Update on rLEDBAT & BBRv2

IICRG, IETF 112

November 8, 2021

Praveen Balasubramanian, Yi Huang, Matt Olson, Daniel Havey

Microsoft

rLEDBAT

- rLEDBAT brings benefits of LEDBAT++ to the receive side of the transport connection
- Use the flow control mechanism to throttle the peer
  - TCP receive window tuning
  - Don’t shrink advertised window
- Why is this important?
  - Software updates use CDNs – most CDNs don’t have LEDBAT(++) support
  - Proxies can prevent effective use of send LEDBAT on end-to-end path
  - Enforce receiver driven preference because it has more information about priority of traffic
- https://tools.ietf.org/html/draft-irtf-iccrg-rledbat-01

Windows TCP implementation

- The same (private) API enables both LEDBAT++ and rLEDBAT
- Includes all the additional mechanisms of LEDBAT++
  - Round trip latency measurements
  - Slower than Reno cwnd increase with adaptive gain factor
  - Multiplicative cwnd decrease with adaptive reduction factor
  - Modified slow start
  - Simplified periodic slowdown: one slowdown period per base delay measurement interval
  - Simplified base delay measurement (as described in draft)
- Require negotiation of timestamps
  - Expose API to app to query rLEDBAT status
  - Currently no action if data packet without TS received after establishment
- RTTs measured might be inflated due to bursts in slow start

Status & Next Steps

- Worked with CDNs to enable timestamps
- Measurements ongoing with Windows Update downloads
  - Aiming to share at next ICCRG

- Should the draft be published as Experimental?

BBRv2 Recap

- BBRv2 is a model-based congestion control algorithm
  - low queue occupancy
  - low loss
  - (bounded) Reno/CUBIC coexistence
- Measure bandwidth, RTT, packet loss, ECN marking
- Notable additions in v2:
  - Adaptive bandwidth-probing time scale
  - Loss and ECN are incorporated into the network model
  - Adapts to loss/ECN even when application-limited
  - Adapt cwnd based on ack aggregation estimation
- Pace at the computed rate

Windows TCP implementation

- Based on https://github.com/google/bbr/blob/v2alpha/net/ipv4/tcp_bbr2.c
- Integrated as a congestion control module
- Available as a experimental knob on Windows 11 Insider builds
- Rate based pacer built into TCP
  - on each send, compute an allowance based on time since last send
  - schedule the pacing timer to send pending data over allowance
- Refactoring has been minimized to enable direct comparison between the ported code and the original
- Not implemented ECN handling yet
- Lack of a draft / spec significantly hindered development

Early data

- WAN test cases
  - Significant improvements in latency – up to 10x in some cases!
  - Some throughout improvements
- Low latency intra-DC test cases
  - Much lower throughput on low latency and loopback: CPU usage bottleneck
  - Interactions between pacing and LSO
- Azure inter-region test
  - 20% throughput improvement, not much difference in latency
- Fairness issues
  - CUBIC dominates BBRv2 across a range of test cases
  - Not incrementally deployable in current form

Status & Next Steps

- Help review and adopt draft
- Resolve Fairness issues when CUBIC shares bottleneck link
- CPU usage optimizations
- Deploy and measure in production

- Big thanks to Neal Cardwell and Yuchung Cheng for their support!

Q&A

WAN Lab results - Throughput

30s_20Mbps_20ms_400pkt_0loss_BBR2	Buf/BDP 12.00	Mbps 18	RttMs 29	LossPkt 0
30s_20Mbps_20ms_400pkt_0loss_CUBIC	Buf/BDP 12.00	Mbps 18	RttMs 211	LossPkt 66
30s_40Mbps_120ms_400pkt_0loss_BBR2	Buf/BDP 1.00	Mbps 33	RttMs 124	LossPkt 204
30s_40Mbps_120ms_400pkt_0loss_CUBIC	Buf/BDP 1.00	Mbps 36	RttMs 198	LossPkt 338
30s_40Mbps_40ms_200pkt_0loss_BBR2	Buf/BDP 1.50	Mbps 35	RttMs 45	LossPkt 15
30s_40Mbps_40ms_200pkt_0loss_CUBIC	Buf/BDP 1.50	Mbps 37	RttMs 82	LossPkt 330
30s_40Mbps_40ms_200pkt_100loss_BBR2	Buf/BDP 1.50	Mbps 7	RttMs 40	LossPkt 271
30s_40Mbps_40ms_200pkt_100loss_CUBIC	Buf/BDP 1.50	Mbps 3	RttMs 40	LossPkt 130
30s_40Mbps_4ms_200pkt_0loss_BBR2	Buf/BDP 15.00	Mbps 37	RttMs 6	LossPkt 0
30s_40Mbps_4ms_200pkt_0loss_CUBIC	Buf/BDP 15.00	Mbps 37	RttMs 52	LossPkt 184
30s_40Mbps_80ms_30pkt_0loss_BBR2	Buf/BDP 0.11	Mbps 17	RttMs 80	LossPkt 419
30s_40Mbps_80ms_30pkt_0loss_CUBIC	Buf/BDP 0.11	Mbps 14	RttMs 80	LossPkt 61
3MB_20Mbps_20ms_400pkt_0loss_BBR2	Buf/BDP 12.00	Mbps 17	RttMs 25	LossPkt 0
3MB_20Mbps_20ms_400pkt_0loss_CUBIC	Buf/BDP 12.00	Mbps 15	RttMs 163	LossPkt 30
3MB_40Mbps_120ms_400pkt_0loss_BBR2	Buf/BDP 1.00	Mbps 17	RttMs 129	LossPkt 0
3MB_40Mbps_120ms_400pkt_0loss_CUBIC	Buf/BDP 1.00	Mbps 14	RttMs 144	LossPkt 126
3MB_40Mbps_40ms_200pkt_0loss_BBR2	Buf/BDP 1.50	Mbps 19	RttMs 48	LossPkt 27
3MB_40Mbps_40ms_200pkt_0loss_CUBIC	Buf/BDP 1.50	Mbps 26	RttMs 67	LossPkt 172
3MB_40Mbps_40ms_200pkt_100loss_BBR2	Buf/BDP 1.50	Mbps 11	RttMs 41	LossPkt 35
3MB_40Mbps_40ms_200pkt_100loss_CUBIC	Buf/BDP 1.50	Mbps 3	RttMs 40	LossPkt 27
3MB_40Mbps_4ms_200pkt_0loss_BBR2	Buf/BDP 15.00	Mbps 37	RttMs 11	LossPkt 0
3MB_40Mbps_4ms_200pkt_0loss_CUBIC	Buf/BDP 15.00	Mbps 36	RttMs 49	LossPkt 79
3MB_40Mbps_80ms_30pkt_0loss_BBR2	Buf/BDP 0.11	Mbps 11	RttMs 80	LossPkt 67
3MB_40Mbps_80ms_30pkt_0loss_CUBIC	Buf/BDP 0.11	Mbps 10	RttMs 80	LossPkt 23

WAN Lab results - Fairness

share_30s_20Mbps_20ms_400pkt_0loss_BBR2_BBR2	Buf/BDP 12.00	Fairness 7	Mbps1 9	Mbps2 8
share_30s_20Mbps_20ms_400pkt_0loss_BBR2_CUBIC	Buf/BDP 12.00	Fairness 0	Mbps1 1	Mbps2 16
share_30s_20Mbps_20ms_400pkt_0loss_CUBIC_CUBIC	Buf/BDP 12.00	Fairness 8	Mbps1 8	Mbps2 9
share_30s_40Mbps_120ms_400pkt_0loss_BBR2_BBR2	Buf/BDP 1.00	Fairness 7	Mbps1 18	Mbps2 16
share_30s_40Mbps_120ms_400pkt_0loss_BBR2_CUBIC	Buf/BDP 1.00	Fairness 4	Mbps1 10	Mbps2 25
share_30s_40Mbps_120ms_400pkt_0loss_CUBIC_CUBIC	Buf/BDP 1.00	Fairness 8	Mbps1 19	Mbps2 16
share_30s_40Mbps_40ms_200pkt_0loss_BBR2_BBR2	Buf/BDP 1.50	Fairness 8	Mbps1 19	Mbps2 16
share_30s_40Mbps_40ms_200pkt_0loss_BBR2_CUBIC	Buf/BDP 1.50	Fairness 2	Mbps1 7	Mbps2 29
share_30s_40Mbps_40ms_200pkt_0loss_CUBIC_CUBIC	Buf/BDP 1.50	Fairness 8	Mbps1 17	Mbps2 18
share_30s_40Mbps_40ms_200pkt_100loss_BBR2_BBR2	Buf/BDP 1.50	Fairness 7	Mbps1 9	Mbps2 8
share_30s_40Mbps_40ms_200pkt_100loss_BBR2_CUBIC	Buf/BDP 1.50	Fairness 3	Mbps1 9	Mbps2 3
share_30s_40Mbps_40ms_200pkt_100loss_CUBIC_CUBIC	Buf/BDP 1.50	Fairness 10	Mbps1 3	Mbps2 3
share_30s_40Mbps_4ms_200pkt_0loss_BBR2_BBR2	Buf/BDP 15.00	Fairness 9	Mbps1 18	Mbps2 18
share_30s_40Mbps_4ms_200pkt_0loss_BBR2_CUBIC	Buf/BDP 15.00	Fairness 0	Mbps1 1	Mbps2 35
share_30s_40Mbps_4ms_200pkt_0loss_CUBIC_CUBIC	Buf/BDP 15.00	Fairness 9	Mbps1 18	Mbps2 18
share_30s_40Mbps_80ms_30pkt_0loss_BBR2_BBR2	Buf/BDP 0.11	Fairness 8	Mbps1 14	Mbps2 15
share_30s_40Mbps_80ms_30pkt_0loss_BBR2_CUBIC	Buf/BDP 0.11	Fairness 4	Mbps1 18	Mbps2 7
share_30s_40Mbps_80ms_30pkt_0loss_CUBIC_CUBIC	Buf/BDP 0.11	Fairness 7	Mbps1 10	Mbps2 10