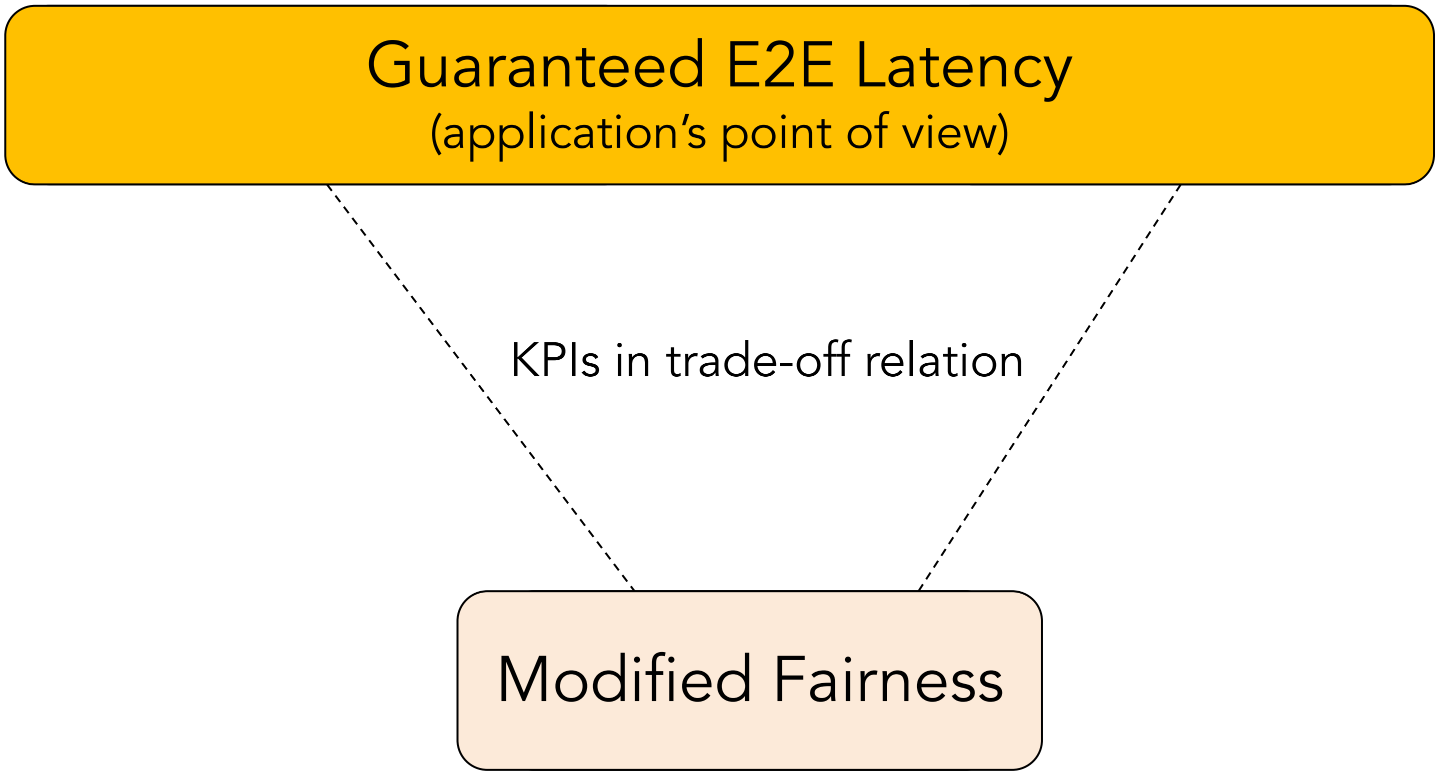
Transport Layer
- High Performance Transport -

Kyunghan Lee
Networked Computing Lab (NXC Lab)
Department of Electrical and Computer Engineering
Seoul National University
https://nxc.snu.ac.kr
kyunghanlee@snu.ac.kr

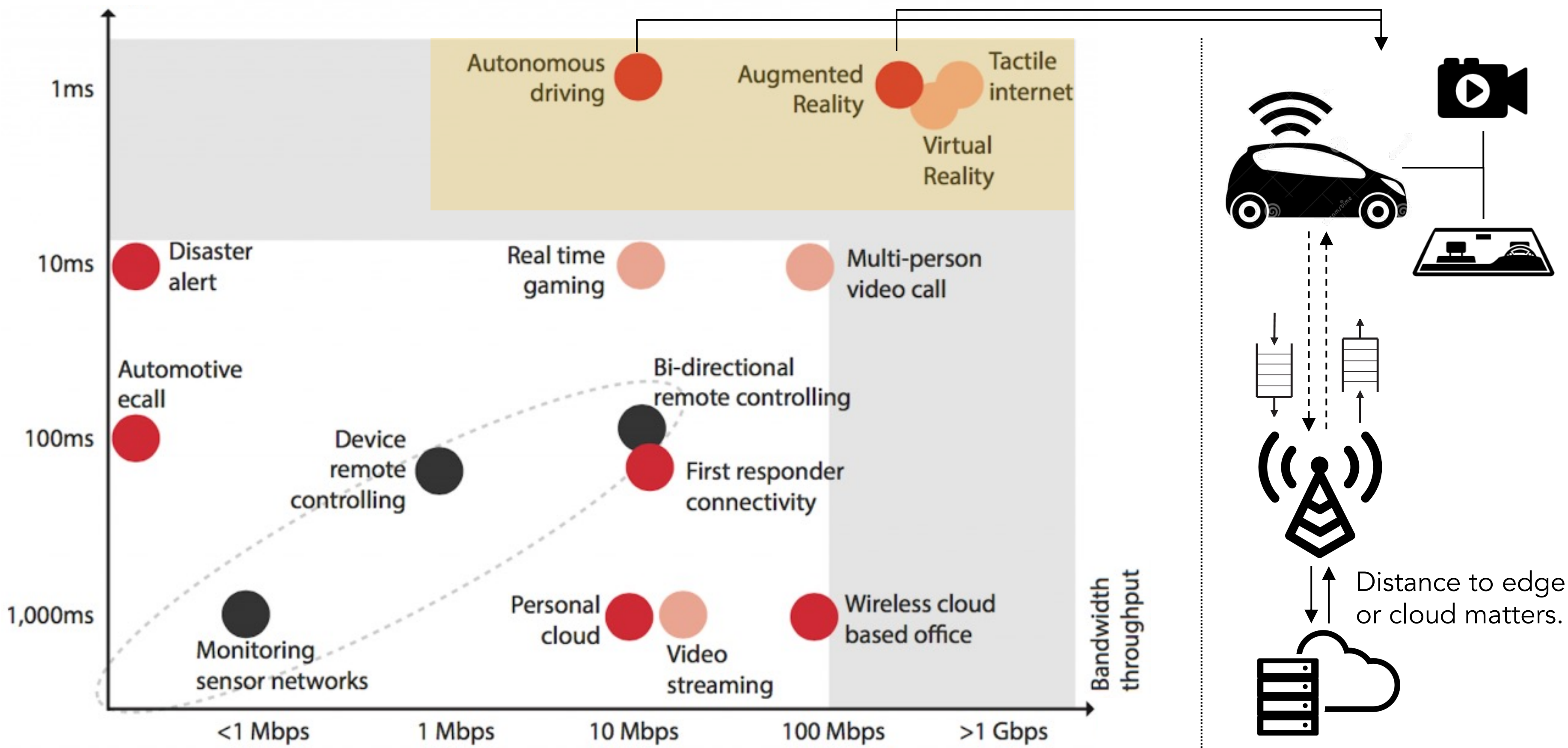
KPIs for Future Transport Layer Protocols



E2E Latency is Everything

E2E Latency = First-byte-Delay + DataVolume/Throughput

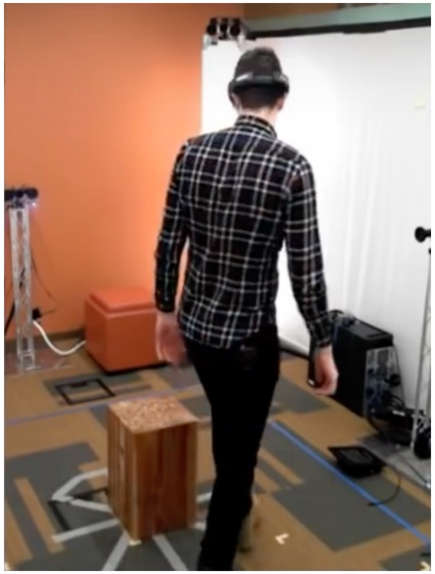
E2E Latency



Mark Glimour, "5G – Latency: New use cases and the need for network slicing," InformaTech reports, Feb. 2017.

Ultimate Service: Low-latency Telepresence

- **Low-latency telepresence** is the idea of providing people the experience as if they are present at a place other than their physical location by making remote interactions indistinguishable from local interactions.

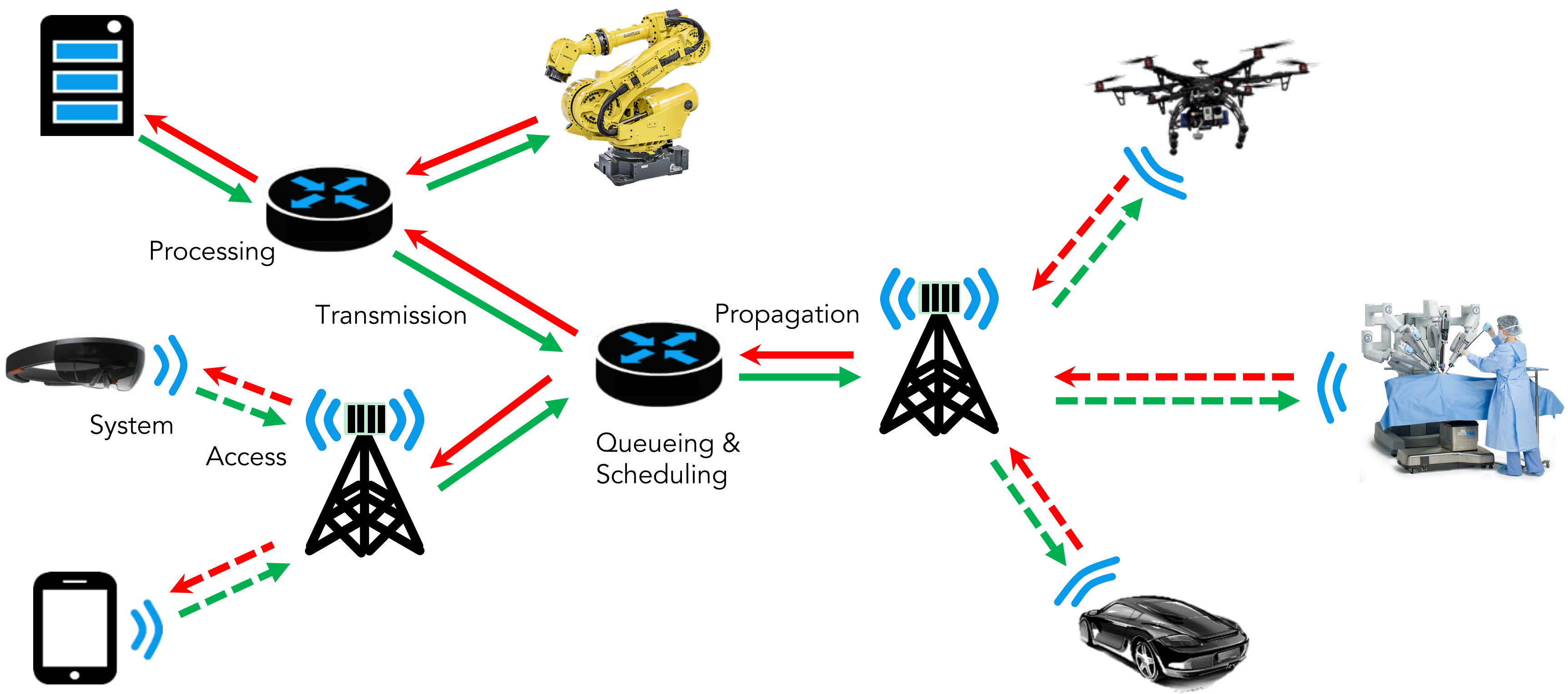


Object data →

within human reaction latency

← Environment data

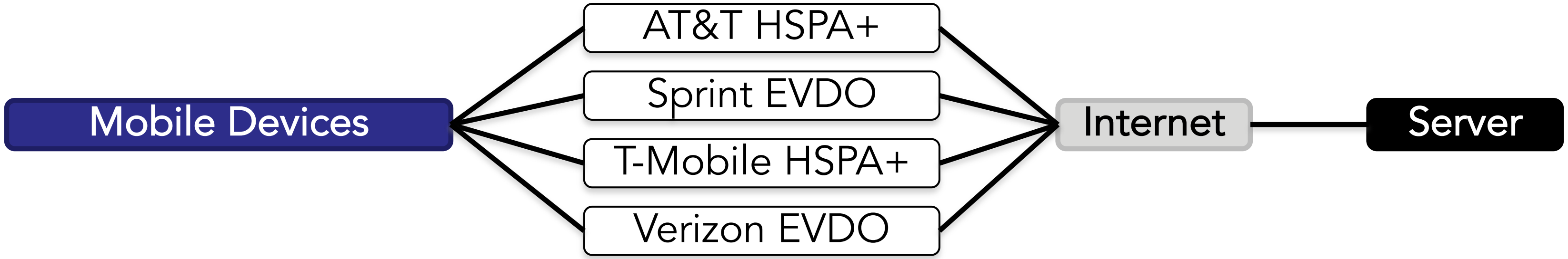
Latency Decomposition



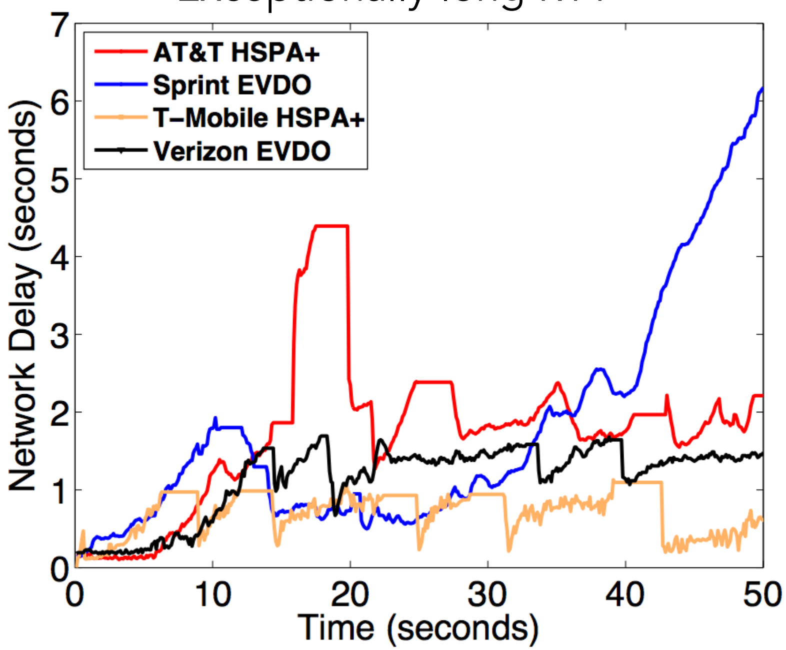
App Delay	Transmission Delay (1 us – 100ms)	Processing Delay (< 1 us)	Scheduling Delay (< 1 ms)	Queueing Delay (< 500 ms)	Propagation Delay (1 ms @ 200km)
Socket control, Encoder control	Redundancy elimination	FPGA, ASIC	Short TTI, Priority sched	Congestion control	Untouchable

Latency Reality in Cellular Networks

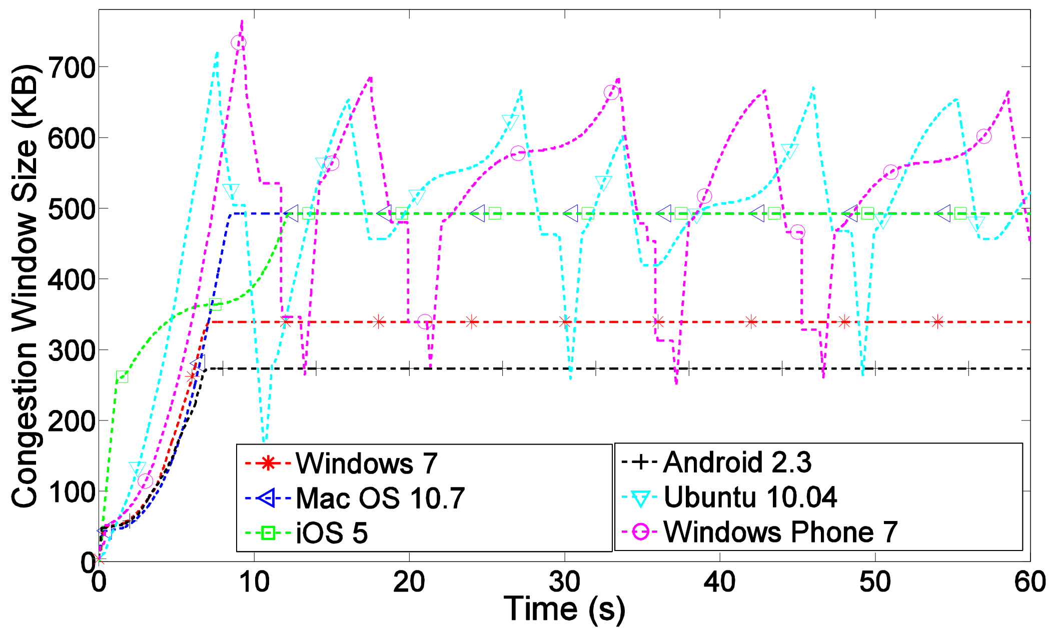
Tackling bufferbloat in 3G/4G networks, ACM IMC 2012



Exceptionally long RTT

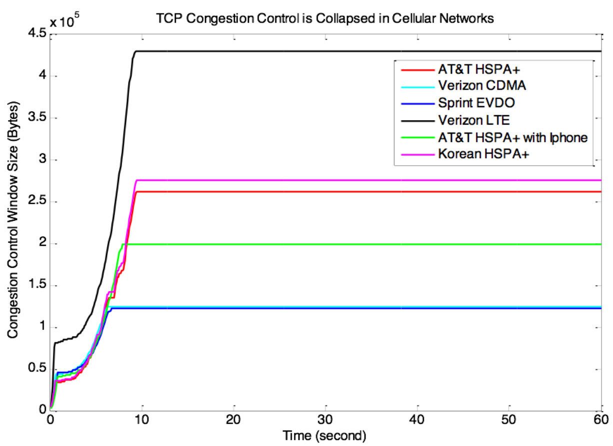


Flat TCP congestion window



Understanding the Problem and the Trick

Packet loss in cellular links → (Link-level recovery) → Putting a large buffer in the BS → Very few packet losses → TCP congestion control failure → Extremely long queue in buffer → Extremely long RTT (latency) → Static limit in rwnd (rmem_max)



$$Tput \sim \frac{rwnd}{RTT}$$

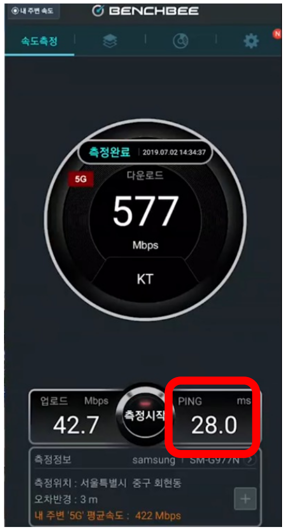
min(cwnd,rwnd)

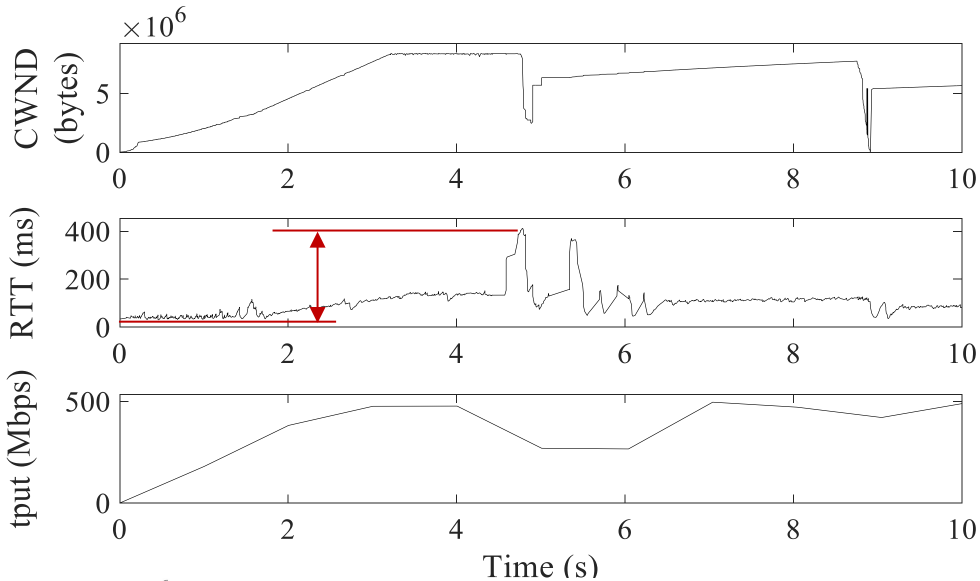
Static limit in rwnd (rmem_max) → A potential solution (delay-based control) → DRWA (Dynamic Receive Window Adjustment)

$$Tput \sim \frac{rwnd(\Delta RTT)}{RTT}$$

Latency Reality in 5G/LTE Networks

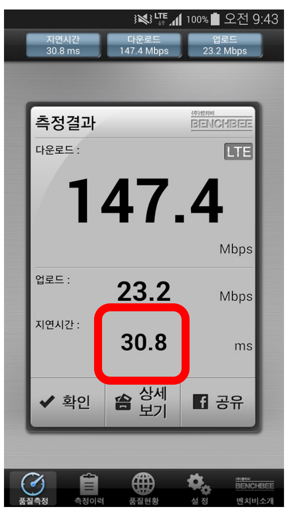
5G
(NSA)

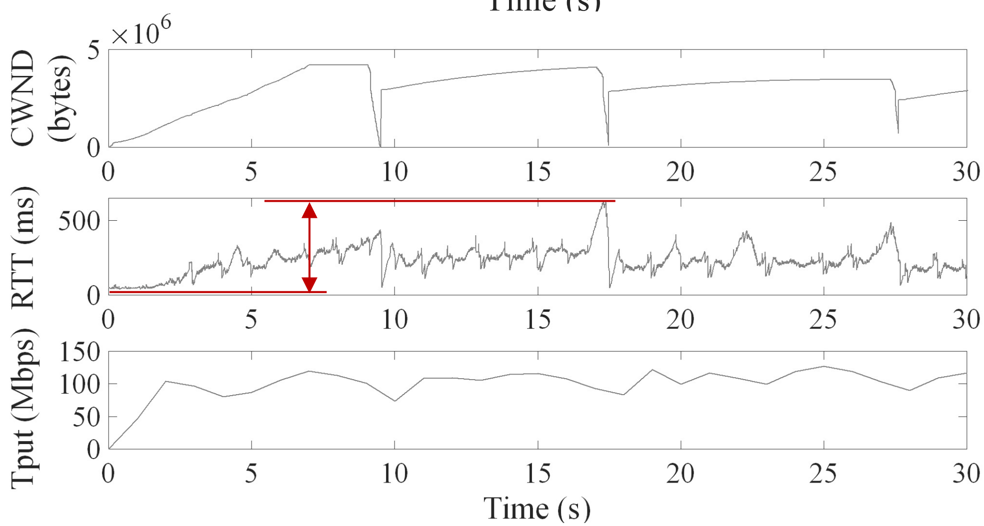




Mean, Min RTT
118.15, 35.28 ms
(~3x inflation)

LTE





Mean, Min RTT
226.10, 37.05 ms
(~6x inflation)

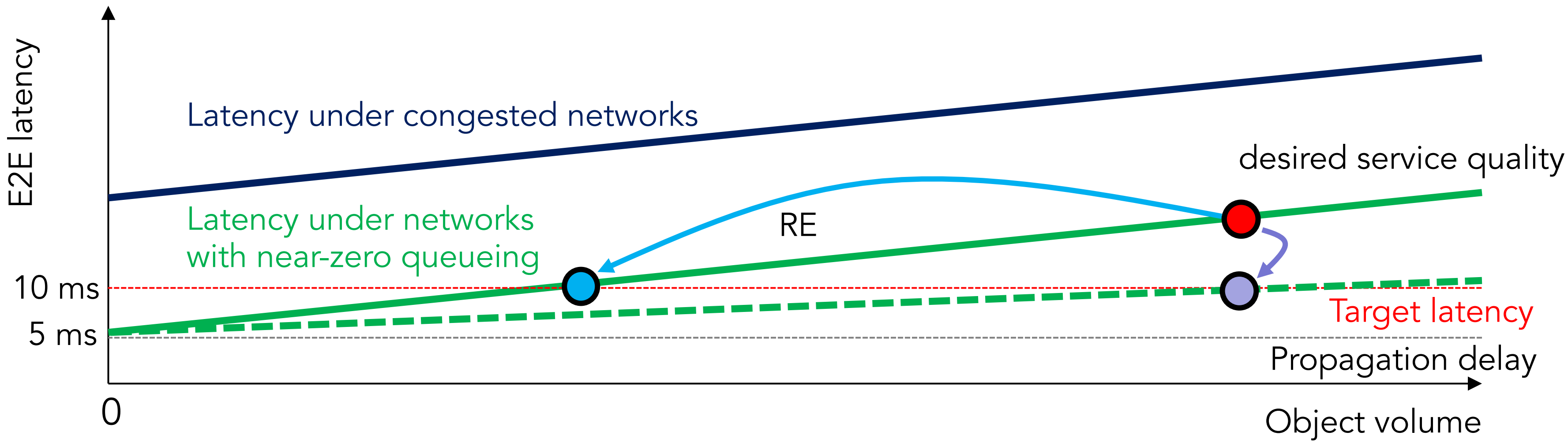
NXC LAB

Core Challenges for Guaranteed E2E Latency

E2E latency: time taken to relocate an information object over a network
(mostly Propagation delay + Queueing delay + Transmission delay + Retransmission delay)

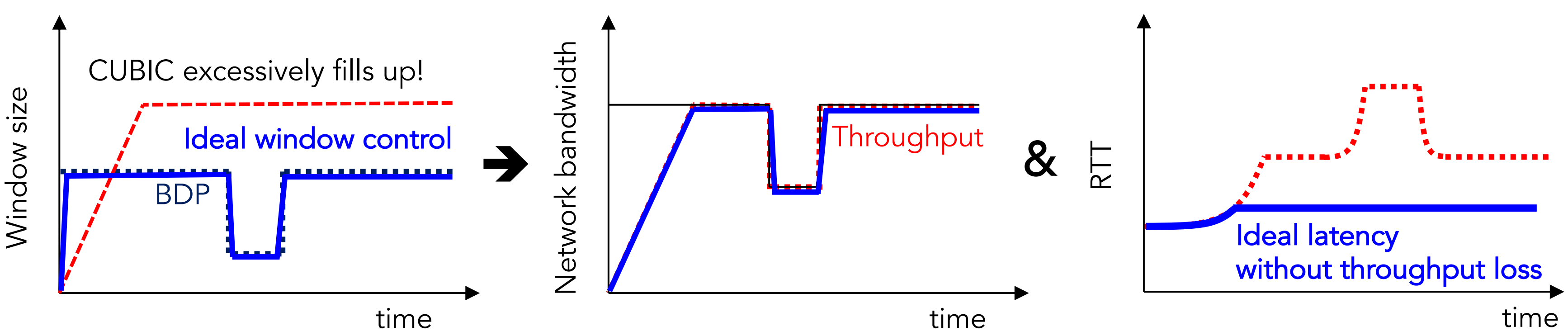
$= \frac{\text{volume} \downarrow}{\text{data rate} \uparrow}$

1. Low-latency transport protocol (congestion control)
2. Redundancy elimination (socket)
3. Fast Convergence (+ PHY evolution)
4. Zero retransmission (socket)

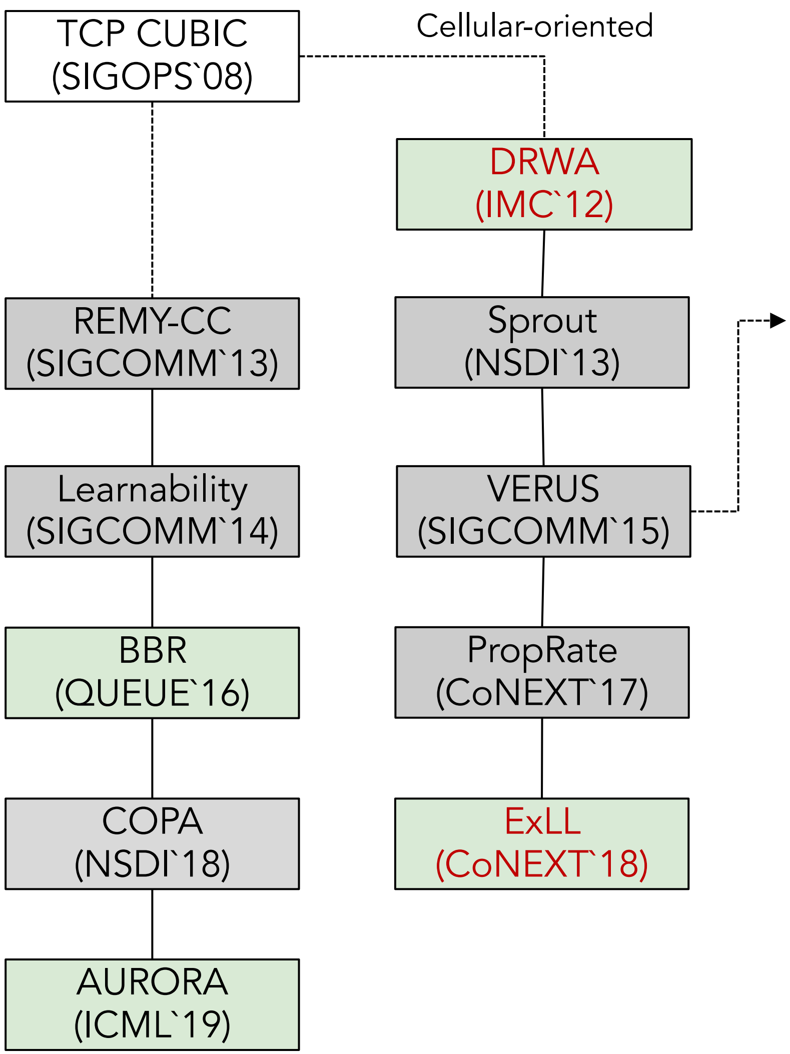


Challenge 1+3: Low-Latency CC (+ Fast Convergence)

RTT in a cellular network with fluctuating bandwidth

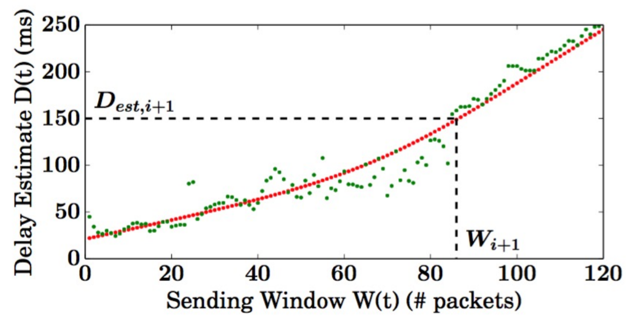


Low-Latency Transport Protocols

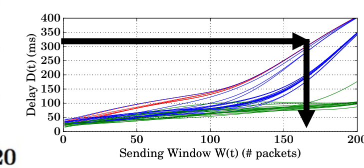


VERUS

builds delay profiles from its window controls



Rebuild profile in every 1s



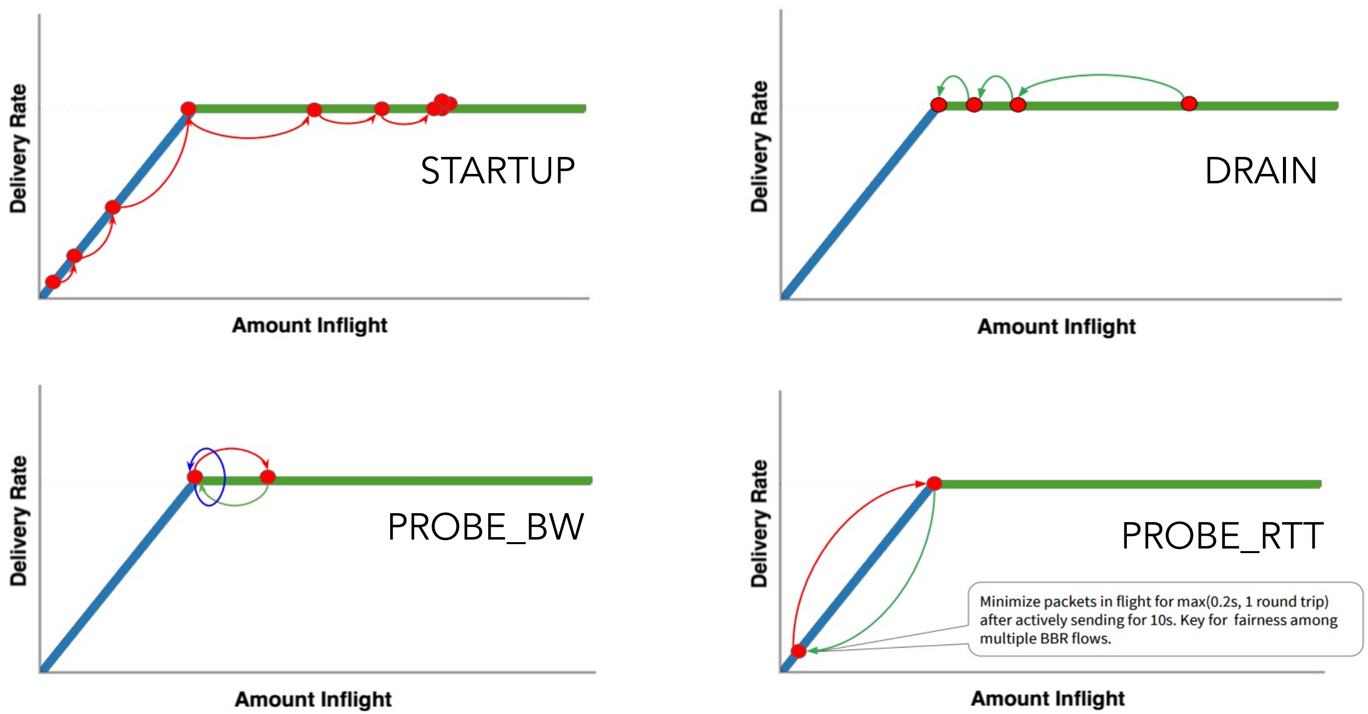
Reduce window if max delay out of a bound

Given more window budget if max delay within the bound

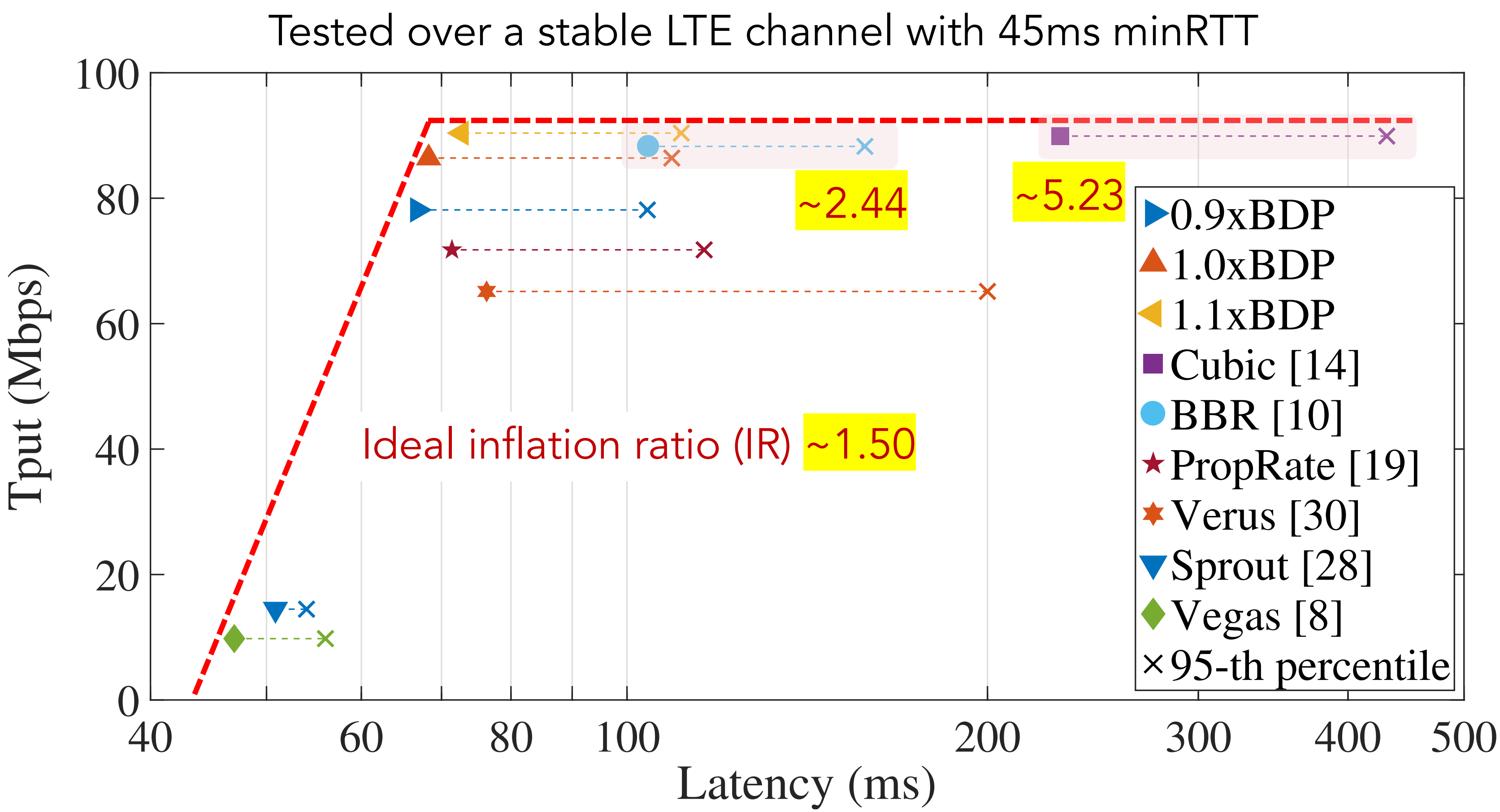
$$D_{est,i+1} = \begin{cases} D_{est,i} - \delta_2 & \text{if } \frac{D_{max,i}}{D_{min}} > R \\ D_{est,i} + \delta_2 & \text{otherwise} \end{cases}$$

BBR (Bottleneck Bandwidth and Round-trip propagation time)

- BBR cycles 4 operations (BBR is in the Linux kernel from 4.9)
  - STARTUP - exponential BW search
  - DRAIN - drain the queue created during startup
  - PROBE_BW - explore max BW, drain queue, cruise
  - PROBE_RTT - guarantee fairness among multiple BBR flows

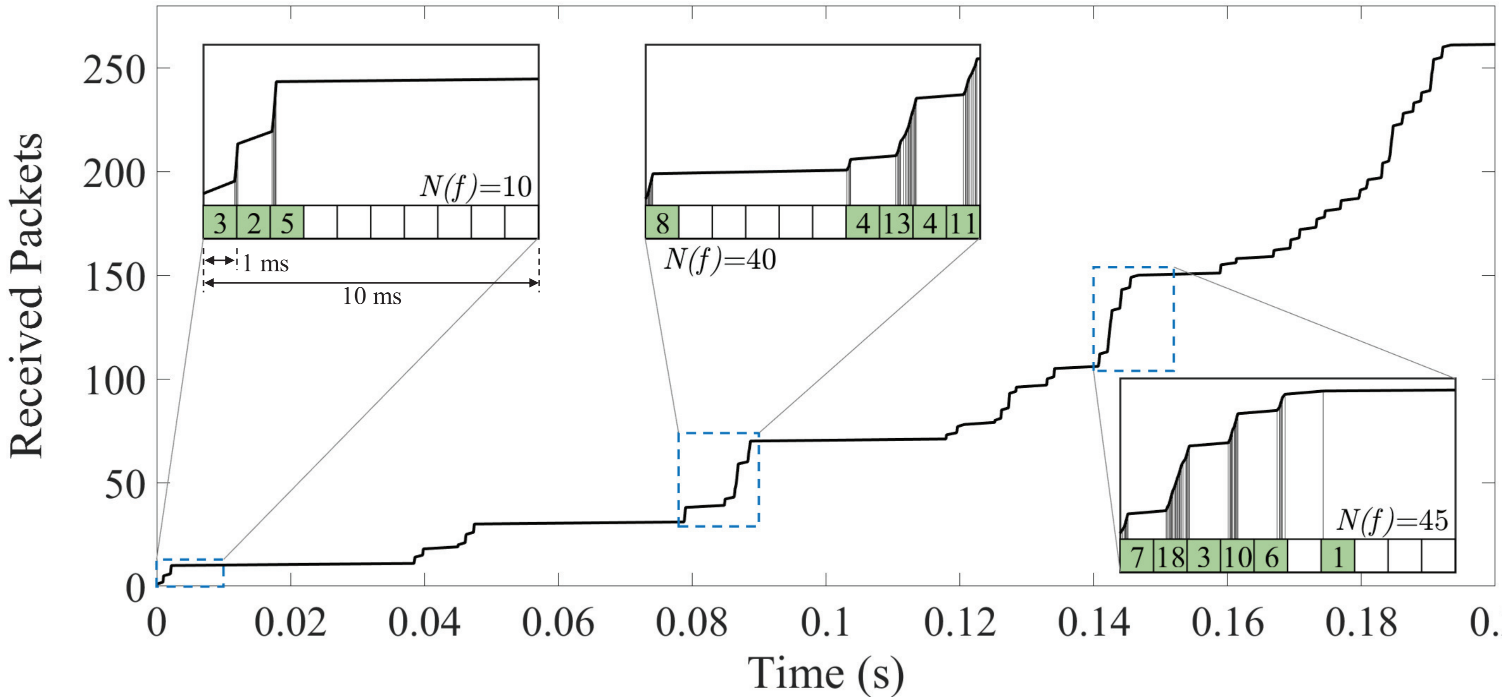


Latency Limit without Throughput Loss



Eliminating Probing Inefficiency

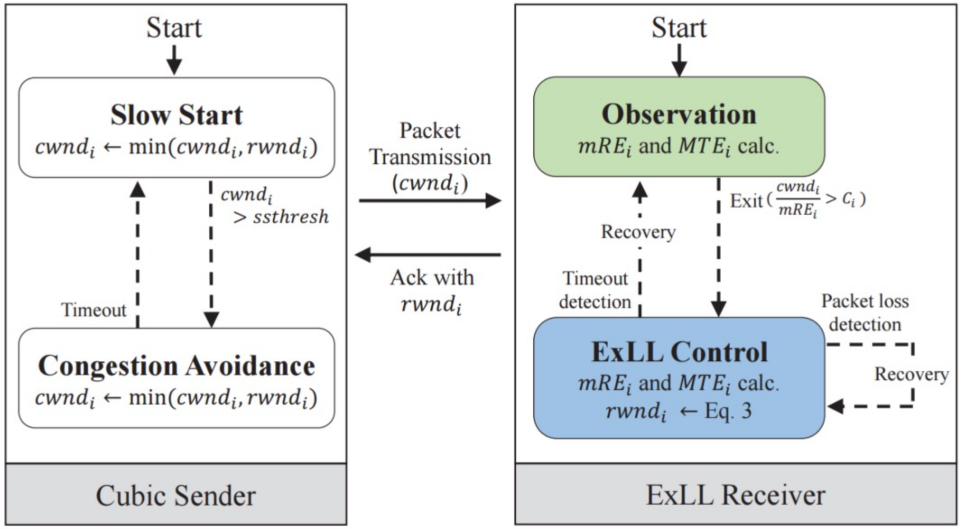
- Letting the cellular receiver infer the available throughput as well as minimum achievable RTT for tracking BDP
  - Packet reception pattern in the receiver has abundant information
    - Consecutive packet interval → available PHY rate (at the chosen MCS)
    - Per-radio frame reception → available PHY rate + user scheduling situation



Window Control Implementation

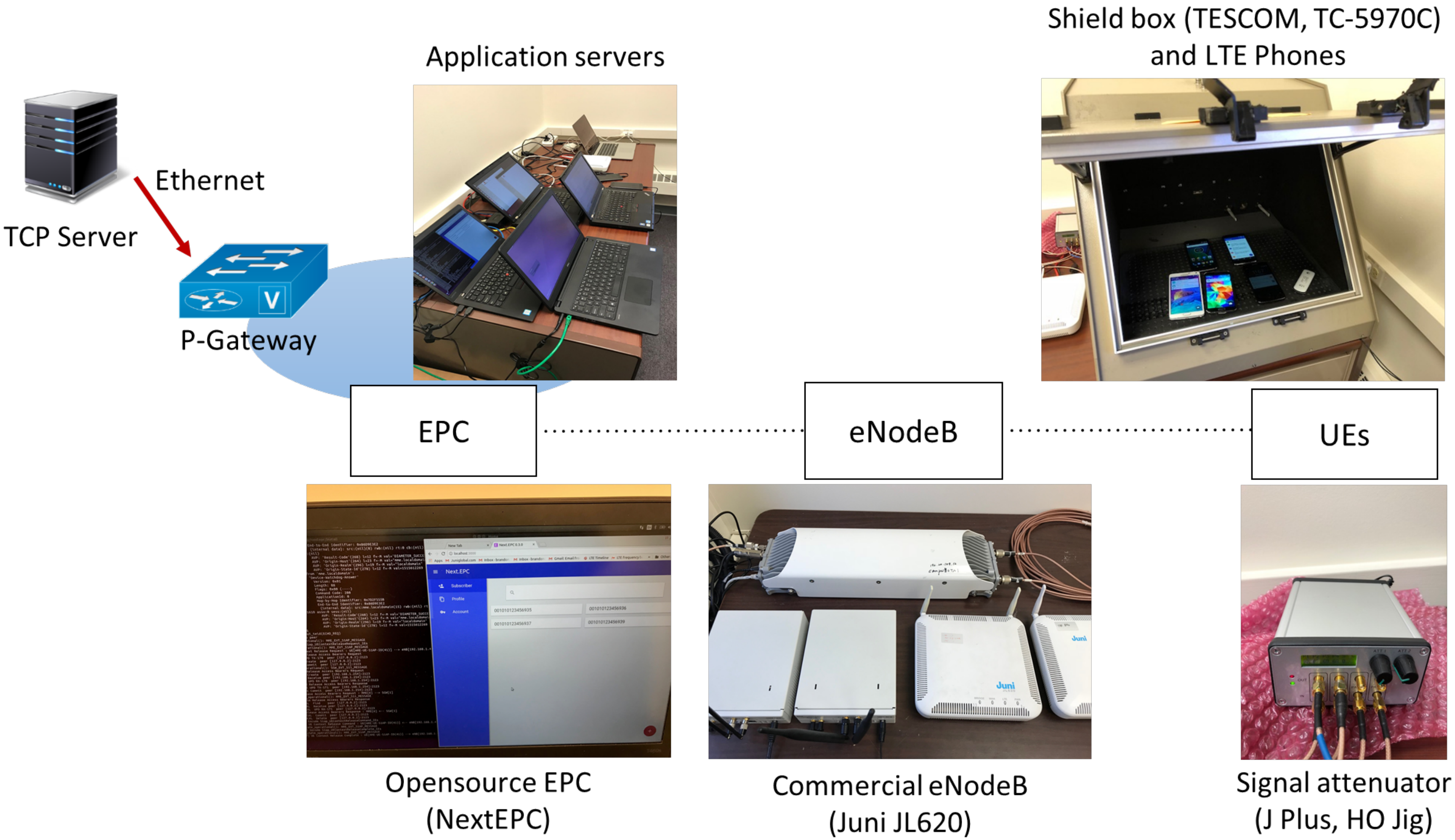
- ExLL: window control with MTE and mRE
  - Converges to the solution of a NUM (network utilization maximization): $\max_{x \geq 0} \sum_i \alpha_i \log x_i$

$$w_{i+1} = (1-\gamma) \cdot w_i + \gamma \left(\underbrace{\frac{mRE_i}{R_i} \cdot w_i}_{\text{Probes minRTT}} + \underbrace{\alpha \left(1 - \frac{T_i}{MTE_i}\right)}_{\text{Probes max throughput, goes to 0 when MTE is achieved}} \right)$$

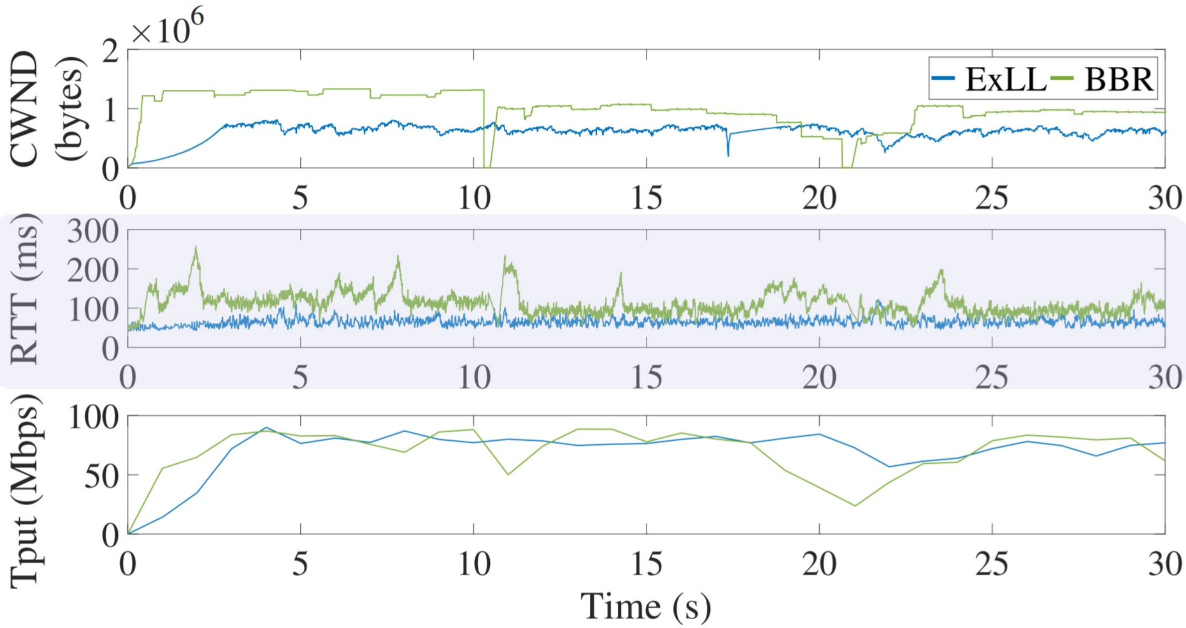


ExLL receiver makes any CUBIC sender to be a low-latency sender.

Testbed with Reproducible Environment

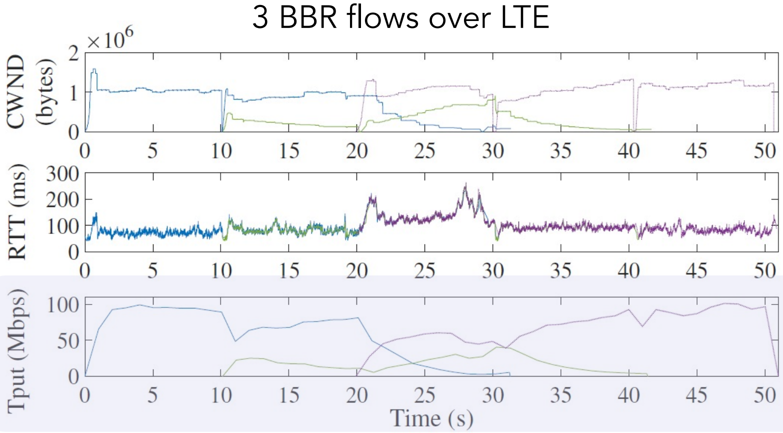


ExLL Provides Near-Ideal Performance

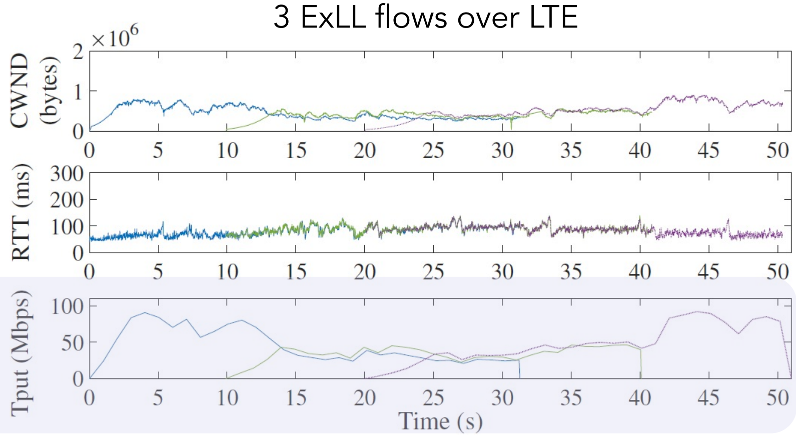


Extremely suppressed latency

3 BBR flows over LTE



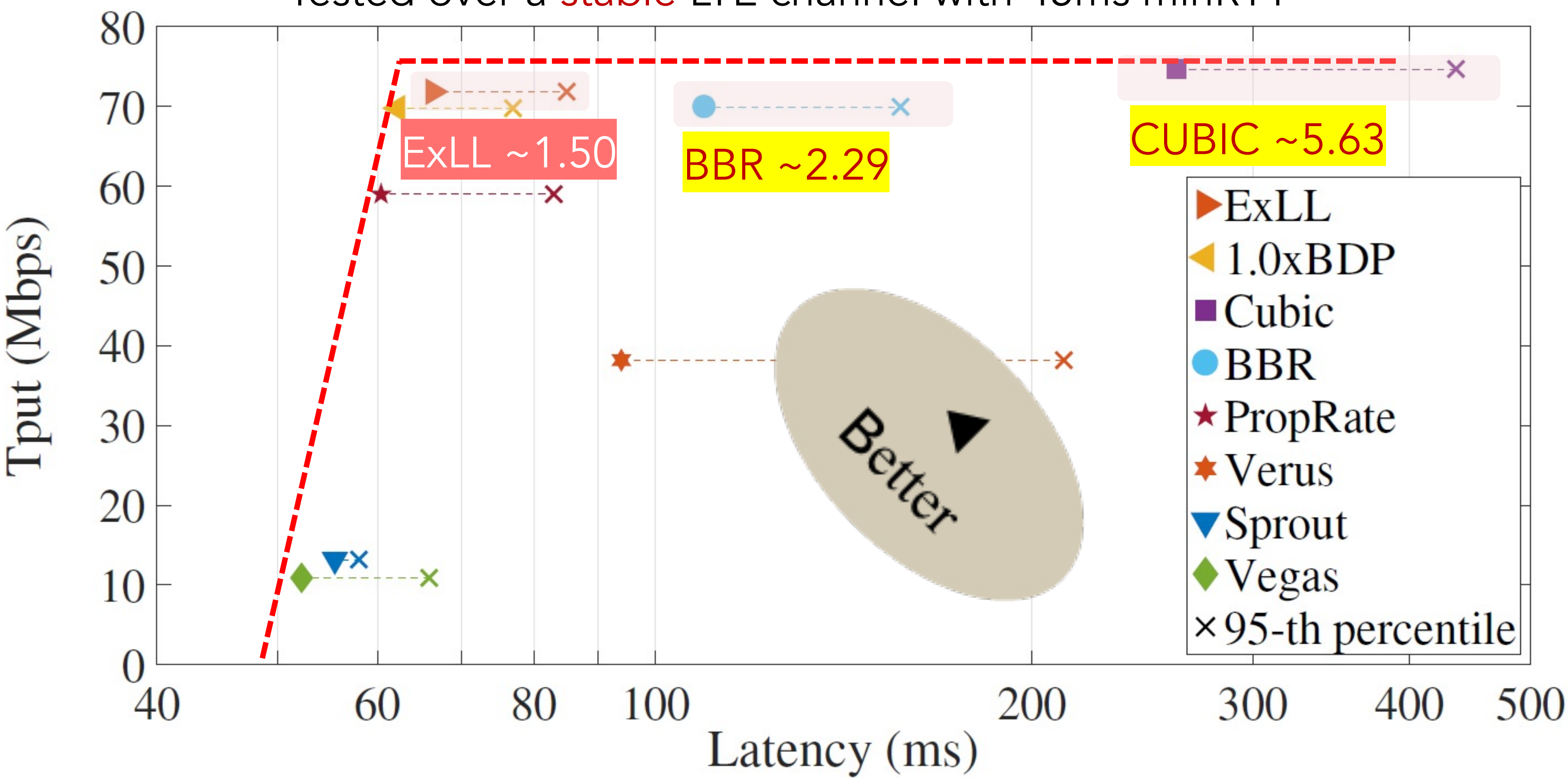
3 ExLL flows over LTE



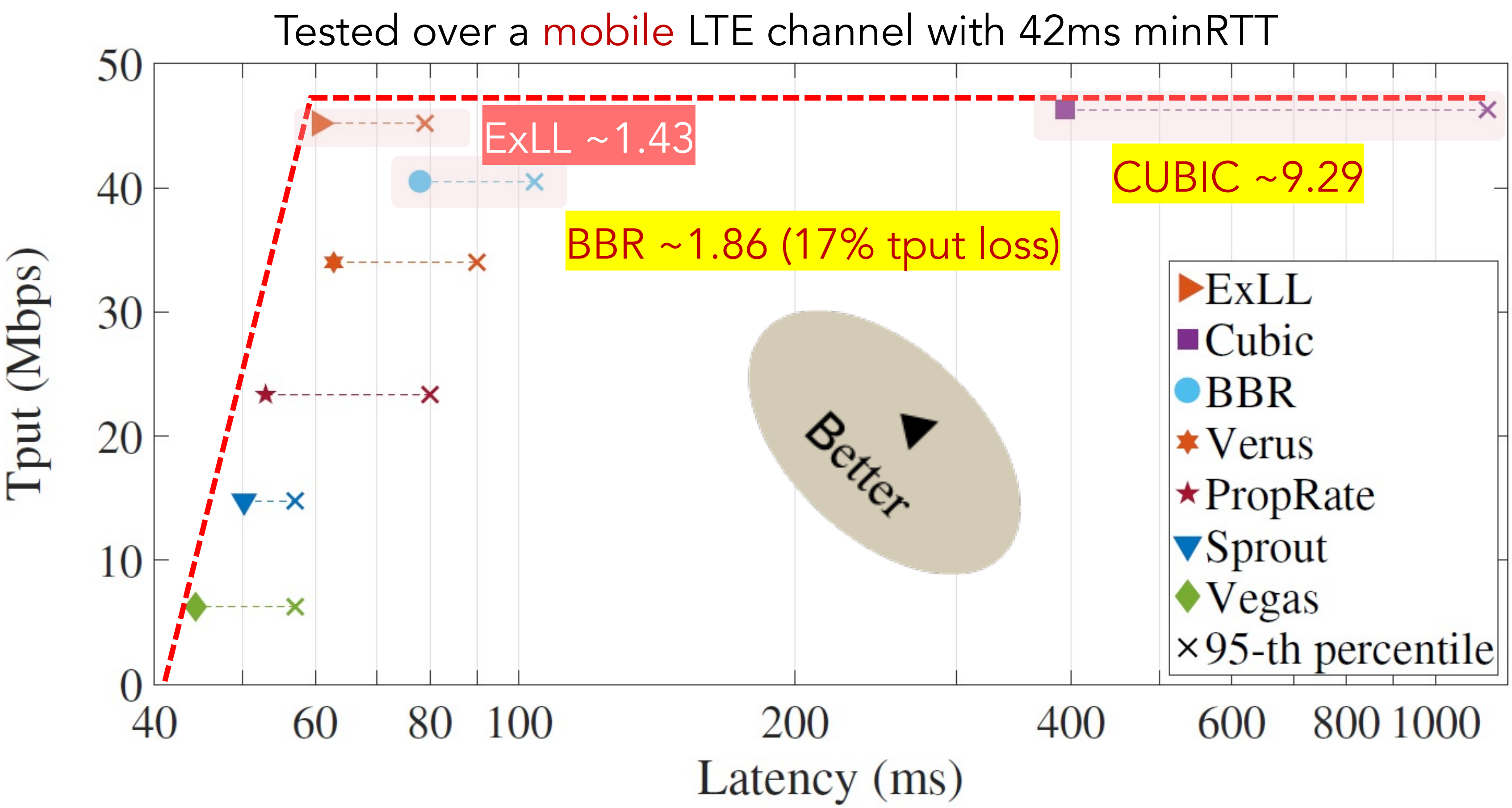
Guaranteed fairness

ExLL Provides Near-Ideal Performance

Tested over a stable LTE channel with 45ms minRTT



ExLL Provides Near-Ideal Performance

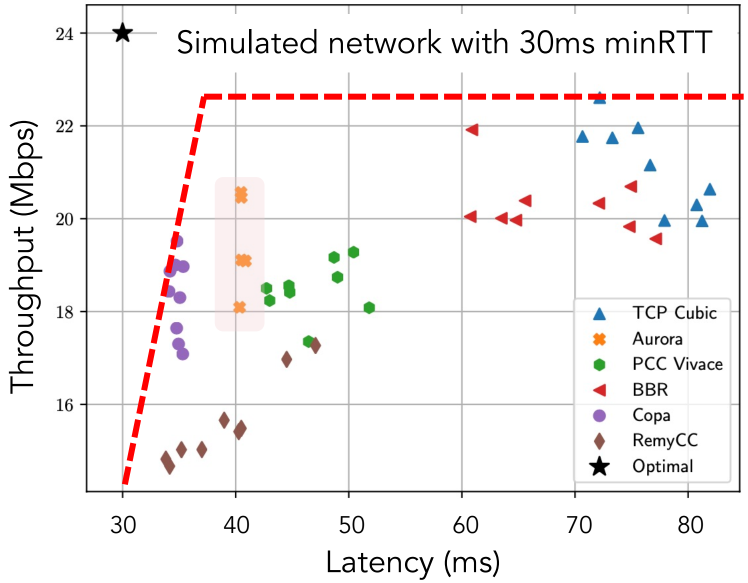


Low-Latency Transport with ML

- People stated to test RL (e.g., Aurora [ICML'19]) for congestion control

$$R_t = \mathbb{E}\left[\sum_t \gamma^t \cdot r_t\right]$$

$$r_t = 10 * throughput - 1000 * latency - 2000 * loss$$

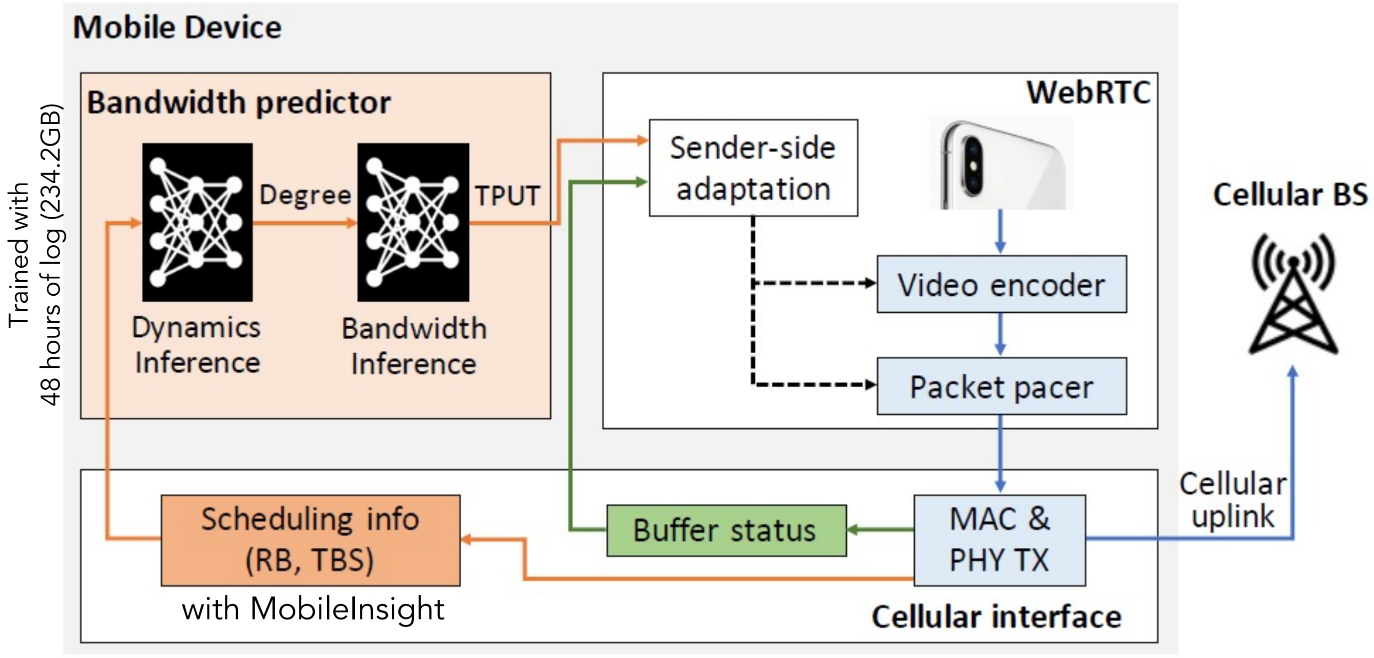


- Training for all possible cases is impossible. (Exploration inefficiency exists.)
- Fairness between flows has not been addressed.

Low-Latency Transport with ML

- Throughput prediction for near-zero queueing
  (under mobility/handoff/user-contention)

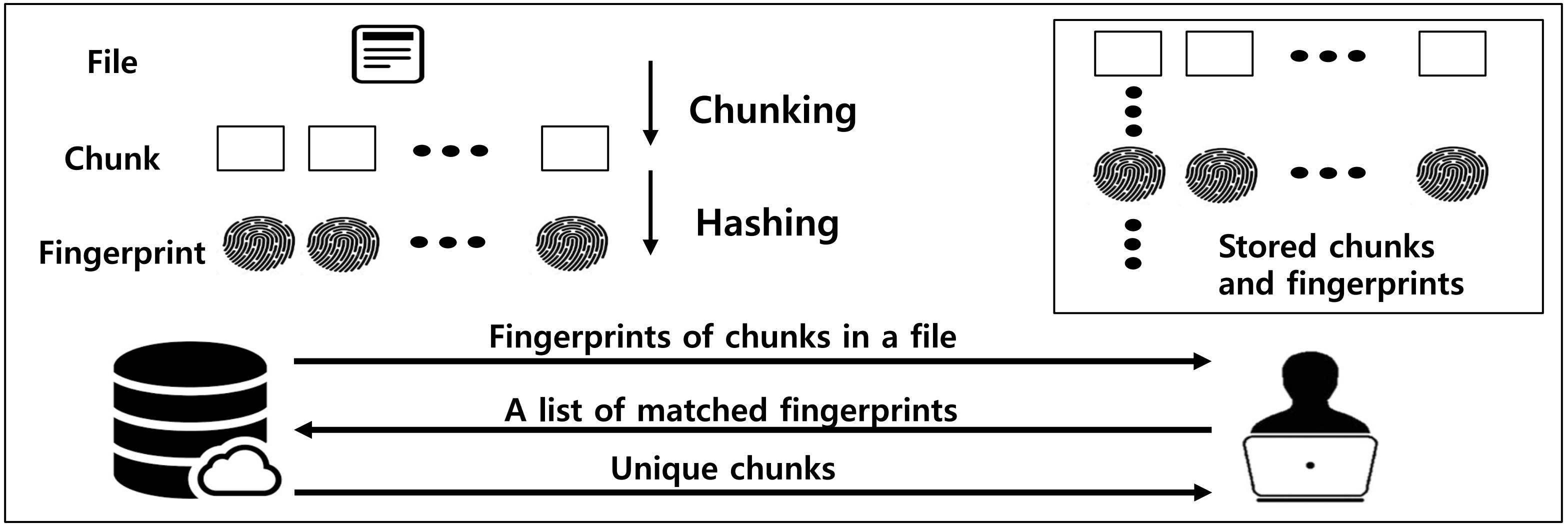
  PERCEIVE [MobiSys'20] 2-stage LSTM on Pixel 3 for cellular uplinks gives
  6.25% tput error @ 224ms
  8.73% tput error @ 22.3ms (with 10x model compression)



Challenge 2: Volume Reduction

- RE is a network (middlebox) technique that can remove duplicate data from within arbitrary network flows for traffic volume reduction.

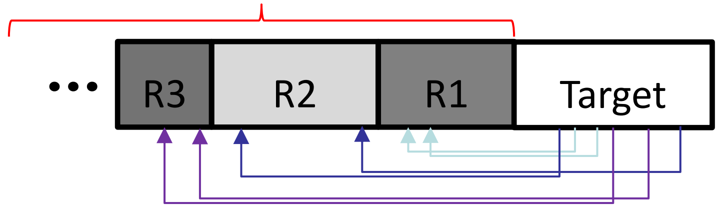
End-host RE ~ Transmission with Dedup



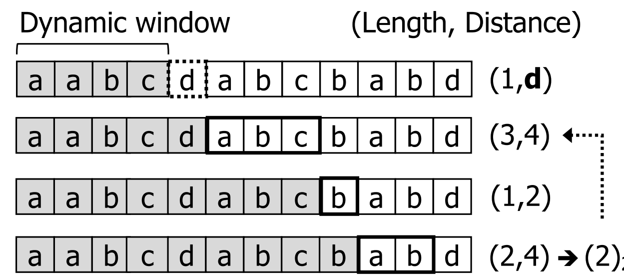
B. Aggarwal, A. Akella, A. Anand, A. Balachandran, P. Chitnis, C. Muthukrishnan, R. Ramjee and G. Varghese, "EndRE: An End-System Redundancy Elimination Service for Enterprises," NSDI 2010.

SyncCoding: RE with Pre-Synced Data

Selected References to encode a Target



Encode longest-matches

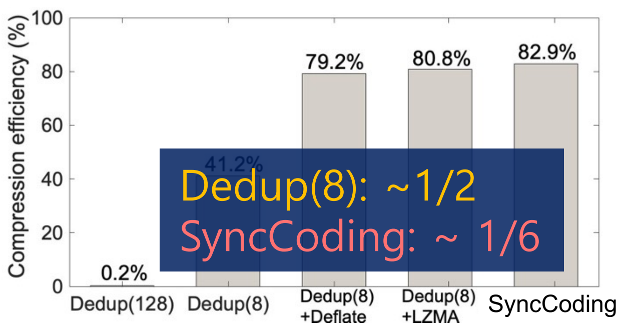


SyncCoding

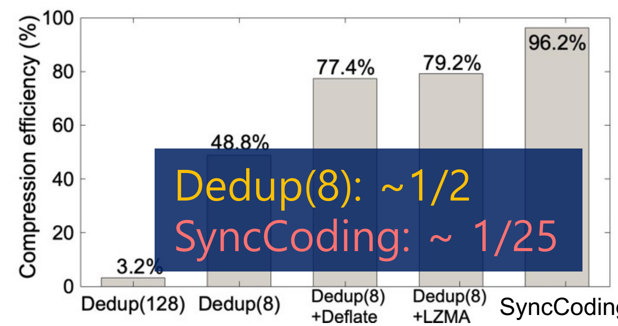
vs.

Dedup(n)

(using n-byte chunk)

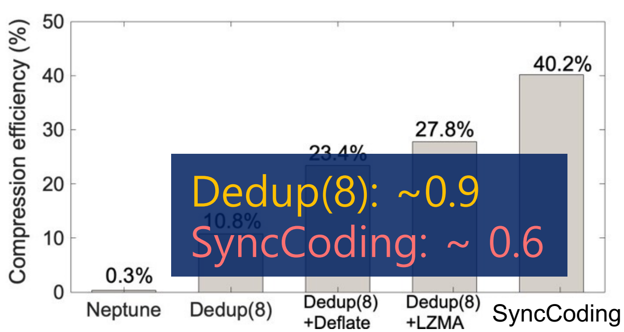


RFC Documents

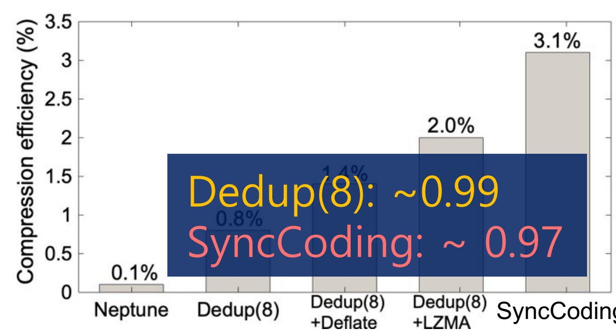


Web Source Codes

Compression efficiency = $\frac{\text{compressed amount}}{\text{original size}}$

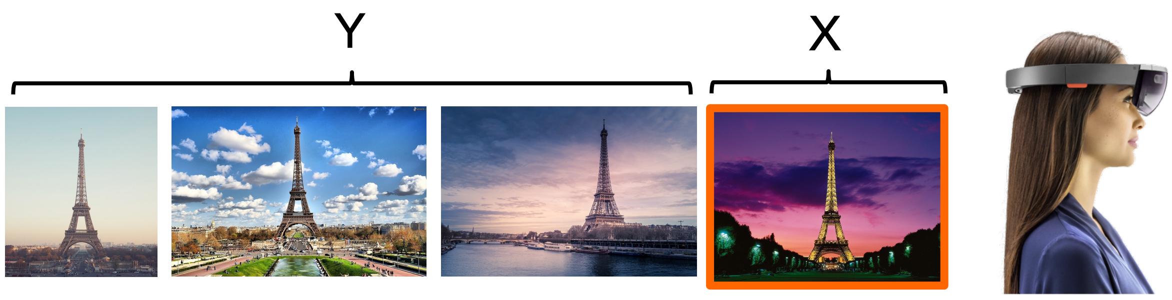


PDF Files



JPEG Images

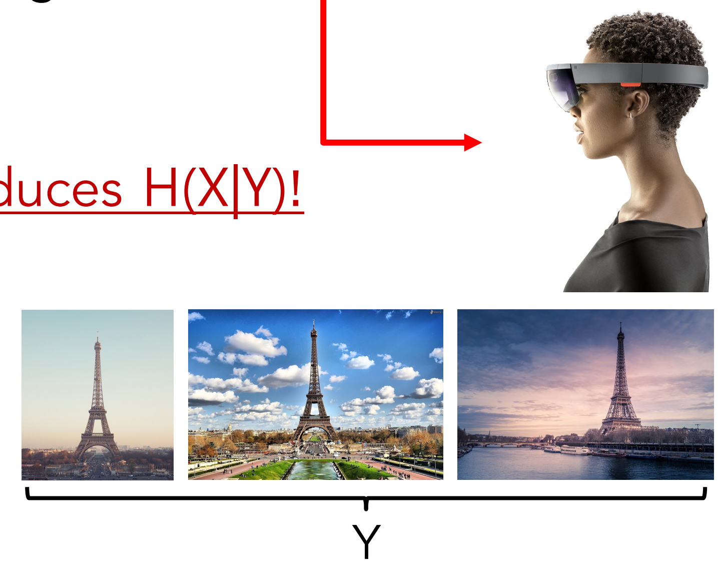
Ultimate RE inside Transport Layer



~~Transmitting H(X)~~

Transmitting H(X|Y)

<u>Design Y that extremely reduces H(X|Y)!</u>



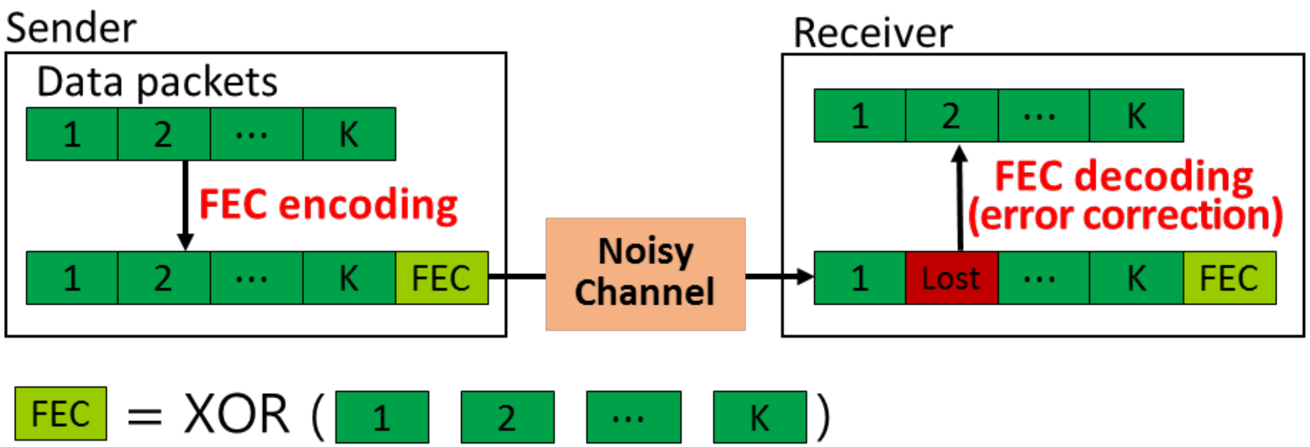
Challenge 3: Retransmission Delay

- HINT: QUIC (Quick UDP Internet Connections)
  - UDP based transport protocol with functions of TCP and TLS.
  - Standardized as HTTP/3 (from 2018.11)

- QUIC advantages
  - Connection establishment latency
  - Forward error correction
  - Improved congestion control
  - Multiplexing without head-of-line blocking
  - Connection migration

QUIC Protocol: FEC

- Forward Error Correction
  - To recover from lost packets without waiting for a retransmission
  - QUIC uses a simple XOR FEC (with session multiplexing)



  - A more efficient FEC can be employed for the same purpose.