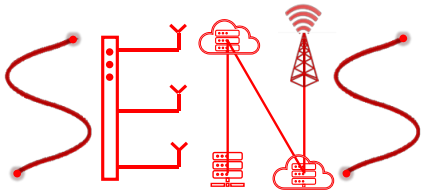
Ultra-Reliable Low-Latency in 5G: A Close Reality or a Distant Goal?

Arman Maghsoudnia[1], Eduard Vlad[2], Aoyu Gong[1],
Dan Mihai Dumitriu[3], Haitham Hassanieh[1]

EPFL[1], RWTH Aachen[2], Pavonis LLC[3]

EPFL



Laboratory of Sensing
& Networking Systems

Ultra-Reliable Low-Latency (URLLC) in 5G

Autonomous driving

Latency: 3 ms [1]

AR/VR

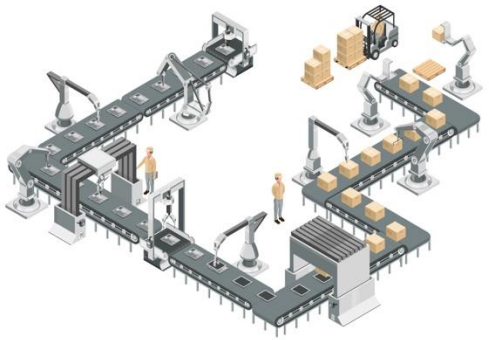
5 ms [1]

Professional Media

1 ms [3]

Smart grids

2 ms [1]



Industrial automation

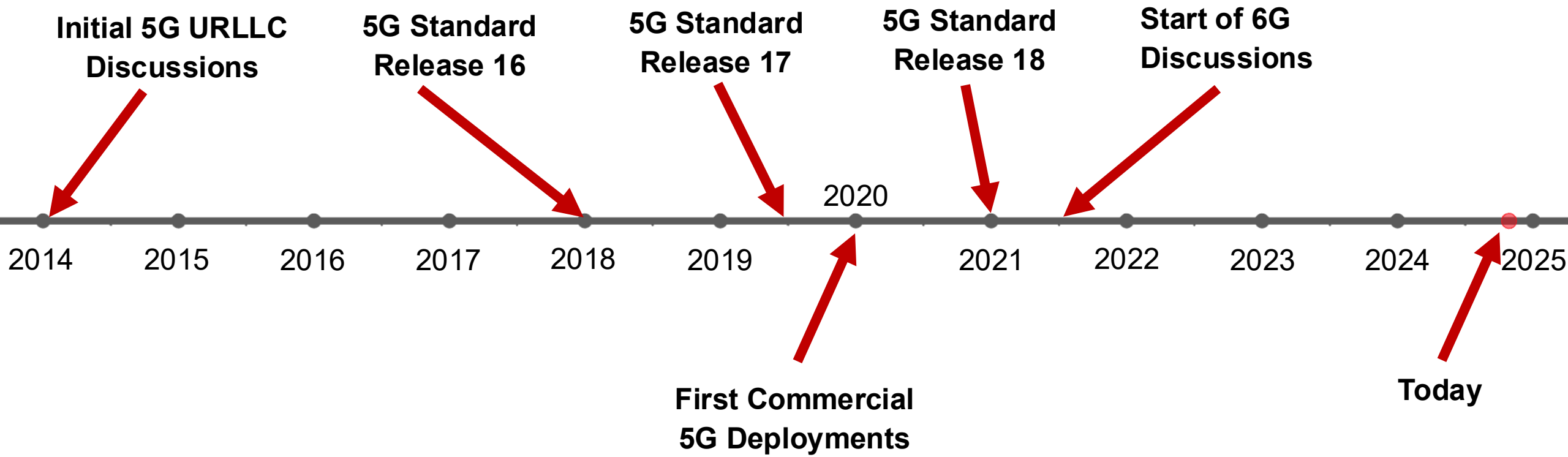
0.5 ms [2]

Latest 5G standard: 0.5ms one-way Latency

6G discussions: 0.1ms

3GPP TR 38.824 V16.0.0 (2019-03) [1], Brown; TR Qualcomm 2018 [2]; Nokia and Sennheiser, white paper 2020 [3]

5G Timeline



No practical URLLC deployment today

Related Work

Theoretical & Simulation-Based Studies

- Overlooking inherent bottlenecks in the system
- Many assuming that reducing the time slot is sufficient

Measurements

On mm-wave:
Only 4.4% sub-millisecond in line-of-sight [1]
Qualcomm in US: 6ms RTT [2]

Campus networks:
6 to 12ms RTT [3, 4, 5]

Practical Implementations

- Professional audio transmission: Minimum latency of 0.8ms for a single user line of sight [6]
- Hexapod Robot by Ericsson: 5ms RTT [7]

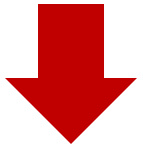
Fezeu et al, PAM 2023 [1]; Qualcomm, TR 2022 [2]; Lackner et al, CIRP 2022 [3]; Rischke et al, IEEE Access 2021 [4]; Rischke et al, IEEE WoWMoM 2022 [5]; Nokia and Sennheiser, white paper 2020 [6]; Ericsson, MWC 2019

Are the latency and reliability requirements of URLLC achievable in practical 5G systems?

Are the latency and reliability requirements of URLLC achievable in practical 5G systems?

At sub-ms latencies → Every source of latency becomes a bottleneck

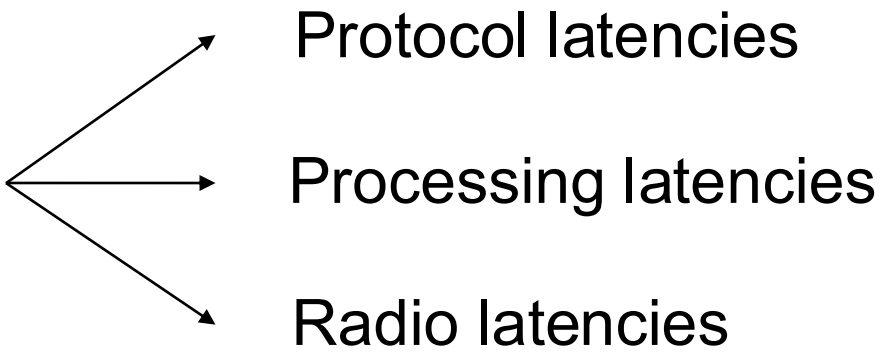
Our Goal: Uncover the bottlenecks in 5G through a system-level analysis



System designs that can meet the latency requirements

In this paper

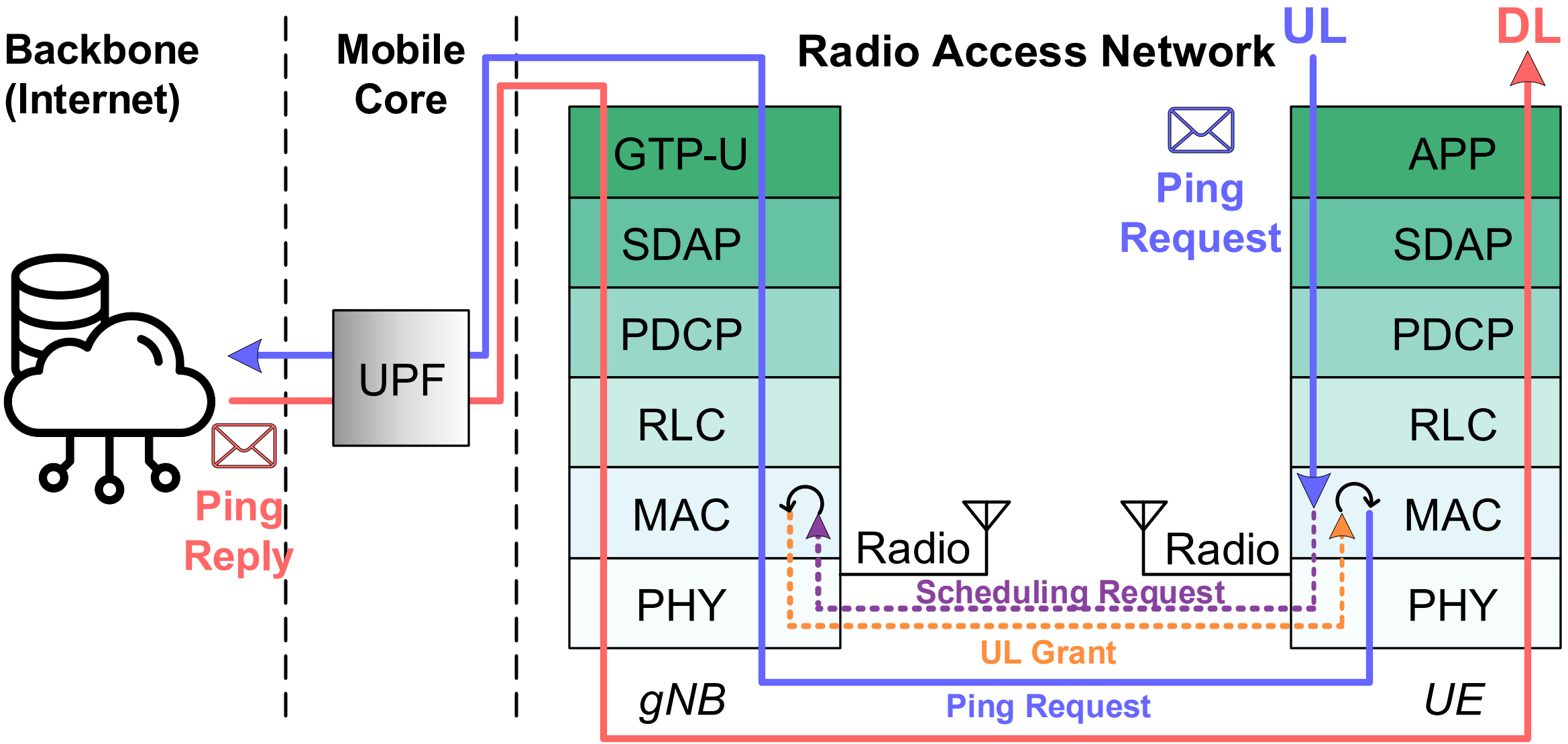
- Highlight the different latency sources



- Analyze the interdependency between latency sources and their impact on the overall system latency
- Validate our analysis on a real-world 5G testbed

While URLLC in principle is possible, it is only feasible under very few 5G configurations, some of which being not practical or scalable.

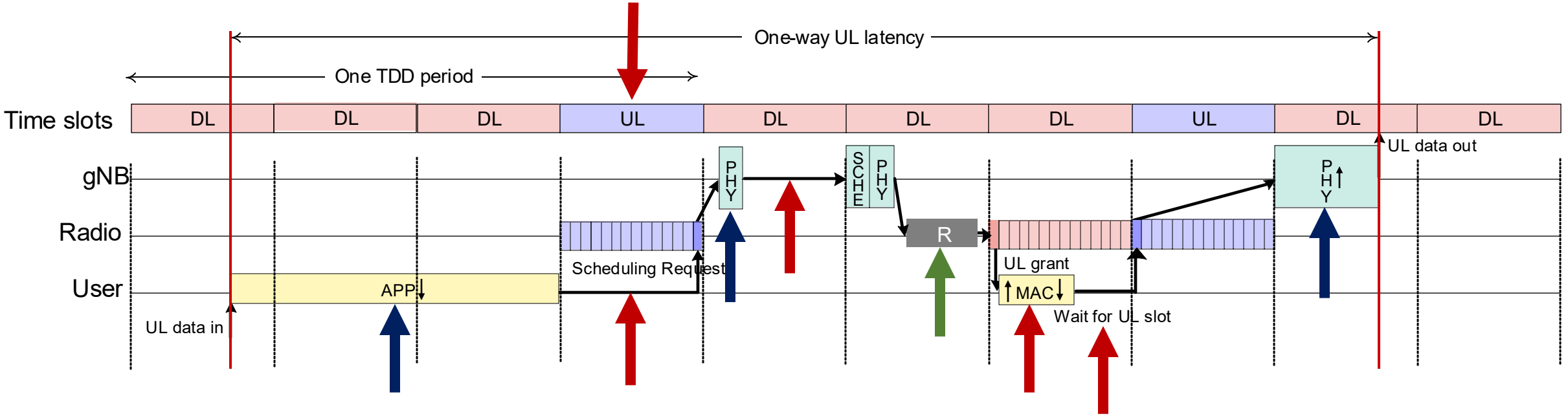
Journey of a Packet



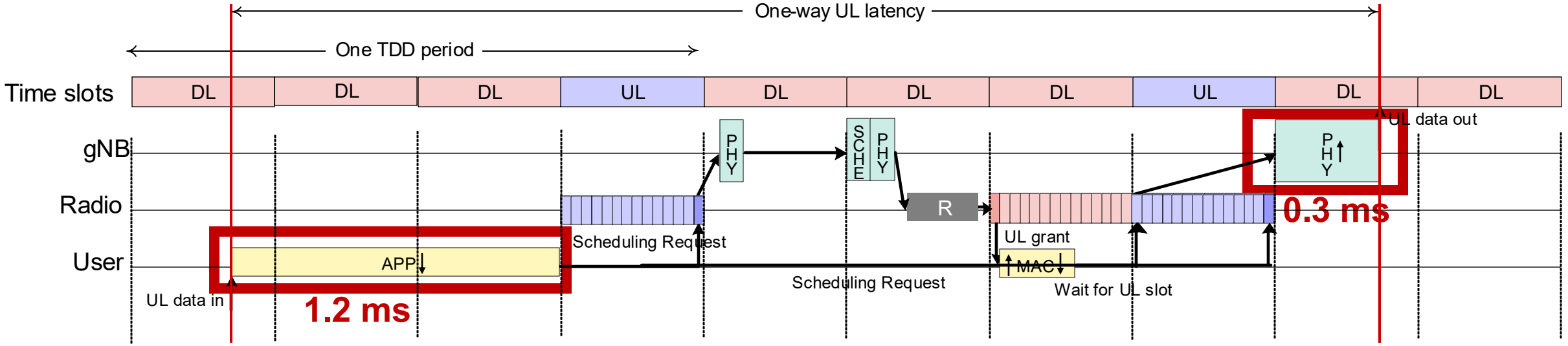
Latency Sources & Their interactions

Configuration

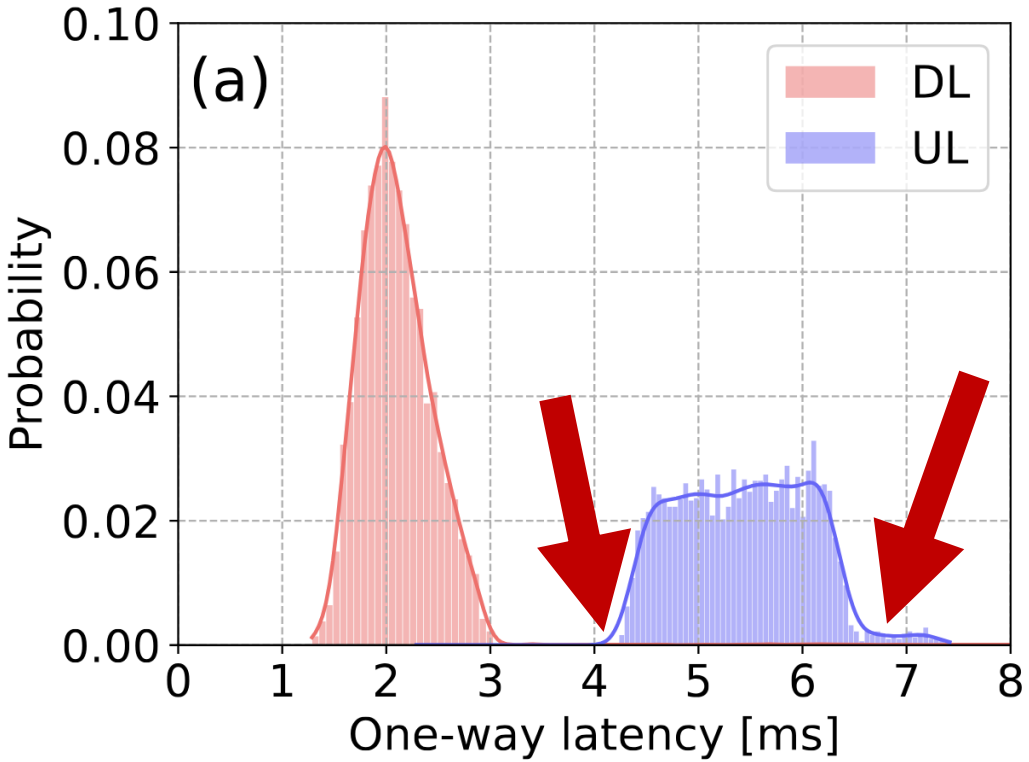
- Time Division Duplexing DL-DL-DL-UL pattern
- Slot duration: 0.5ms

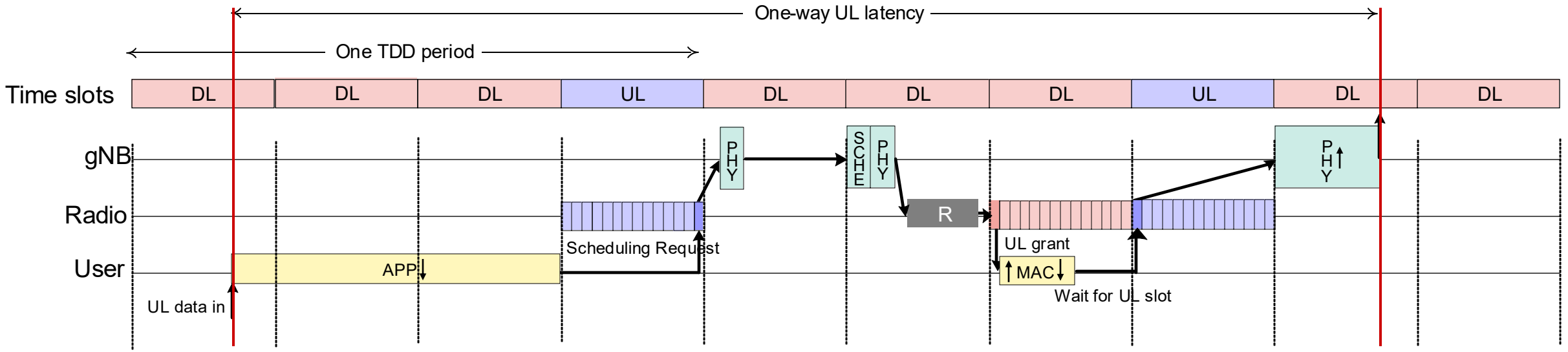


Protocol-based latency Processing-based latency Radio-based latency

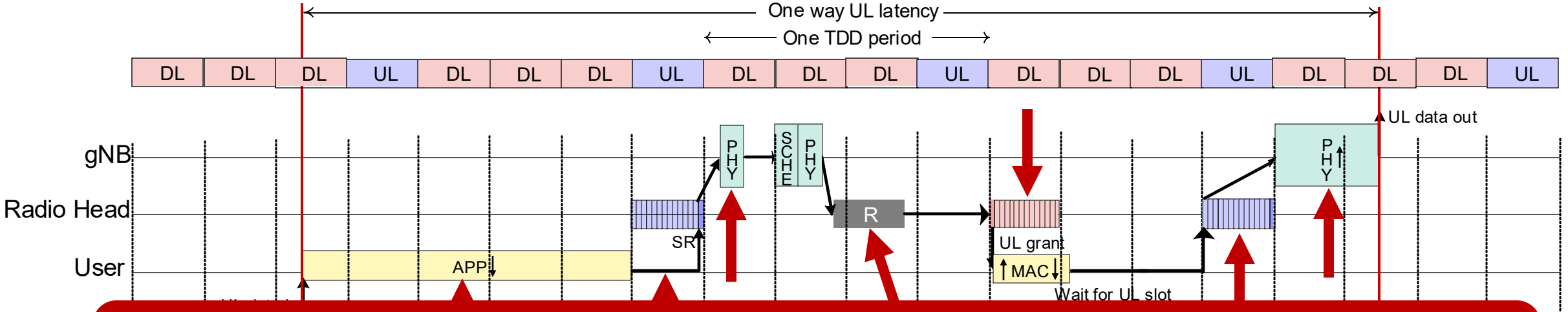

1.2 ms + 5 * 0.5 ms + 0.3 ms = 4 ms
One-way UL latency
One TDD period
Time slots
DL
DL
DL
UL
DL
DL
DL
UL
DL
DL
gNB
Radio
User
PHY
SCHE
PHY
R
PHY
UL data out
0.3 ms
Scheduling Request
UL grant
APP
MAC
Scheduling Request
Wait for UL slot
UL data in
1.2 ms


gNB logs
Core logs
5G Core
Radio Head
gNB
5G UE


(a)
DL
UL
Probability
0.10
0.08
0.06
0.04
0.02
0.00
0
1
2
3
4
5
6
7
8
One-way latency [ms]

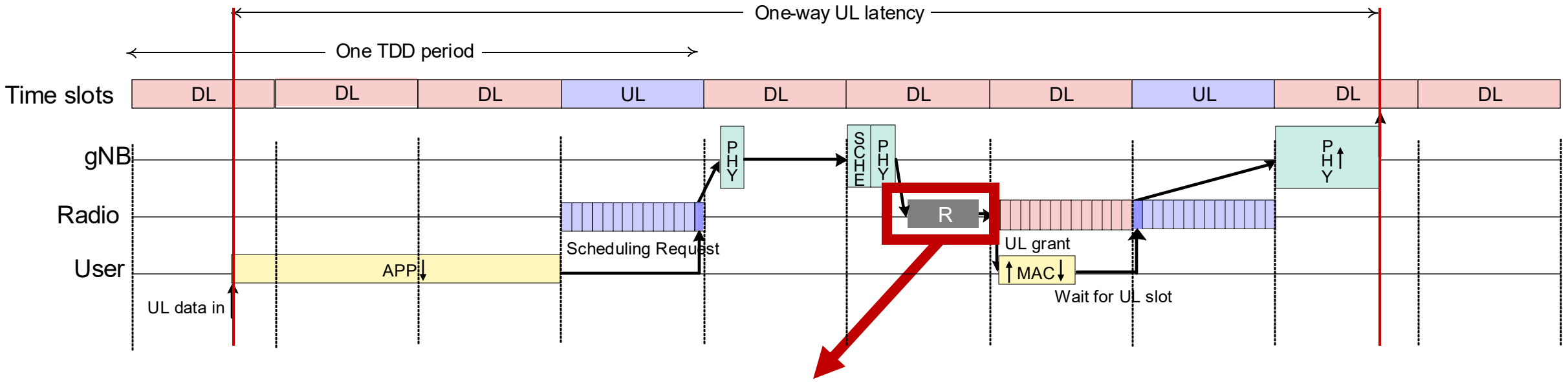


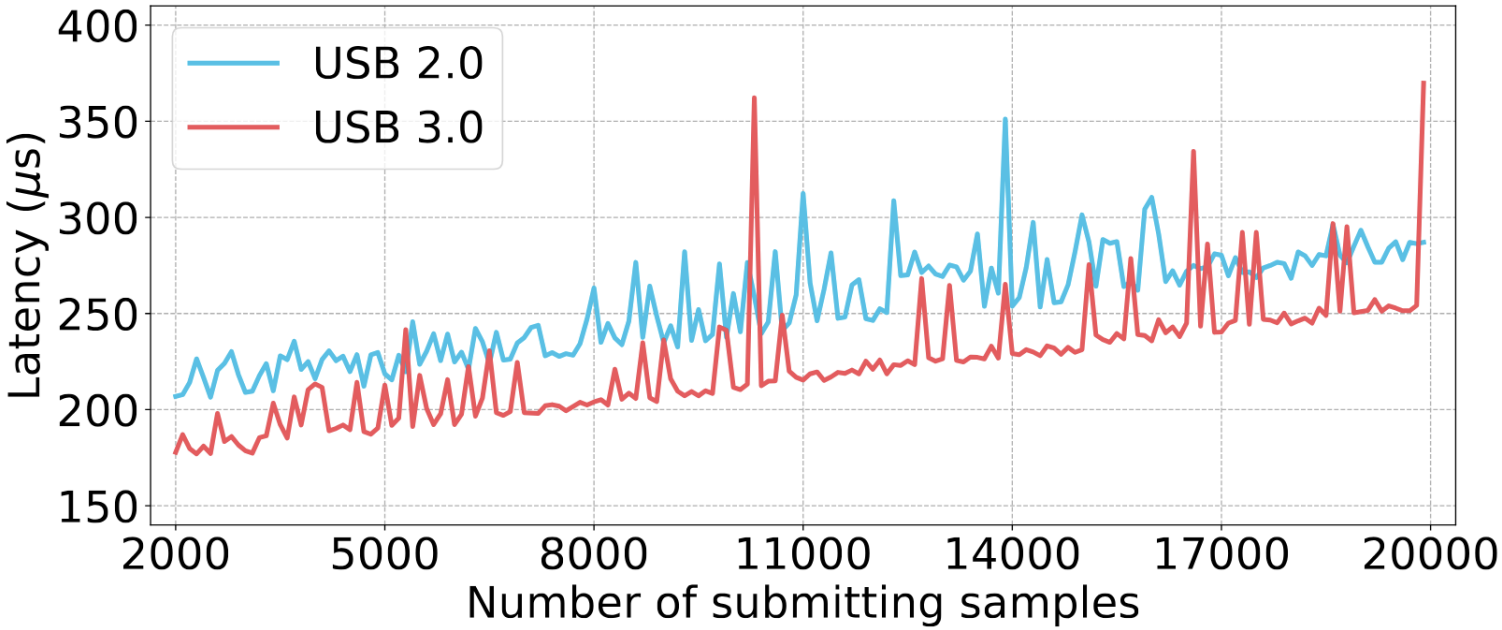
Slot duration halved



Simply reducing slot duration without accurate system-design does not noticeably reduce latency

Non-determinism of Latency Sources





Configuration Choices

- FDD: only sub 2.6 GHz, mostly taken by public operators, inefficiency
- Sub 6 GHz: Shortest slot duration 0.25ms
- TDD mm-wave: poor reliability

Feasibility: Only a few out could theoretically achieve the URLLC requirements.

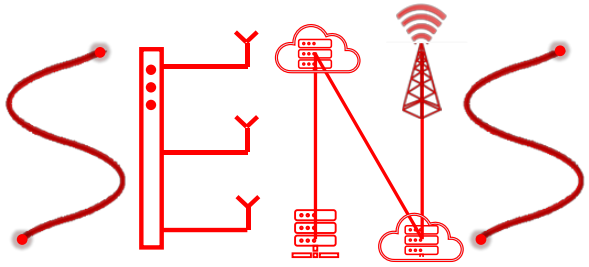
Practicality: Scalability, availability, and reliability issues.

Conclusion & Future directions

- **In our work**
  - For the first time, a system-level analysis of low-latency in 5G
  - Finding the bottlenecks in the system and their interdependency
  - Feasibility and practicality of configuration choices

- **Future directions**
  - Extend to multiple users, and reliability
  - Scalability in URLLC

- **6G URLLC**
  - 0.1ms Latency?



Laboratory of Sensing and Networking @ EPFL