

Low latency networking in the IETF

IEEE 802 / IETF session @ IETF99

July 15, 2017, Prague

Mirja Kühlewind (transport AD)

Note on latency in networking

Approaches may optimize for two different kind of latency:

- Total (file) transmission time
e.g. request/response like in HTTP
- Per-packet latency and latency variation (jitter)
e.g. interactive communication like WebRTC

Low Latency Applications and the Internet Architecture

- <https://tools.ietf.org/html/draft-arkko-arch-low-latency-01>
 - Applications with Special Focus on Low Latency
 - Role of Low-Latency vs. Other Communications
 - Selected Improvements to Communications Latency
 - Architectural Considerations & Implications

Application/Security

- **tls**: 0-RTT session resumption in TLS1.3
- **rtcweb**: use of DiffServ codepoints to request network support
- **alto**: service to provide and request information for server selection

Transport and Congestion Control

- **tcpm:**
 - TCP's Initial Window of 10 (IW10)
 - TCP Fast Open (TFO) enables data on TCP SYN
 - TCP Rack fast loss detection for short transmission with tail loss
 - Datacenter TCP (DCTCP) Congestion Control for datacenters to avoid high loss and delay due to incast
- **QUIC:** new, encrypted transport protocol that also provides a 0-RTT transport handshake on session resumption
- **rmcat:** RTP Media Congestion Avoidance Techniques that aims to minimize self-induced latency
- **mptcp:** aims to utilize the path with the best performance if multiple paths with different characteristics are available

AQM and queuing

- **aqm**: new Active Queue Management (AQM) schemes that are optimized to keep queuing delay low, e.g. CoDel and PIE (wg finished and will be closed soon)
- **tsvwg**: L4S (Low Latency, Low Loss, Scalable Throughput) services provide a separate queue for scalable congestion control

Networking and Routing

- **detnet**: Provisioning of paths that can provide bounds on latency, loss, and packet delay variation (jitter).