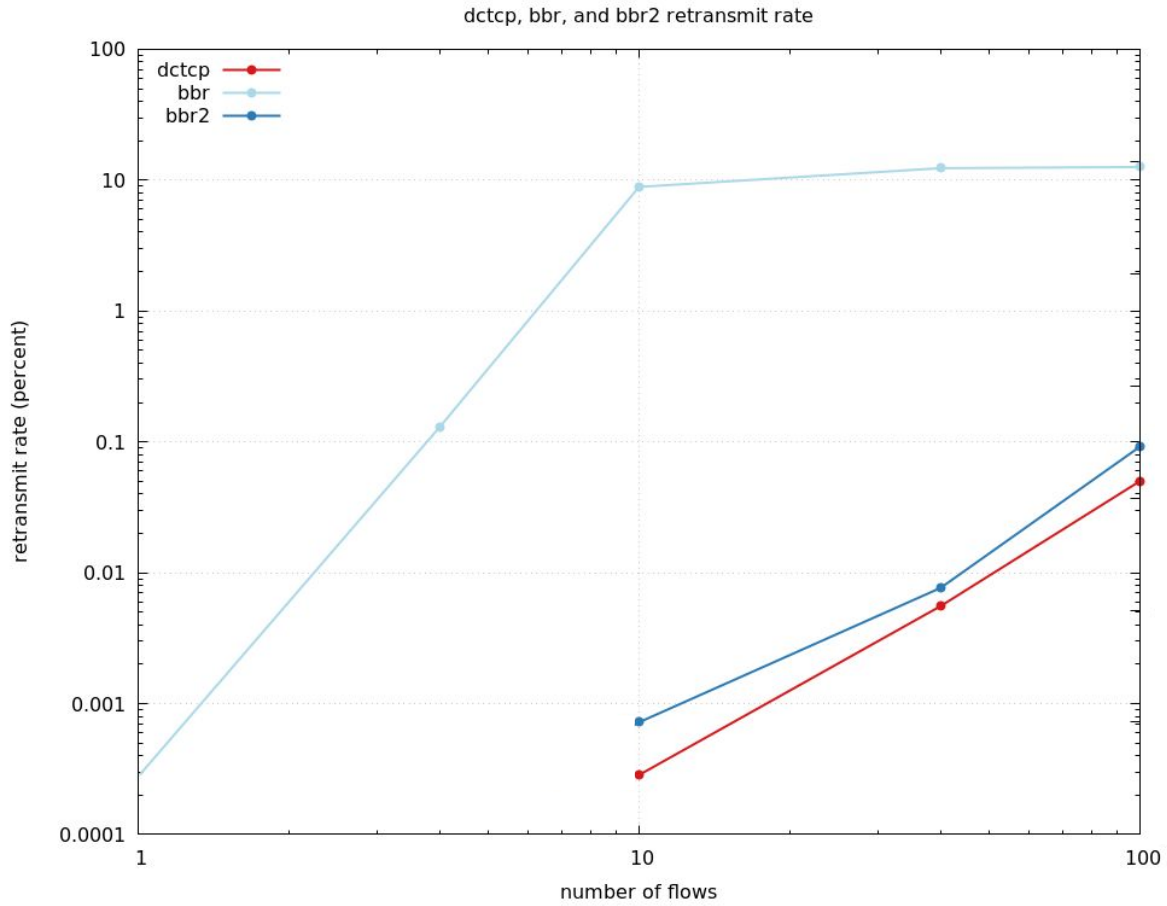
buf = 1000 packets (12ms)

10 sec. netperf TCP stream

ECN CE mark iff packet had more than 242us sojourn time (i.e. 20-packet queue).

Y axis: p50 of p50 of 10 trials; srtt shows impact of queuing delay.

Low losses using DCTCP/L4S-style ECN signals (2/2)



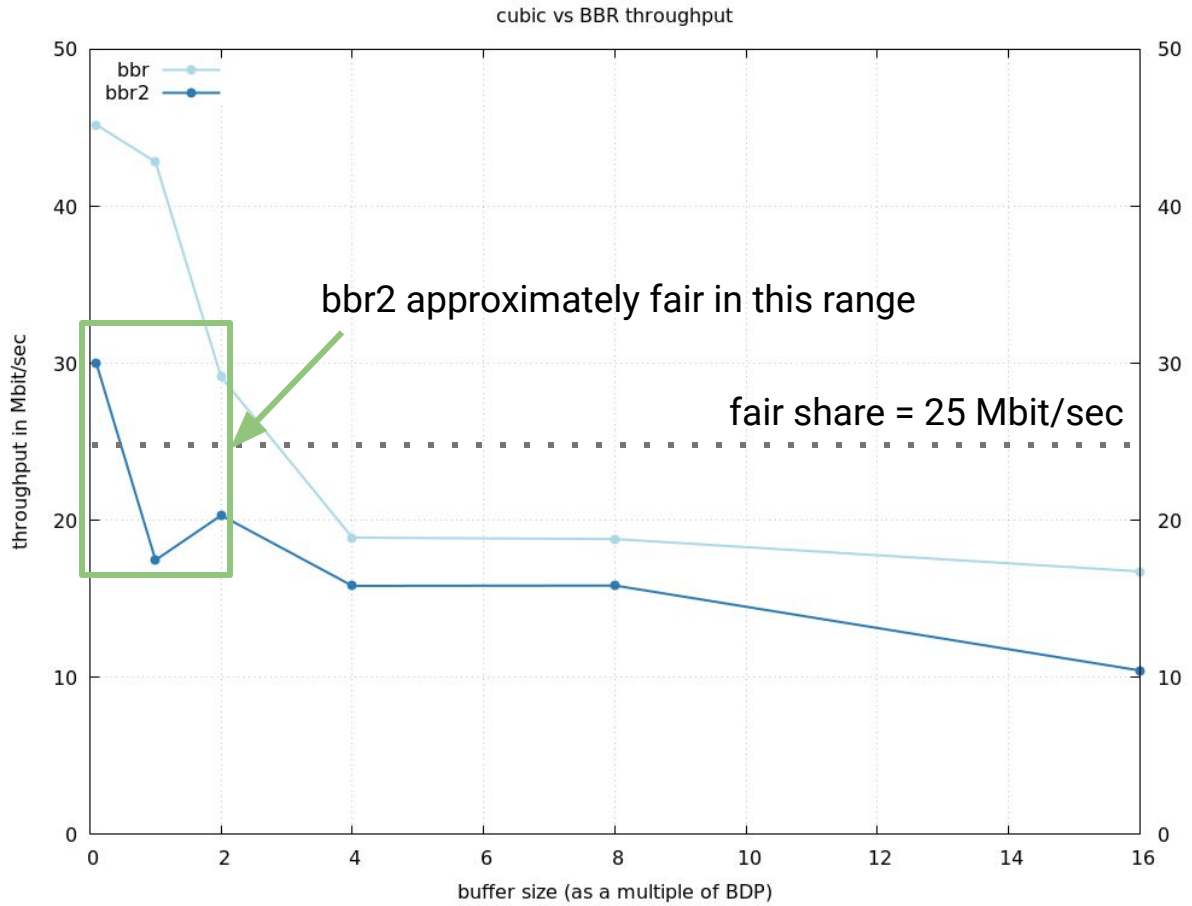
Losses from bulk flows w/ ECN

Same experiment as previous slide

Y axis: p50 of retransmit rate of 10 trials (log scale); loss rate shows impact of queuing pressure

(The bbr2 and dctcp cases with num_flows=1 are not depicted because they had no losses, and y=0.)

Coexistence with usable throughput for CUBIC



Bulk throughput

1 cubic sharing w/ 1 bbr or bbr2

bw = 50Mbit/sec, min_rtt = 30ms

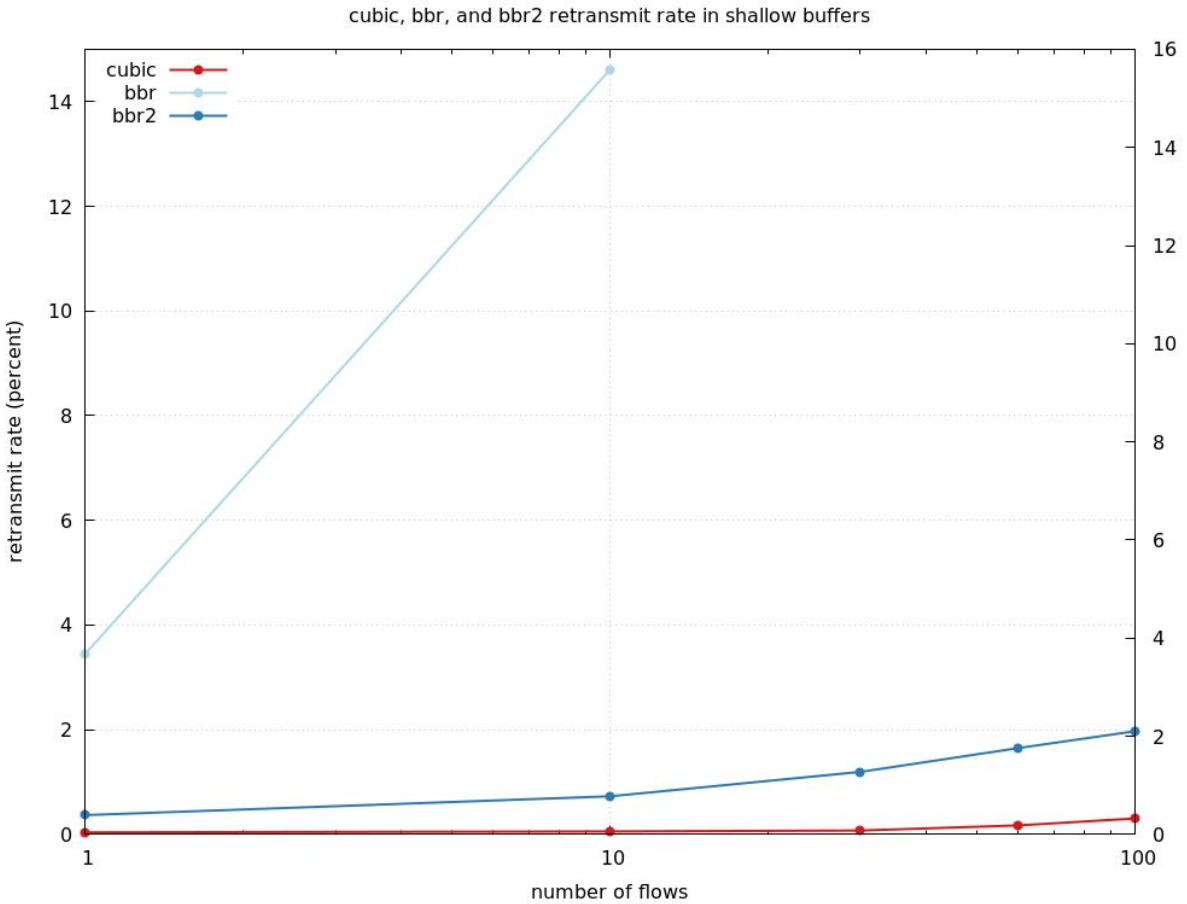
3 min. netperf TCP stream

cubic at t=0, bbr/bbr2 at t=2s

buf = {.1, 1, 2, 4, 8, 16}xBDP

ECN is disabled

Losses caused in shallow buffers



Retransmits from bulk flows

N cubic, bbr, or N bbr2

num_flows = {1, 10, 30, 60, 100}

bw = 1Gbit/sec, min_rtt = 100ms

BDP = 8256 packets

5 min. netperf TCP stream

buffer = .02*BDP

(*bbr v1 tests with 30 or more flows failed due to netperf setup timeouts)

BBR v2 status

BBR v2 algorithm status

- The known remaining issues in the BBRv2 algorithm:
  - Flows that experience ECN or loss early on, but never thereafter, sometimes don't reach their full fair share
  - Queue pressure higher than desired for large aggregates of BBRv2 flows
  - ECN response not tuned well for long RTTs
  - ECN response not tuned well for cases with more flows than slots in the BDP
- We're continuing to refine the algorithm...

BBR v2 deployment status

- YouTube: deployed for a small percentage of TCP users
  - Reduced queuing delays: RTTs lower than BBR v1 and CUBIC
  - Reduced packet loss: loss rates closer to CUBIC than BBR v1
- Internal: experiments between and within some Google data-centers
  - BBRv2 has lower tail latency compared to Google-DCTCP
  - Fixed a major performance issue with DCTCP-ECN and Linux delayed ACKs
    - The receiver may not ACK quickly under continuous CE marking
    - Caused high RPC latency under severe network congestion
    - The issue affected both DCTCP and BBRv2
- Continuing to iterate using production experiments and lab tests

Conclusion

- First BBR v2 "alpha/preview" release is now ready for research experiments
  - We invite researchers to share...
    - Ideas for test cases and metrics to evaluate
    - Test results
    - Algorithm/code ideas
  - Always happy to see patches or look at packet traces...
- Work on BBR v2 continues...
  - Actively working on BBR v2 at Google
  - Work under way for BBR in FreeBSD TCP @ Netflix as well

Q & A

https://groups.google.com/d/forum/bbr-dev

Internet Drafts, paper, code, mailing list, talks, etc.

Special thanks to Eric Dumazet, Nandita Dukkipati, C. Stephen Gunn, Kevin Yang, Jana Iyengar, Pawel Jurczyk, Biren Roy, David Wetherall, Amin Vahdat, Leonidas Kontothanassis, and {YouTube, google.com, SRE, BWE} teams.

Backup slides...

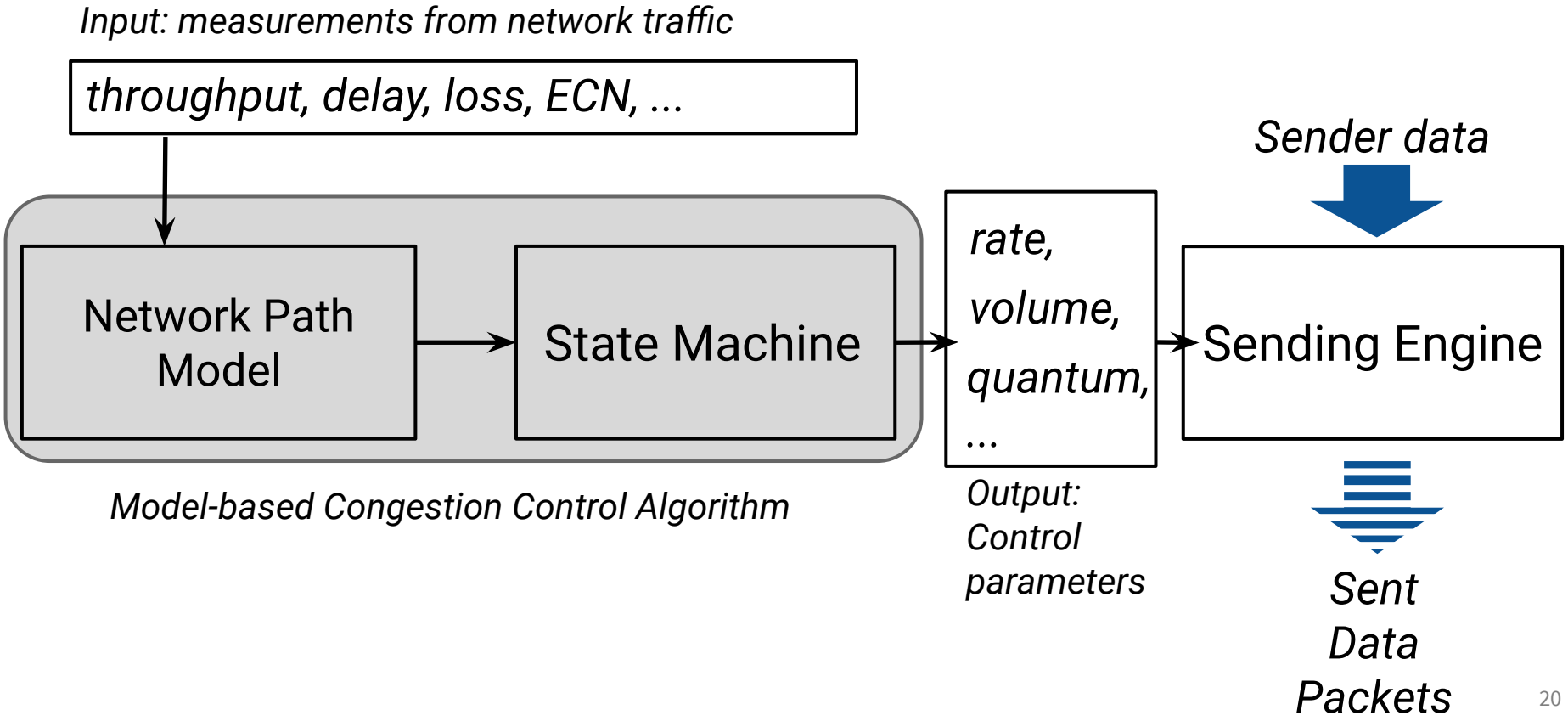
BBR v1 status: deployment, release, documentation

- BBR v1 used for TCP/QUIC on Google.com/YouTube, Google WAN backbone
  - Better performance than CUBIC for web, video, RPC traffic
- BBR v1 code open source in Linux TCP (dual GPLv2/BSD), Chromium QUIC (BSD)
- BBR v2 **preview** code available: Linux TCP (dual GPLv2/BSD), Chromium QUIC (BSD)
- Active BBR work under way for BBR in FreeBSD TCP @ Netflix
- BBR v1 Internet Drafts are out and ready for review/comments:
  - Delivery rate estimation: draft-cheng-iccrg-delivery-rate-estimation
  - BBR congestion control: draft-cardwell-iccrg-bbr-congestion-control
- IETF presentations: 97 | 98 | 99 | 100 | 101 | 102 | 104 (v2 design overview)
- BBR v1 Overview in Feb 2017 CACM

What's new in BBR v2: a summary

	CUBIC	BBR v1	BBR v2
Model parameters to the state machine	N/A	Throughput, RTT	Throughput, RTT, max aggregation, max inflight
Loss	Reduce cwnd by 30% on window with any loss	N/A	Explicit loss rate target
ECN	RFC3168 (Classic ECN)	N/A	DCTCP-inspired ECN
Startup	Slow-start until RTT rises (Hystart) or any loss	Slow-start until tput plateaus	Slow-start until tput plateaus or ECN/loss rate > target

BBR congestion control: the big picture



BBR v2: the network path model