TCP for low RTTs

draft-bagnulo-tcpm-tcp-low-rtt-00
IETF96 – Berlin
Marcelo Bagnulo
Koen De Schepper
Glenn Judd

Networks with Low RTT

- Networks with low RTTs: RTTs between a few nanosecs and hundreds of nanosecs
  - Common in Data center environment
- Lots of operational experience
- E.g. documented in
  - Judd, G., "Attaining the promise and avoiding the pitfalls of TCP in the Datacenter", NSDI 2015, 2015.
  - V. Vasudevan et al.,"Safe and Effective Fine-grained TCP Retransmissions for Datacenter Communication", SIGCOMM 2009
- Few recommendations to TCP to perform well in these environment

Minimum RTO

- RFC6298 states:
  - Whenever RTO is computed, if it is less than 1 second, then the RTO SHOULD be rounded up to 1 second.
- Current implementations use RTOmin between 200ms and 400ms
- Experience shows that such values make TCP perform poor in low RTT networks
  - Especially in incast situations
- Recommendation to use 1 ms
- Useful to provide BCP to qualify RFC6298 SHOULD in low RTT networks?

Delay for delayed ACKs

- RFC5681 states that delayed ACKs
  - MUST be generated within 500 ms of the arrival of the first unacknowledged packet
- Current implementations use ranges between tens and hundreds of ms
- Results in performance penalties
- Better performance using 1 ms or lower
- No need to change the RFC
- BCP recommendation needed?

Minimum CWND

- RFC5681 states that:
  - When the third duplicate ACK is received, a TCP MUST set ssthresh to no more than the value given in equation (4)
  - ssthresh = max (FlightSize / 2, 2*SMSS) (4)
- RFC5681 does not specifies how to calculate a smaller CWND
- Implementations set it to 2SMSS
- In low RTTs, a CWND of 2 MSSs results in large rates, below which TCP is unresponsive
- New spec to calculate and use CWND smaller than 2 MSSs

Proposed next steps

- Identify other issues (if any) for low RTT networks
  - Comments welcome
- Document BCP for low RTTs
  - This document?
- New spec describing how to calculate CWND smaller than 2MSSs