Device-Based LTE Latency Reduction at the Application Layer

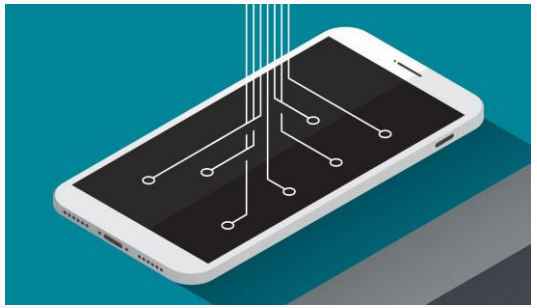
Zhaowei Tan, Jinghao Zhao, Yuanjie Li,
Yifei Xu, Songwu Lu

UCLA WiNG Wireless Networking Group

Latency-Sensitive Mobile Applications

Mobile VR

Mobile Gaming



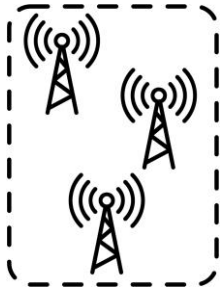
Mobile Sensing

- Emerging mobile apps have stringent latency requirement
- Small, frequent, yet regular uplink data

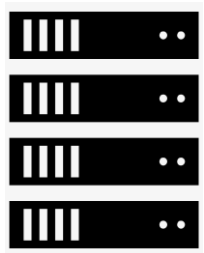
Mobile Apps in 4G LTE

These apps typically run on 4G LTE networks, the only large-scale infrastructure for “anywhere, anytime” Internet services

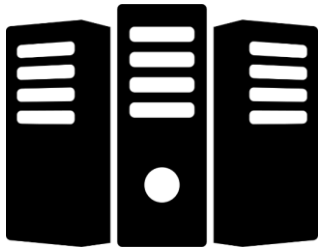
Phone Devices



Base stations



Core Network

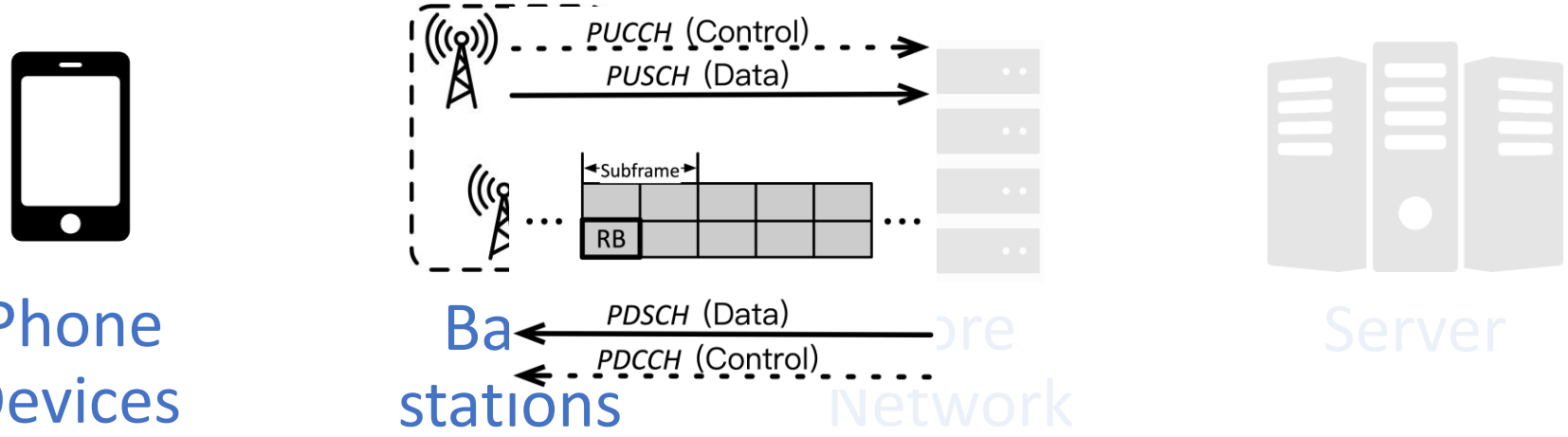


Server

Latency at Access Network

This work: Reducing LTE access latency for every data packet at the active state, without root privilege

- Complementary to app-specific optimization and control plane latency reduction
- Application-layer solution without root or hardware/firmware change



Outline

This work: Reducing LTE access latency for every data packet at the active state, without root privilege

This talk:

1. What incurs long latency for mobile apps?
2. How to reduce the latency components?
3. Can we design a solution at the app layer?

1. Roadblocks for low latency

Methodology for Analysis

Analyze traces from a showcase VR application and PUBG

Trace Collection

10-month empirical study
4 US mobile carriers

Data & Standard Analysis

Trace analysis with MobileInsight
3GPP standard compliant

VR Game

Who is the Bottleneck?

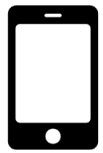
Intuition:

DL incurs high latency for the apps

Reality: Uplink latency is a major latency component

- 66-78% of the app network latency is from UL
- Sufficient bandwidth with new PHY technologies

Why High Uplink Latency?



Packet arrives in buffer

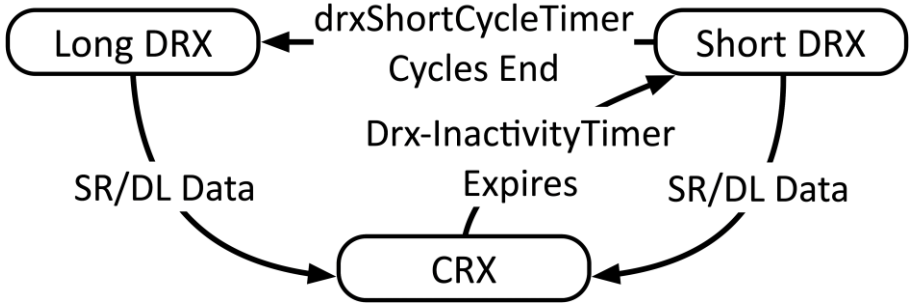
DRX Related

T_{drx_doze}

DRX OFF

DRX ON

DRX (Discontinuous Reception)

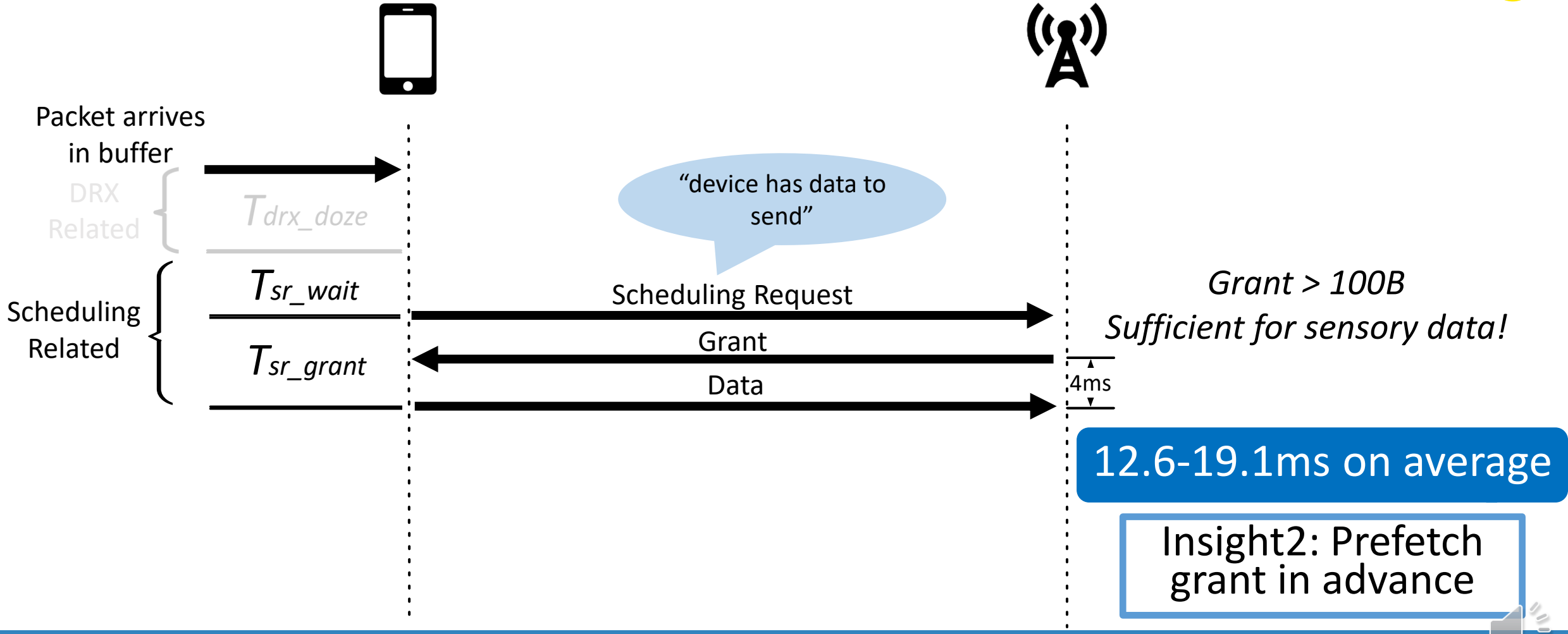


Different from DL DRX latency

28.3-31.9ms on average

Insight 1: Wake up DRX in advance

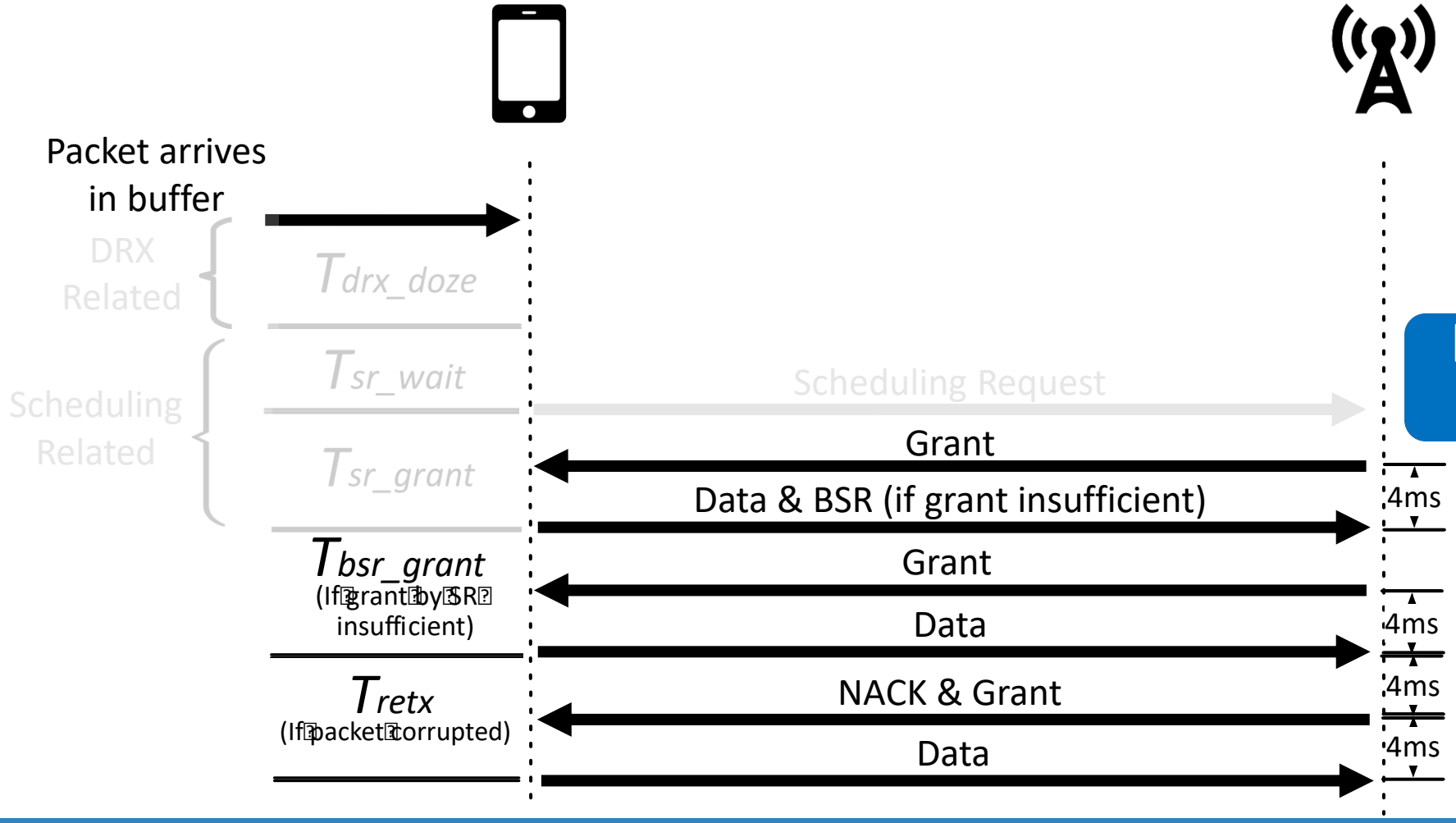
Why High Uplink Latency?



12.6-19.1ms on average

Insight2: Prefetch grant in advance

Why High Uplink Latency?



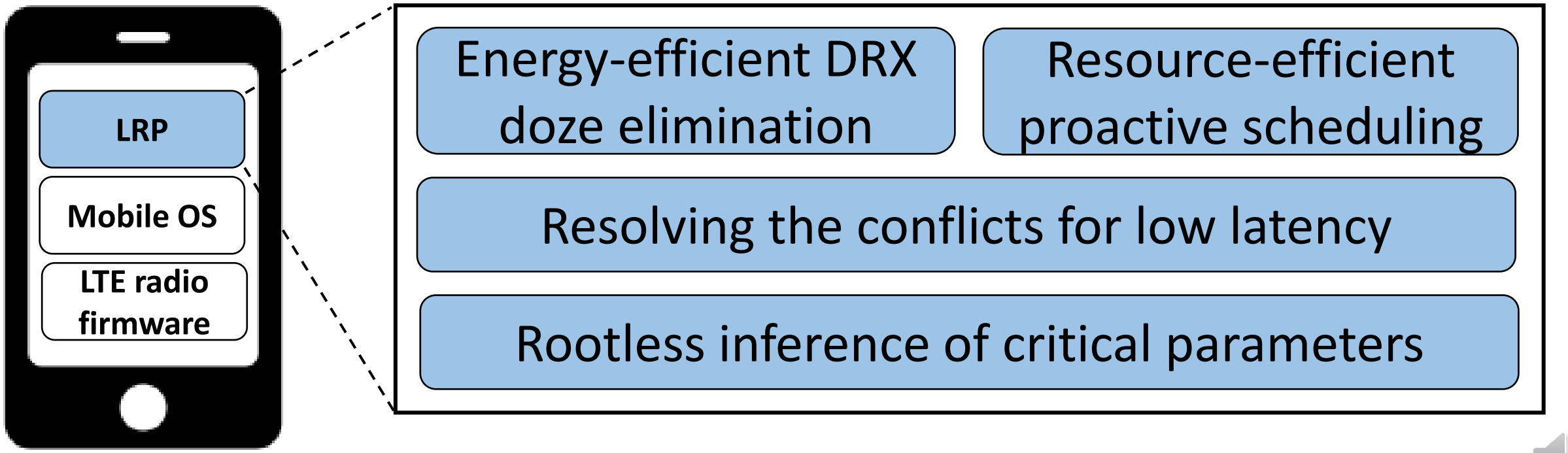
BSR and Retx Latency: <1ms on average

Insight 3: Negligible BSR and ReTx Latency

2. Solution to latency reduction

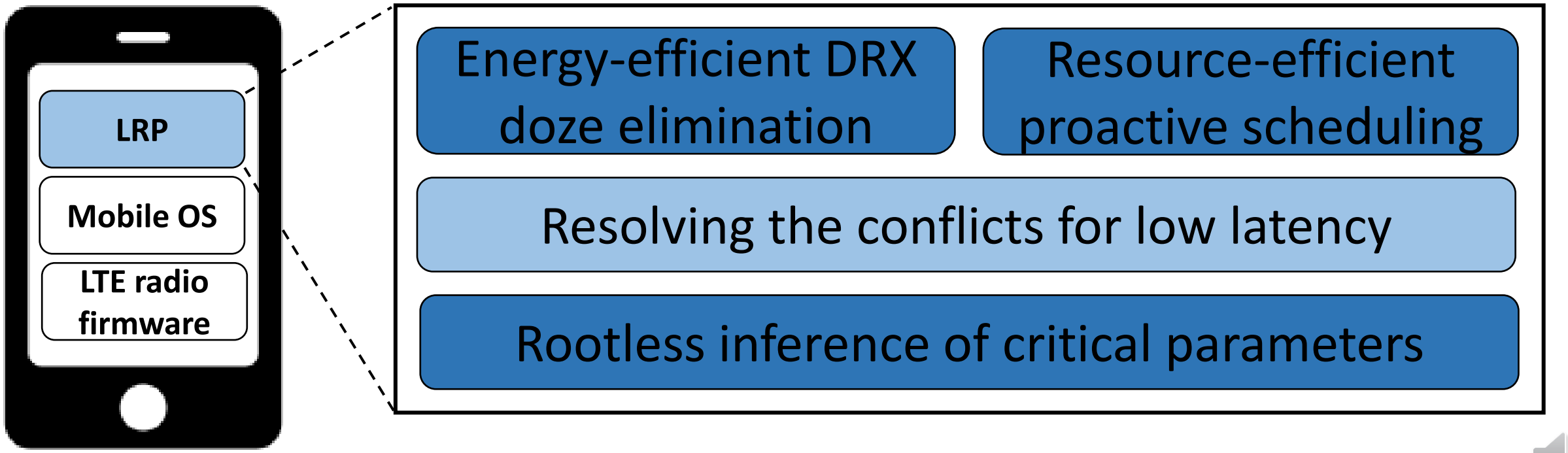
LRP Overview

In-device, rootless software solution to
4G/5G latency reduction for mobile apps



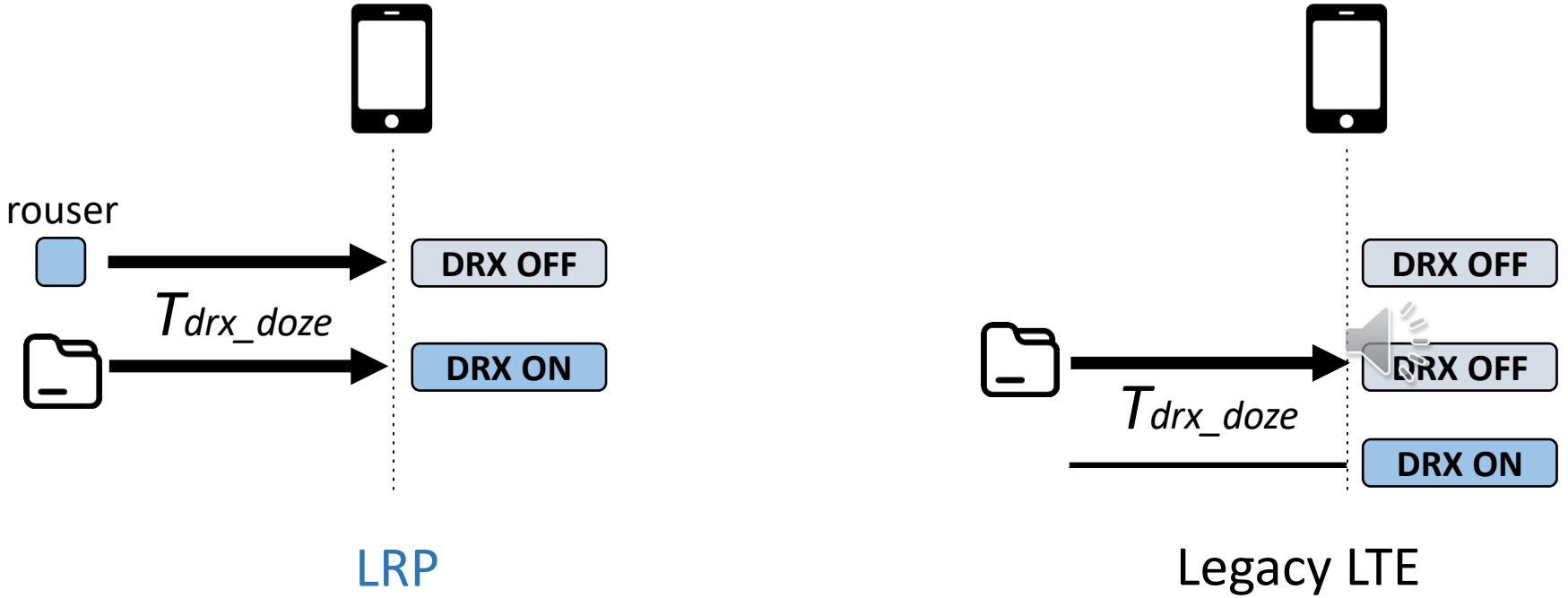
LRP Overview

In-device, rootless software solution to
4G/5G latency reduction for mobile apps



Energy-Efficient DRX Doze Elimination

- Idea: Send a rouser in advance
  - Data packet arrives at DRX ON → no doze latency



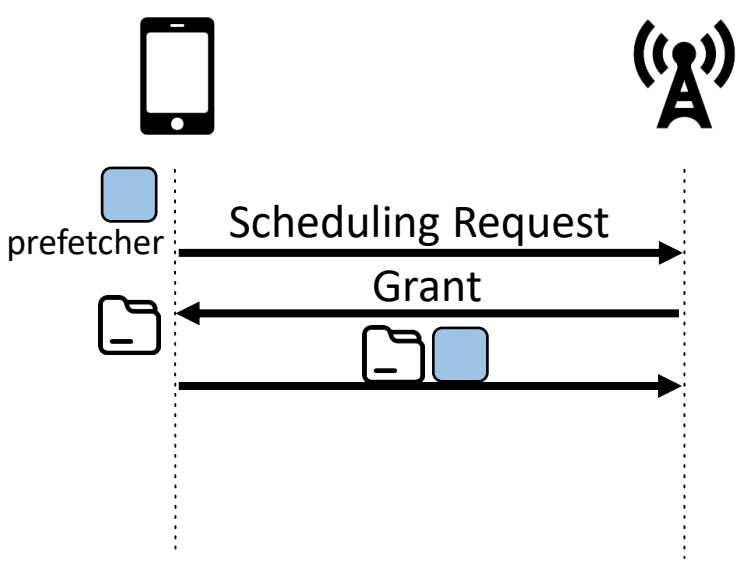
Energy-Efficient DRX Doze Elimination

- Idea: Send a rouser in advance
  - Data packet arrives at DRX ON → no doze latency

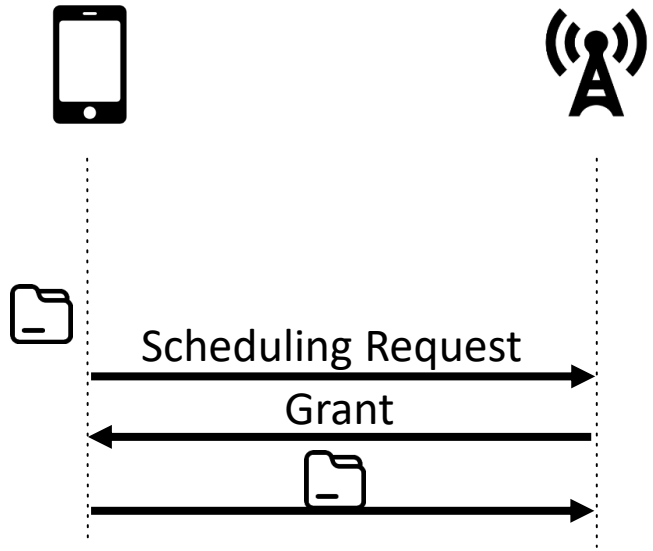
- Issue: Early rouser incurs high energy overhead
- Solution: Timing control for rouser
  - Send the rouser T_{drx_doze} before data packet
  - Data periodicity makes this possible; need to know T_{drx_doze}

Resource-Efficient Proactive Scheduling

- Idea: Send a prefetcher in advance
  - Prefetcher triggers SR and asks for a grant



LRP



Legacy LTE

Resource-Efficient Proactive Scheduling

- Idea: Send a prefetcher in advance
  - Prefetcher triggers SR and asks for a grant
- Issue 1: An early/late prefetcher misses latency reduction or wastes requested resource
- Solution: Timing control on prefetcher
  - Send the rouser T_{sr_grant} before data packet
- Issue 2: Insufficient grant for prefetcher + data
  - Rare occurrence and limited impact

3. Inferring key parameters from the application layer

Rootless Inference of Critical Parameters

- Design LRP as a software daemon *without* root privilege
  - <10% of mobile devices are rooted

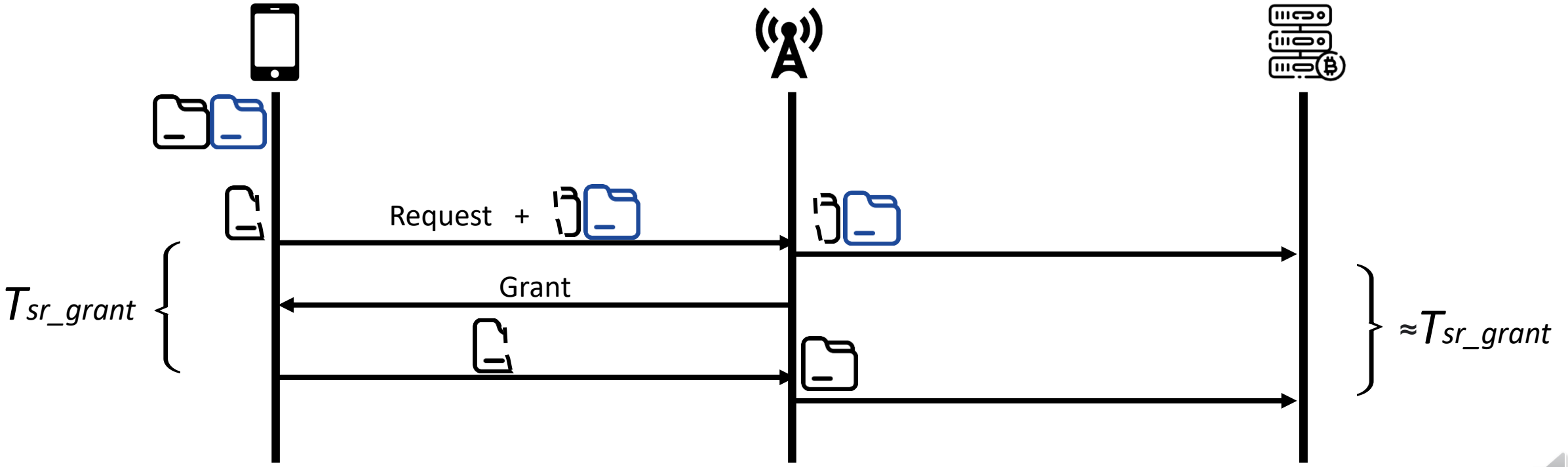
Challenge: Infer LTE-specific parameters;

How to single out access network latency?

Solution idea: Send a pair of controlled packets and observe their receiving interval

Rootless Inference of Critical Parameters

Solution idea: Send a pair of controlled packets and observe their receiving interval



Rootless Inference of Critical Parameters

Solution idea: Send a pair of controlled packets and observe their receiving interval

Alt: Send DNS requests and measure interval of responses

Premise: The receiving interval dominated by UL LTE

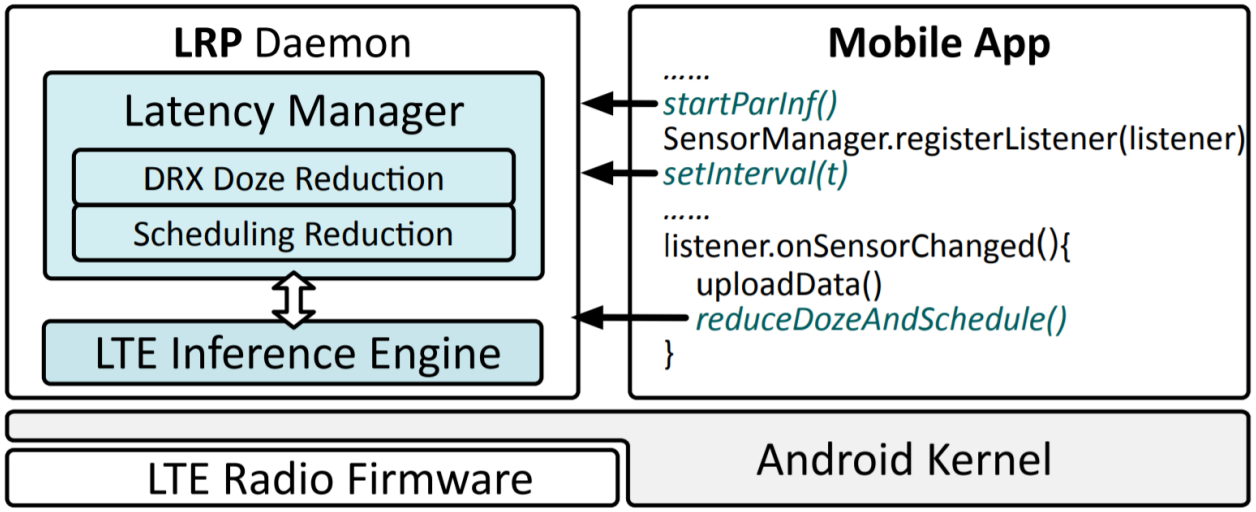
- Core network and LTE DL: Little impact on the packet pair interval

Discussion on LRP Design

- LRP applies to 5G
- LRP still helps reduce latency, if
  - there is background traffic
  - the packet arrival is not strictly regular or predictable
- LRP has little impact on the network side
- LRP does not affect those not using LRP

Implementation

- Implement LRP on Android
  - Work as a standalone user-space daemon
  - Provide APIs for the applications



Evaluation

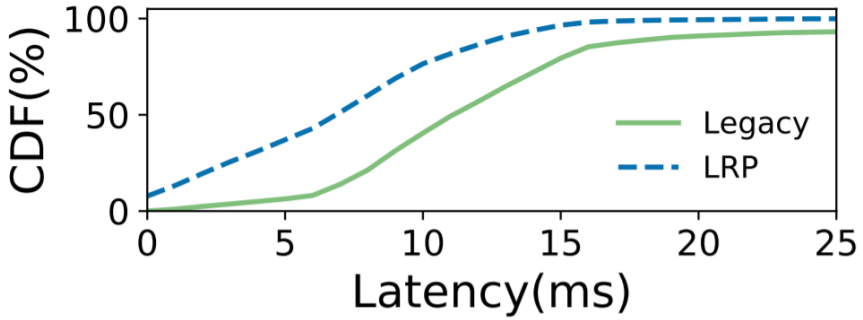
- Can LRP reduce latency for mobile apps?
- How much overhead does LRP occur?
- Can LRP benefit apps in 5G?

Latency Reduction for Mobile Apps

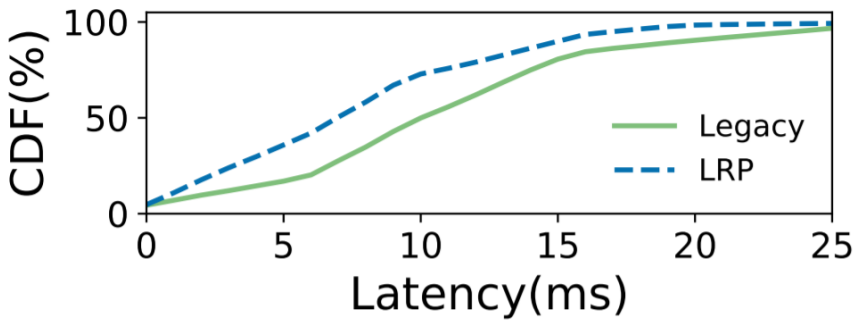
- Cover 375 cells, 5 operators, and 2 countries
- 4 mobile applications
- Experiments under different mobility

0.3-7.4x median LTE network latency reduction
Up to 3.5x 95-th percentile LTE network latency reduction

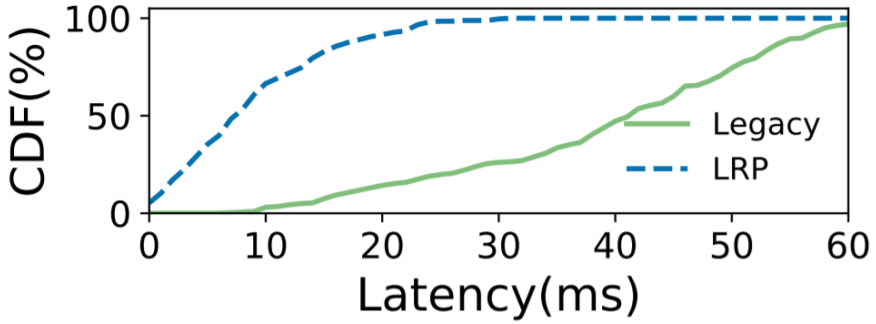
Latency Reduction for Mobile Apps



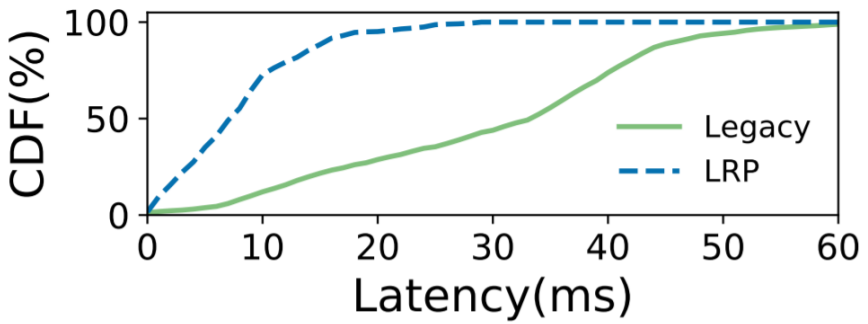
(a) Mobile VR



(b) Gaming



(c) Localization



(d) Object Detection

Latency Reduction: VR Demo

Micro-Benchmarks

- DRX doze Latency

  21-41ms median latency reduction
  40-57ms 95-th percentile latency reduction

- Scheduling latency

  0.3-2.5x median latency reduction
  Up to 1.7x 95-th percentile latency reduction

- Accuracy for rootless inference

  1.3-3.2% rootless inference error rate

LRP Overhead

Small overhead!

0.05-0.33KB/s data overhead
Up to 4.3% extra messages
1.0-2.5% extra battery cost

LRP exploits timing control to reduce its overhead

Can LRP be Used in 5G?

Yes!

- Evaluate LRP in AT&T 5G networks
  - Measure RTT without access to fine-grained logs

Reduce RTT by 4.3-20.5ms for apps

Summary

- LTE UL poses as a roadblock for latency-sensitive apps
- LRP: A rootless mobile app latency solution
  - Unveil and reduce LTE UL latency elements
  - Infer LTE parameters without root
- Applicable to both 4G and 5G

Thank you!

For code release:
http://metro.cs.ucla.edu/lrp.html