HotRFC
Fast Congestion Response in Data Centers

draft-even-fast-congestion-response-00

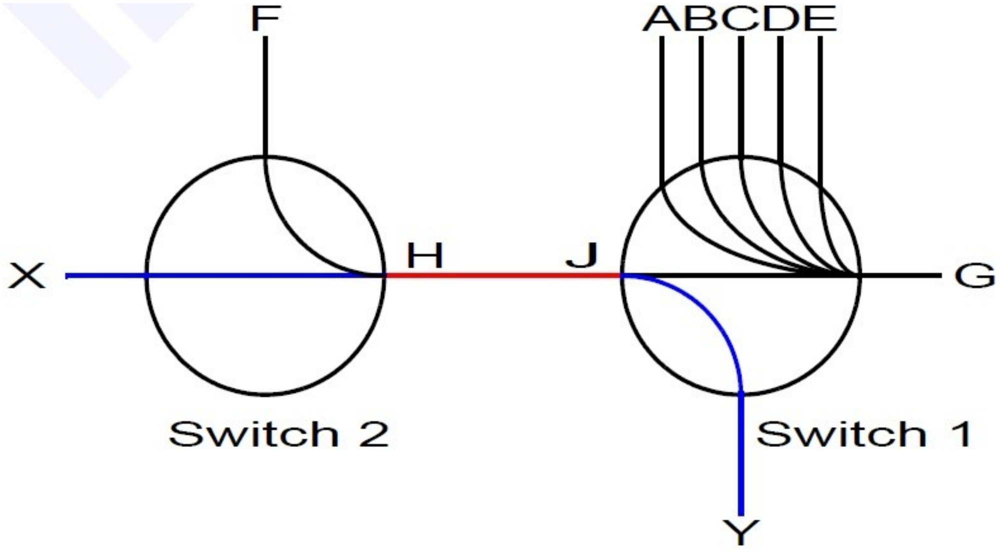
Roni Even

Problem statement

- The high link speed (100Gb/s) in Data Centers (DC) are making network transfers complete faster and in fewer RTTs. Short data bursts requires low latency while longer data transfer require high throughput.
- Current congestion control using ECN is re-active. On congestion the Switches mark ECN bits and the receiver notify the sender. The sender reduce the transmit rate and increase it back based on pre-defined policy.
  - DCTCP extends the ECN processing to estimate the fraction of bytes that encounter congestion and scales the TCP congestion window based on this estimate. QUIC and RTP report back the number of ECN marked packets.
  - Data Centers internal communication protocol is RoCEv2 (InfiniBand over UDP)

Problem statement

- Link-Layer Flow- Control IEEE 802.1Qbb(PFC) is used to provide a lossless network in the Data Center but has problem with Head of Line blocking.
  - Traffic from F to G may block traffic from X to Y if switch 1 send a Pause to port J when port G is congested. (the drawing is from InfiniBand architecture specification)



New Direction

- Define a proactive congestion control that will provide a faster convergence time.
  - Currently the Switches only monitor congestion state. The data sender analyze and react based on partial information.
  - The switches can provide better congestion information allowing the sender to react faster.
    - This direction can be specified for intra-data-center environment where both endpoints and the switching fabric are under a single administrative domain.

Thank you!

Catch me during the week or email me at roni.even@huawei.com if you are interested.