20170058 Keonwoo Kim · October 18, 2020
CS341 Introduction to Computer Networks

Reading Assignment #3

Data center TCP is a variant of TCP protocol, which is appropriate for large data centers. There are a lot of network switches connecting servers in large data centers. The applications running on those servers generate often a mix of short and long network flows, where short flows need low latency and long flows need high congestion tolerance and high utilization. They are seemingly contradicting, as the queue must be short enough to ensure low latency while the queue must be long enough to buffer long data entirely and not make excessive packet loss.

In data centers, shallow buffered switched are used since they are cheap. But it affects to the response time (due to increased timeouts or added delay) and makes queue buildup impairment occurs (to reduce latency caused by queueing, the size of the queues should be reduced), and that queue build up reduces the shared memory so remaining buffer space.

To resolve those problems in TCP protocol, the authors designed a new protocol called DCTCP. It achieves the goals by controlling the TCP congestion window depending on the extent of the congestion, where the vanilla TCP reacts to just the existence of the congestion.

1. First, mark an arriving packet with the CE (congestion experienced) codepoint if the queue occupance is greater than K upon when the packet is received. The authors recommend to choose $K > (RTT \cdot C)/7$, where C is the link rate in packets per second.

2. Second, echo congestion information on the receiver. Unlike TCP protocol in which a receiver sets the ECN-Echo flag in a series of ACK packets until receiving confirmation from the sender, in DCTCP, the receiver tries to accurately transmit the marked packets back to the sender, by ACK every packet, setting the ECN-Echo flag iff the packet has marked as CE.

3. Finally, process echoed congestion indications on the sender. The parameter α, meaning an estimate of the fraction of marked packets, is updated once for every window of data as follows:

   $$\alpha \leftarrow (1 - g)\alpha + g \cdot F,$$

   where F is the fraction of marked packets in the last window, and $0 < g < 1$ is a predefined weight, where the authors suggest to choose $g < 1.386/\sqrt{2(C \cdot RTT + K)}$. After that, update cwnd as follows:

   $$cwnd \leftarrow cwnd \cdot (1 - \alpha/2).$$

   When α is small (low congestion), cwnd is only slightly reduced, while TCP always reduces it to half. In DCTCP, cwnd becomes half iff $\alpha = 1$, that is, the congestion is in a very high level.