Towards URLLC with Open-Source 5G Software

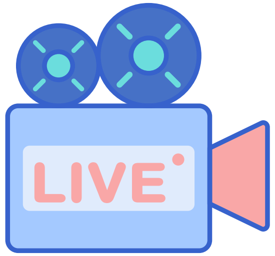
Aoyu Gong[†], Arman Maghsoudnia[†], Raphael Cannatà[†], Eduard Vlad[¶], Néstor Lomba Lomba[†], Dan Mihai Dumitriu[♣], Haitham Hassanieh[†]

EPFL[†], RWTH Aachen[¶], Pavonis LLC[♣]

Motivation

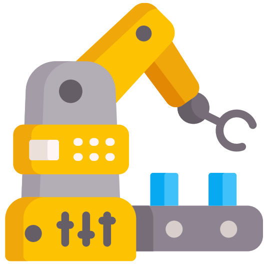
Next-Generation Cellular Networks

Rescue Services
(5 – 20 ms)



Autonomous Vehicles
(5 – 10 ms)

Live Production
(5 – 10 ms)



Industrial Automation
(1 – 2 ms)

Extended Reality
(4 – 7 ms)

NextG applications require low-latency communications

Motivation

Ultra-Reliable Low-Latency Communication

5G : URLLC
0.5 ms for UL/DL

6G : HRLLC
0.1 ms for UL/DL

3GPP TR 38.913 Release 18

Discussion

① Where does latency come from in real 5G systems?

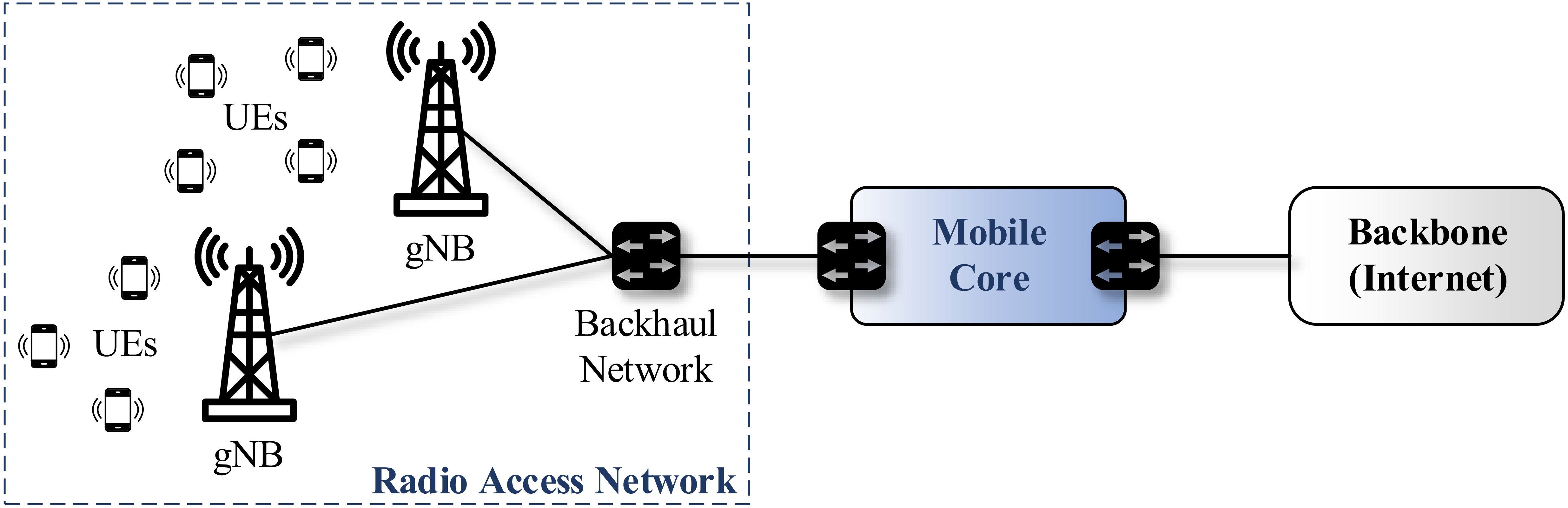
② How do these latency sources interact?

③ What bottlenecks do theorical and simulation work overlook?

④ How can open-source 5G testbeds help reach low latency?

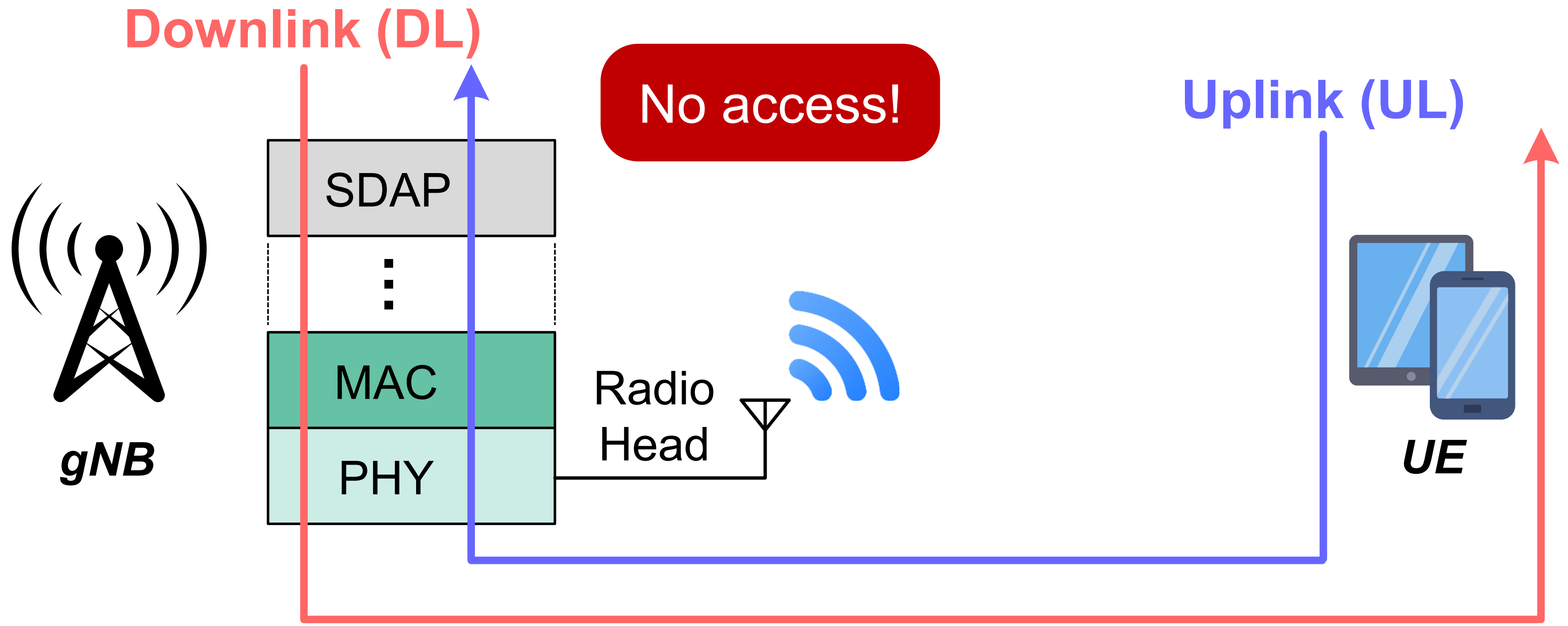
Background

5G Network



Background

5G Stack



- **Layer 2** : Medium Access Control (MAC)
- **Layer 1** : Physical Layer (PHY) & Radio Head

Background

Open-Source 5G Software

Mobile Core :

Radio Access Network :

Full-stack programmability + Ability to experiment on real-world setup

Background

Open-Source 5G Software

Mobile Core :

free5GC   Open5GS

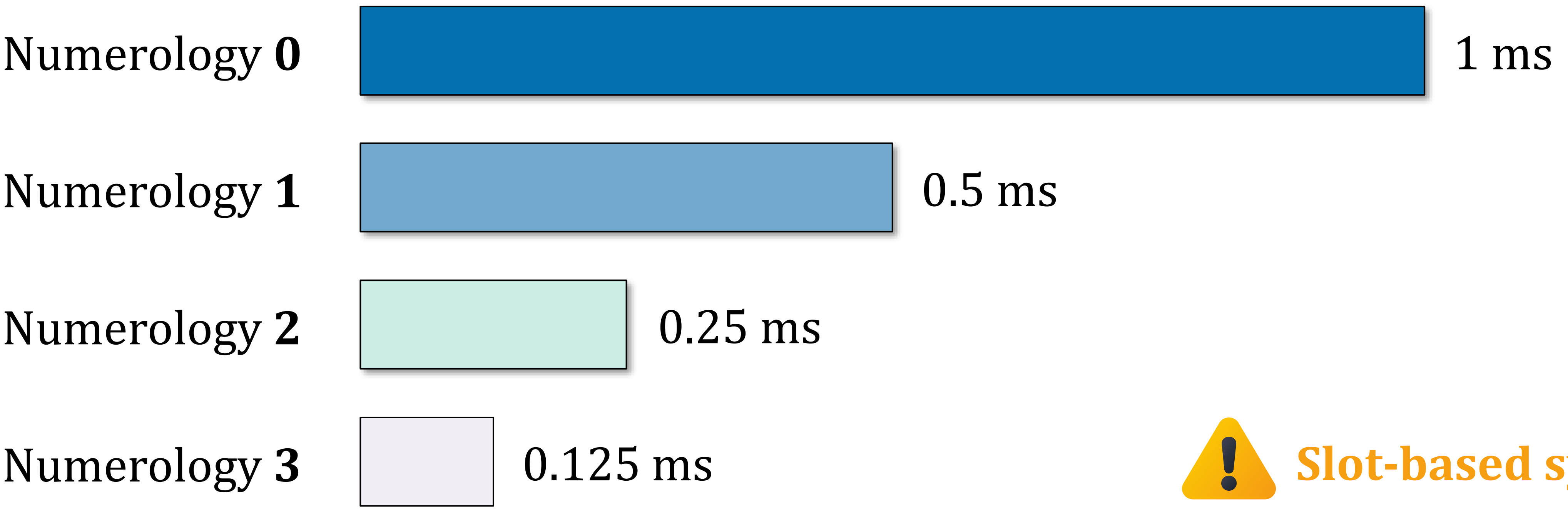
Radio Access Network :

SRSRAN PROJECT   OPEN AIR INTERFACE

Full-stack programmability + Ability to experiment on real-world setup

Background

5G Time Slots

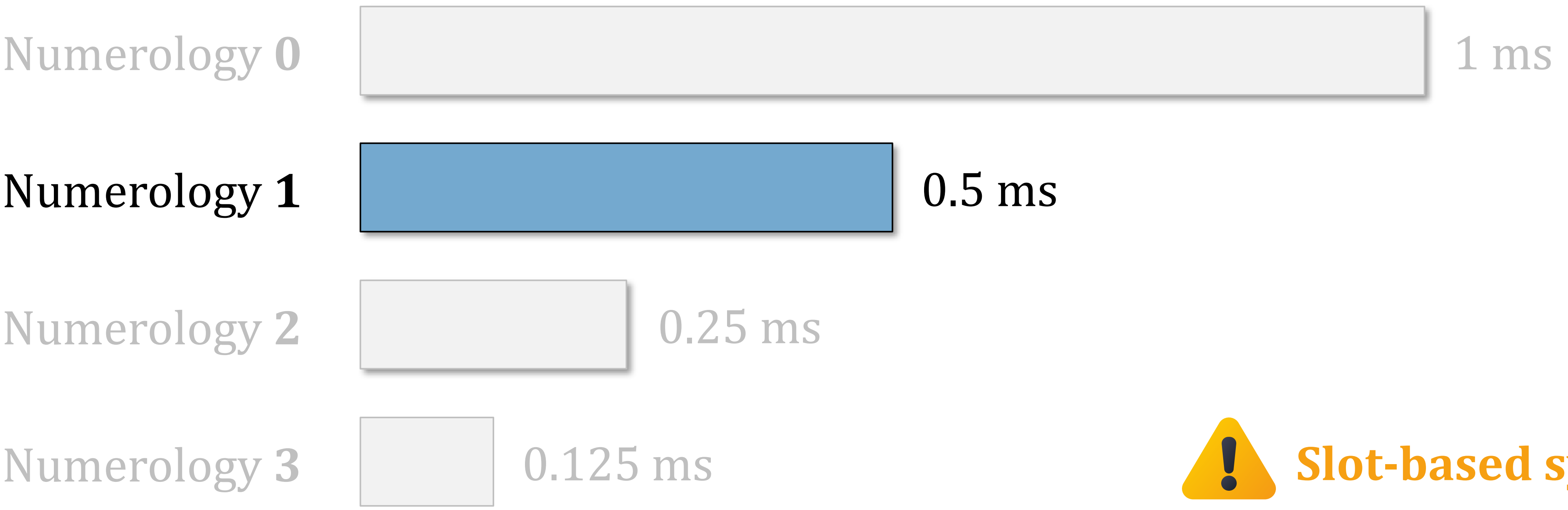


Slot-based system

More flexible than 4G → **Numerology** → Slot duration

Background

5G Slots

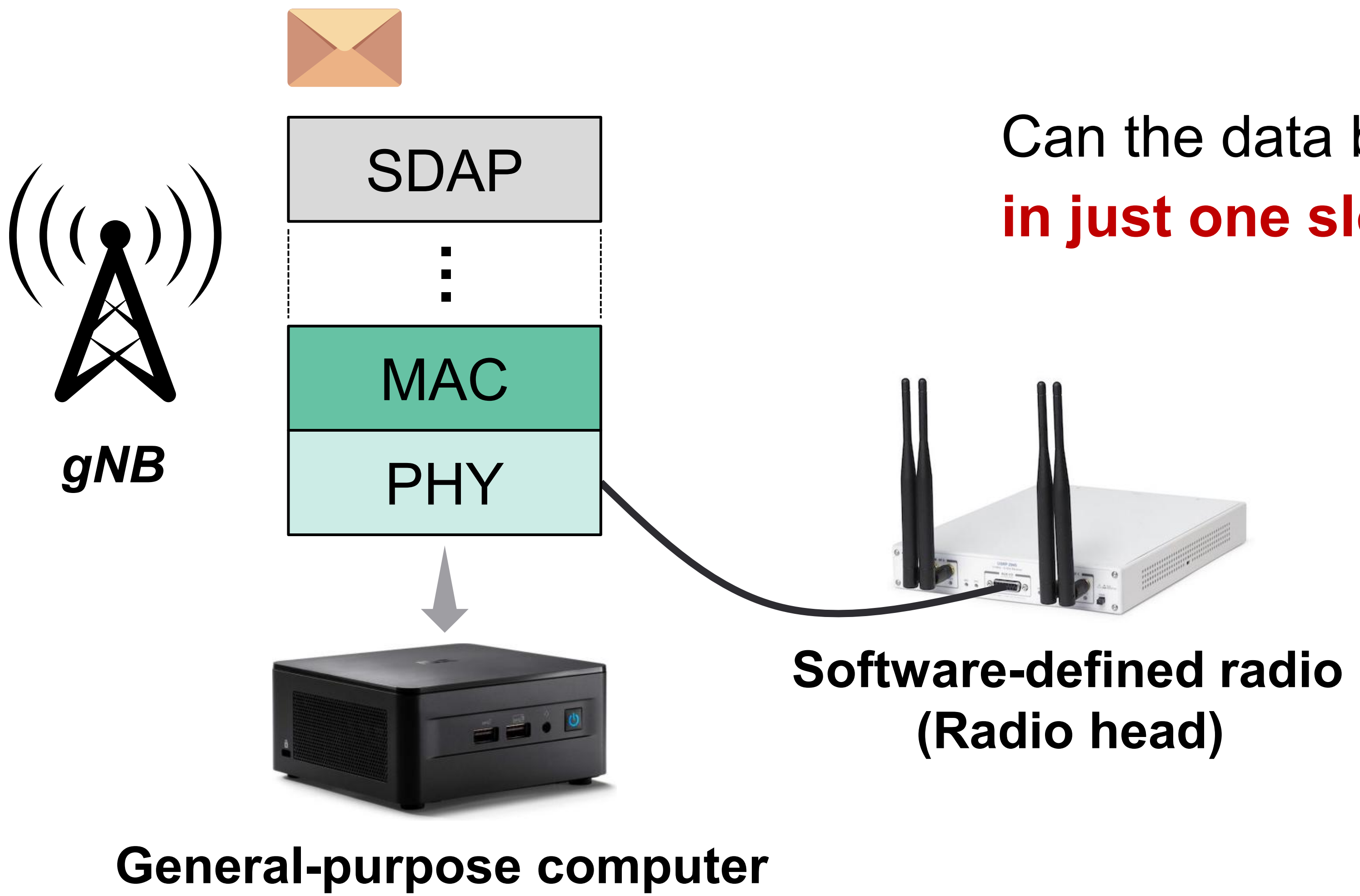


Slot-based system

More flexible than 4G → **Numerology** → Slot duration

Latency Analysis

Slot-Based System

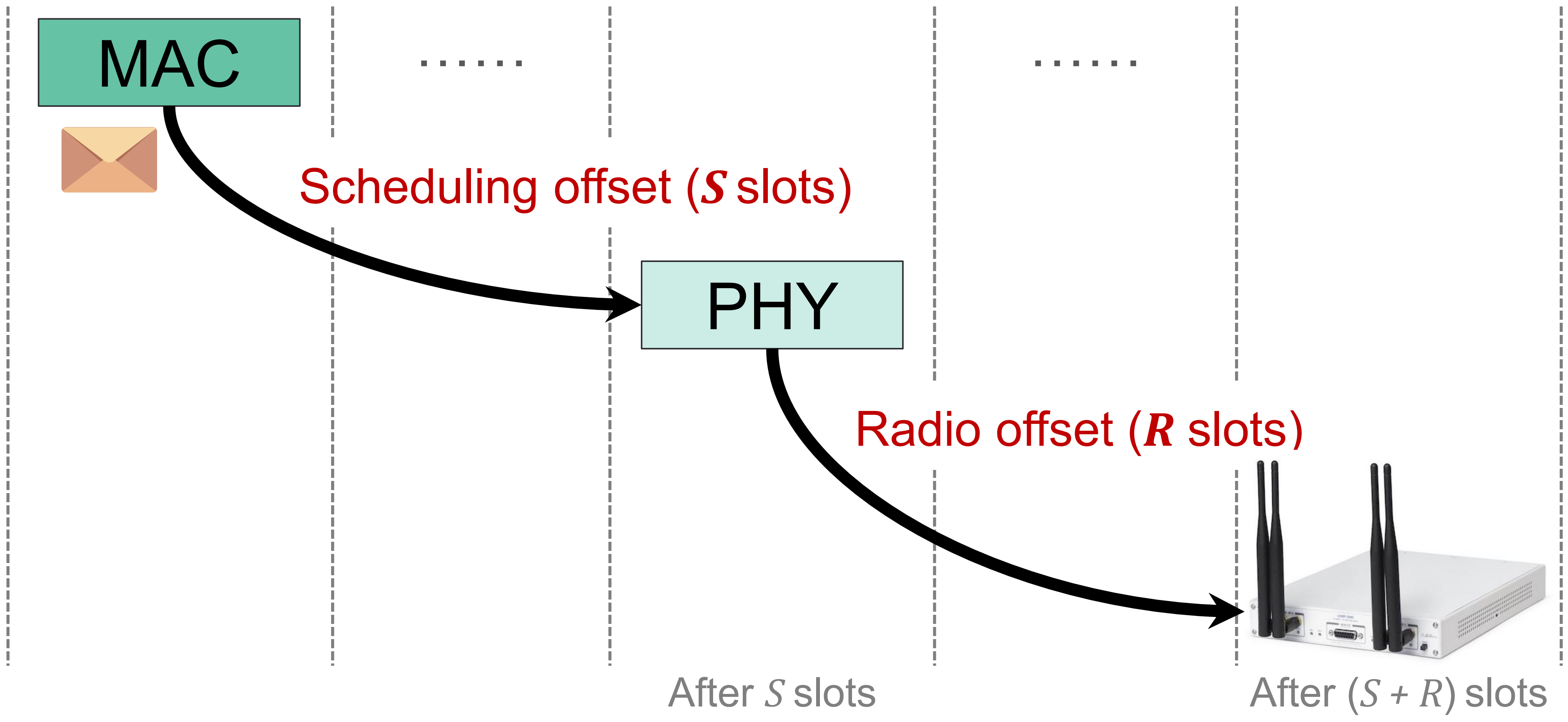


Can the data be delivered to the radio **in just one slot**?

NO

Latency Analysis

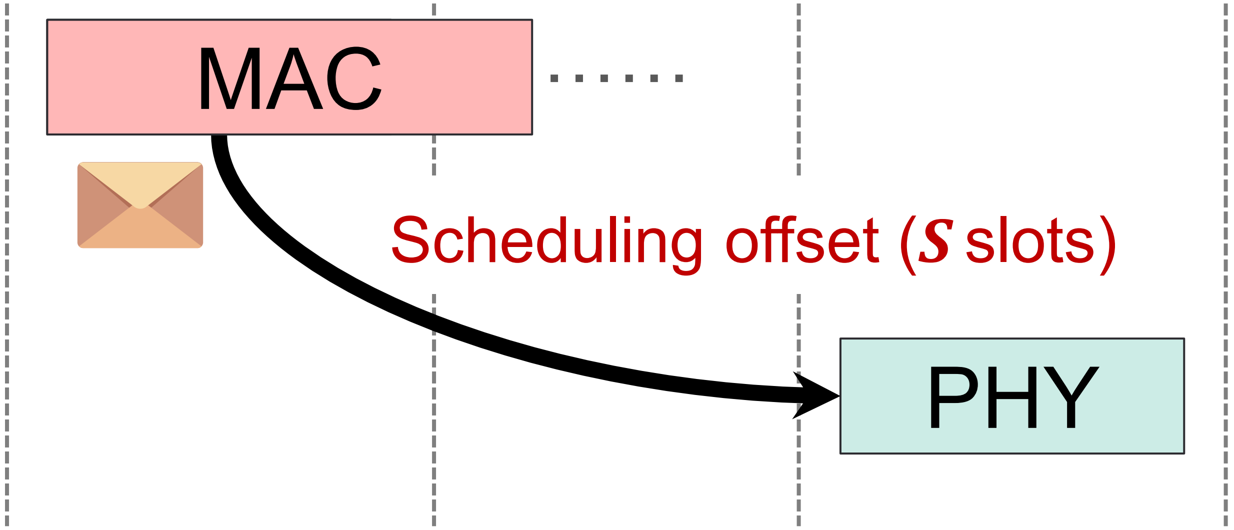
Slot Offsets Across Layers



Higher layers always process and forward downlink data in advance

Latency Analysis

Scheduling Offset (*S* slot)



Default : *S* = 1 slot

Provide a safety margin

Wait 0.5 ms

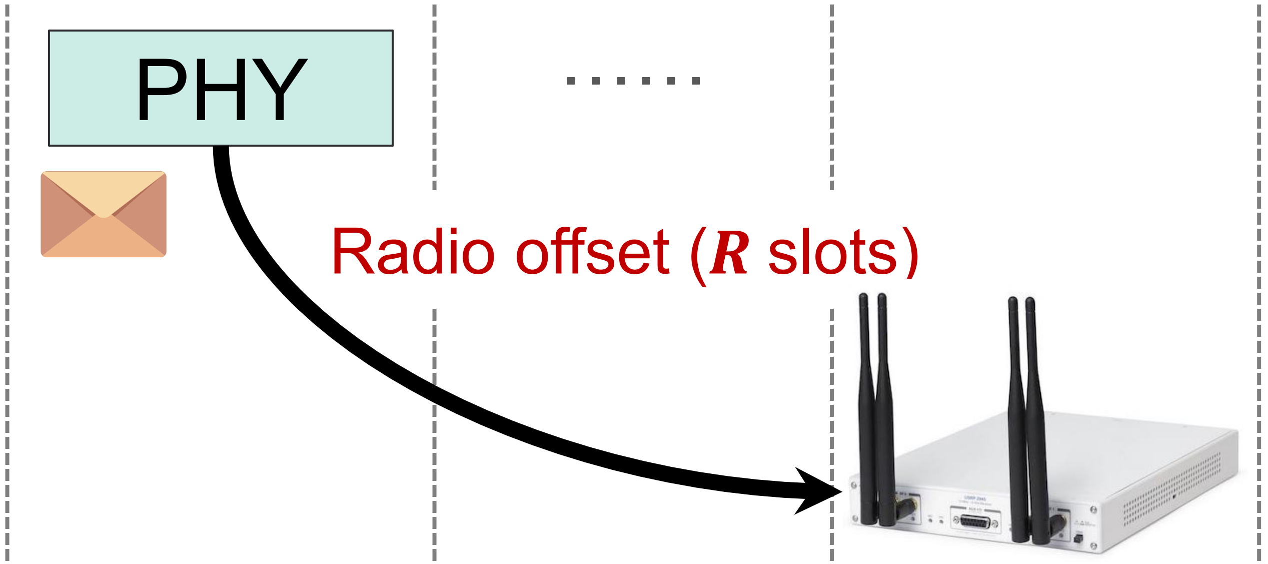
Give extra time in case the execution is occasionally slower than expected

Latency Analysis

Radio Offset (*R* slot)

- **Sufficiently large to support different hardware**
- **Wait 1.5 ms**

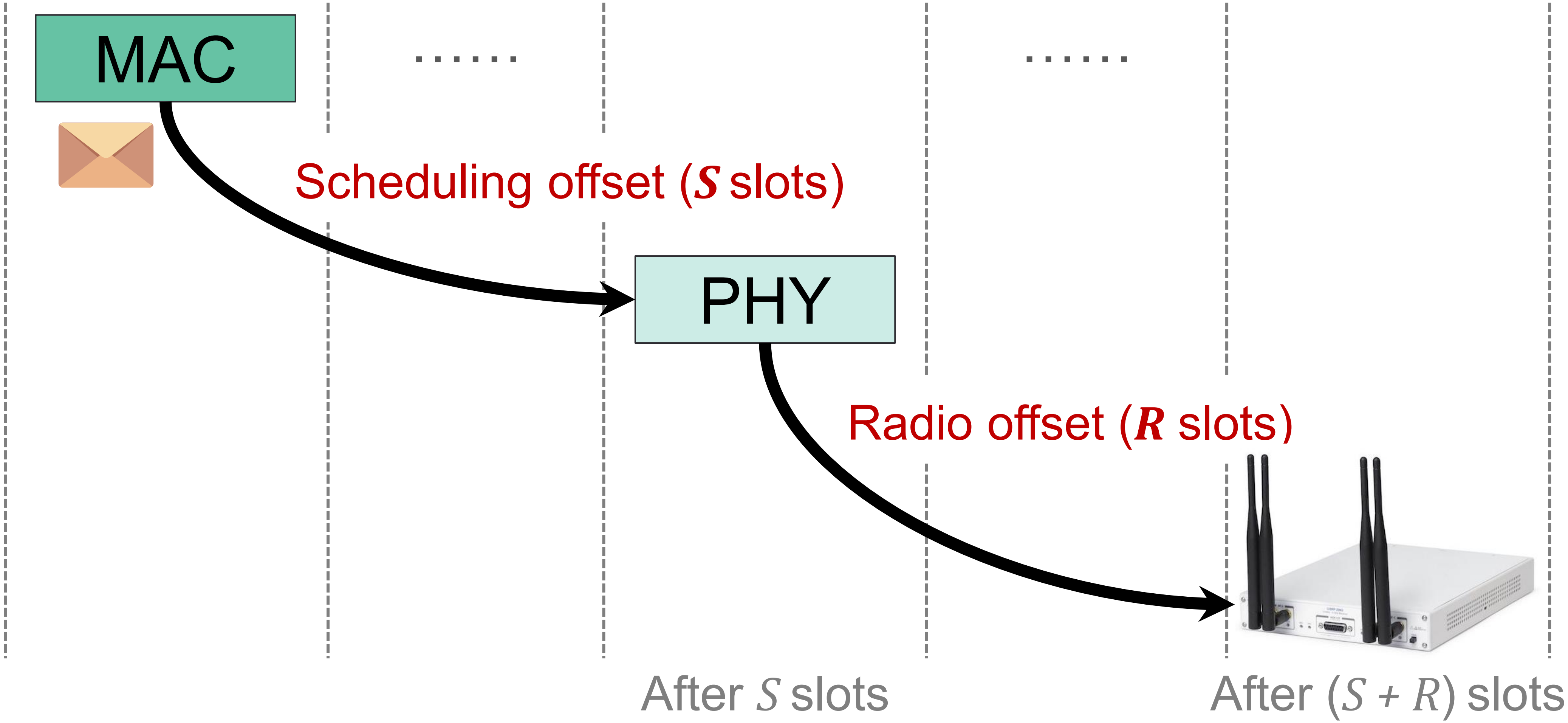
Default : R = 3 slots



Give extra time for wired transmission (e.g., USB) and radio processing

Latency Analysis

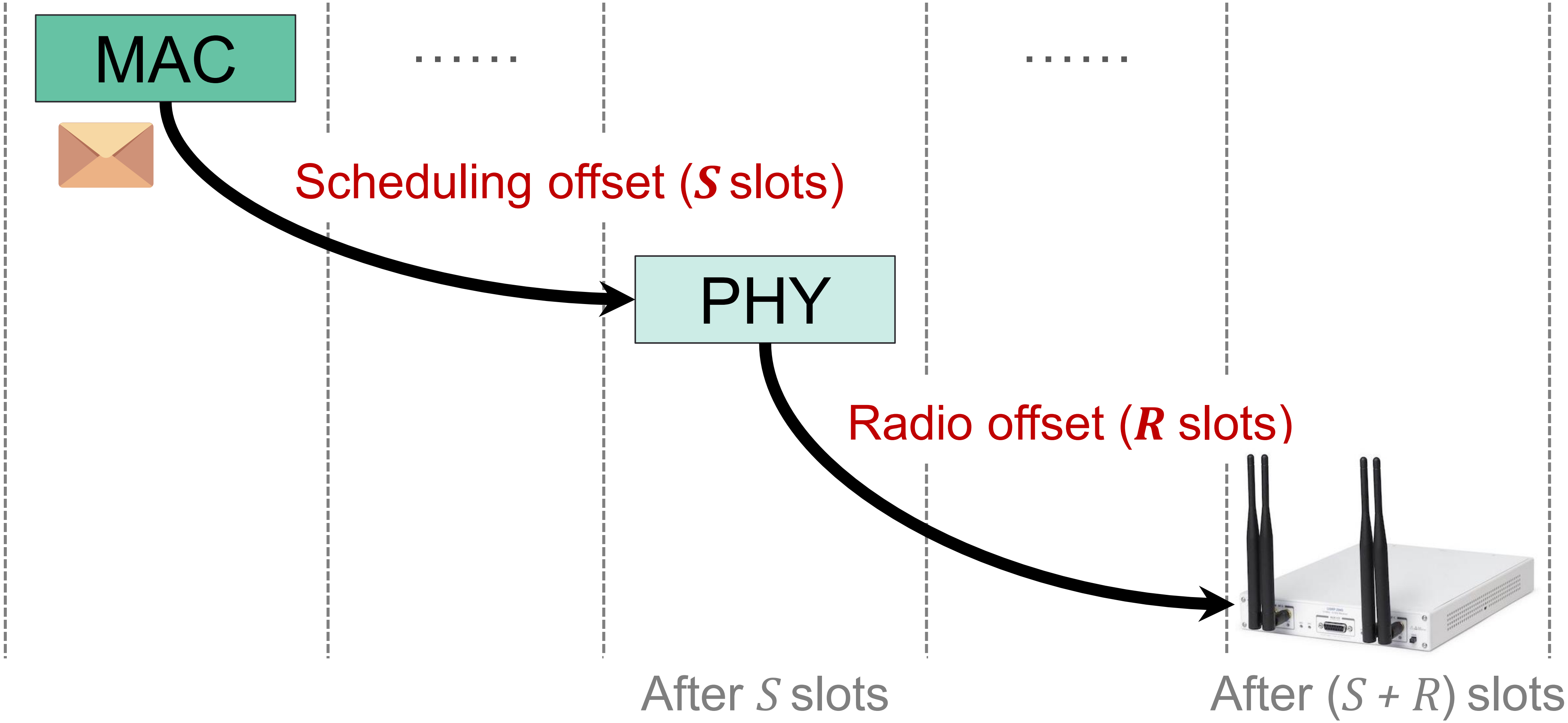
Implementation-Level Latency



- **Slot offsets → Significant latency ($S + R$ = 4 slots = 2 ms)**

Latency Analysis

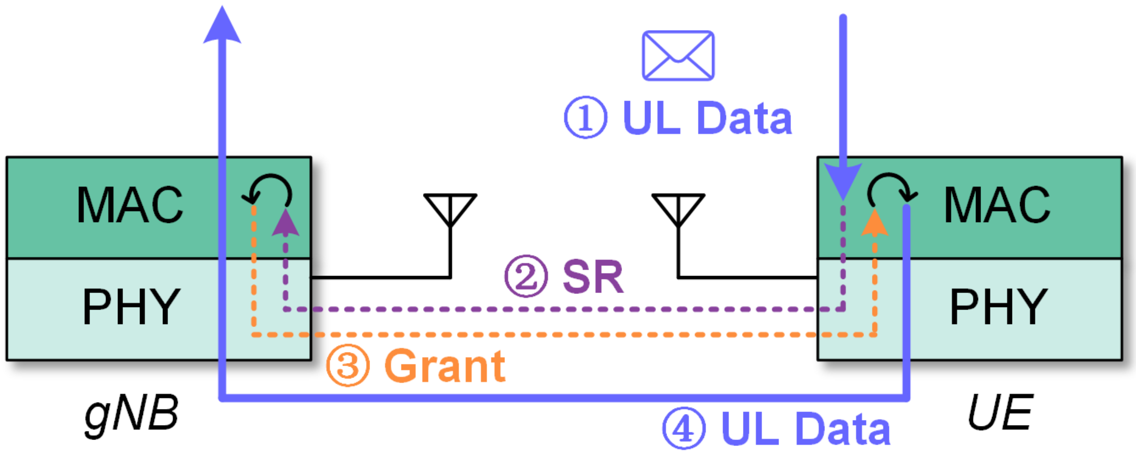
Implementation-Level Latency



Also affect uplink scheduling

Latency Analysis

Specification-Level Latency

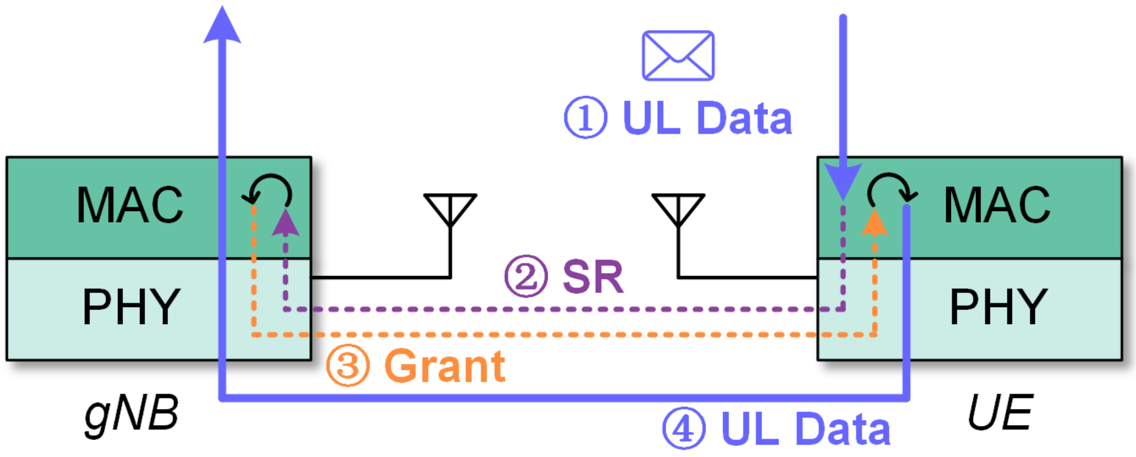


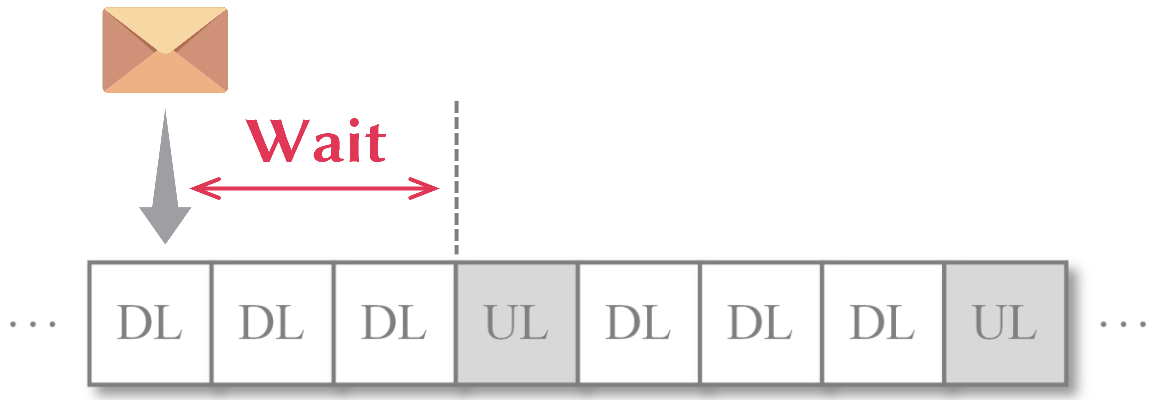
Grant-based access

Send Scheduling Request (SR)

Latency Analysis

Specification-Level Latency





Grant-based access

Time Division Duplex (TDD)

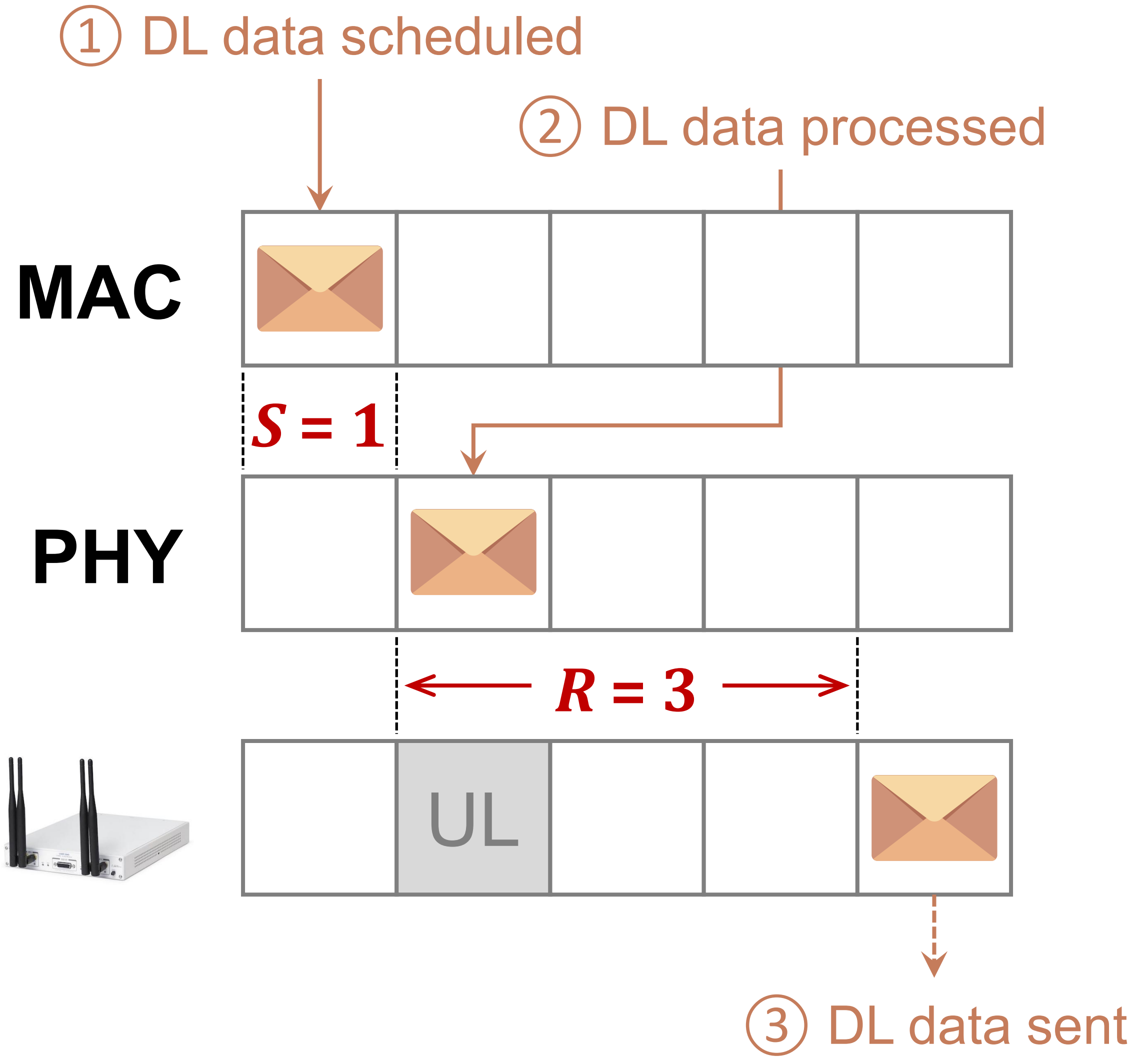
⚠ **Send Scheduling Request (SR)**

⚠ **Wait during downlink slots**

Latency Analysis

Latency Breakdown (Downlink Scheduling)

Numerology **1** / 1 slot = 0.5 ms

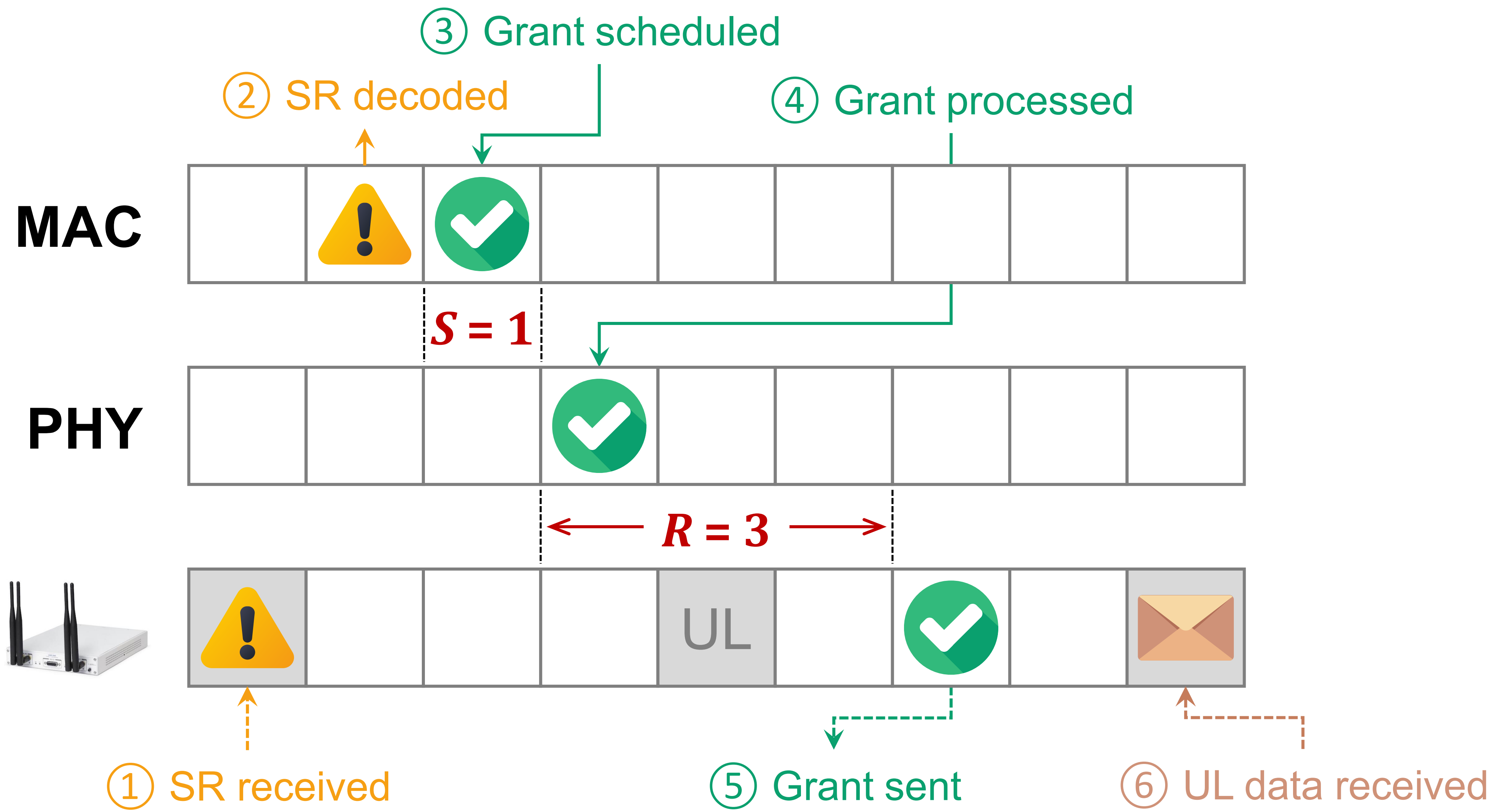


Steps ① → ③

DL > 5 slots = 2.5 ms

Latency Analysis

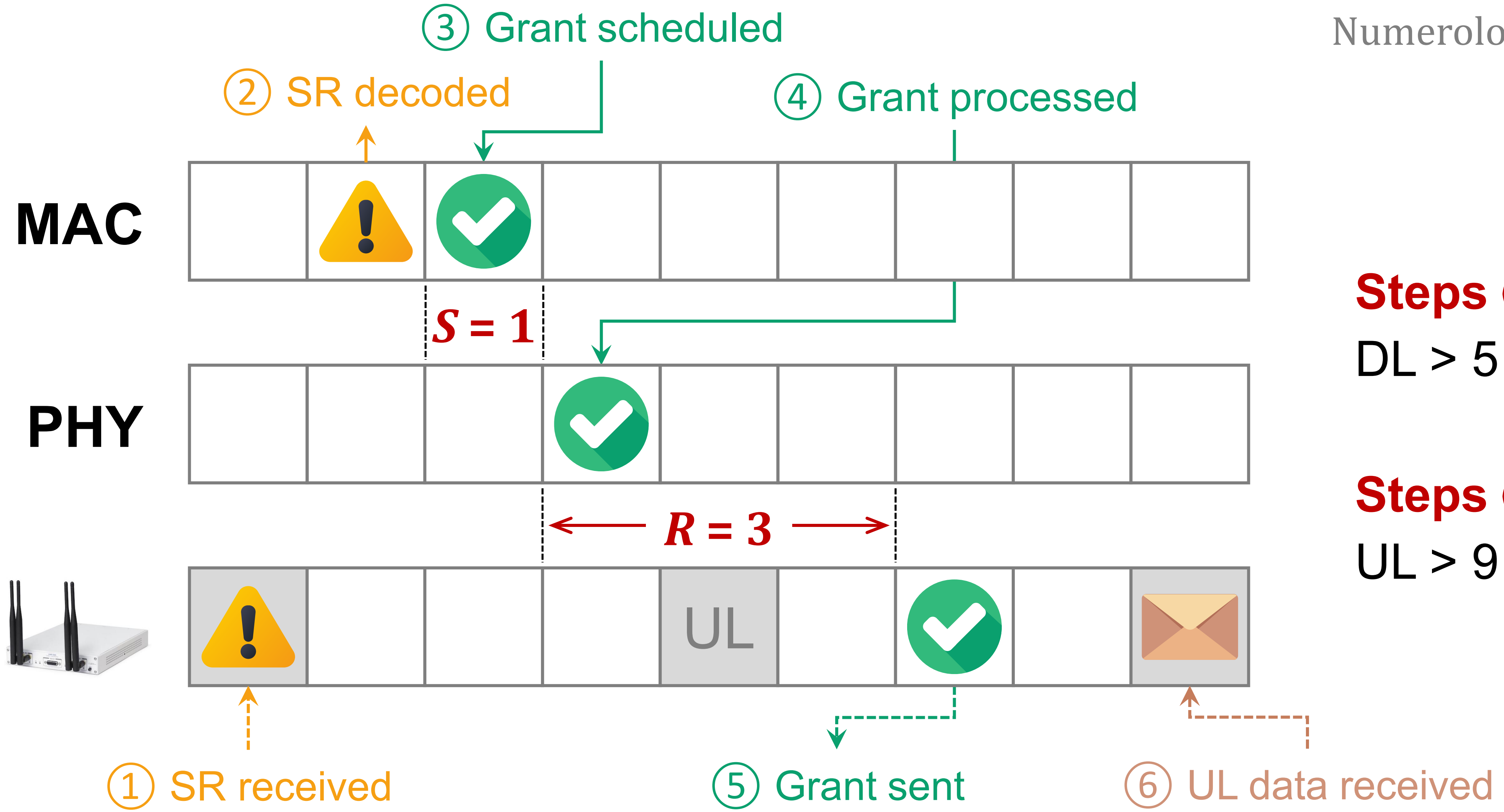
Latency Breakdown (Uplink Scheduling)



Latency Analysis

Latency Lower Bounds

Numerology **1** / 1 slot = 0.5 ms



Steps ③ → ⑤

DL > 5 slots = 2.5 ms

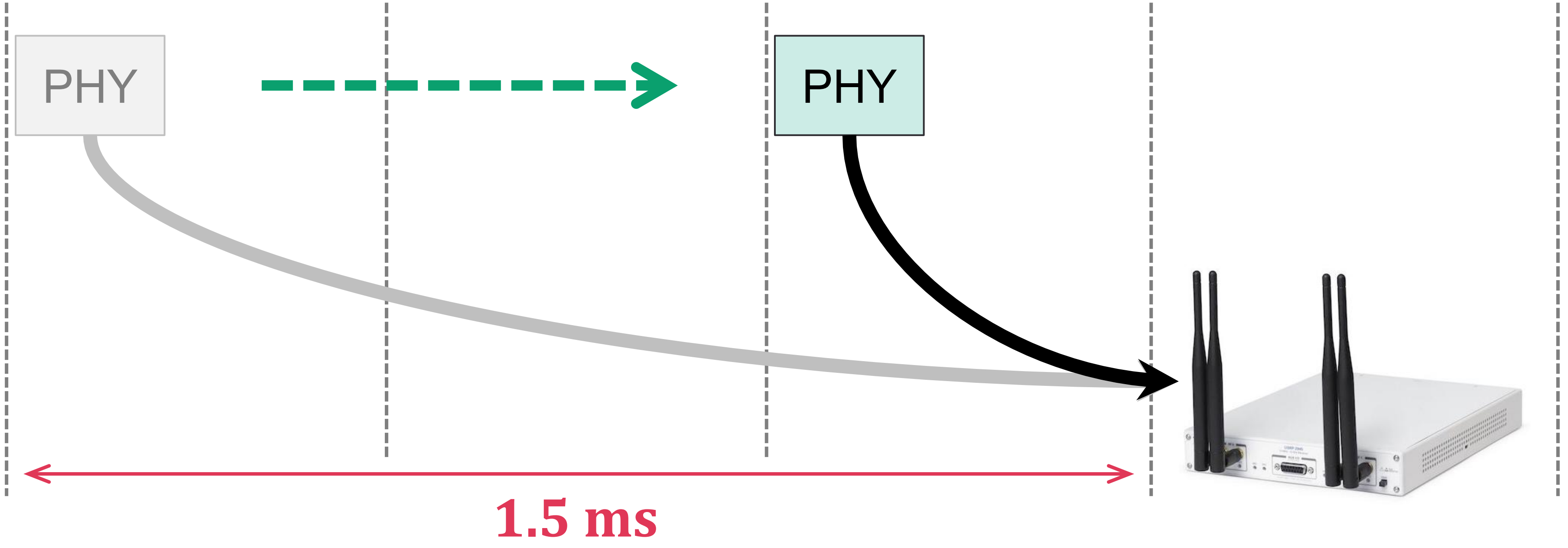
Steps ① → ⑥

UL > 9 slots = 4.5 ms

Latency Improvements

Improvement 1 (I1) : Reduce Radio Offset (R)

Minimize **radio offset** to the lowest safe value



R = 3 (1.5 ms)

R = 1 (0.5 ms)

Latency Improvements

Improvement 2 (I2) : Reduce Scheduling Offset (S)

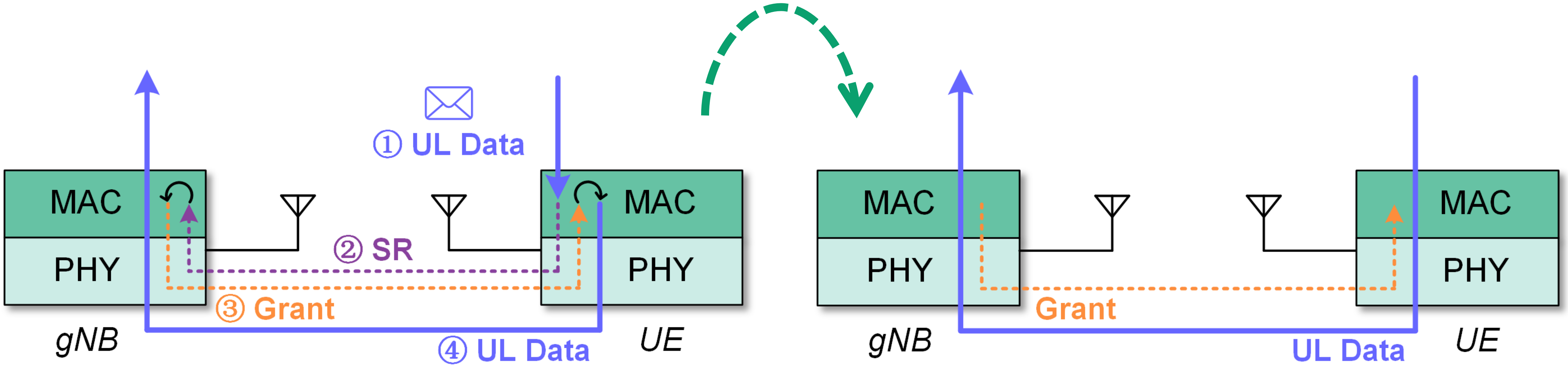
Remove **scheduling offset** by letting MAC trigger PHY directly

S = 1 (0.5 ms)

S = 0 (0 ms)

Improvement 3 (I3) : Scheduling-Request-Free Access

Send grants before every uplink slot

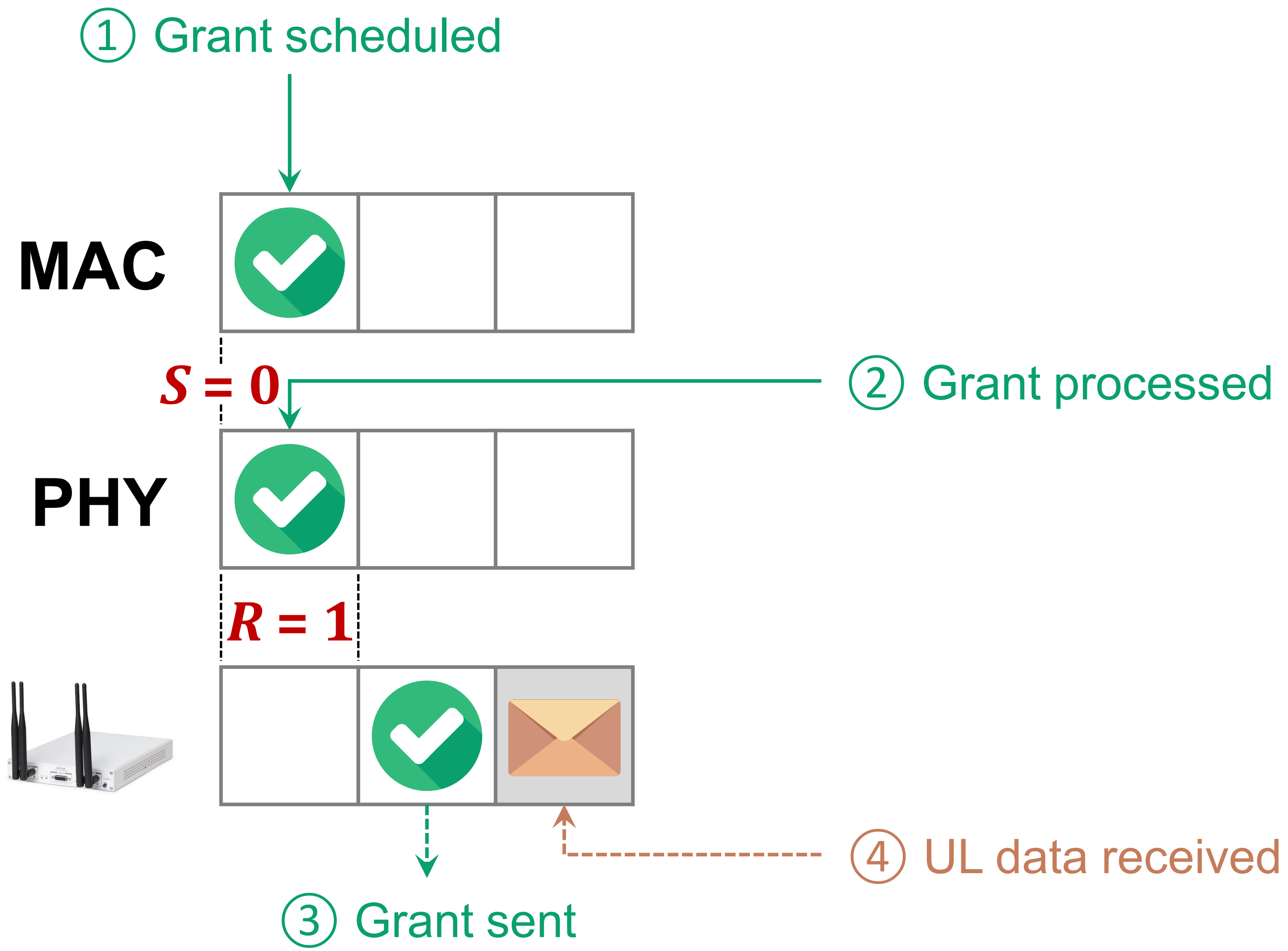


Grant-based access

SR-free access

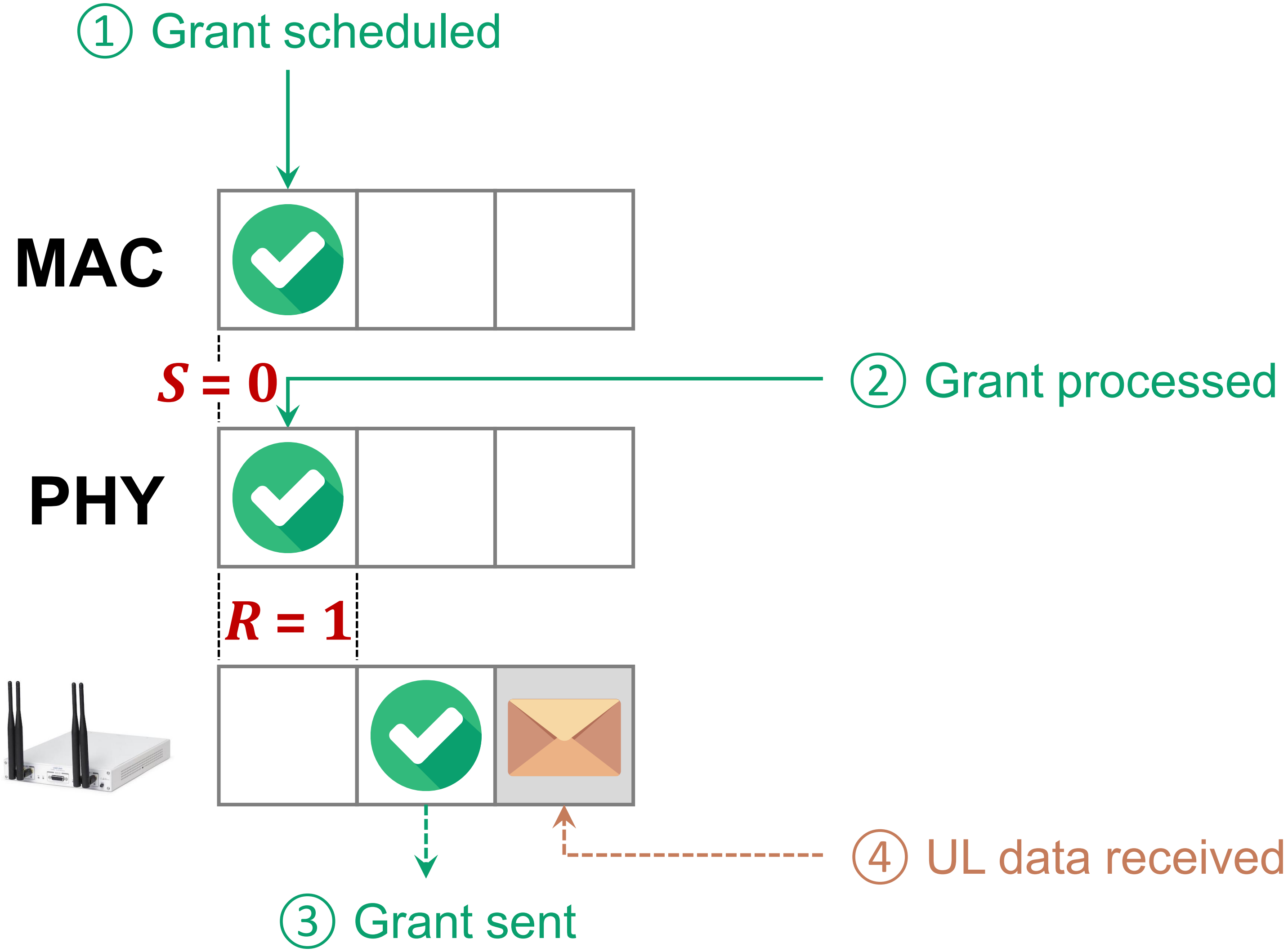
Latency Improvements

Latency Breakdown (I1 + I2 + I3)



Latency Improvements

Latency Lower Bounds (I1 + I2 + I3)



Steps ① → ③

DL > 2 slots = 1 ms

Reduced by 1.5 ms

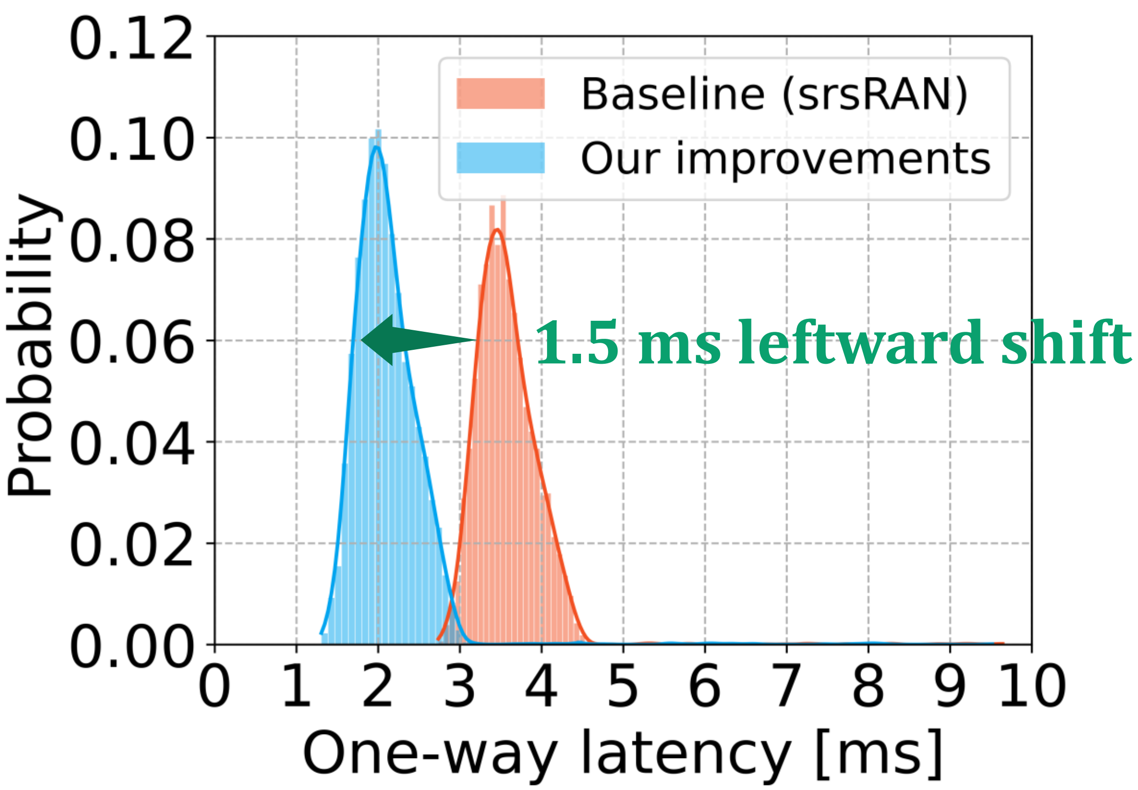
Step ④

UL > 1 slot = 0.5 ms

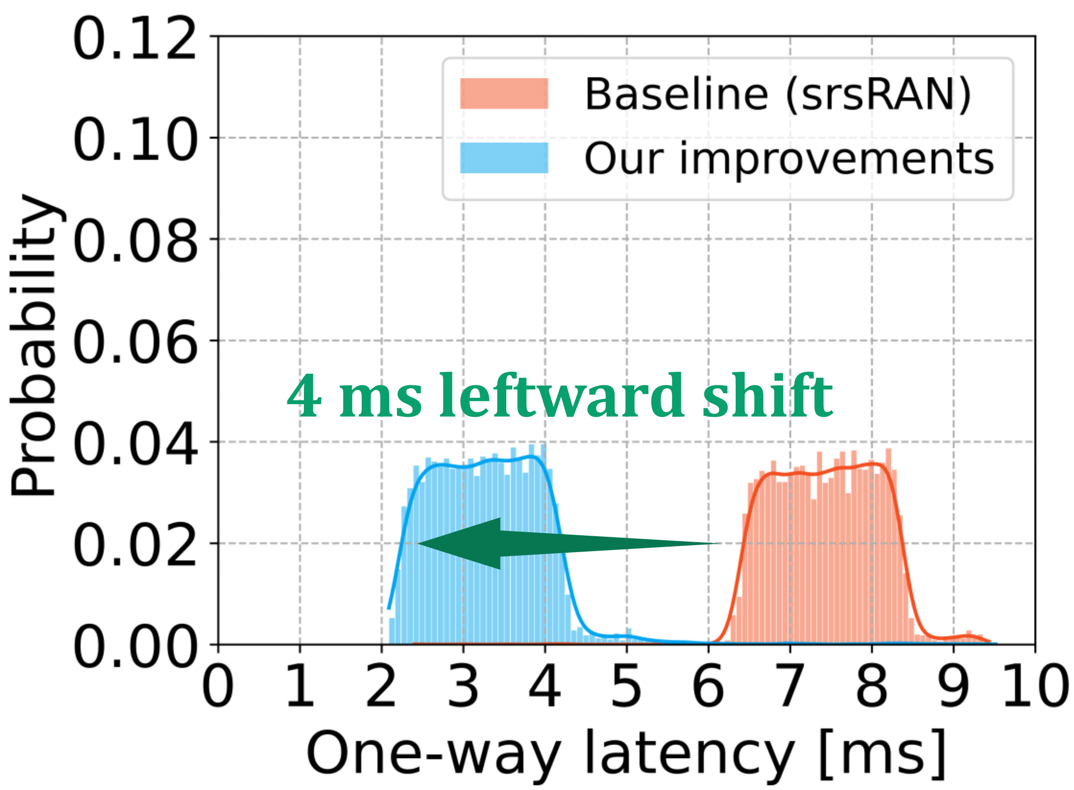
Reduced by 4 ms

Evaluation

Comparing Latency Distributions



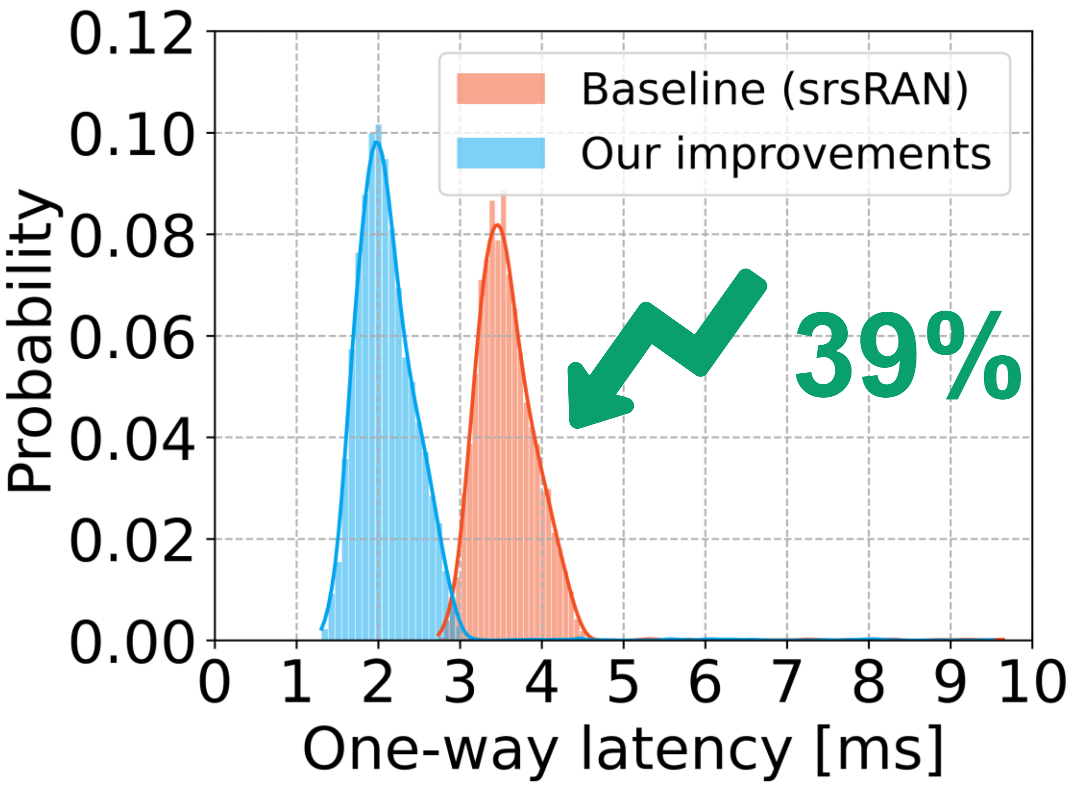
Downlink latency

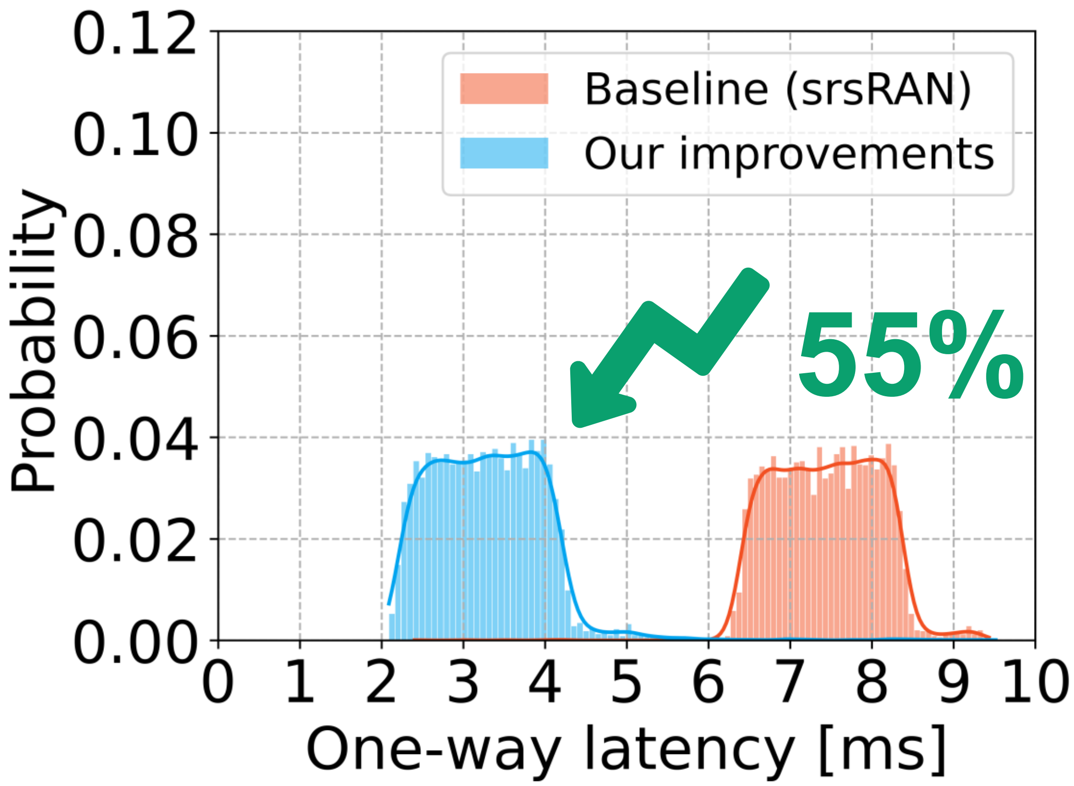


Uplink latency

Evaluation

Comparing Average Latency



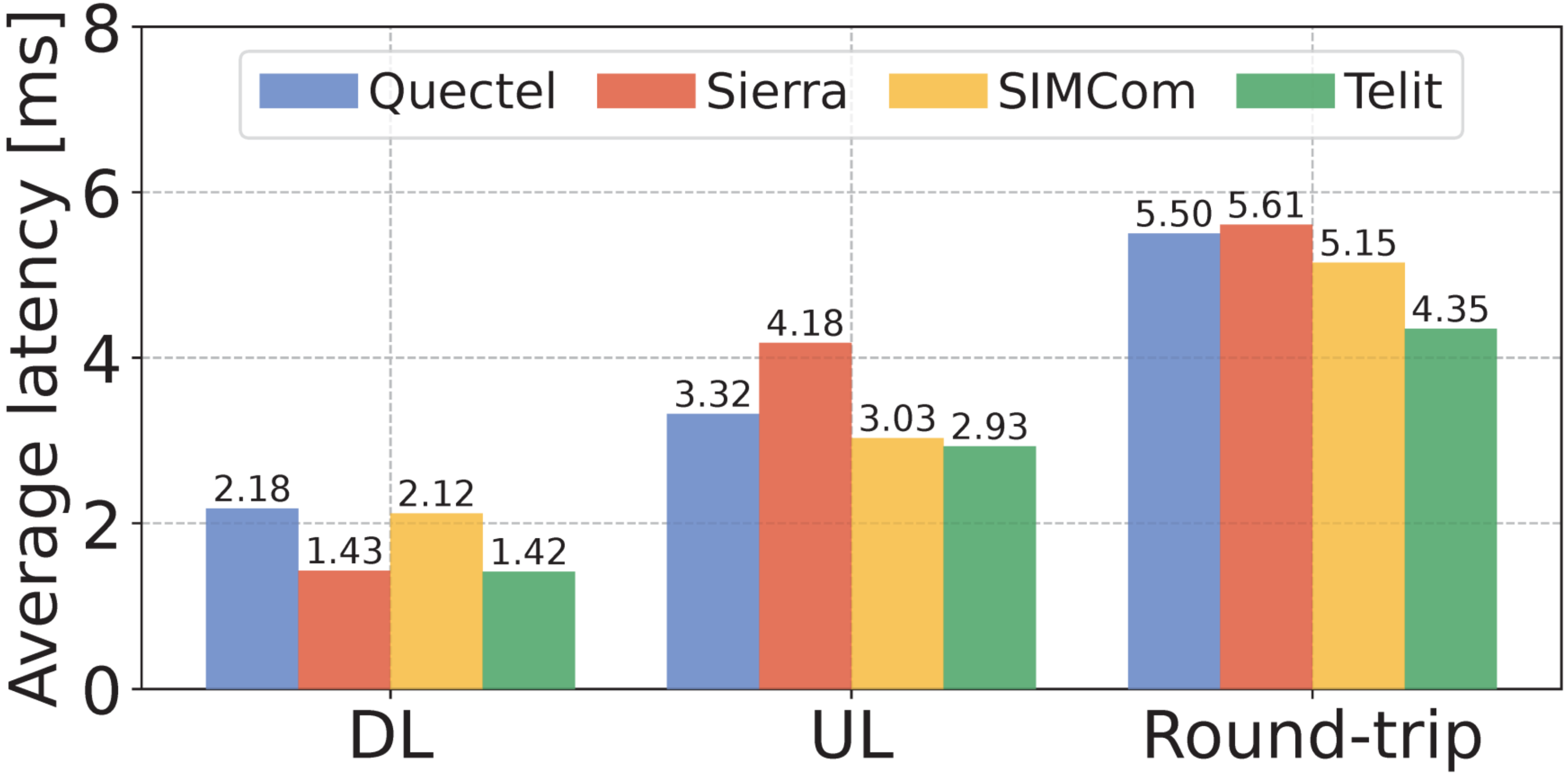


DL: 3.6 ms → 2.2 ms

UL: 7.4 ms → 3.3 ms

Evaluation

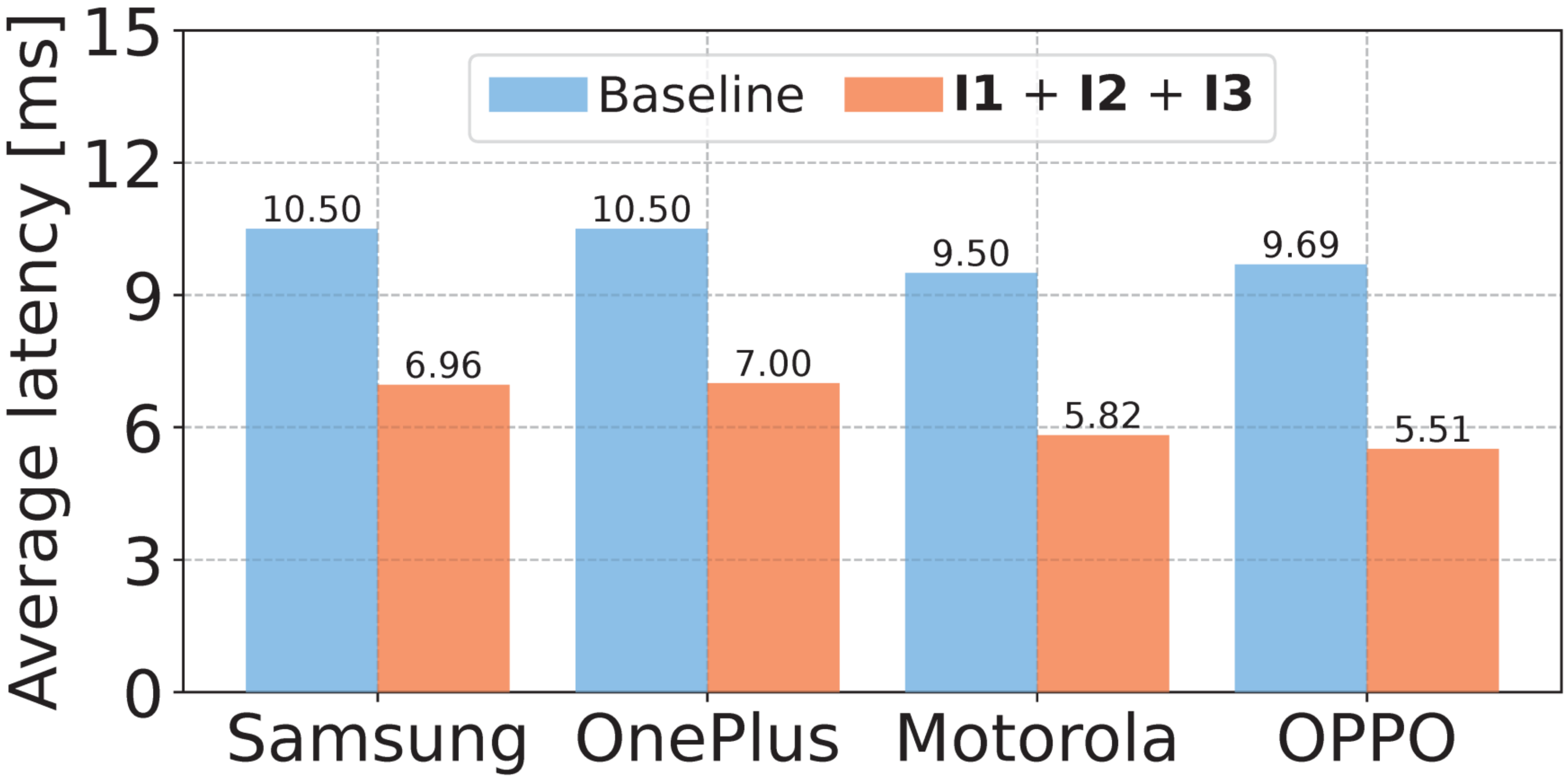
Evaluating Commercial 5G Modules



Consistently deliver low-latency performance across diverse 5G modules

Evaluation

Evaluating Commercial 5G Phones



Demonstrate robustness and scalability in a realistic multi-UE setup

Explore More!

Visit our GitHub Repo

srsRAN_Project_Low_Latency