A Machine Learning Based Approach to Mobile Network Analysis

Zengwen Yuan[1], Yuanjie Li[1], Chunyi Peng[2], Songwu Lu[1], Haotian Deng[2], Zhaowei Tan[1], Taqi Raza[1]

[1] UCLA [2] Purdue University

Overview

Overview

Background

Overview

Background

Why machine learning for mobile network analysis

Overview

Background

Why machine learning for mobile network analysis

Mobile network analysis: state-of-the-art and our approach

Overview

Background

Why machine learning for mobile network analysis

Mobile network analysis: state-of-the-art and our approach

Case study: analyzing latency for mobile networks

- How mobile apps work over LTE
- How to breakdown app-perceived latency
- Challenges and ML scheme
- Primary results from crowdsourcing

Overview

Background

Why machine learning for mobile network analysis

Mobile network analysis: state-of-the-art and our approach

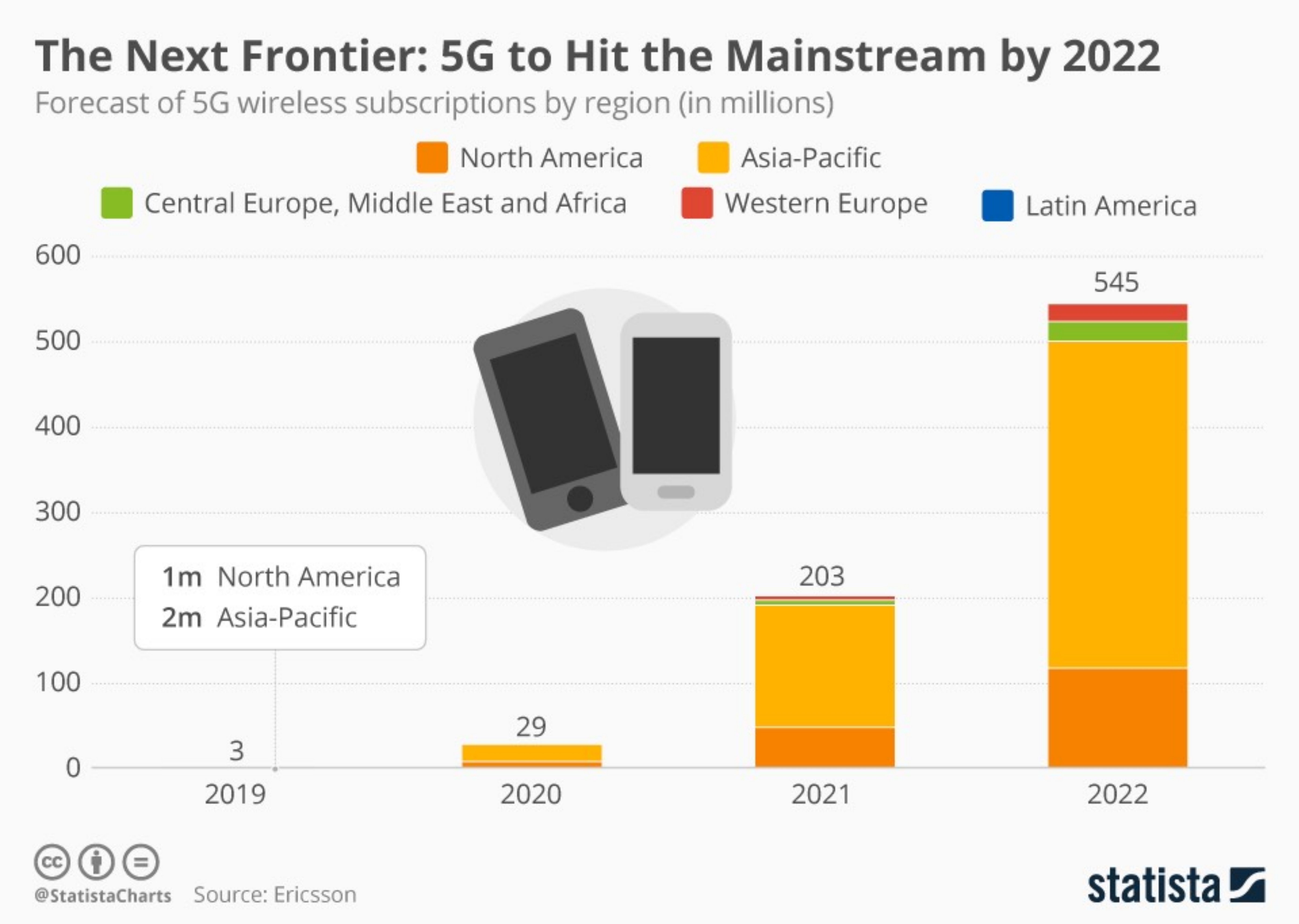
Case study: analyzing latency for mobile networks

- How mobile apps work over LTE
- How to breakdown app-perceived latency
- Challenges and ML scheme
- Primary results from crowdsourcing

Conclusion

Ubiquitous cellular networks connect everyone, everything

The race to 5G opens many new opportunities



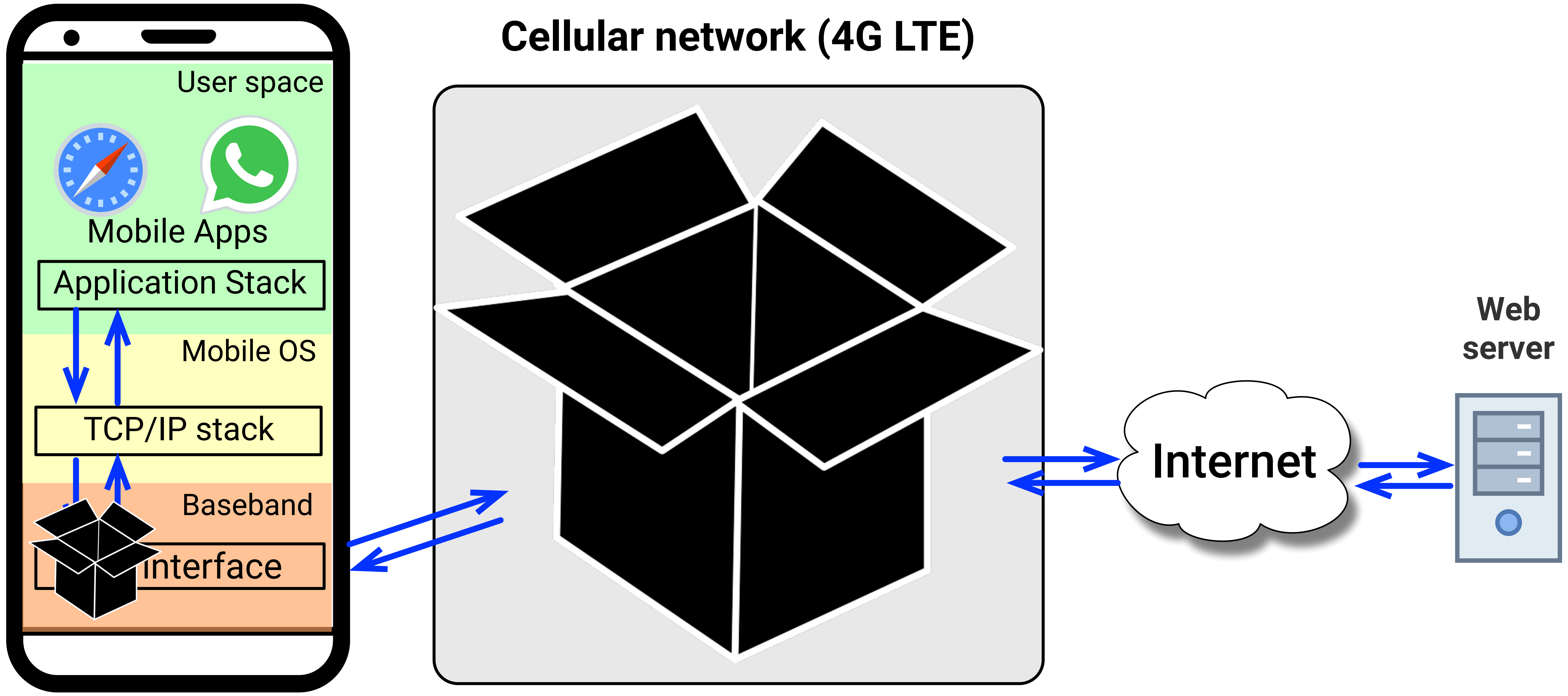
Yet, access to mobile network analytics is barred

What's going on in the 3G/4G/5G network??

Researcher
(you)

Chipset/Mobile OS

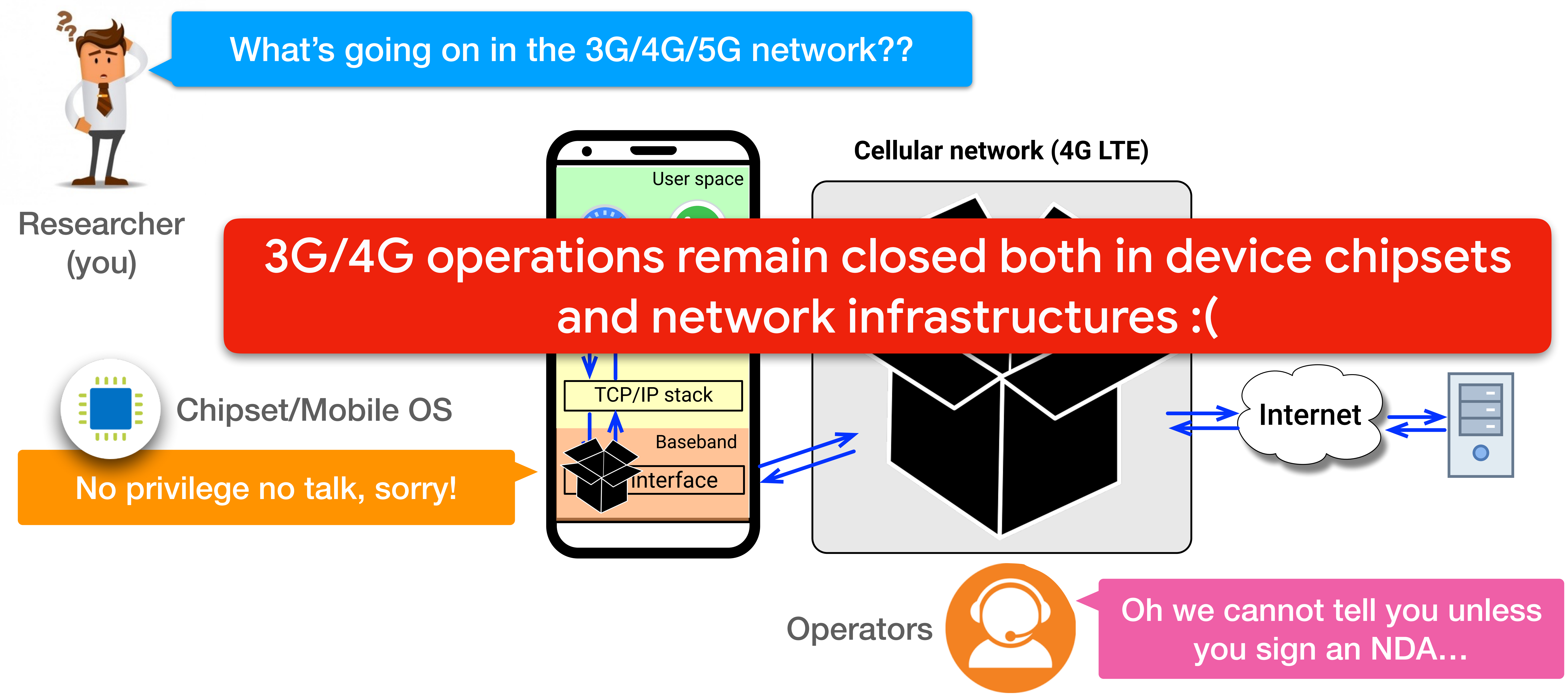
No privilege no talk, sorry!



Operators

Oh we cannot tell you unless you sign an NDA...

Yet, access to mobile network analytics is barred

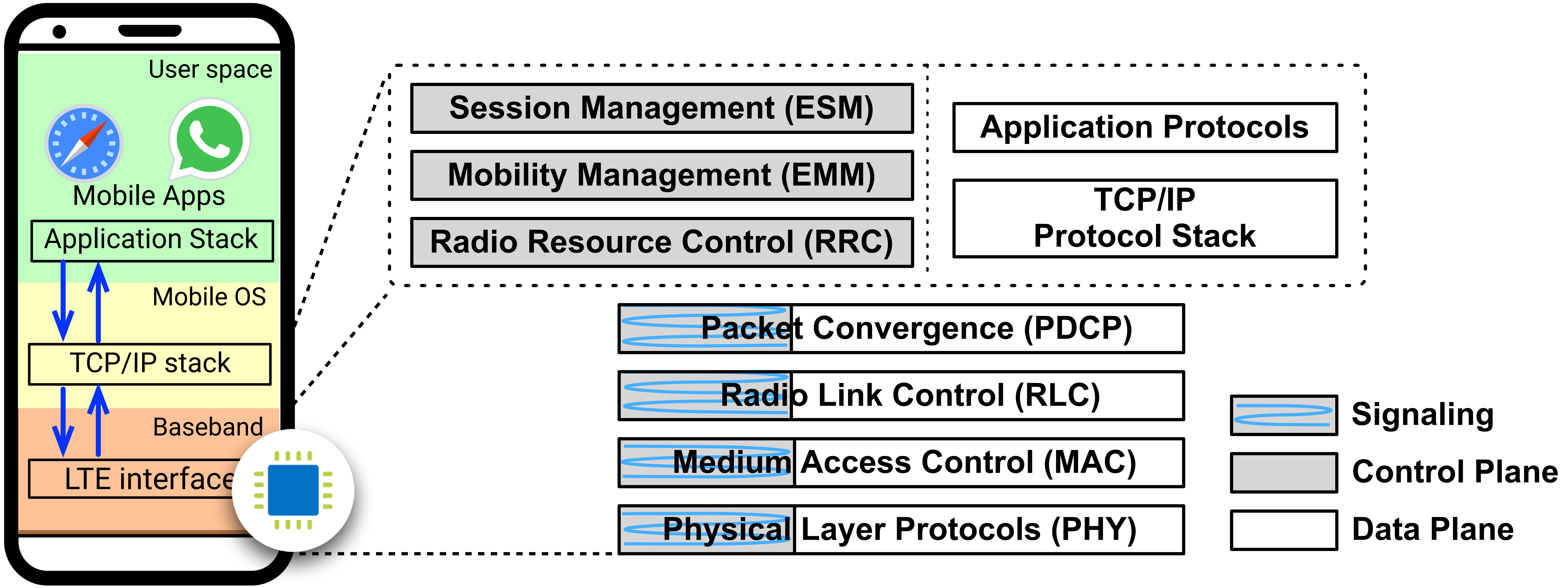


3G/4G operations remain closed both in device chipsets and network infrastructures :(

Plus, mobile networks are complex & distributed

More complex functions on both control and data planes

Operations are distributed across layers



Moreover, analytics tasks are app specific

Moreover, analytics tasks are app specific

Analytics for mobile networks is problem-specific, for example:

Moreover, analytics tasks are app specific

Analytics for mobile networks is problem-specific, for example:

- Web browsers:
  - ✦ Why the time-to-first-byte (TTFB) is so long?
  - ✦ What's the major component of latency?
  - ✦ …

Moreover, analytics tasks are app specific

Analytics for mobile networks is problem-specific, for example:

- Web browsers:
  - ✦ Why the time-to-first-byte (TTFB) is so long?
  - ✦ What's the major component of latency?
  - ✦ …
- Instant message apps:
  - ✦ Does the recipient read my message?
  - ✦ Is my message delivered in time?
  - ✦ …

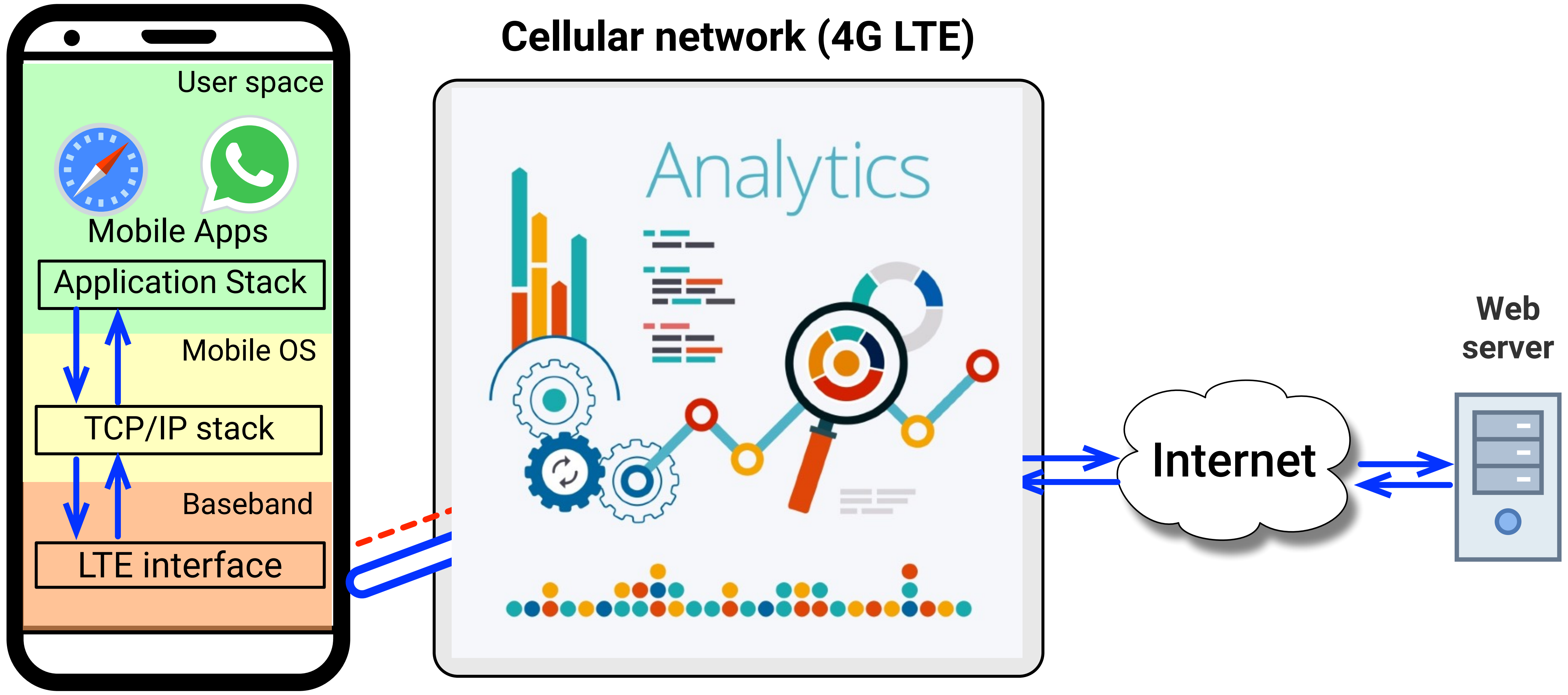
State-of-the-art mobile network analytics

State-of-the-art mobile network analytics

Current 4G network analytics is primarily "infrastructure-based":

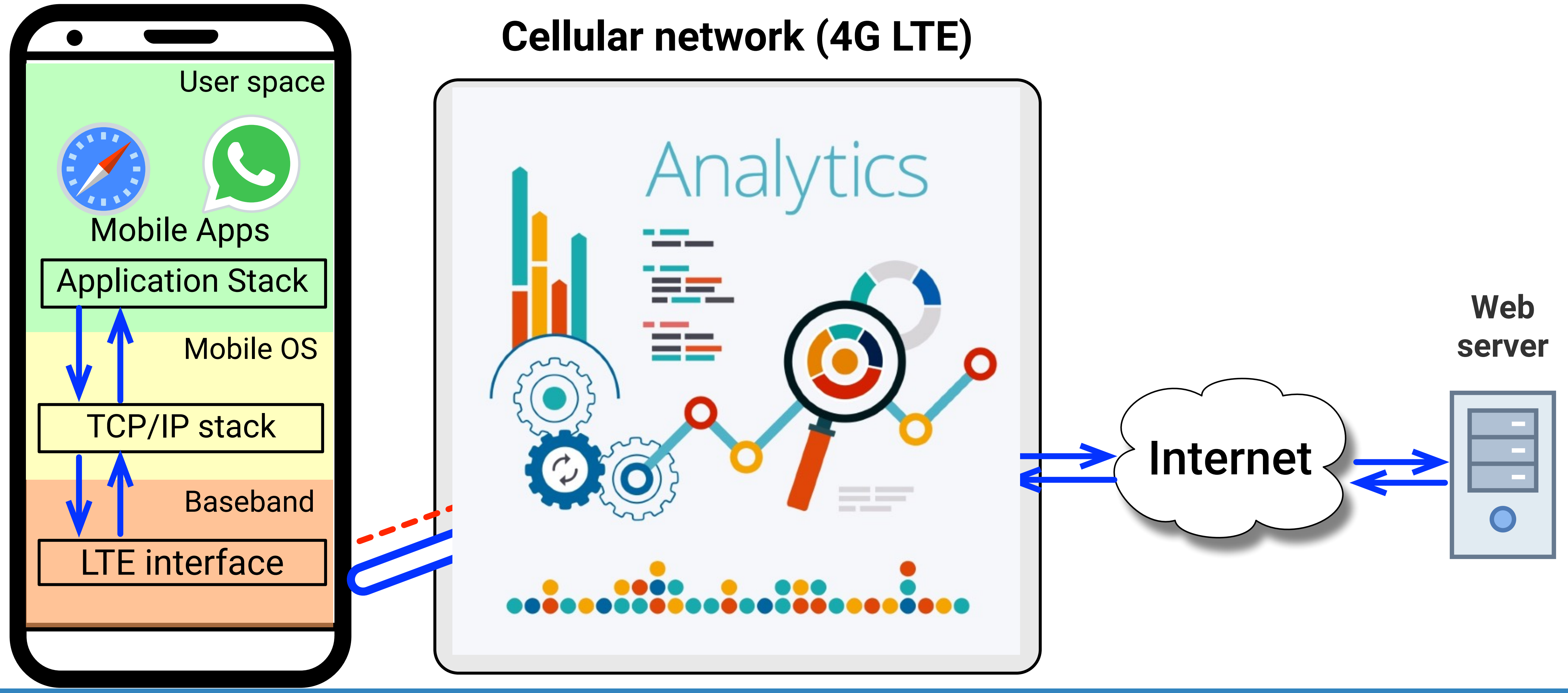
State-of-the-art mobile network analytics

Current 4G network analytics is primarily "infrastructure-based":



State-of-the-art mobile network analytics

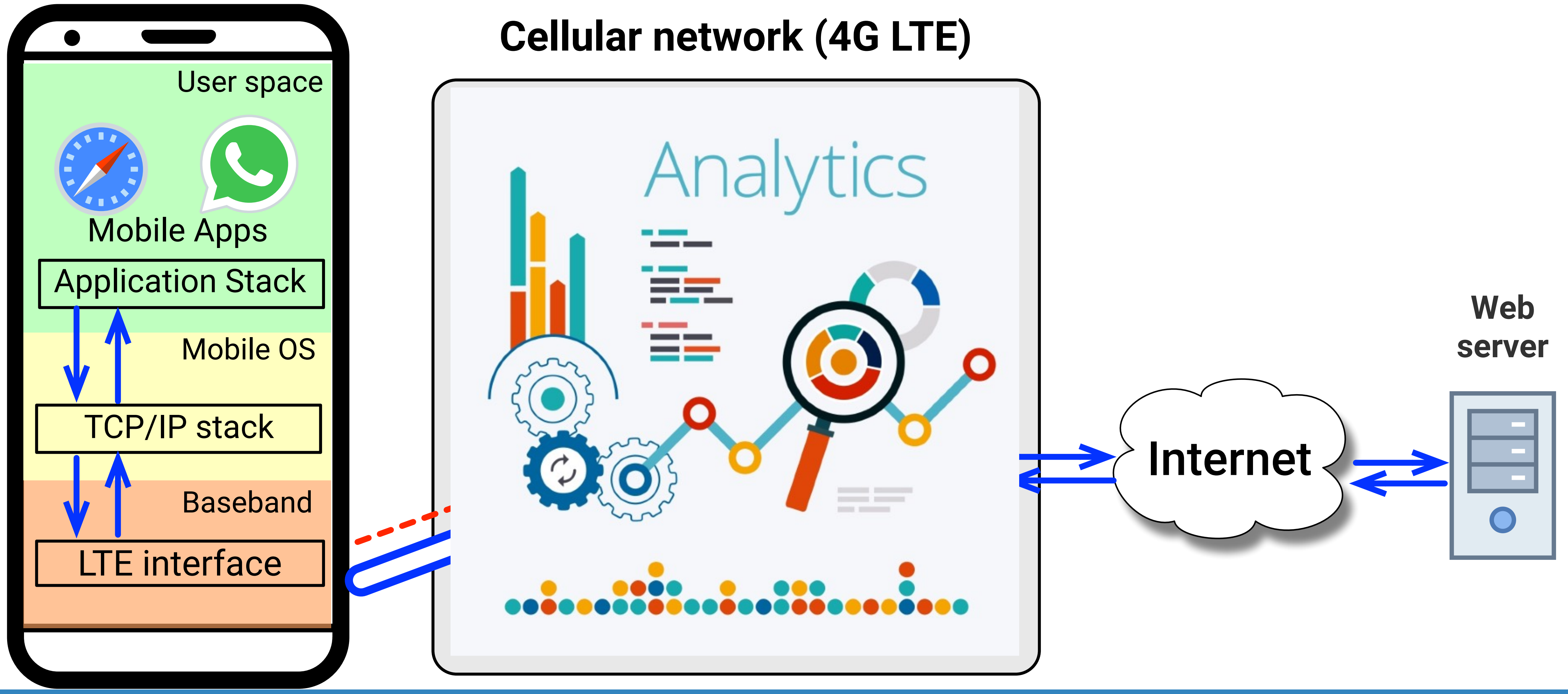
Current 4G network analytics is primarily "infrastructure-based":



State-of-the-art mobile network analytics

Current 4G network analytics is primarily "infrastructure-based":

Not Scalable

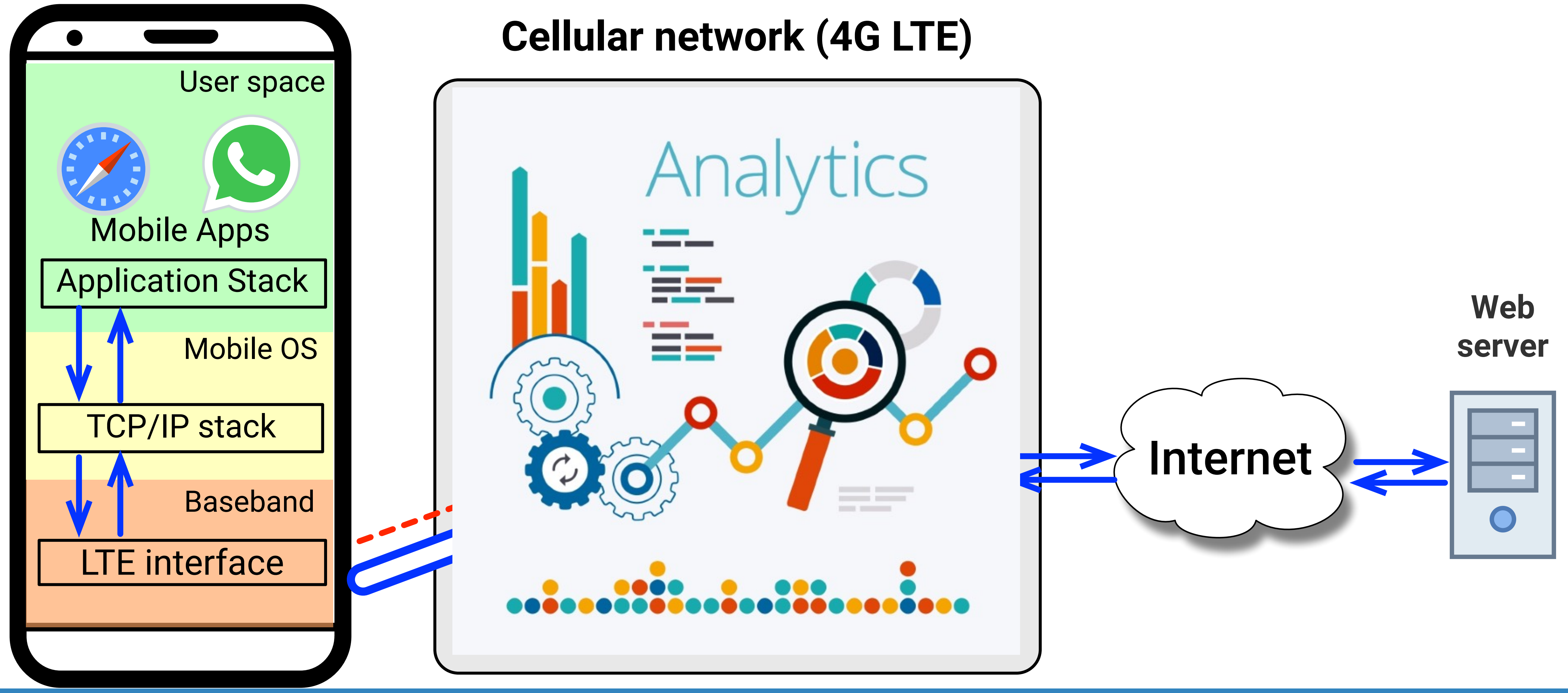


State-of-the-art mobile network analytics

Current 4G network analytics is primarily "infrastructure-based":

Not Scalable

Incomplete View



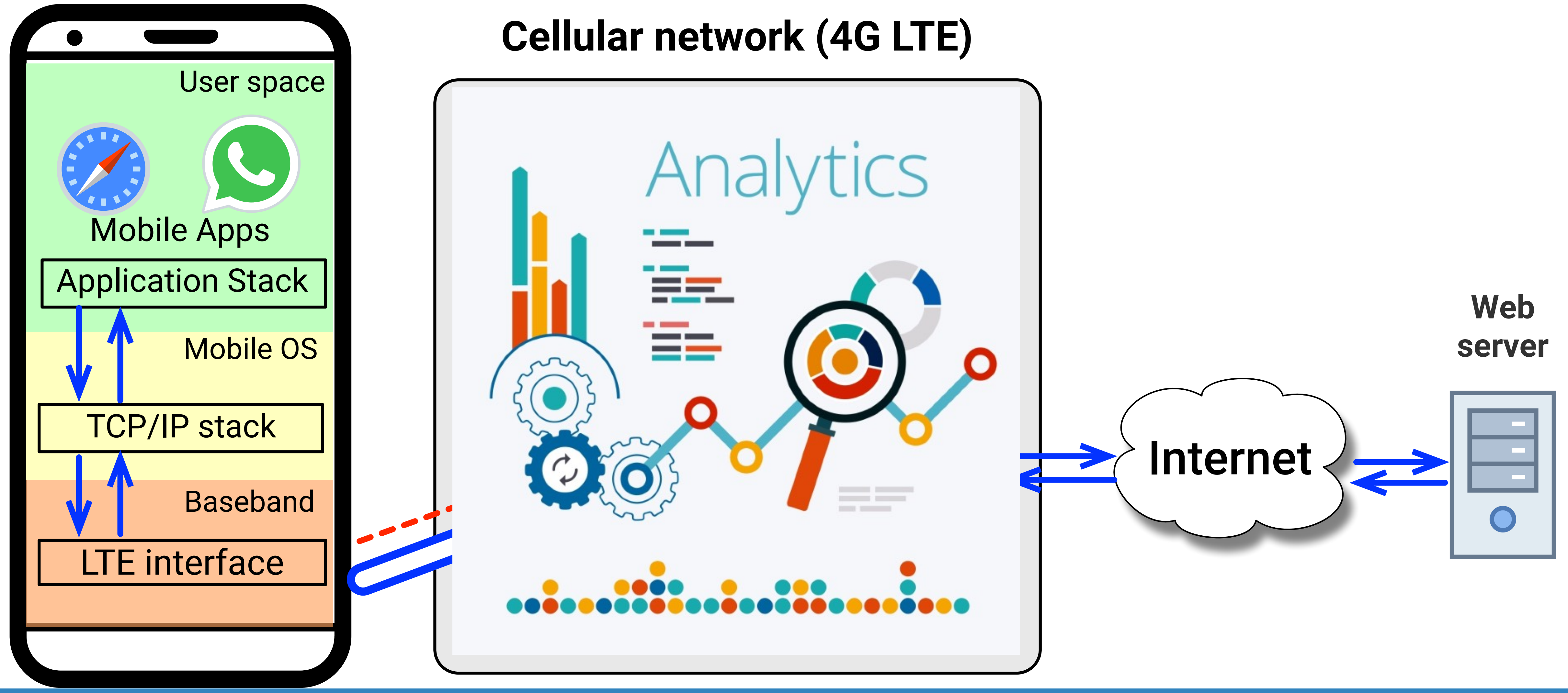
State-of-the-art mobile network analytics

Current 4G network analytics is primarily "infrastructure-based":

Not Scalable

Incomplete View

Opaqueness

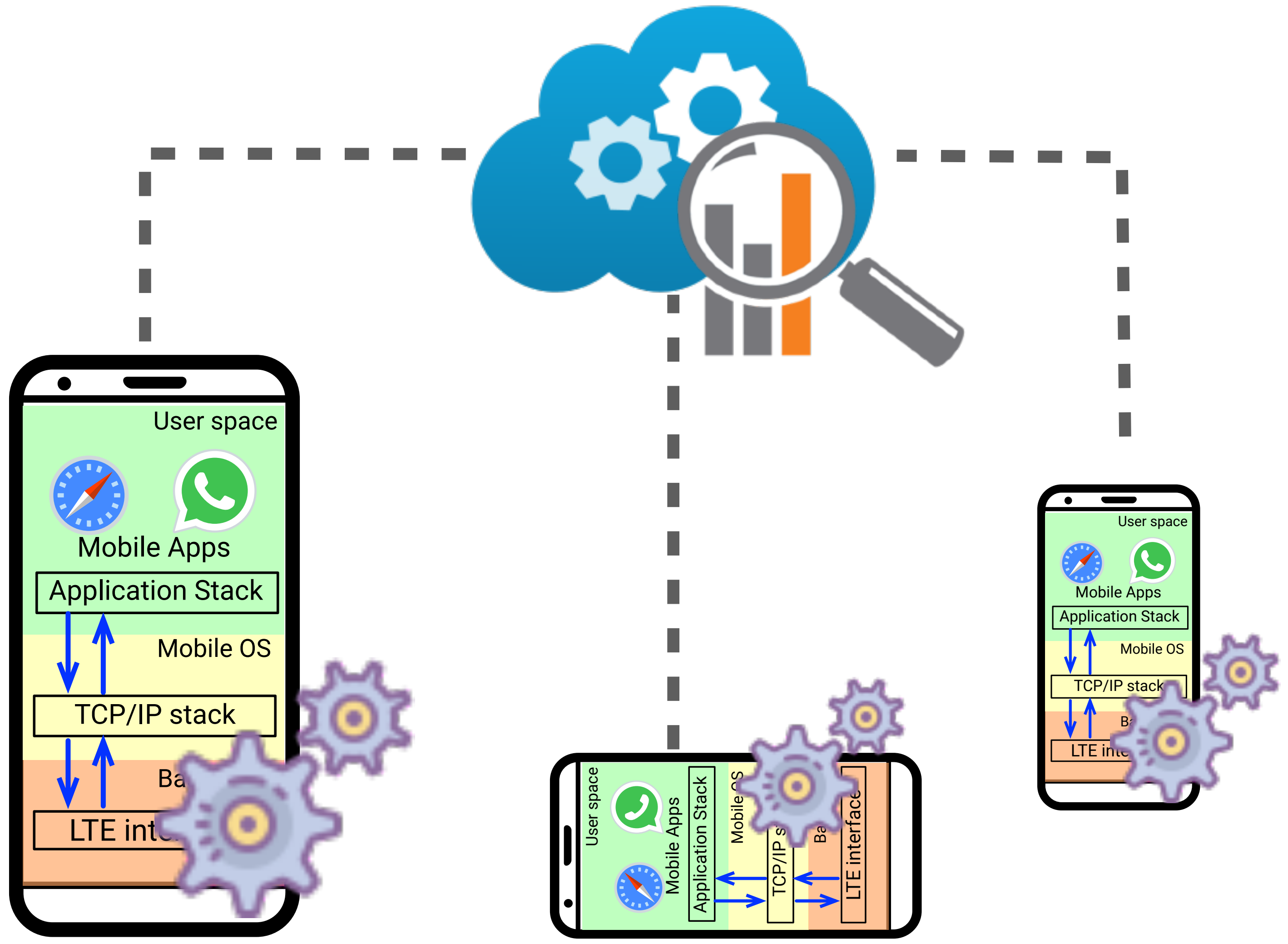


ML-based approach is a must-have feature for mobile network analytics

~~Not Scalable~~

~~Incomplete View~~

~~Opaqueness~~



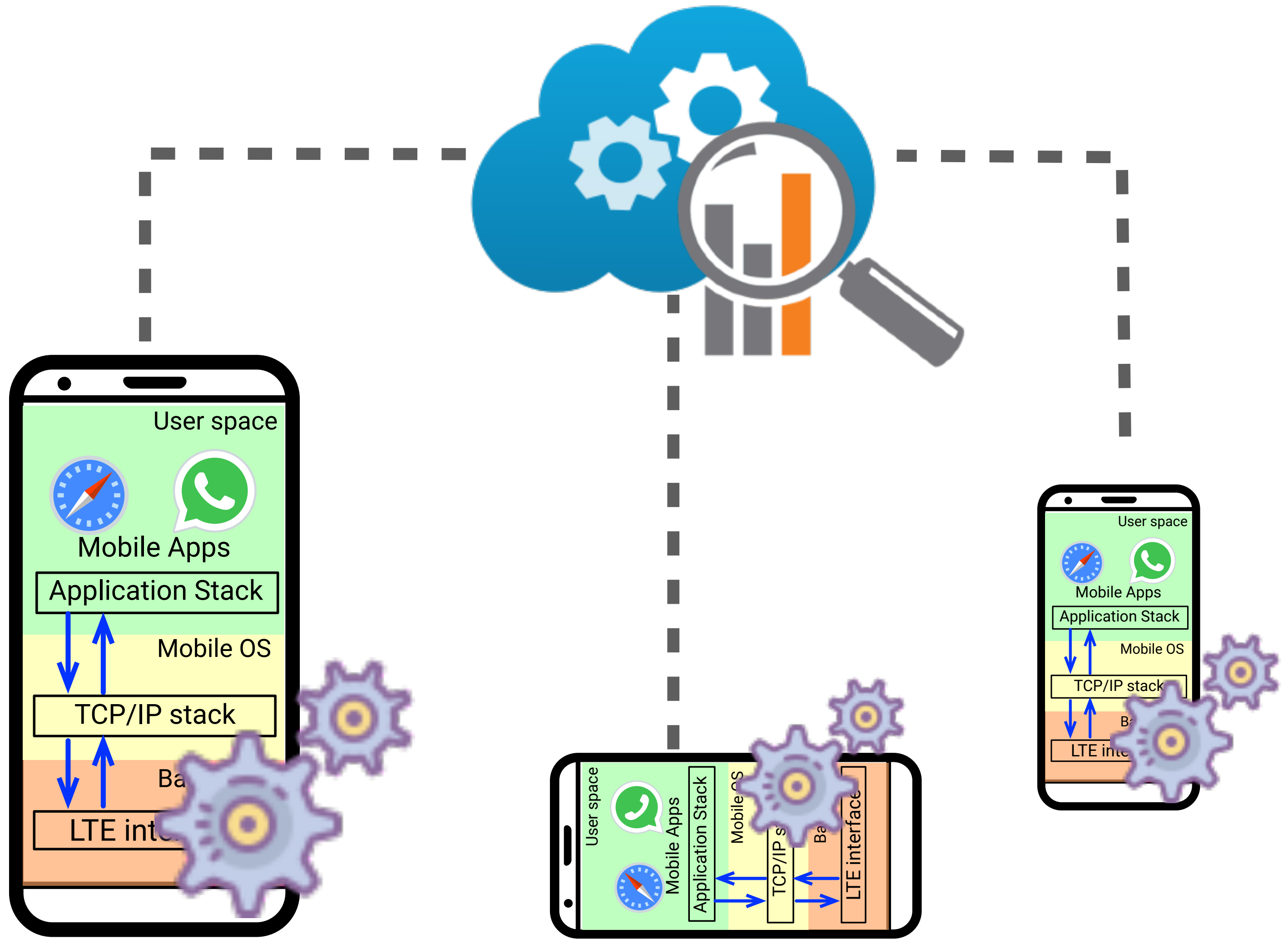
ML-based approach is a must-have feature for mobile network analytics

Device-centric ML approach brings new hope

~~Not Scalable~~

~~Incomplete View~~

~~Opaqueness~~



ML-based approach is a must-have feature for mobile network analytics

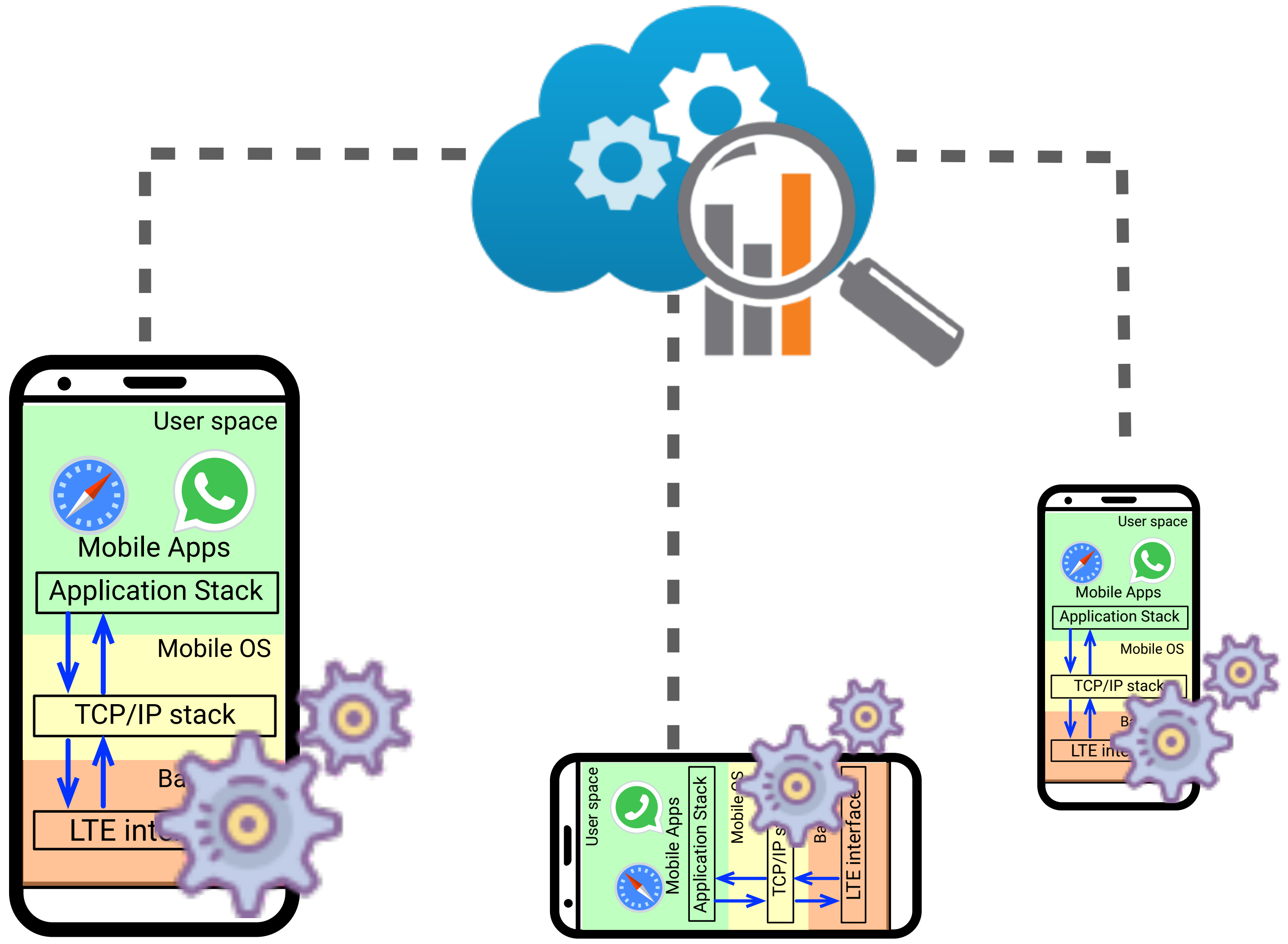
Device-centric ML approach brings new hope

~~Not Scalable~~

~~Incomplete View~~

~~Opaqueness~~

Scalability

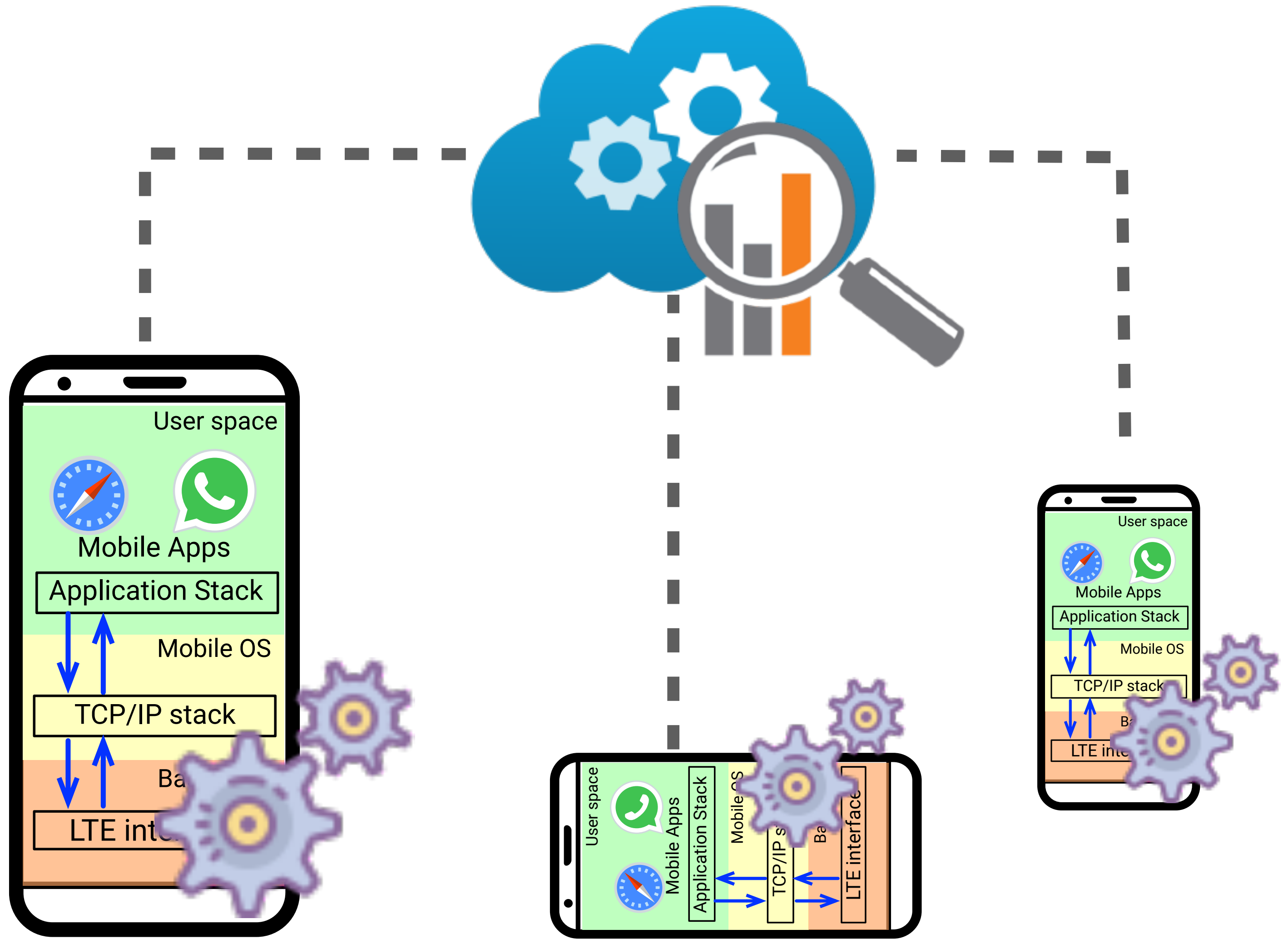


ML-based approach is a must-have feature for mobile network analytics

Device-centric ML approach brings new hope

~~Not Scalable~~ | ~~Incomplete View~~ | ~~Opaqueness~~

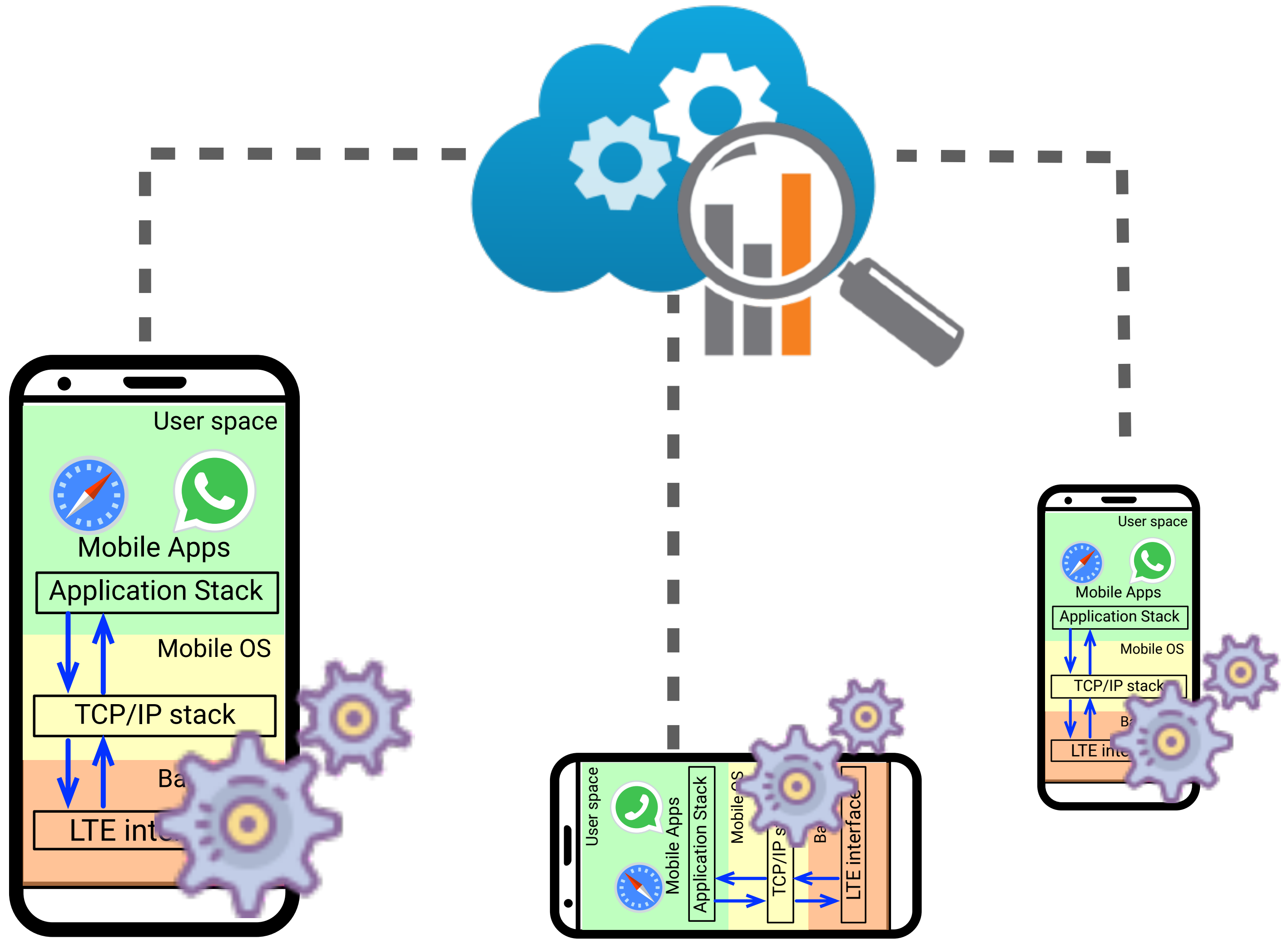
Scalability | Device QoE View



ML-based approach is a must-have feature for mobile network analytics

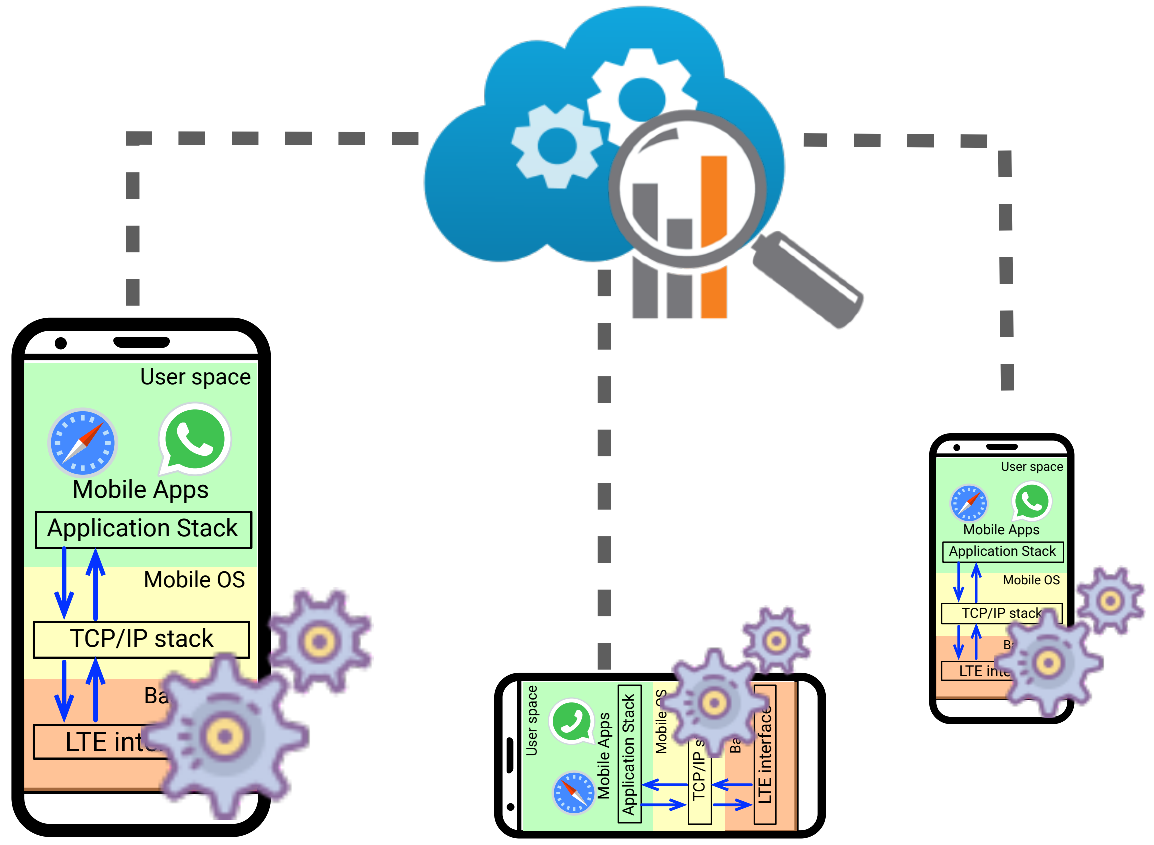
Device-centric ML approach brings new hope

~~Not Scalable~~	~~Incomplete View~~	~~Opaqueness~~
Scalability	Device QoE View	Availability



It is probably true that machine learning is a *must-have* approach, rather than a *nice-to-have* one, to our field for mobile network analysis

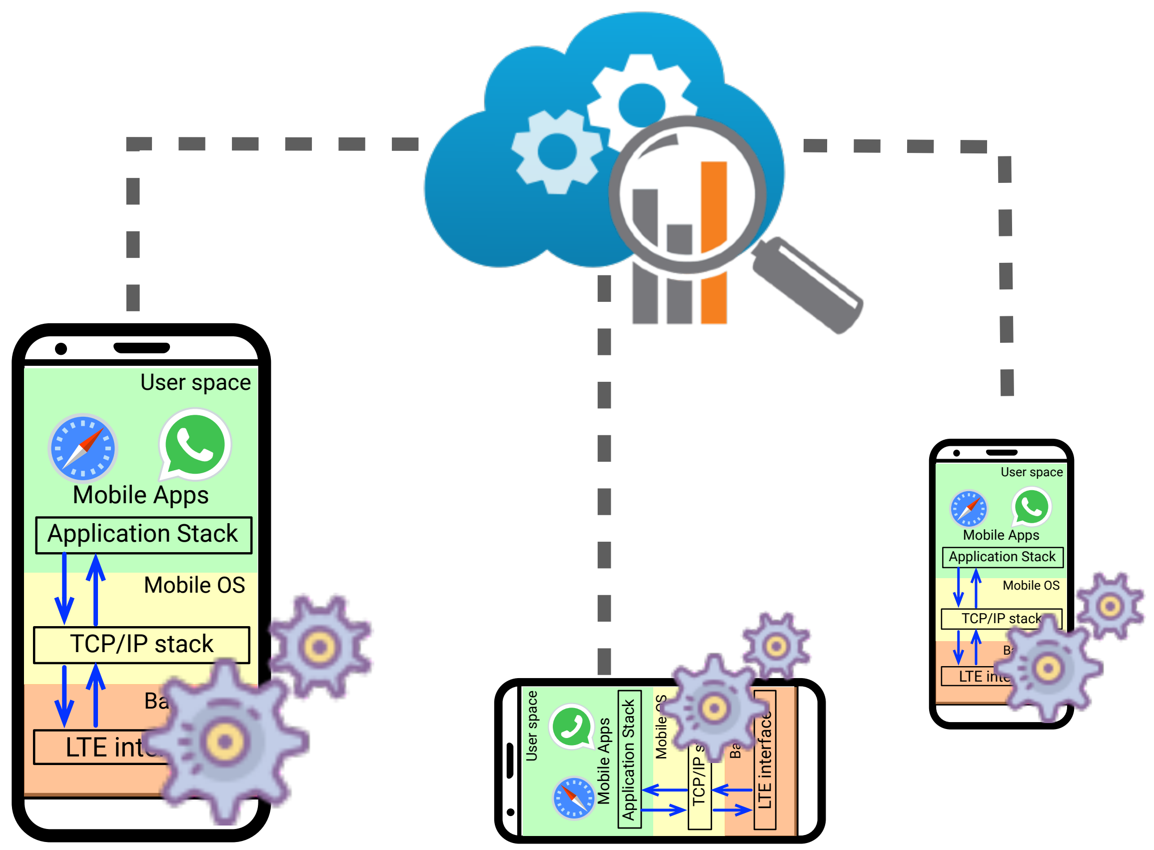
Our proposal: two-level device-centric ML approach



Our proposal: two-level device-centric ML approach

Local level: sensing mobile network data inside each smartphone

- Via hardware-software coordination (e.g. MobileInsight [ACM MobiCom'16])
- Via higher-layer (application/transport/IP) and lower-layer (cellular-specific) integration
- Via ML-assisted data plane prediction from control plane protocol reconstruction



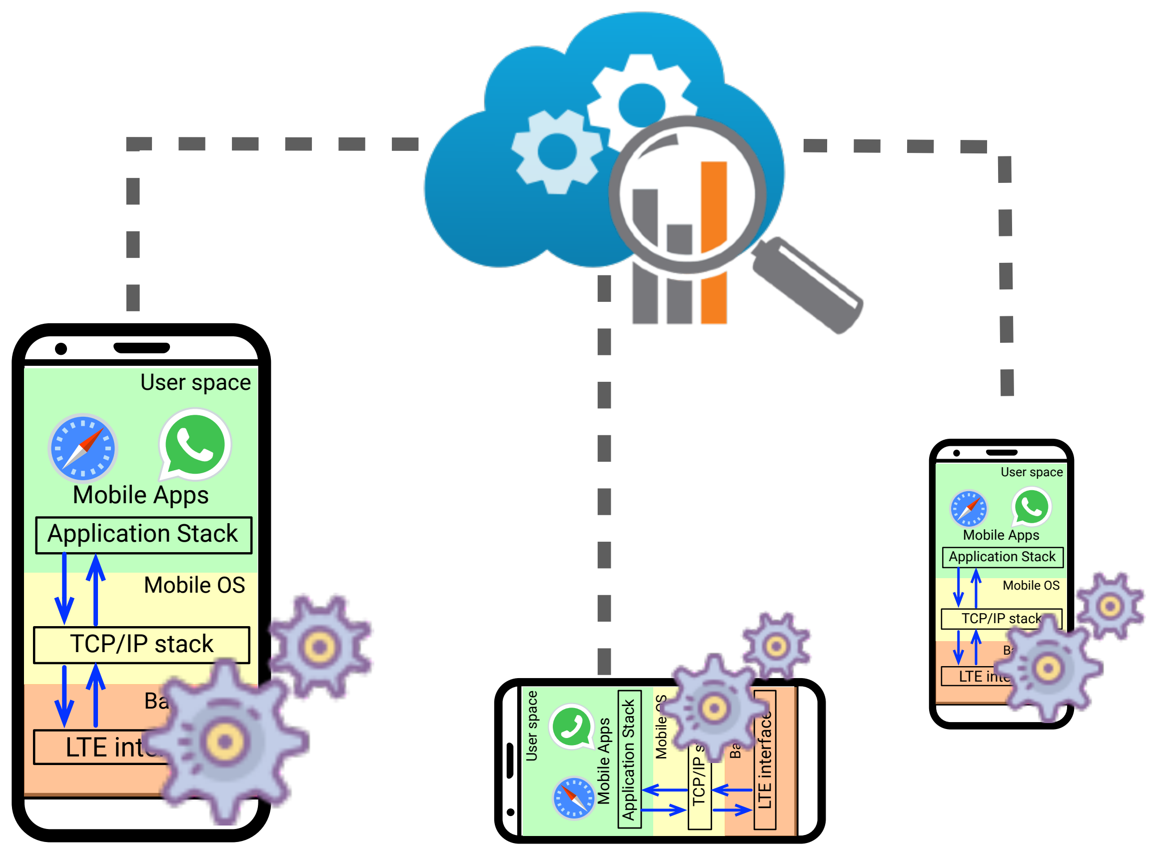
Our proposal: two-level device-centric ML approach

Local level: sensing mobile network data inside each smartphone

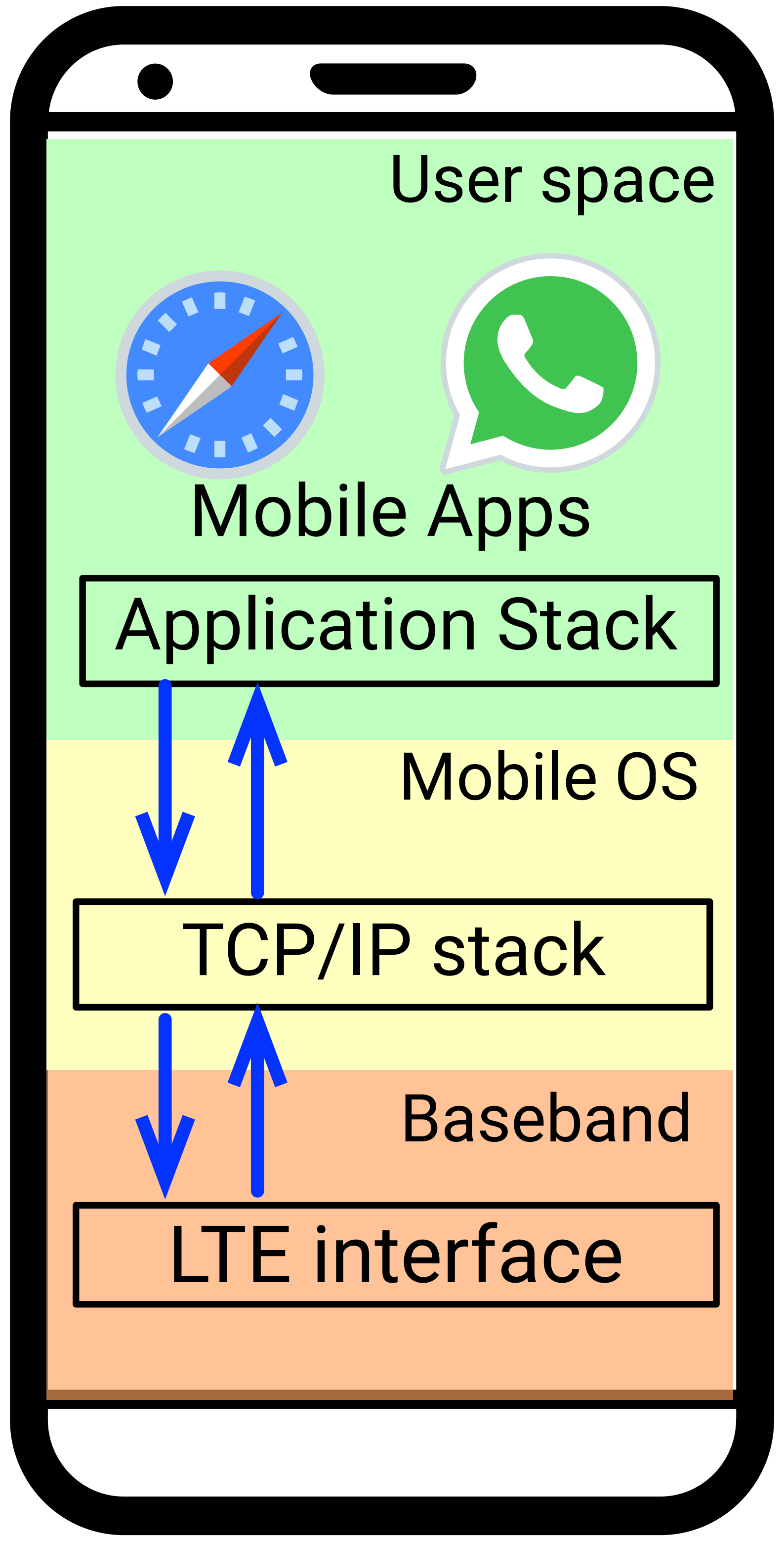
- Via hardware-software coordination (e.g. MobileInsight [ACM MobiCom'16])
- Via higher-layer (application/transport/IP) and lower-layer (cellular-specific) integration
- Via ML-assisted data plane prediction from control plane protocol reconstruction

Global level:

- Crowdsourcing-based dataset
- Cloud-synthesized insights



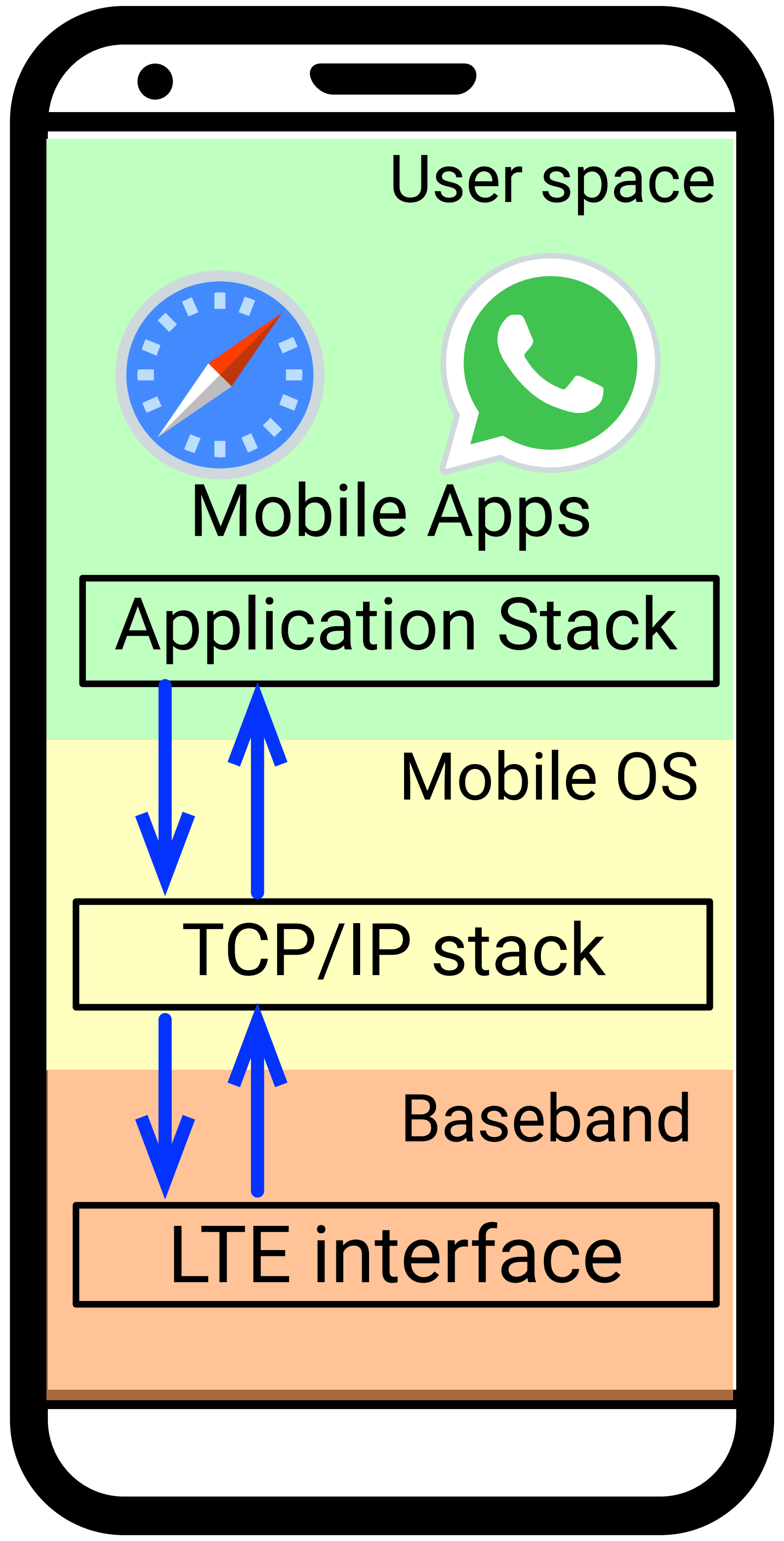
Local analysis



Local analysis

Step 1: open up the “black-box” operations

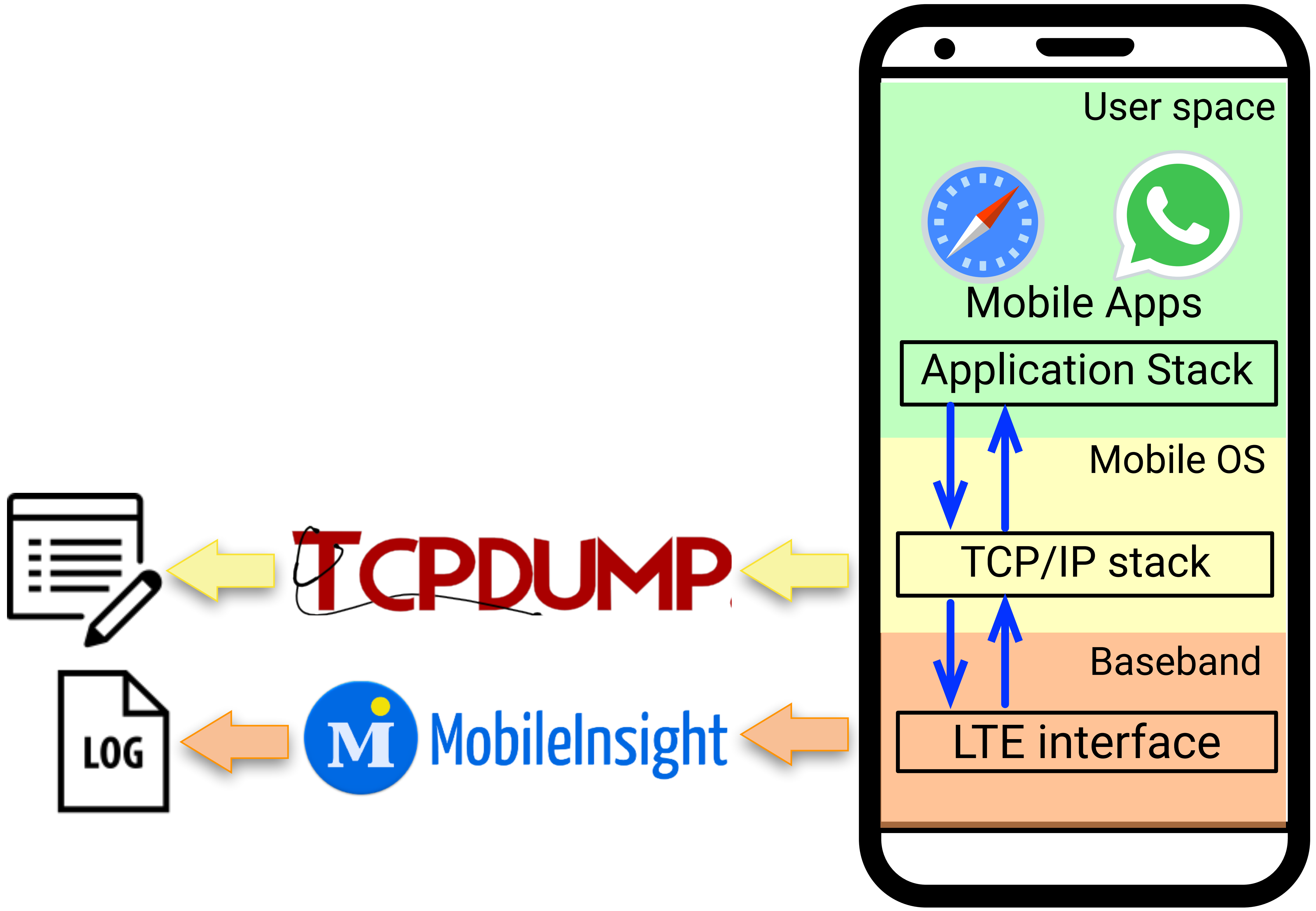
- At/above IP network data: TCPDUMP
- Below IP network data: MobileInsight



Local analysis

Step 1: open up the “black-box” operations

- At/above IP network data: TCPDUMP
- Below IP network data: MobileInsight



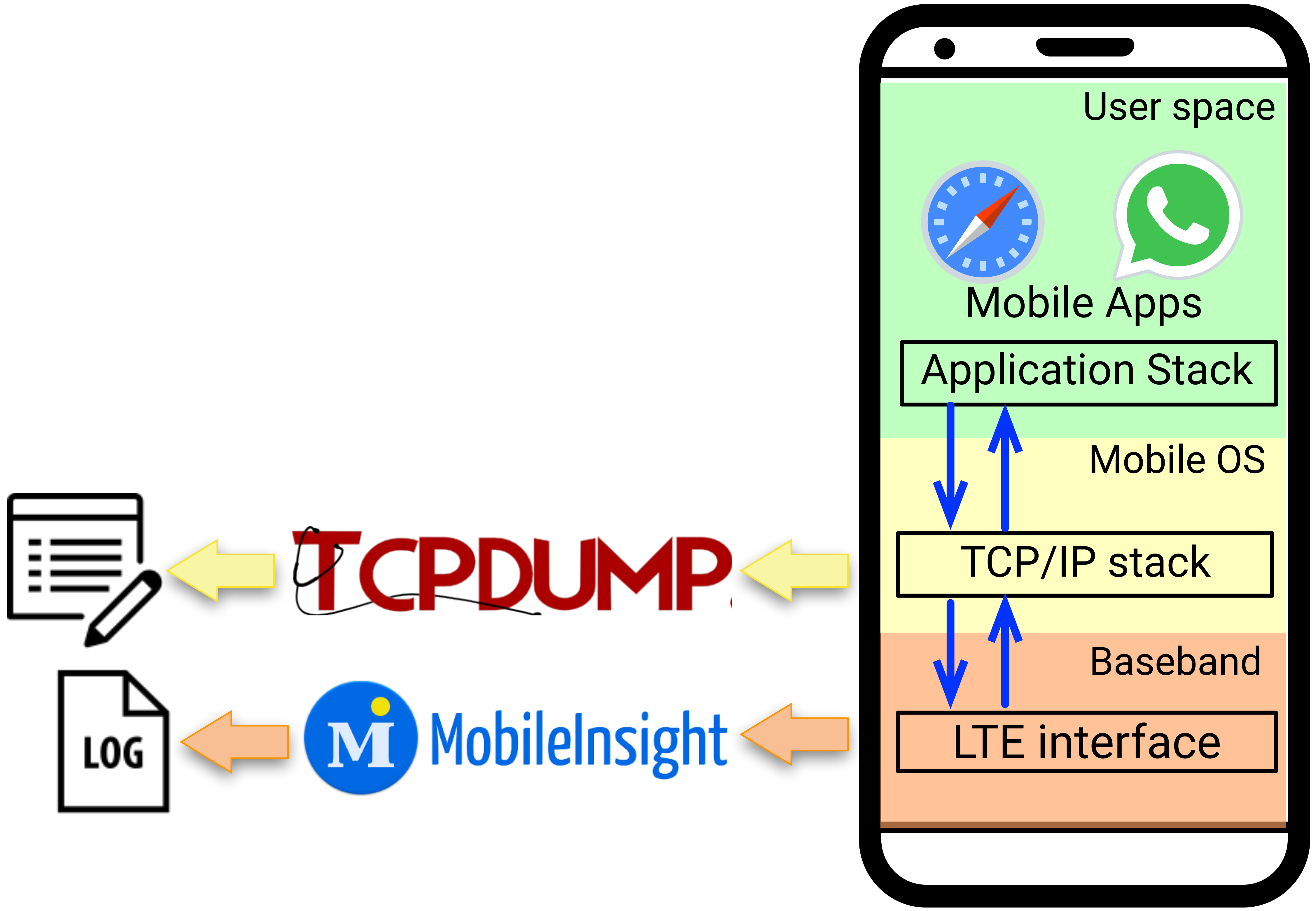
Local analysis

Step 1: open up the "black-box" operations

- At/above IP network data: TCPDUMP
- Below IP network data: MobileInsight

Step 2: automated data preprocessing

- Data cleansing and integration of two sources



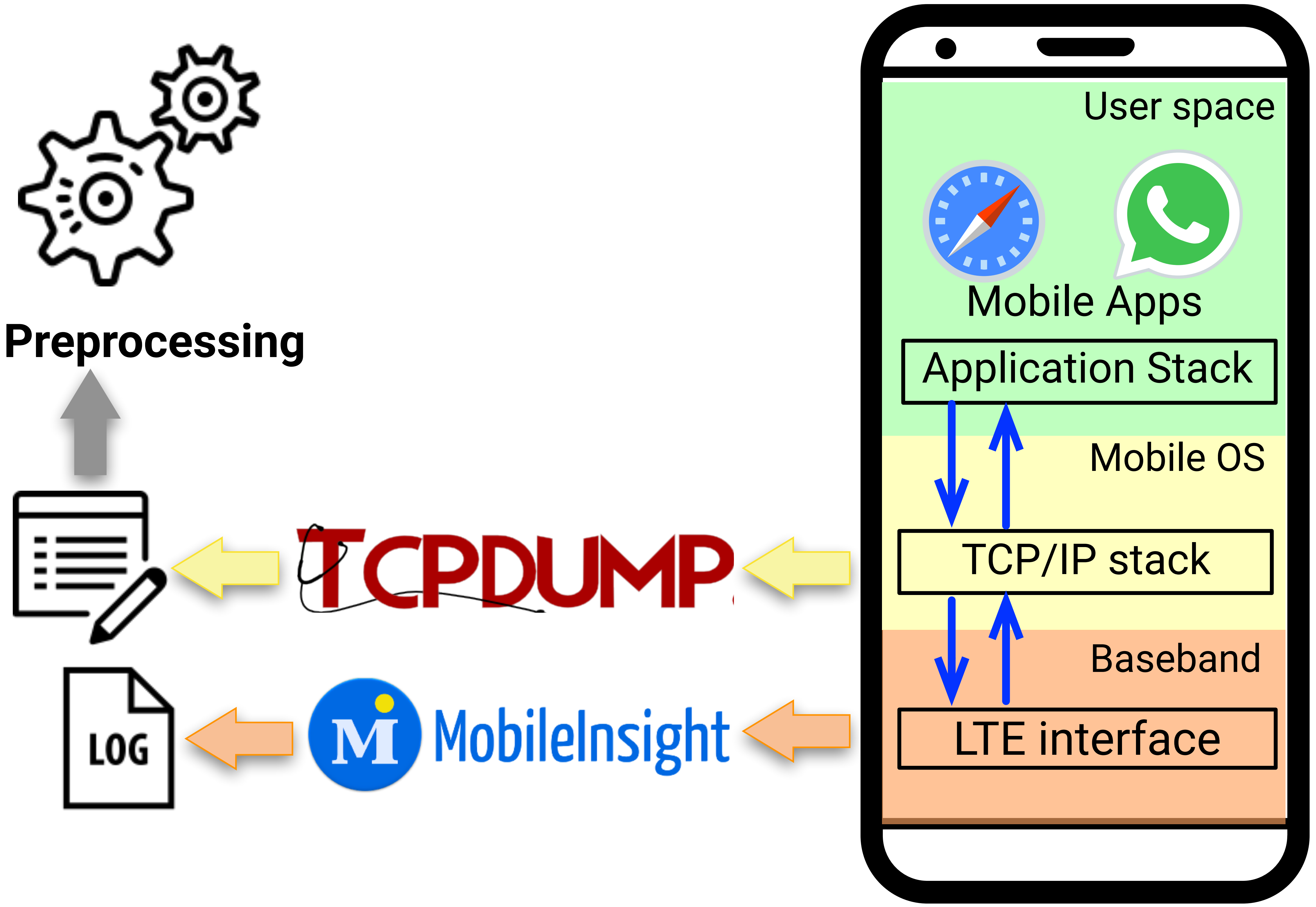
Local analysis

Step 1: open up the “black-box” operations

- At/above IP network data: TCPDUMP
- Below IP network data: MobileInsight

Step 2: automated data preprocessing

- Data cleansing and integration of two sources



Local analysis

Step 1: open up the "black-box" operations

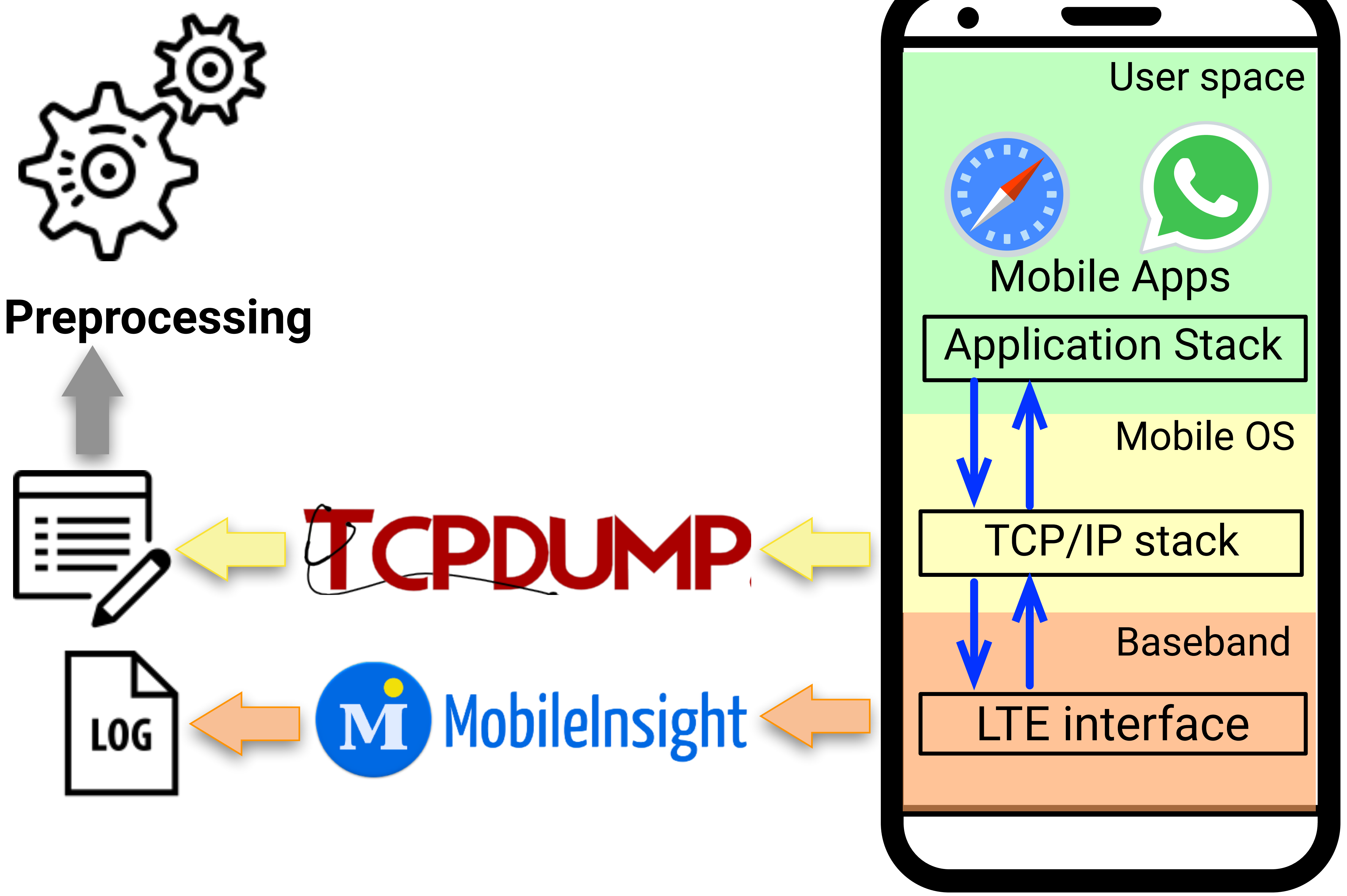
- At/above IP network data: TCPDUMP
- Below IP network data: MobileInsight

Step 2: automated data preprocessing

- Data cleansing and integration of two sources

Step 3: local ML-based analysis

- Control plane for protocol operations
- Data plane for performance



Local analysis

Step 1: open up the "black-box" operations

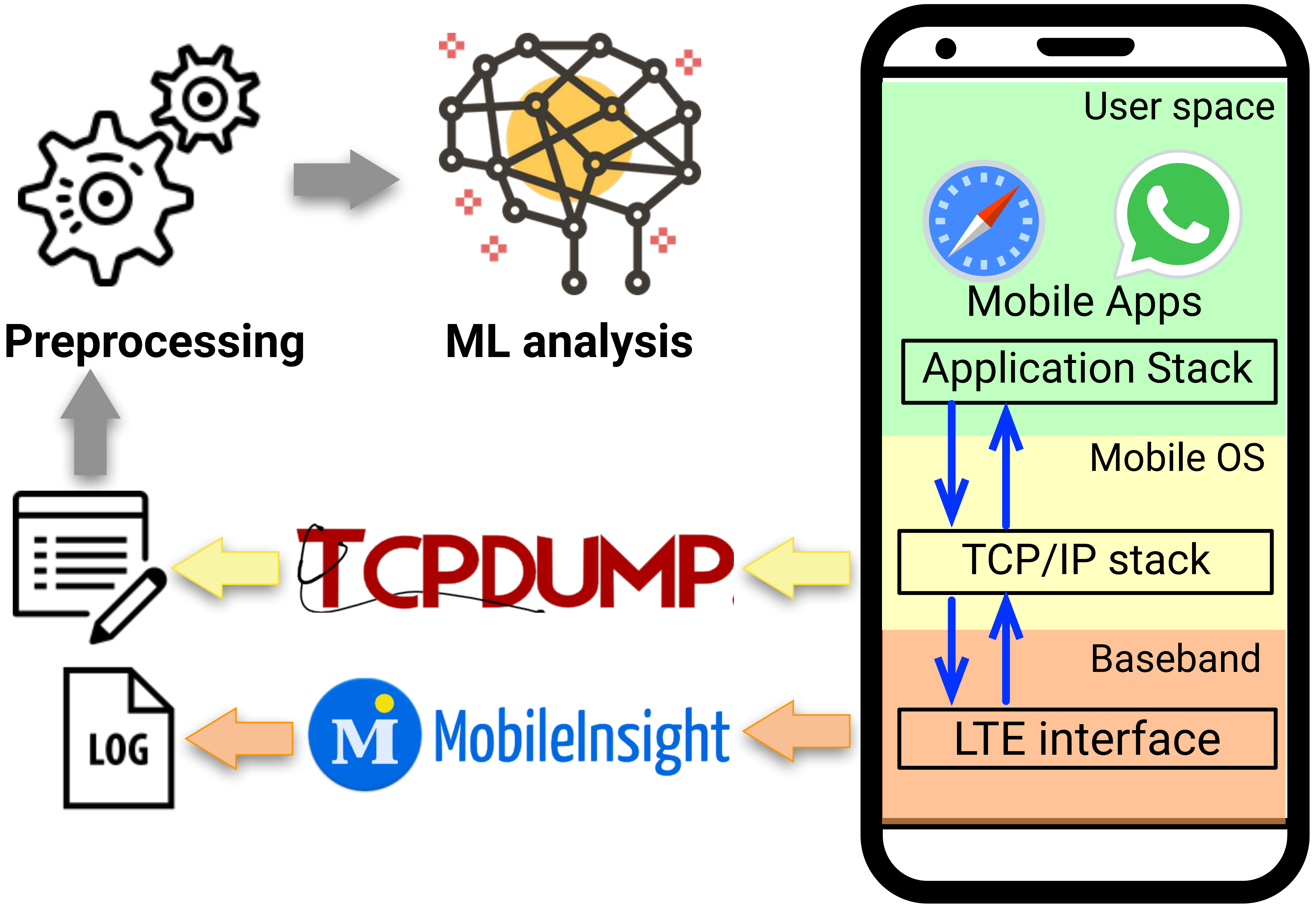
- At/above IP network data: TCPDUMP
- Below IP network data: MobileInsight

Step 2: automated data preprocessing

- Data cleansing and integration of two sources

Step 3: local ML-based analysis

- Control plane for protocol operations
- Data plane for performance

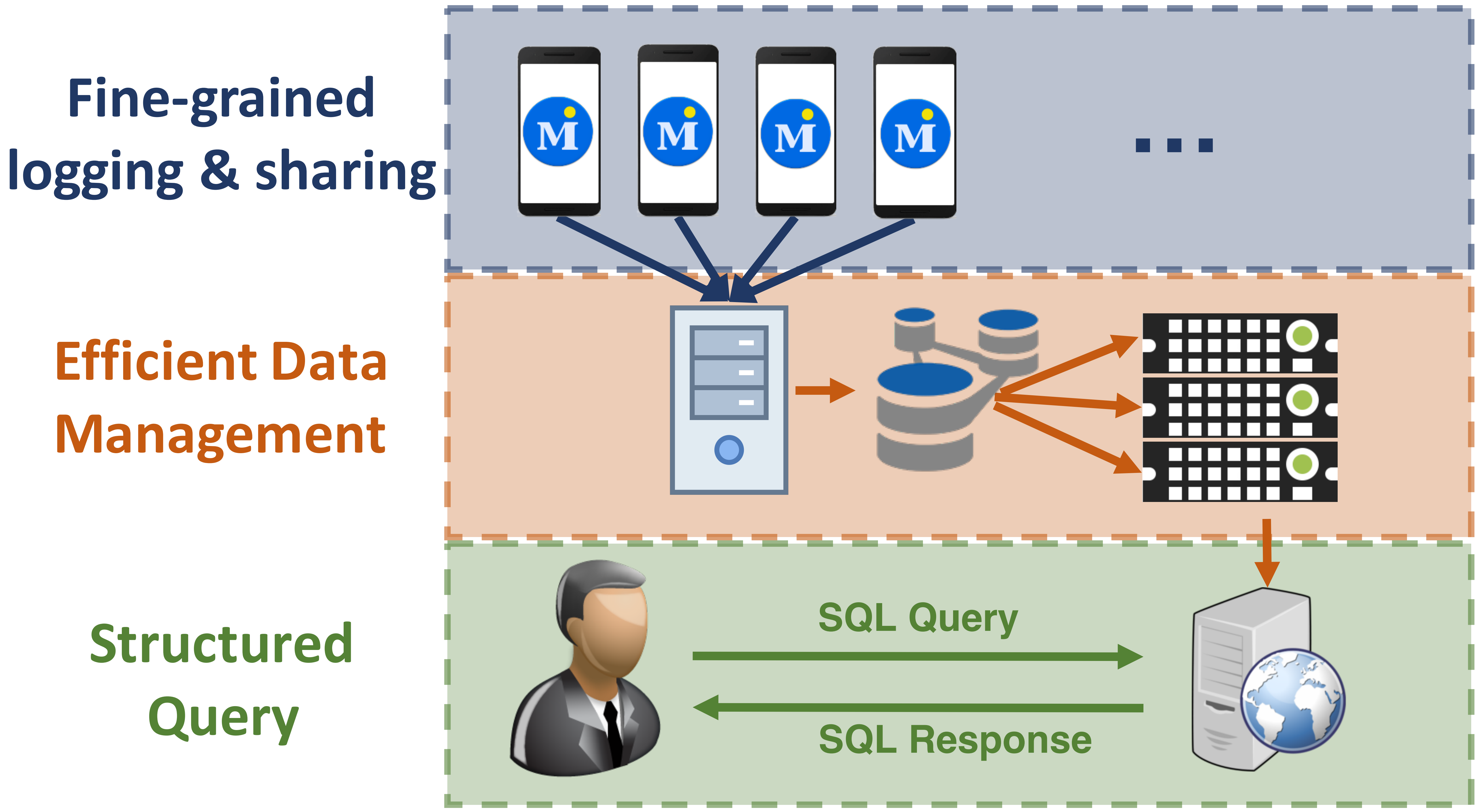


Global analysis

Enabled by cloud-based crowdsourcing (e.g. cniCloud [HotWireless'17])

Analytical Insights for:

- Geographical location
- Operators
- Phone models
- …



Case study: latency analysis in mobile networks

Every millisecond of latency matters!

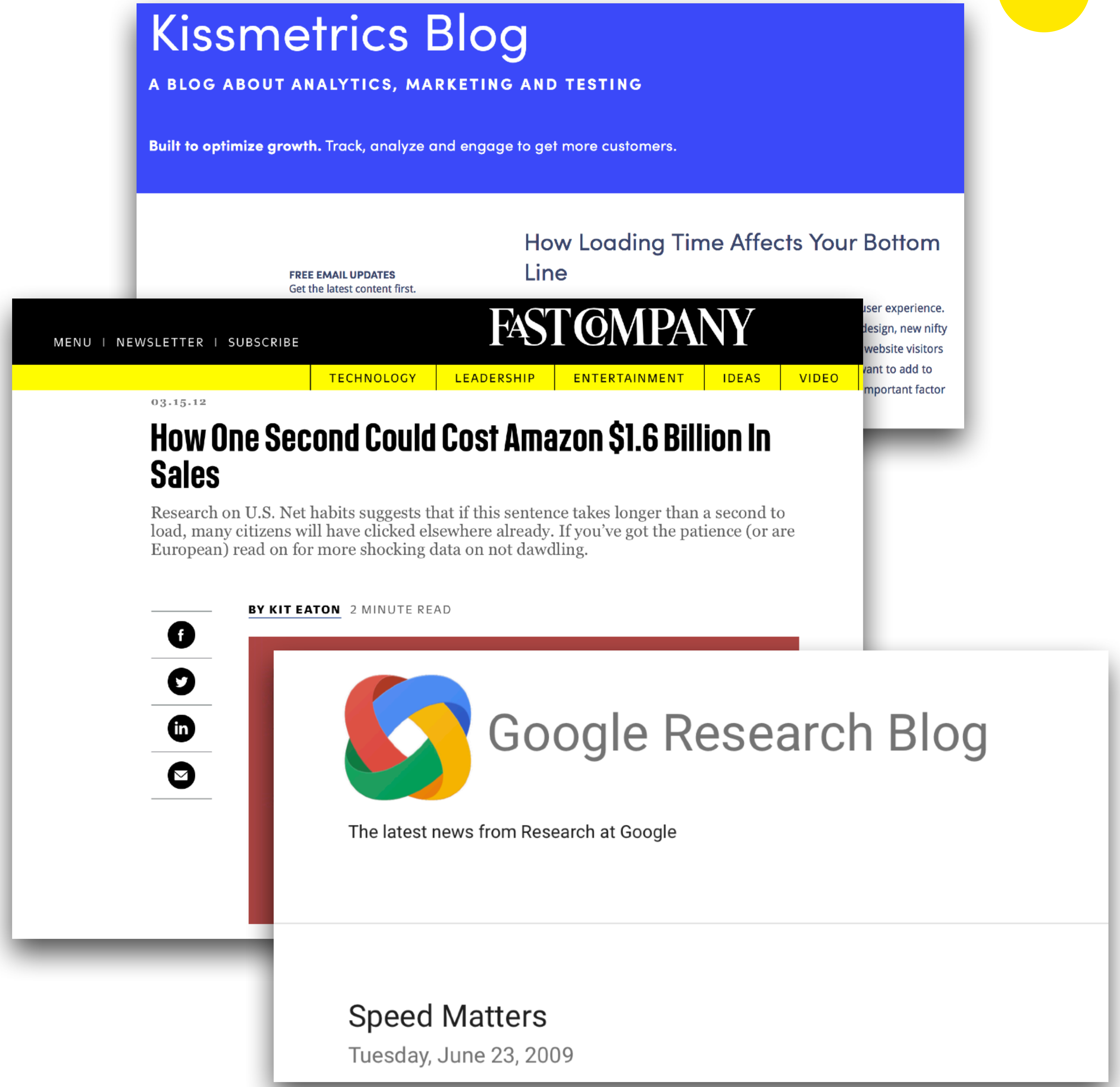
Mobile network users want *fast* access

- 1 second latency in page response → 7% reduction in PageView [KissMetrics 2011]

Developers *lose revenue* due to long latency

- Every **100 ms** costs Amazon **1% ($1.6 bn)** in sales
- An extra **400 ms** latency drops daily Google searches per user by **0.6%**

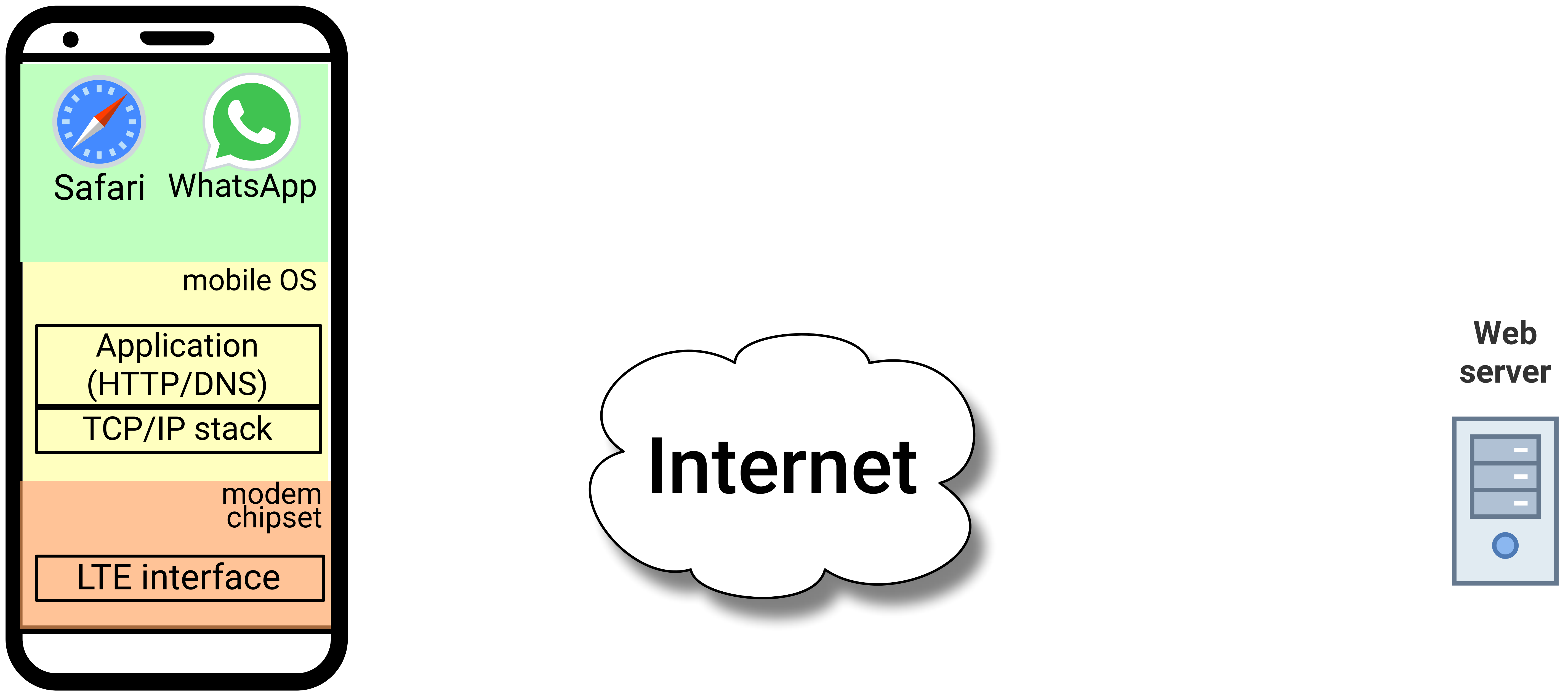
Latency does matter a lot!



Background: how do mobile apps work over 4G LTE?

What happens under the hood?

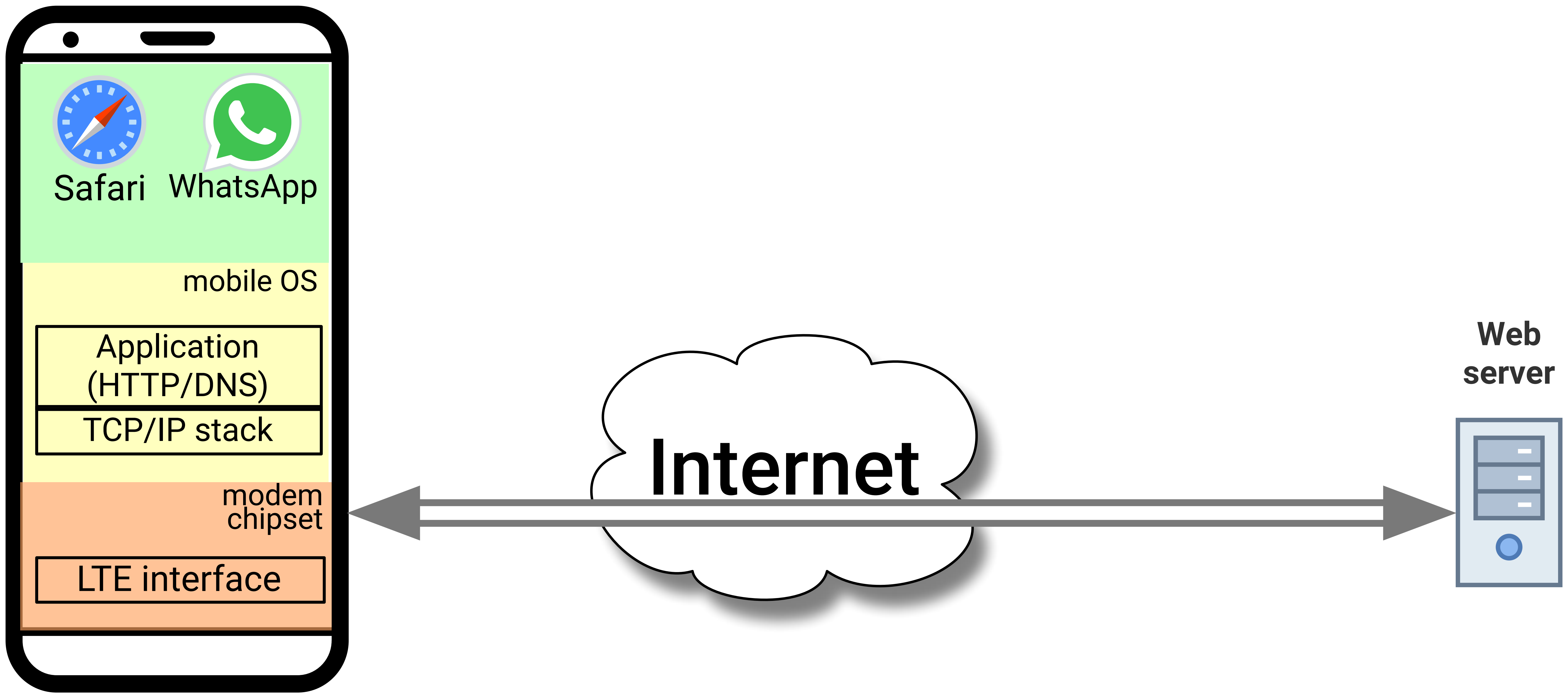
How LTE impacts perceived latency on mobile web/IM app?



Background: how do mobile apps work over 4G LTE?

What happens under the hood?

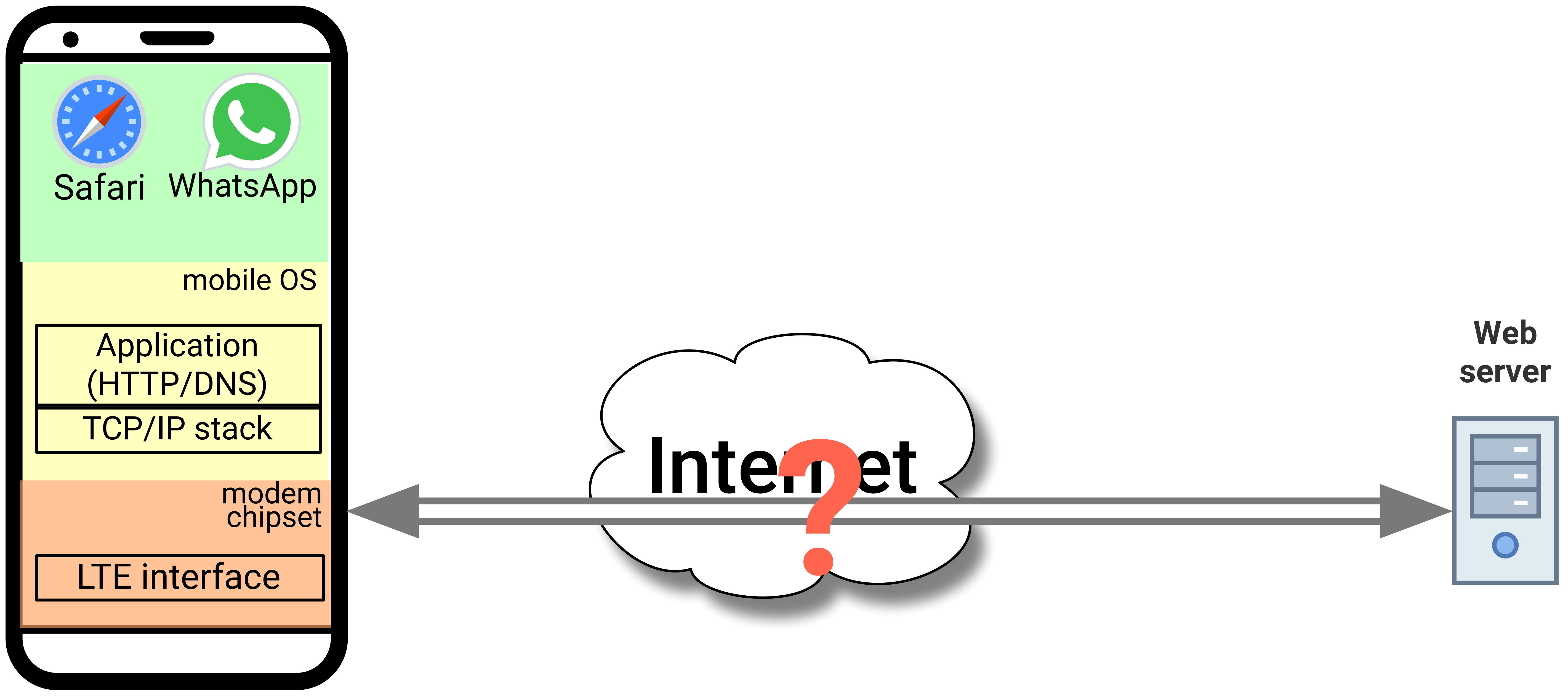
How LTE impacts perceived latency on mobile web/IM app?



Background: how do mobile apps work over 4G LTE?

What happens under the hood?

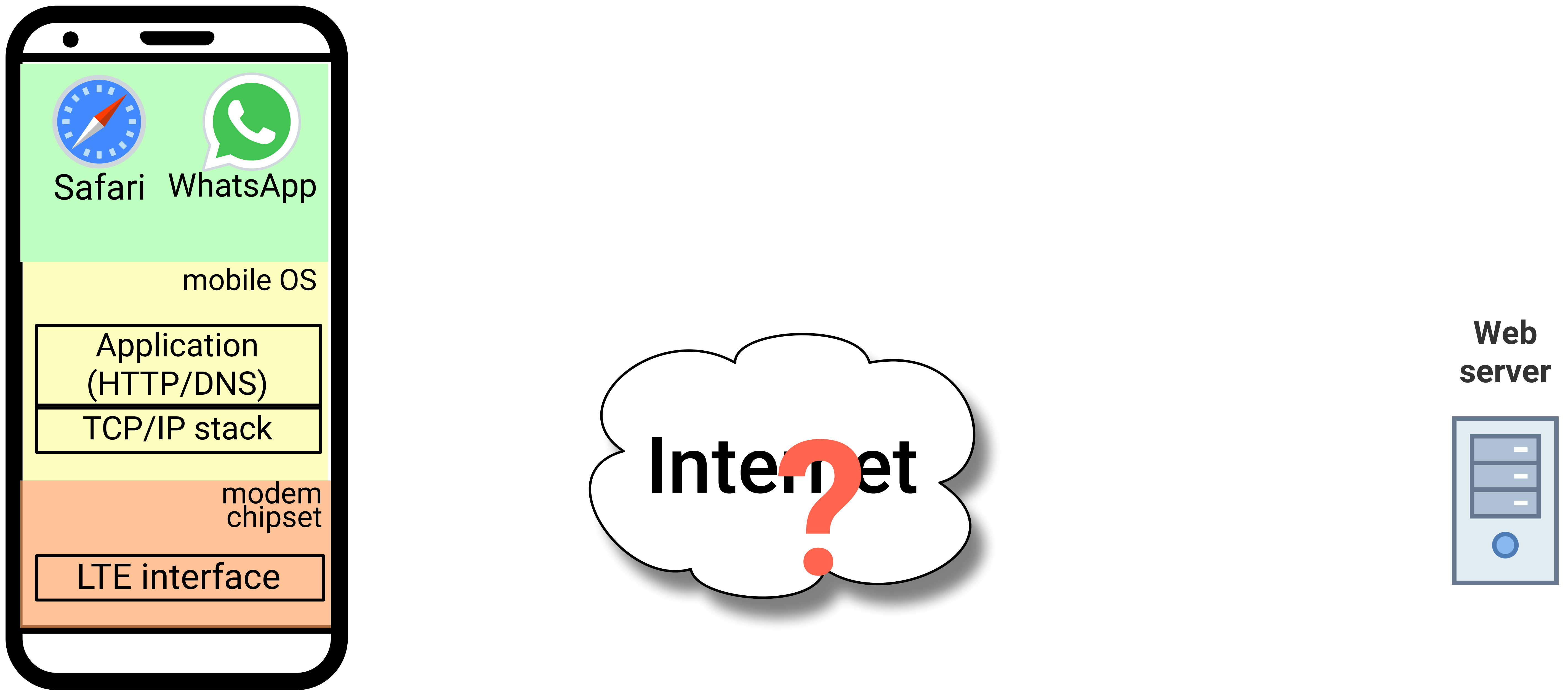
How LTE impacts perceived latency on mobile web/IM app?



Background: how do mobile apps work over 4G LTE?

What happens under the hood?

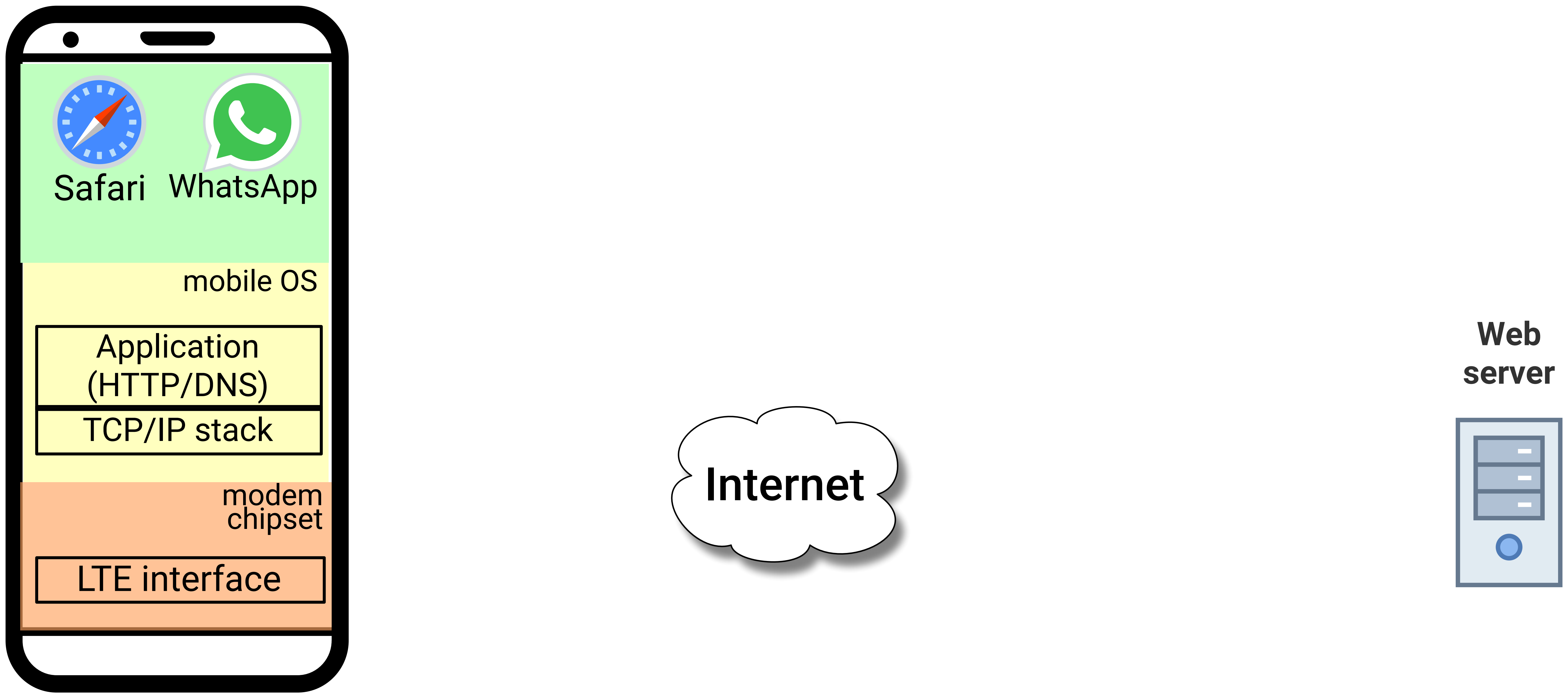
How LTE impacts perceived latency on mobile web/IM app?



Background: how do mobile apps work over 4G LTE?

What happens under the hood?

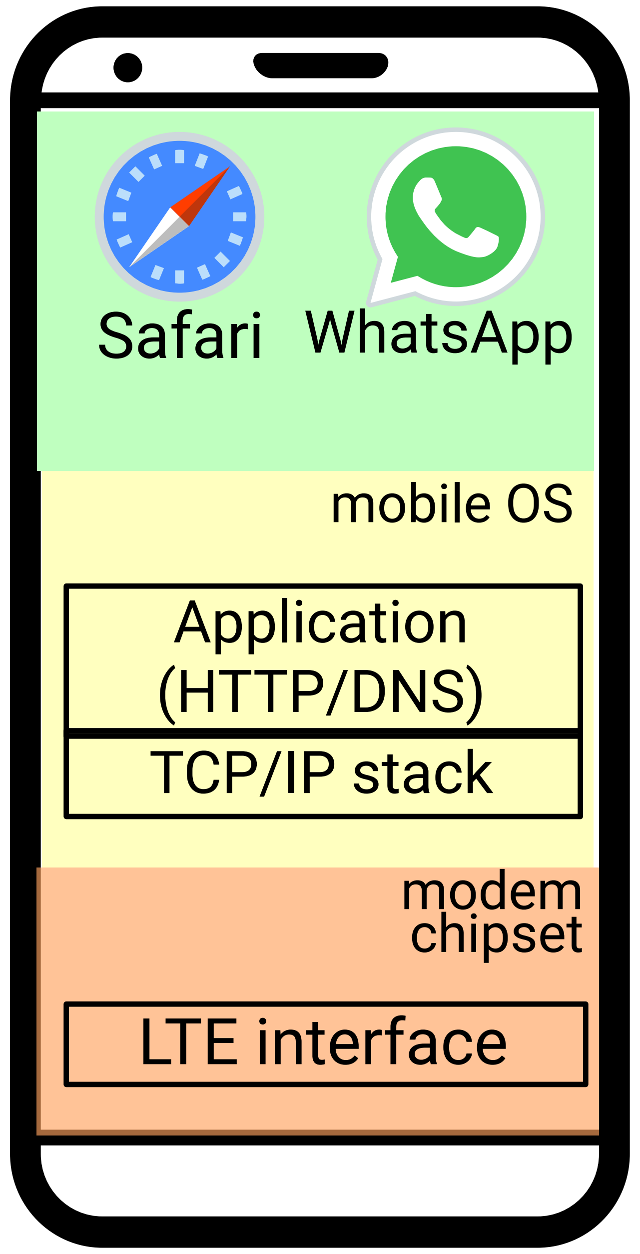
How LTE impacts perceived latency on mobile web/IM app?

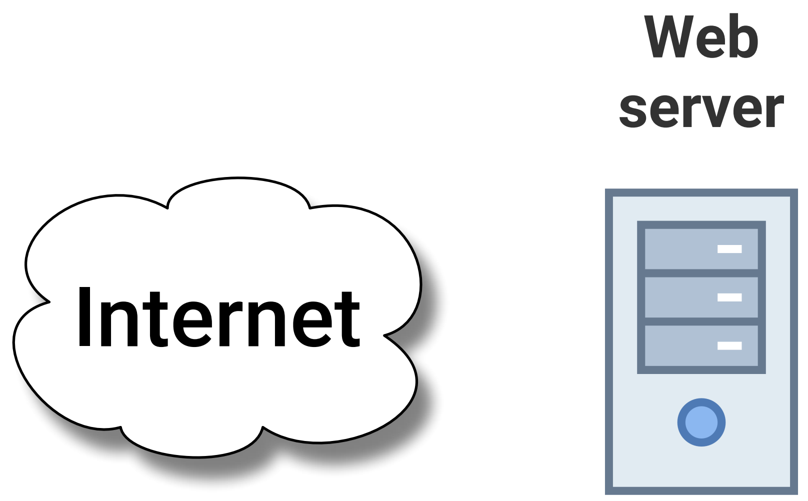


Background: how do mobile apps work over 4G LTE?

What happens under the hood?

How LTE impacts perceived latency on mobile web/IM app?

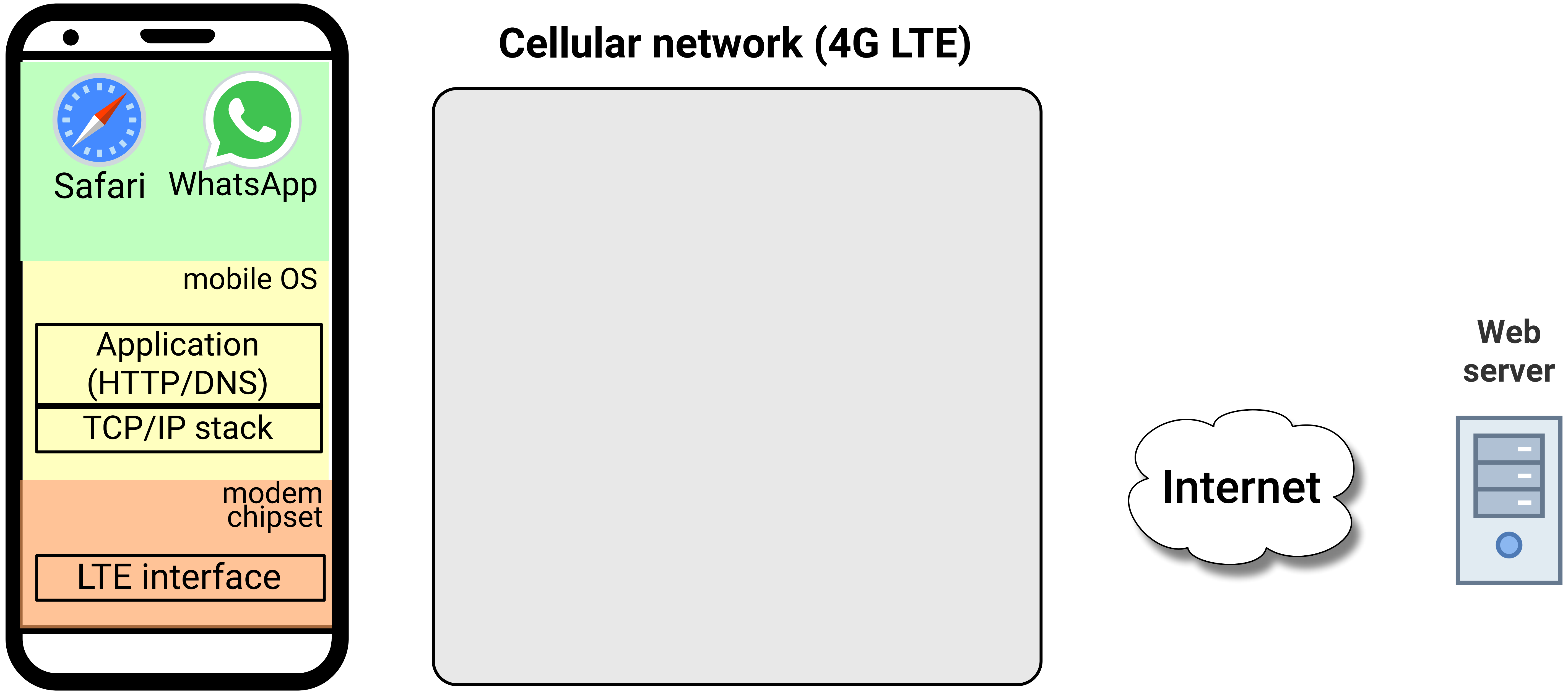




Background: how do mobile apps work over 4G LTE?

What happens under the hood?

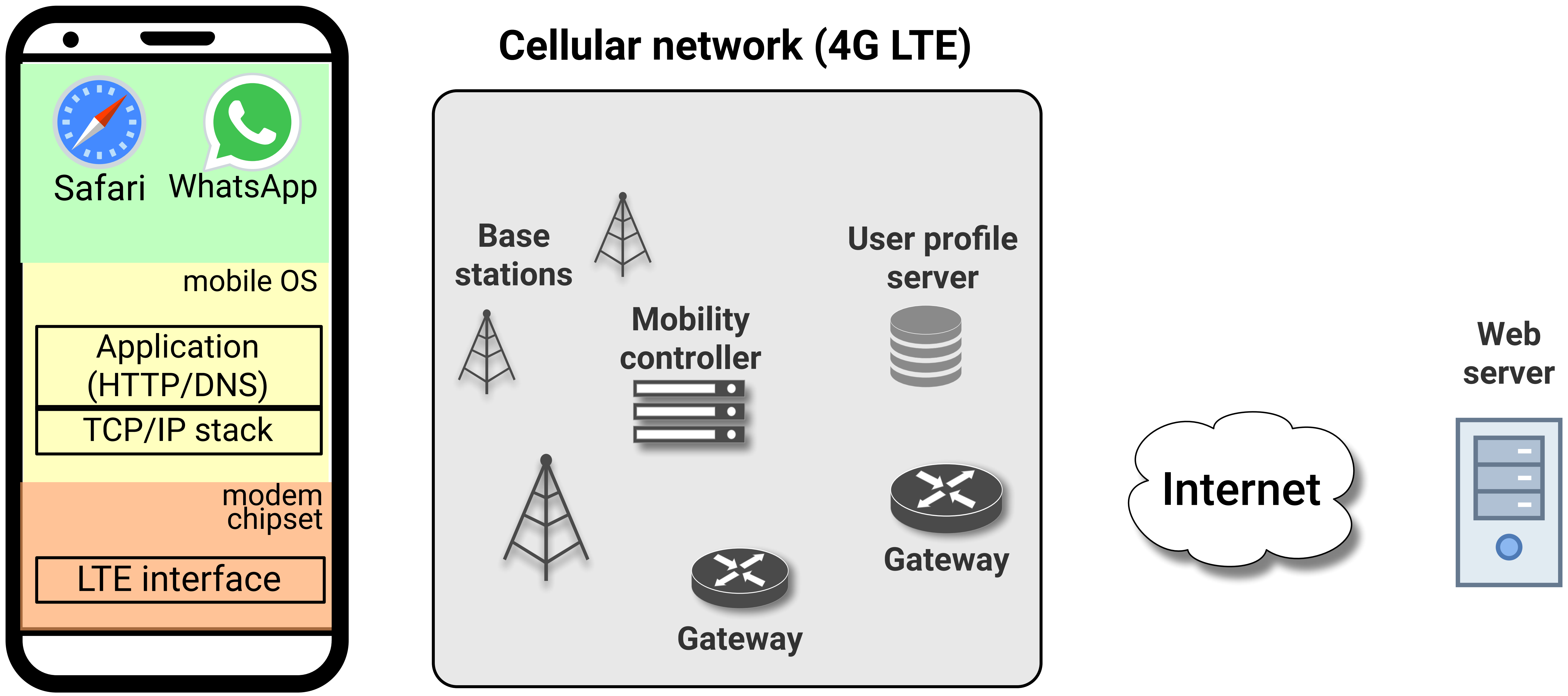
How LTE impacts perceived latency on mobile web/IM app?



Background: how do mobile apps work over 4G LTE?

What happens under the hood?

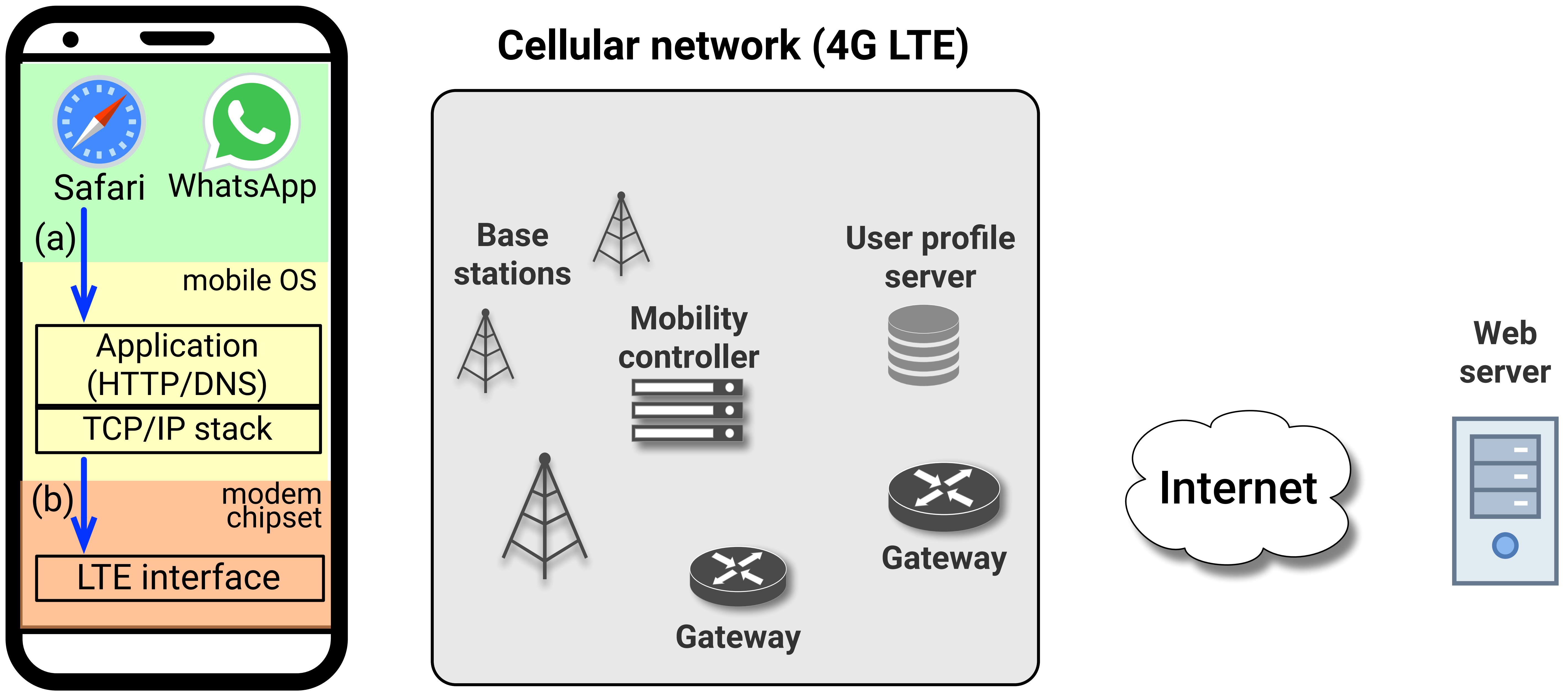
How LTE impacts perceived latency on mobile web/IM app?



Background: how do mobile apps work over 4G LTE?

What happens under the hood?

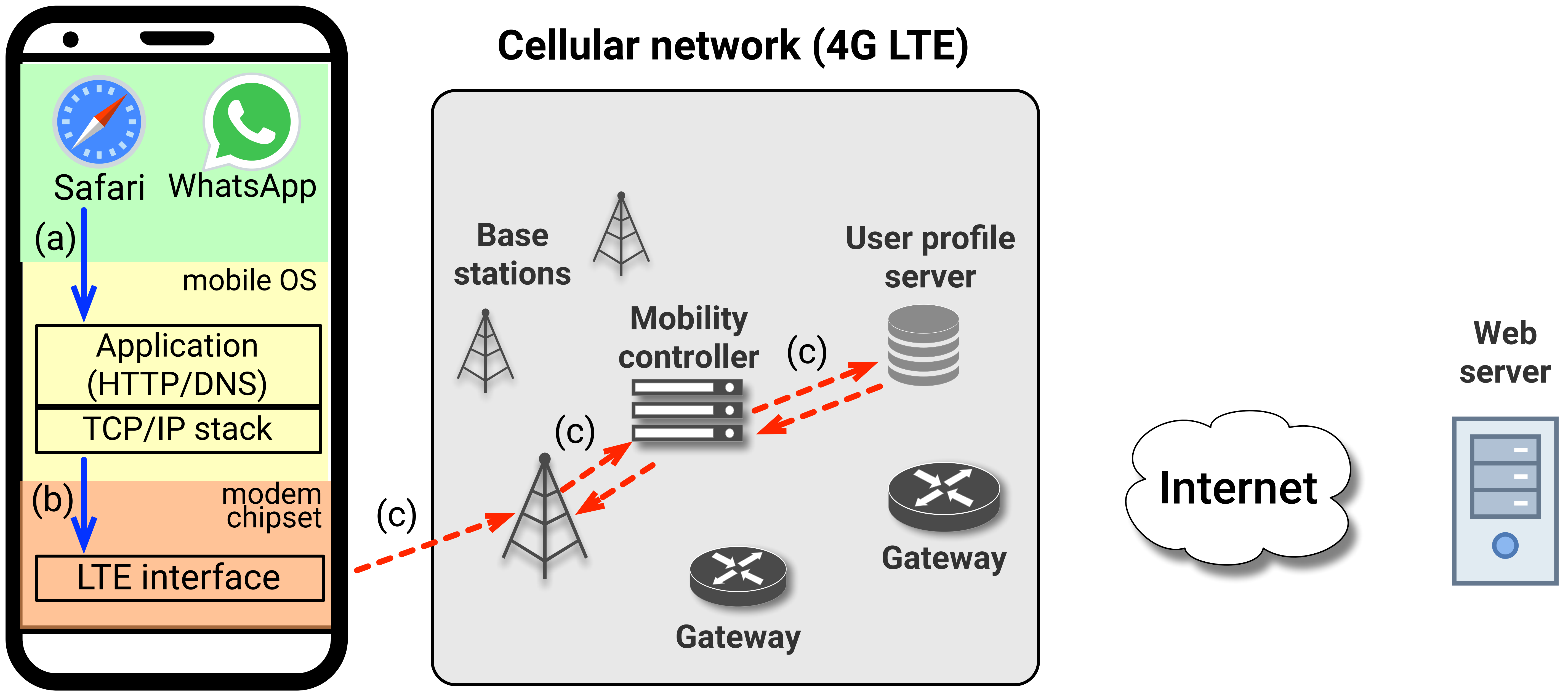
How LTE impacts perceived latency on mobile web/IM app?



Background: how do mobile apps work over 4G LTE?

What happens under the hood?

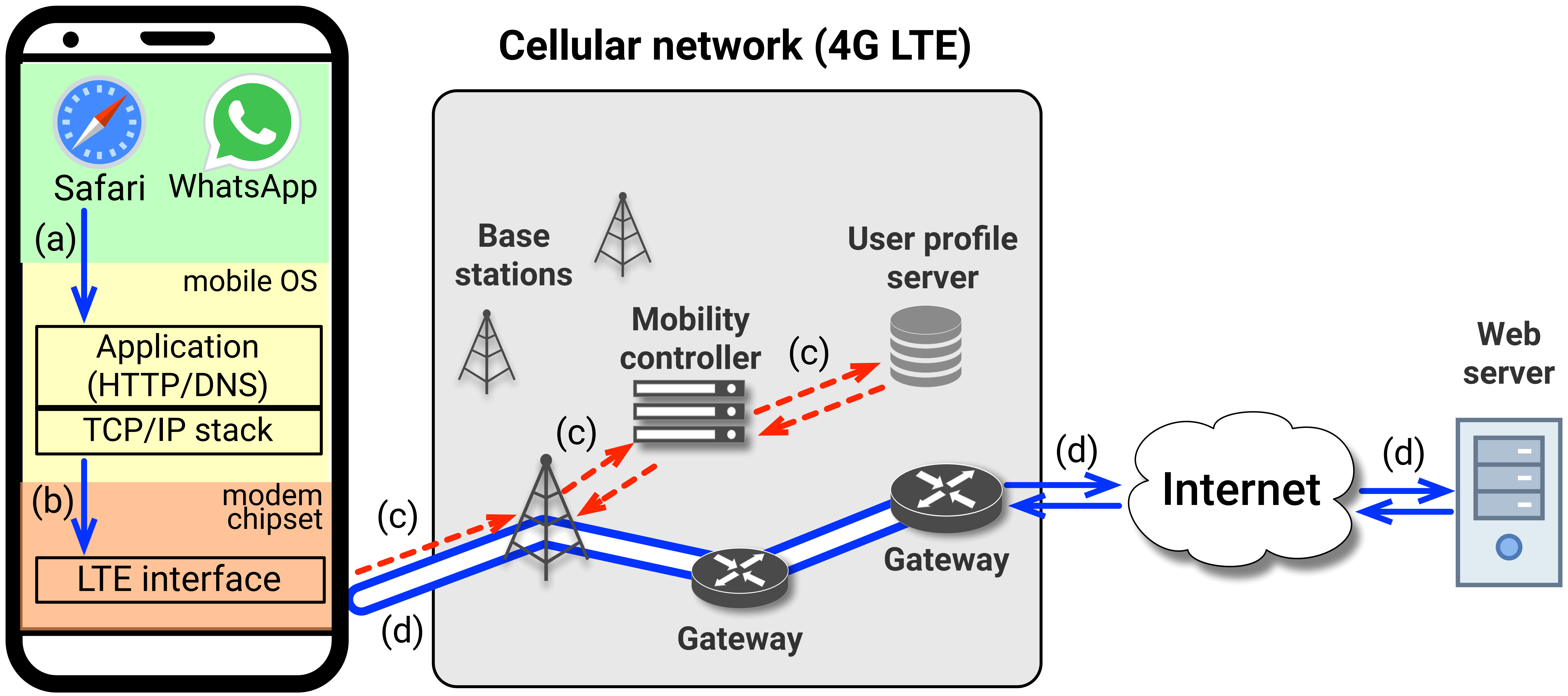
How LTE impacts perceived latency on mobile web/IM app?



Background: how do mobile apps work over 4G LTE?

What happens under the hood?

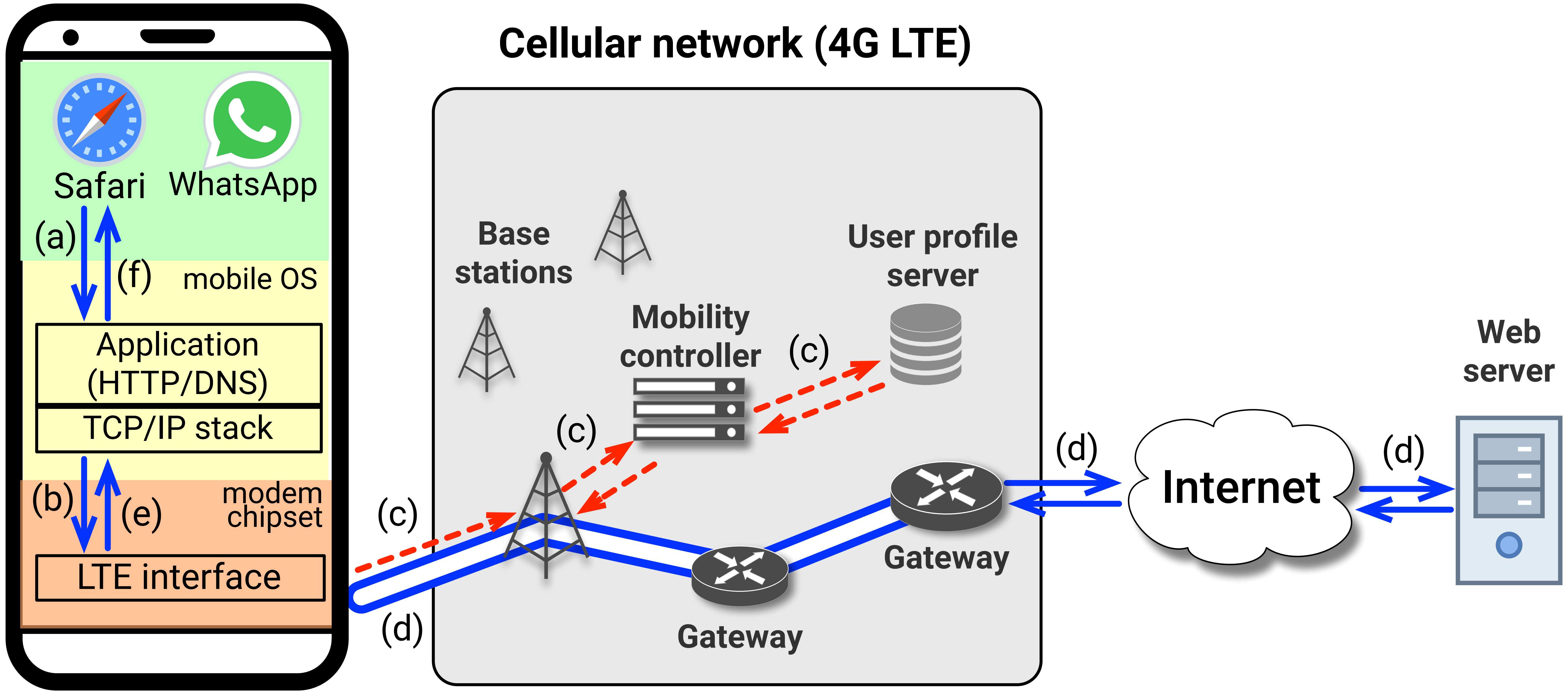
How LTE impacts perceived latency on mobile web/IM app?



Background: how do mobile apps work over 4G LTE?

What happens under the hood?

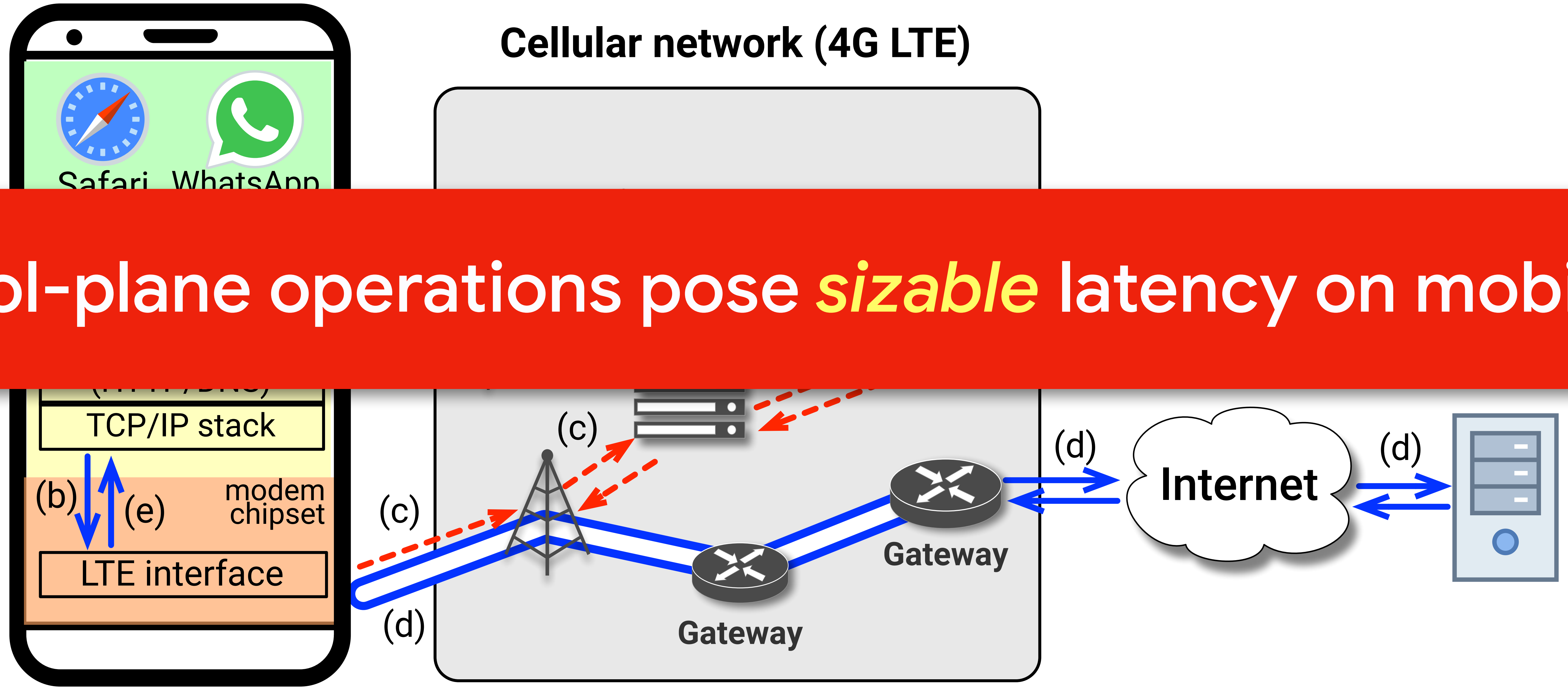
How LTE impacts perceived latency on mobile web/IM app?



Background: how do mobile apps work over 4G LTE?

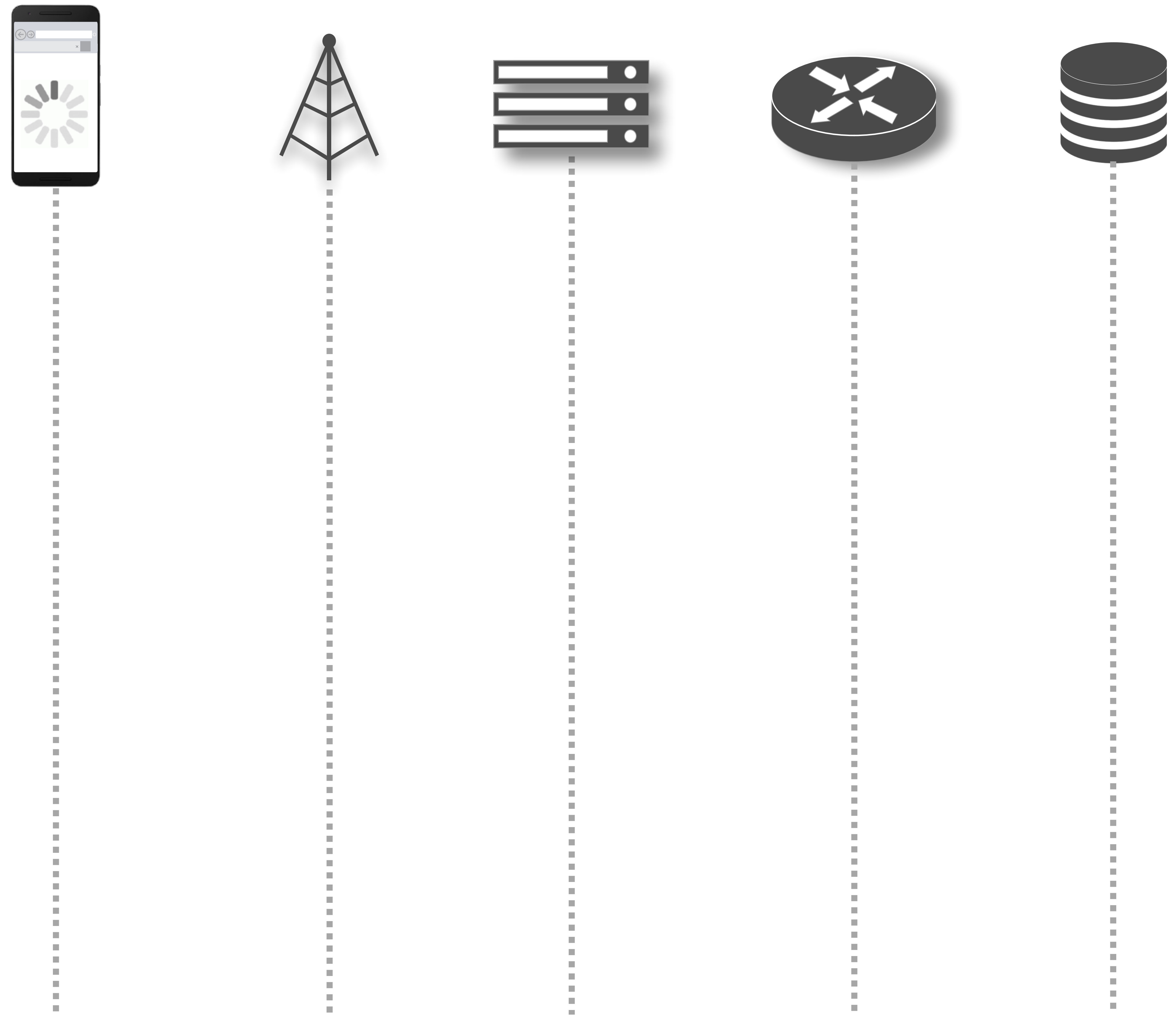
What happens under the hood?

How LTE impacts perceived latency on mobile web/IM app?

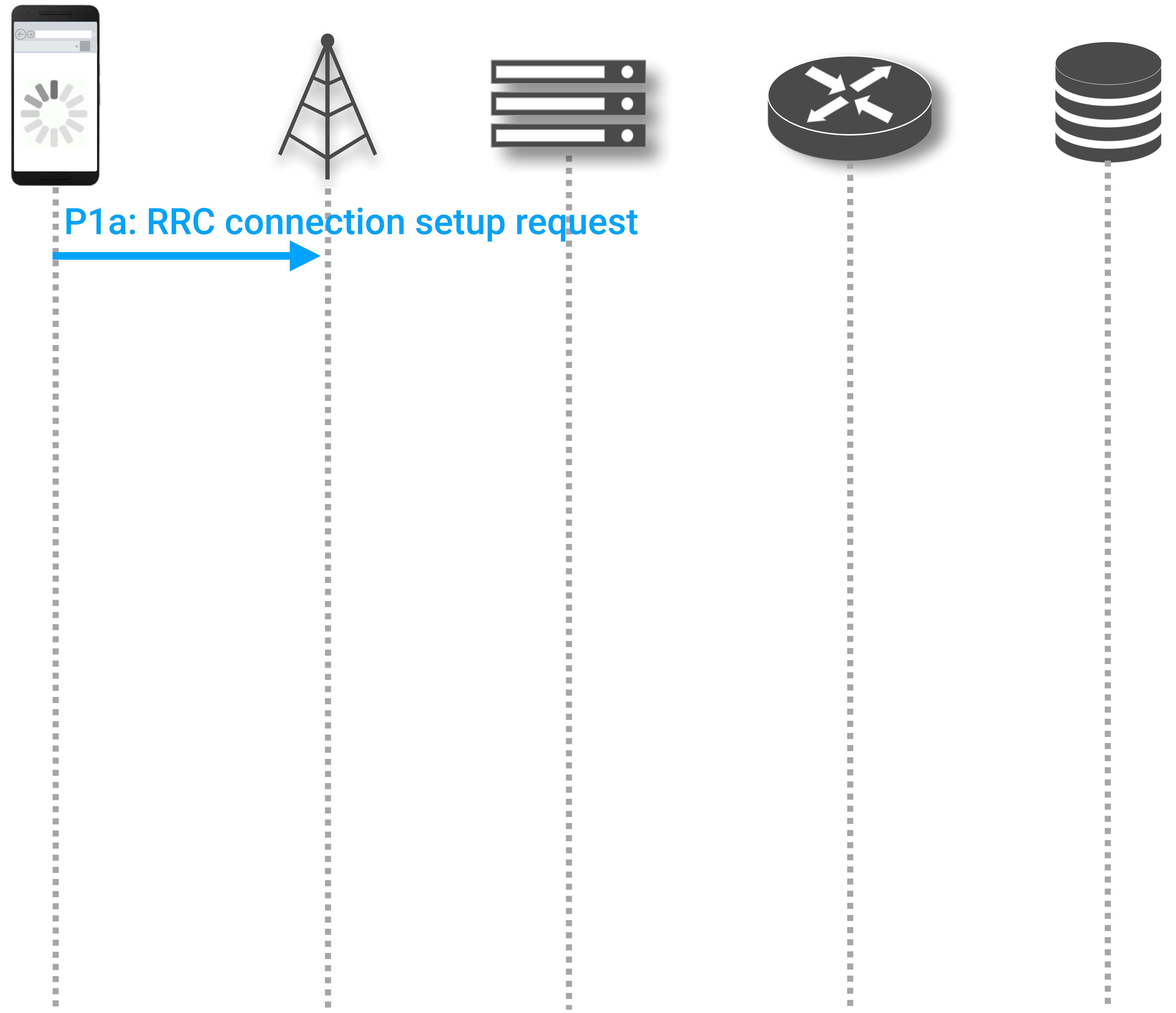


LTE control-plane operations pose *sizable* latency on mobile apps

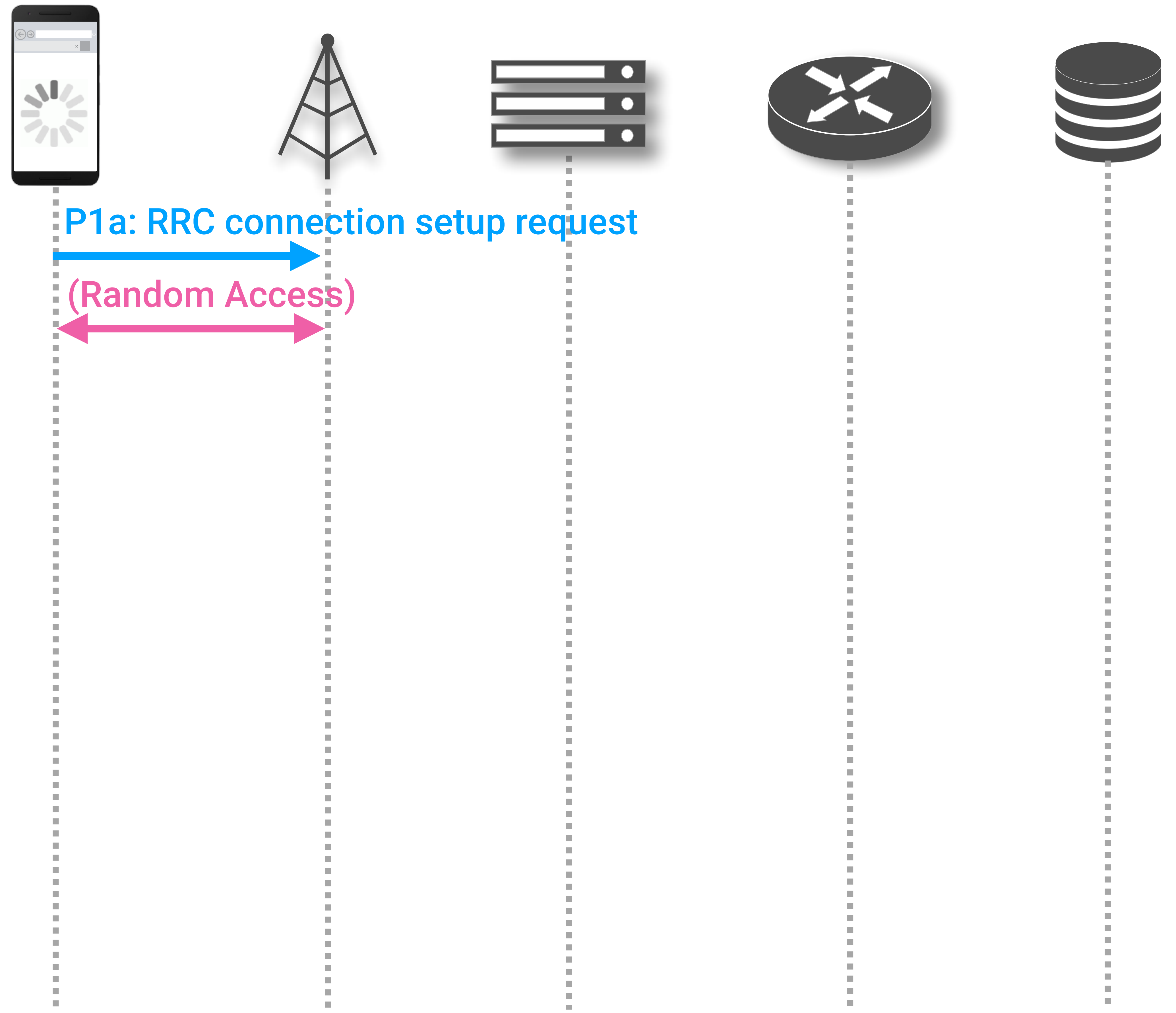
Timing breakdown of control plane operations



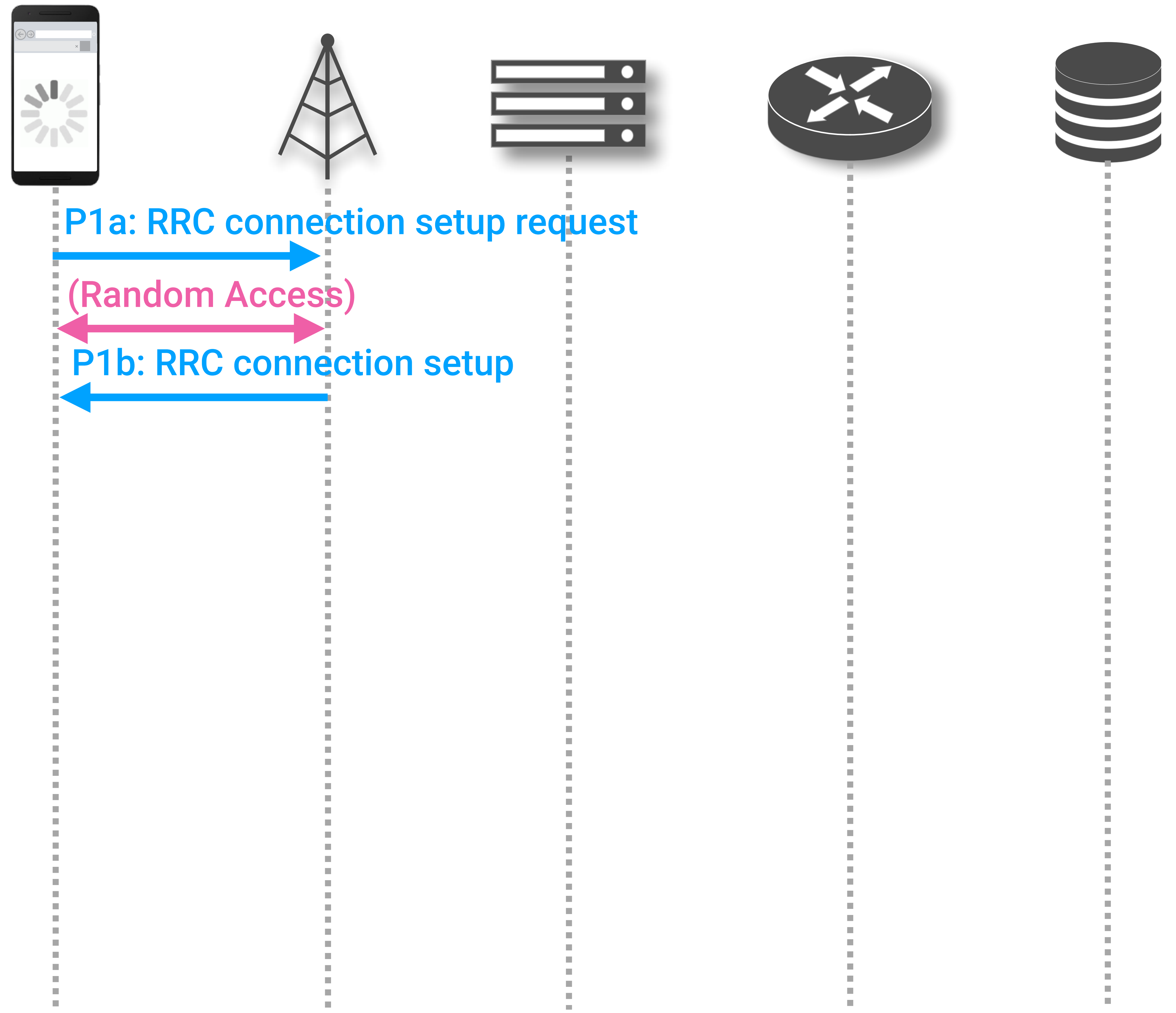
Timing breakdown of control plane operations



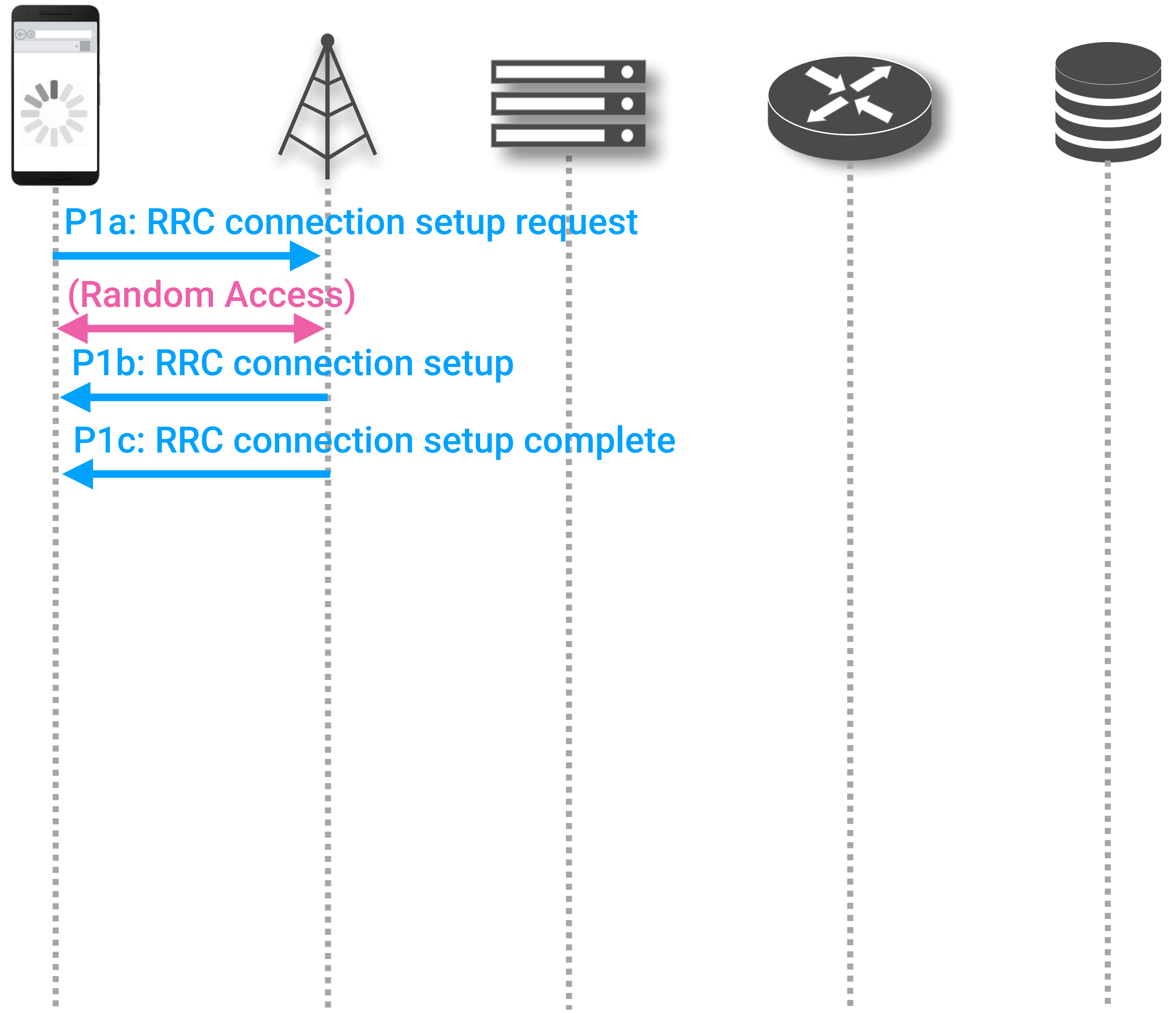
Timing breakdown of control plane operations



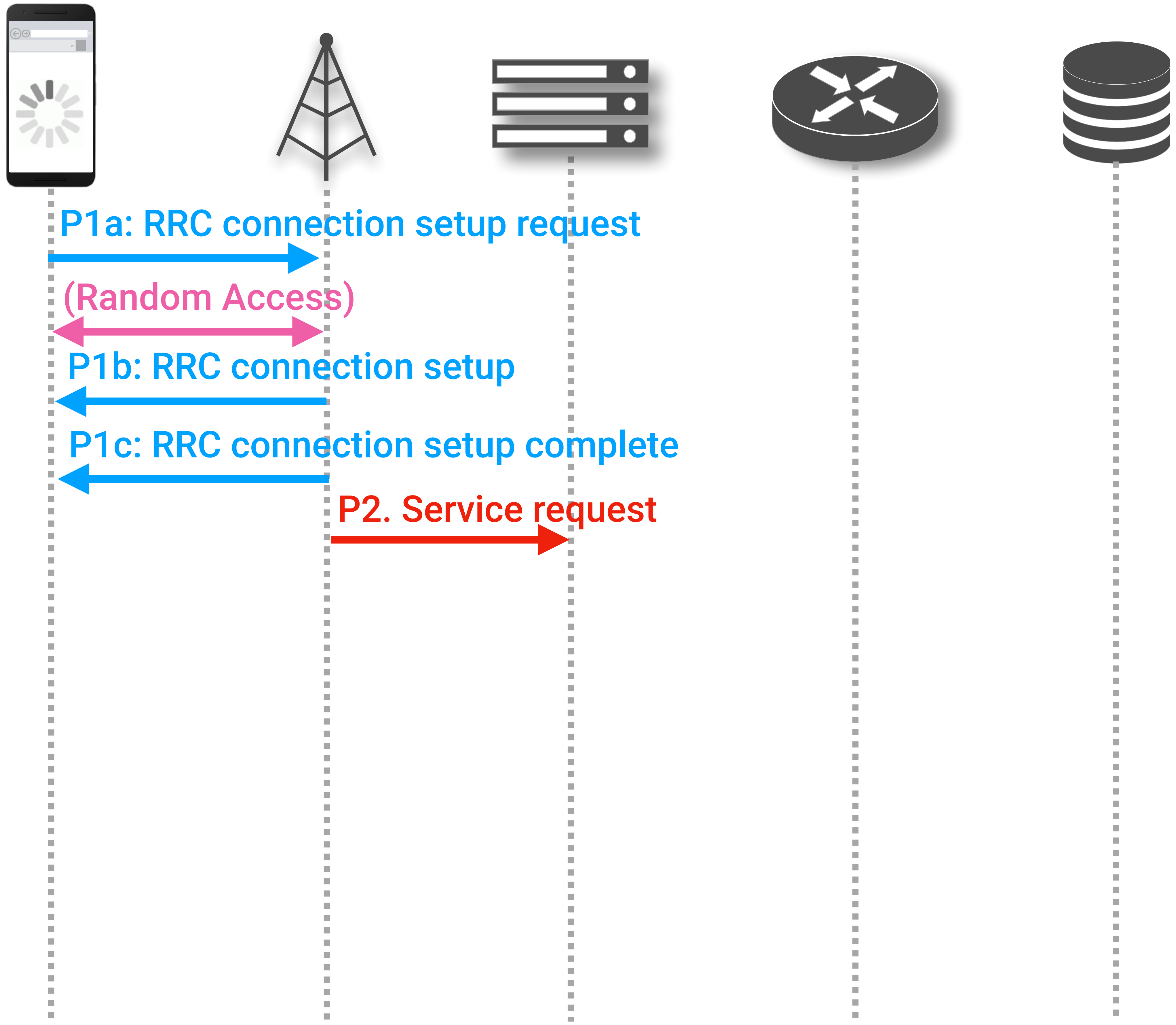
Timing breakdown of control plane operations



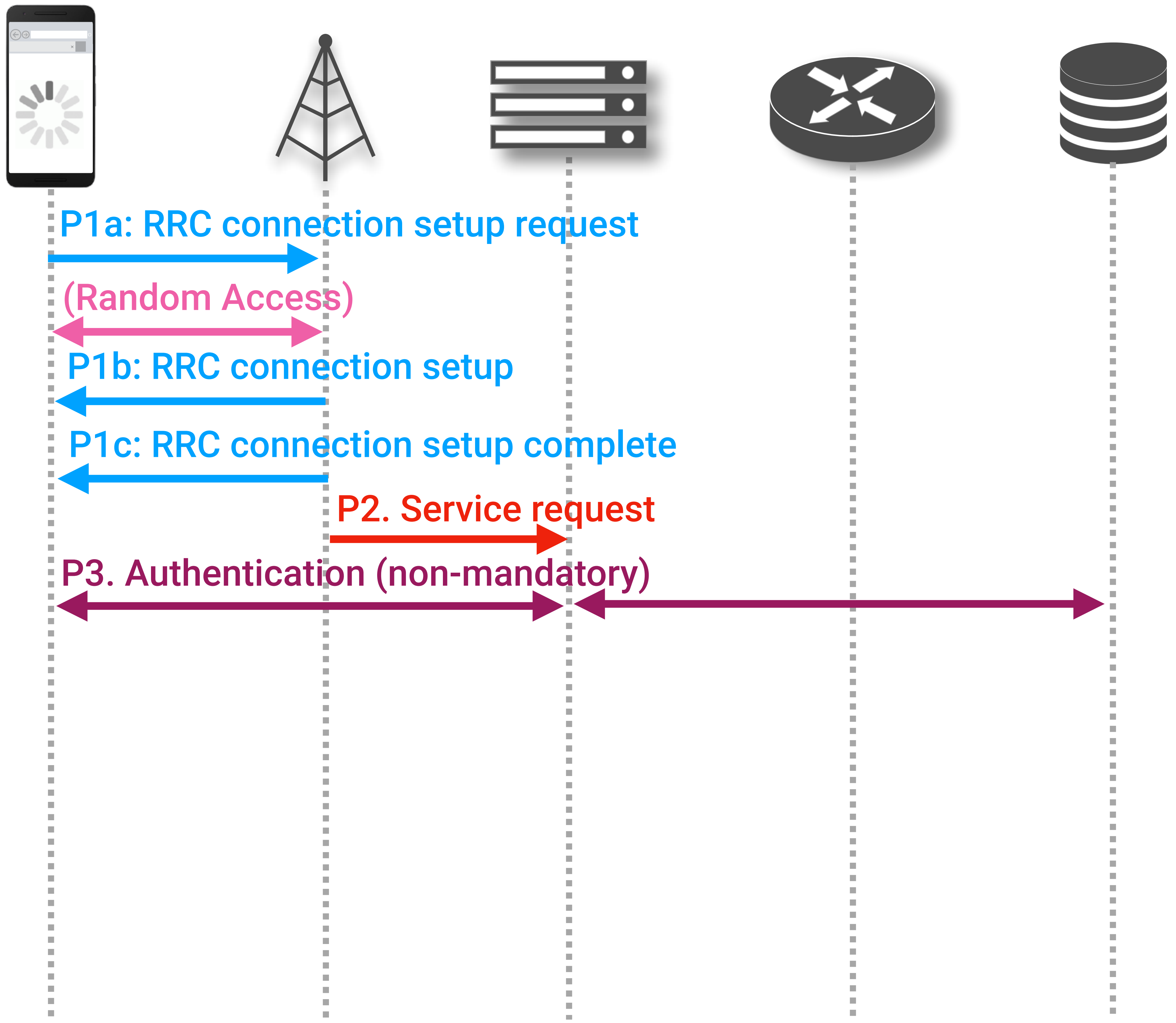
Timing breakdown of control plane operations



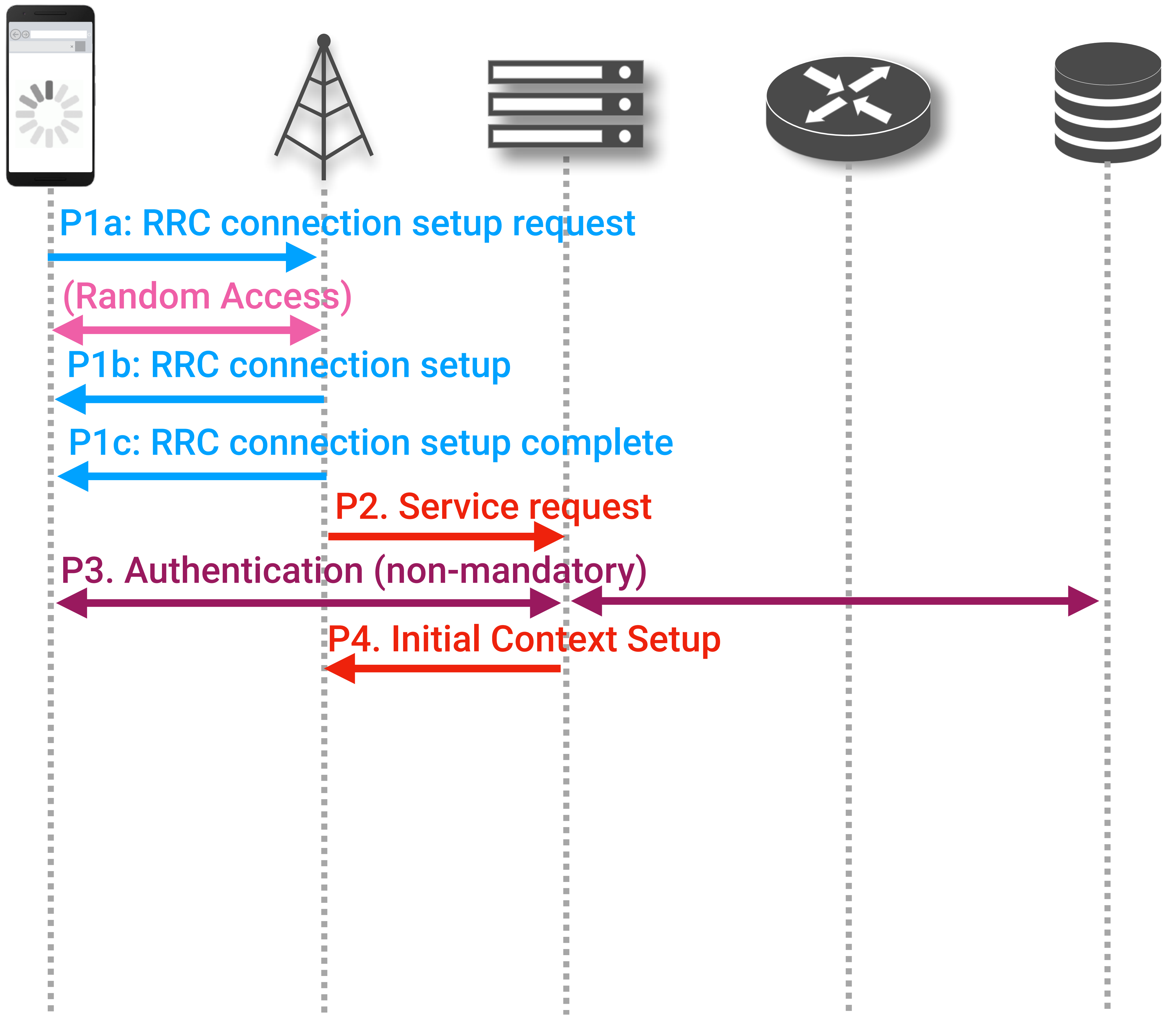
Timing breakdown of control plane operations



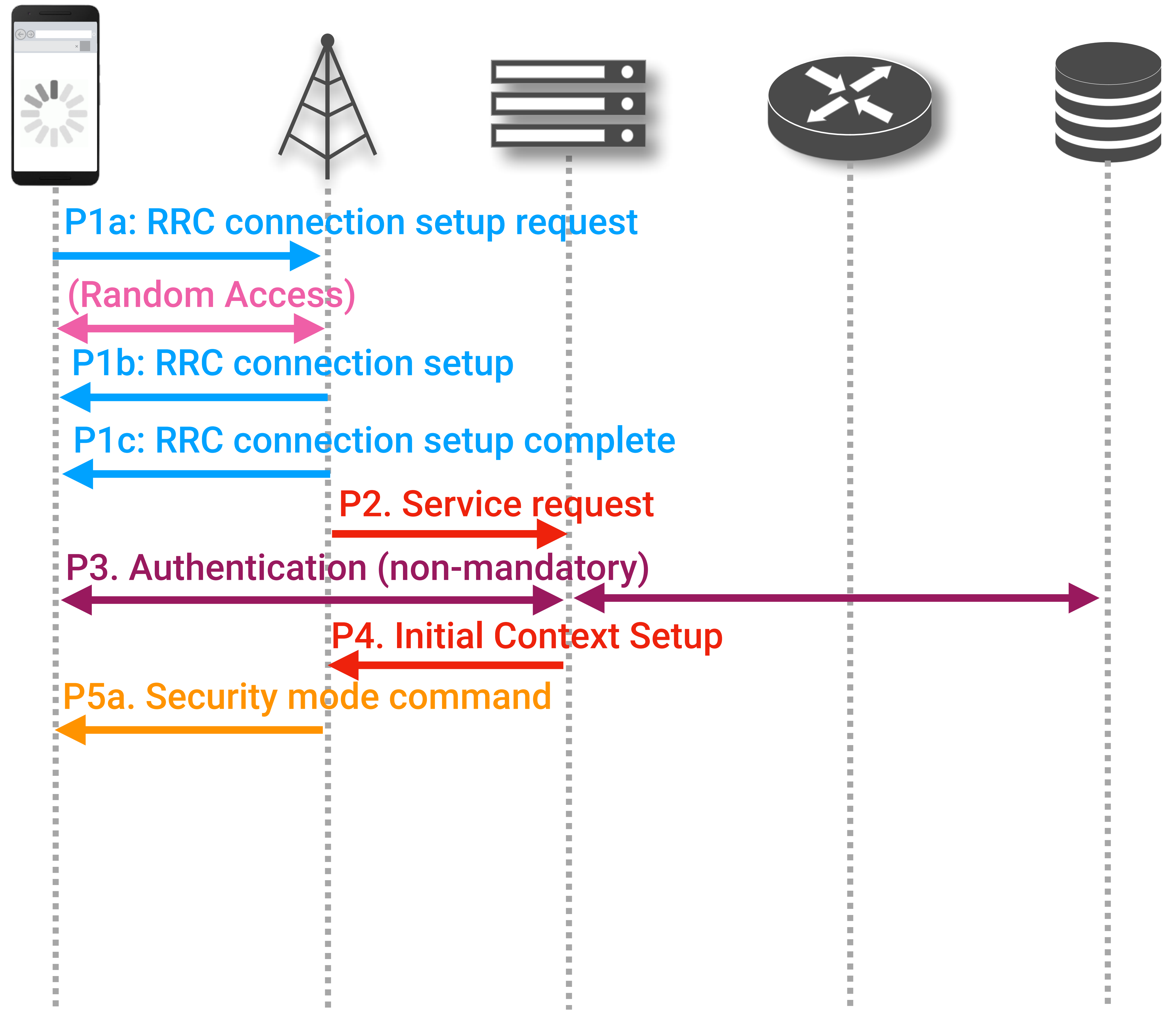
Timing breakdown of control plane operations



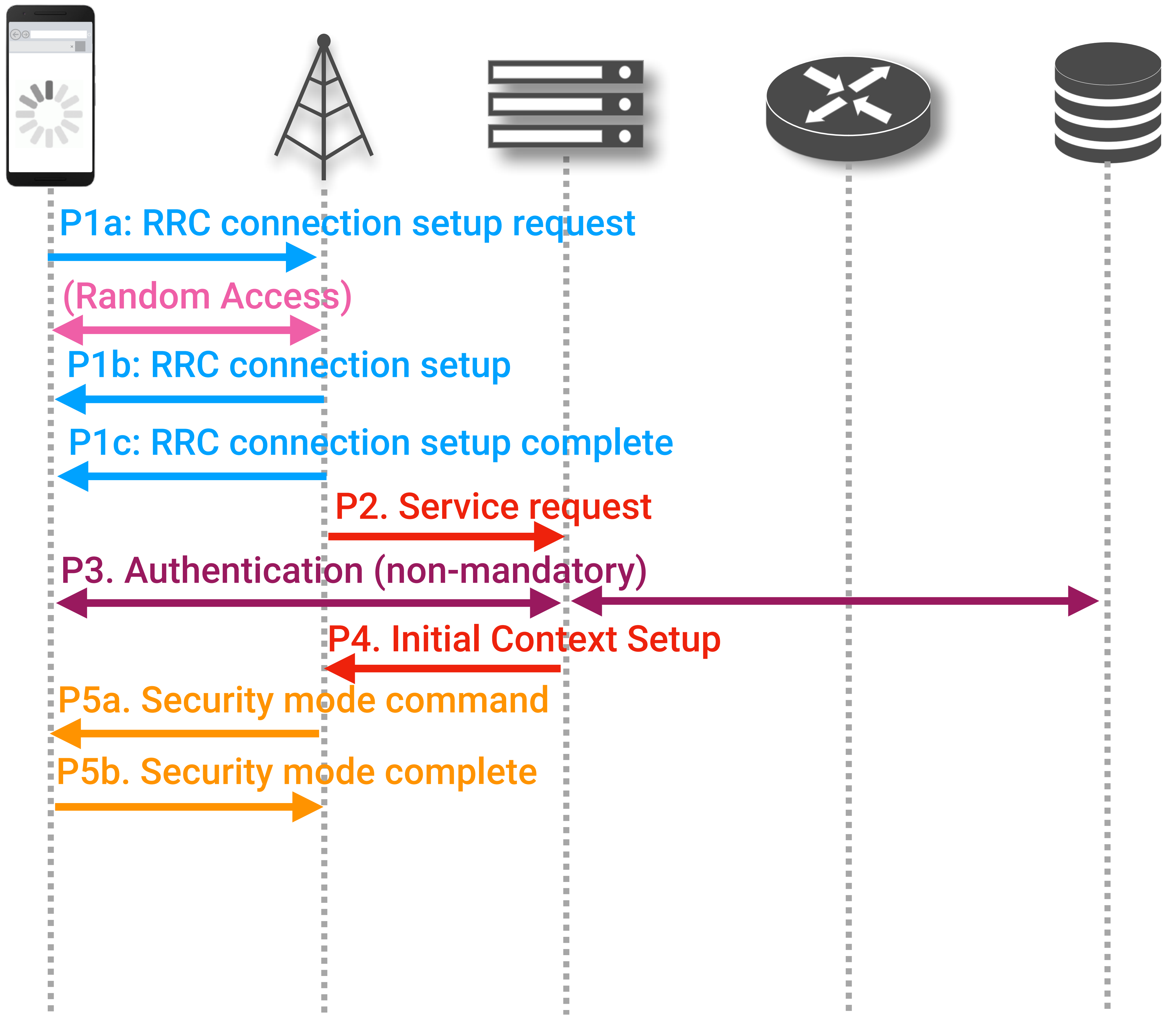
Timing breakdown of control plane operations



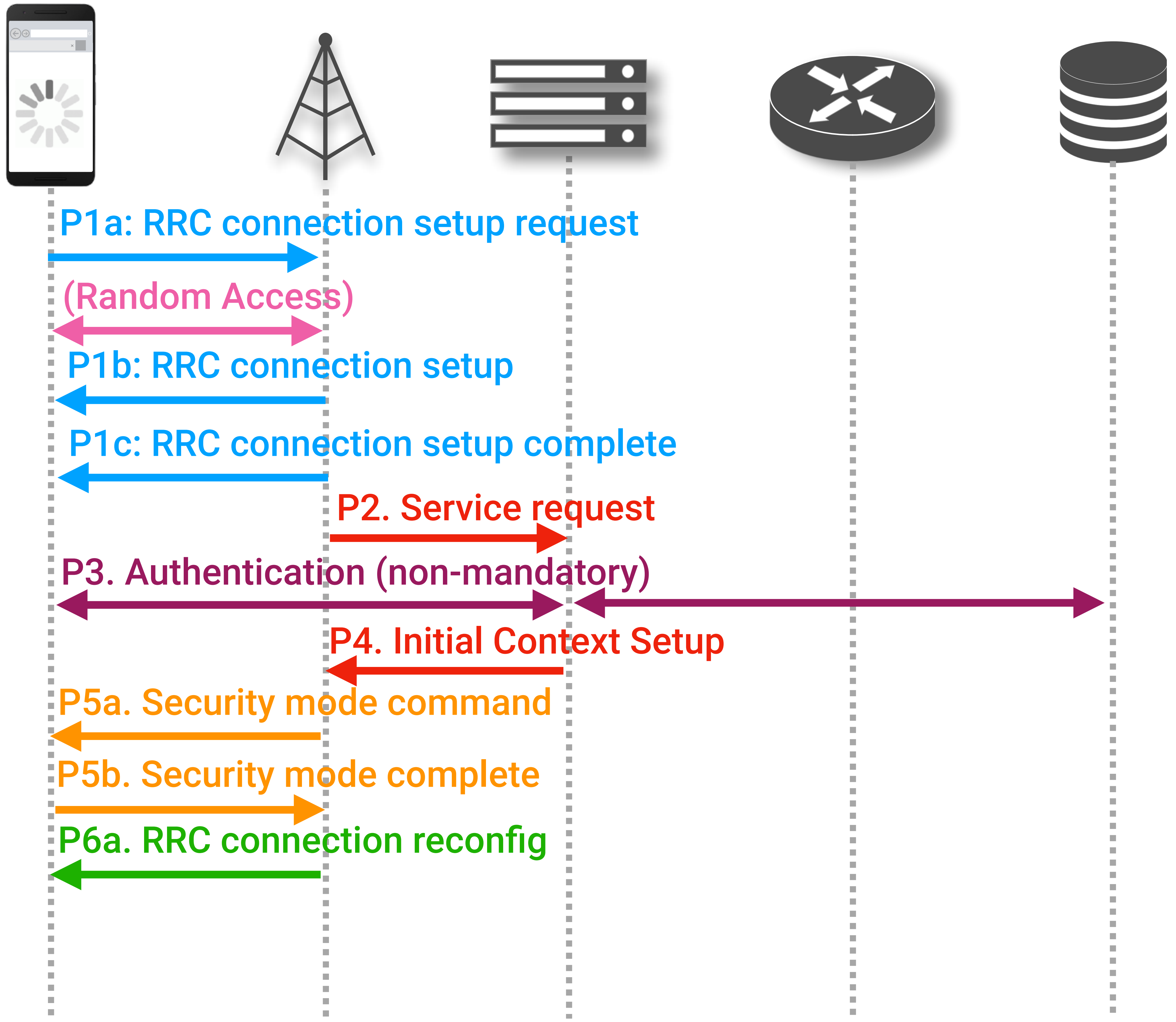
Timing breakdown of control plane operations



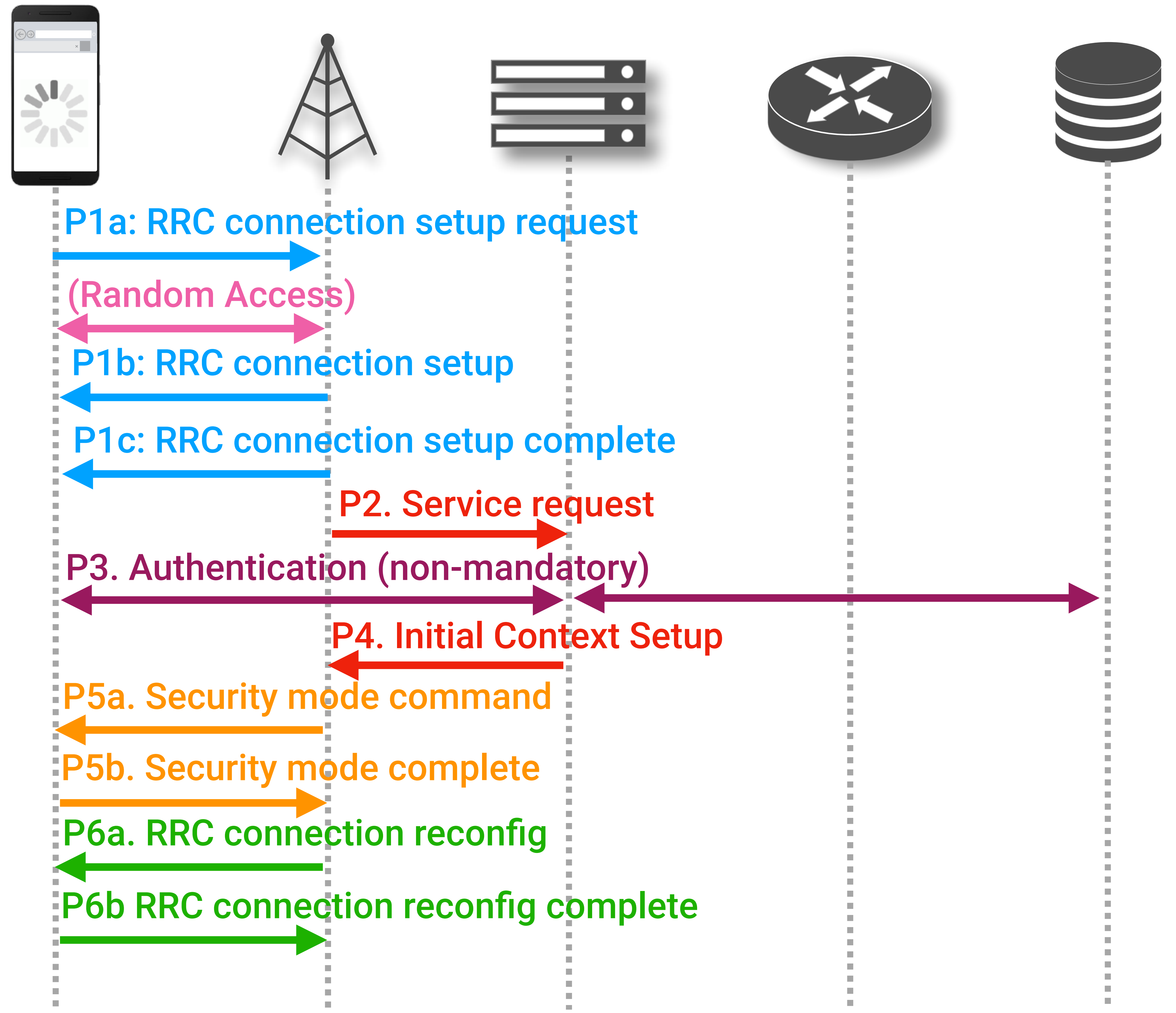
Timing breakdown of control plane operations



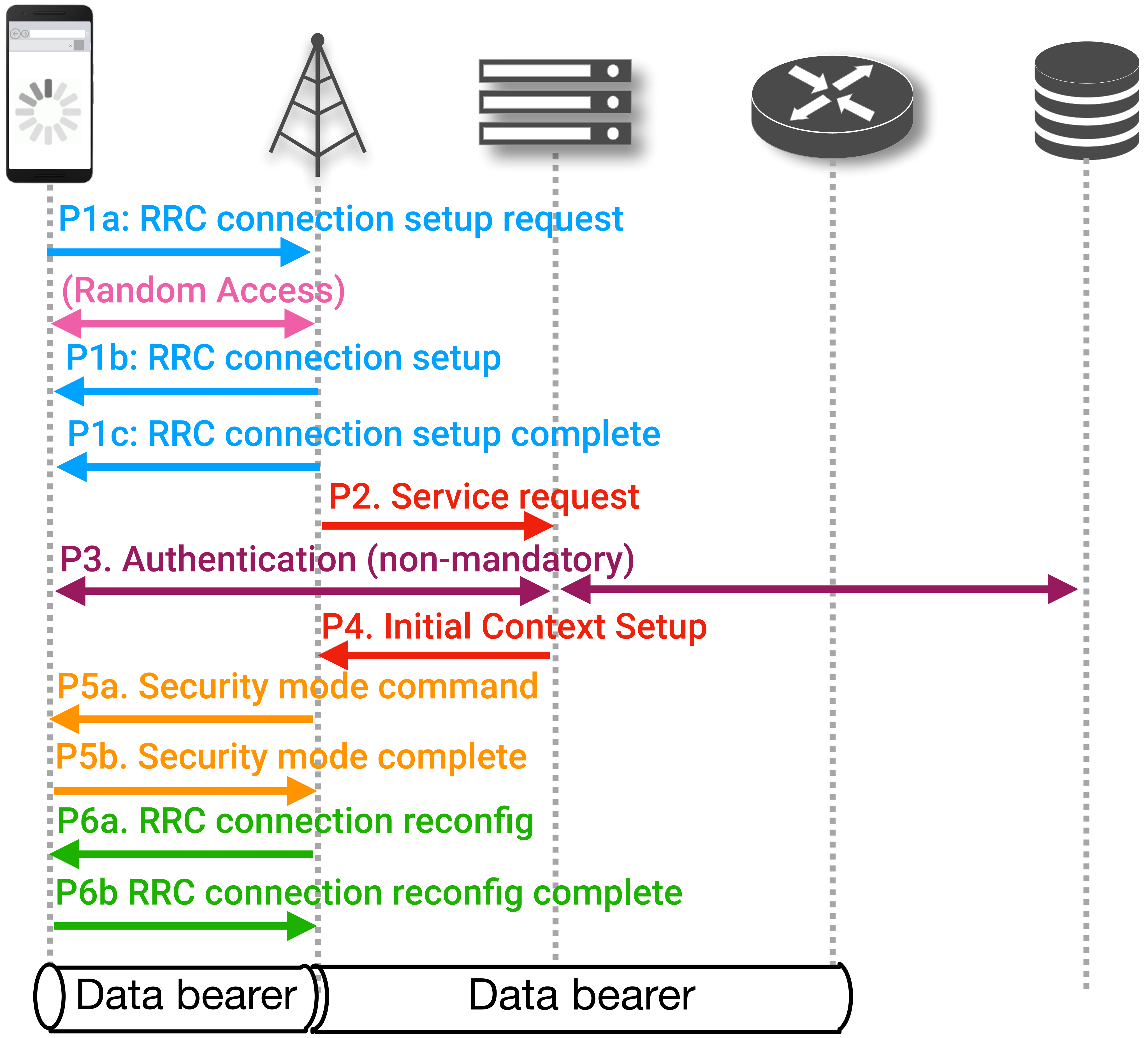
Timing breakdown of control plane operations



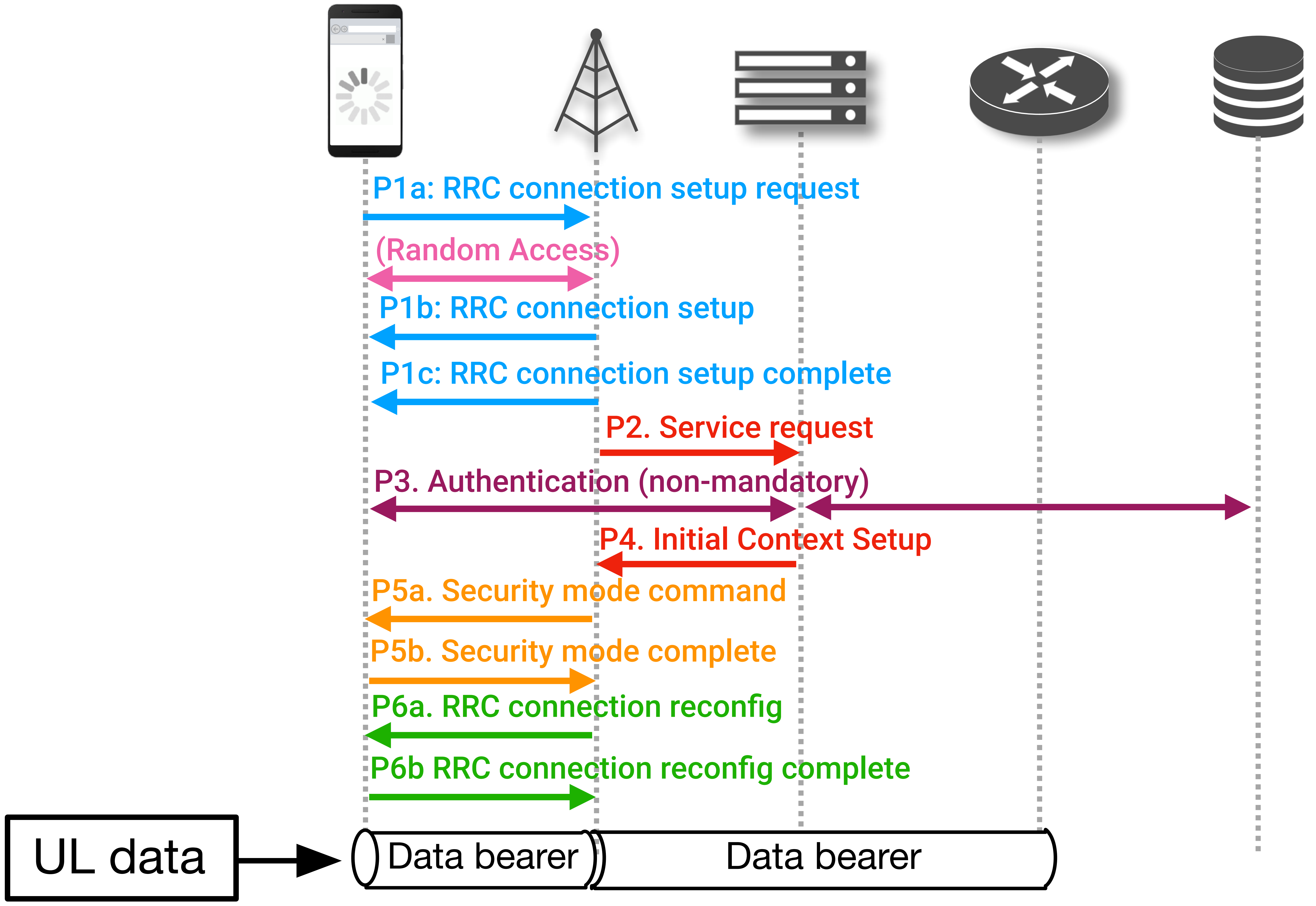
Timing breakdown of control plane operations



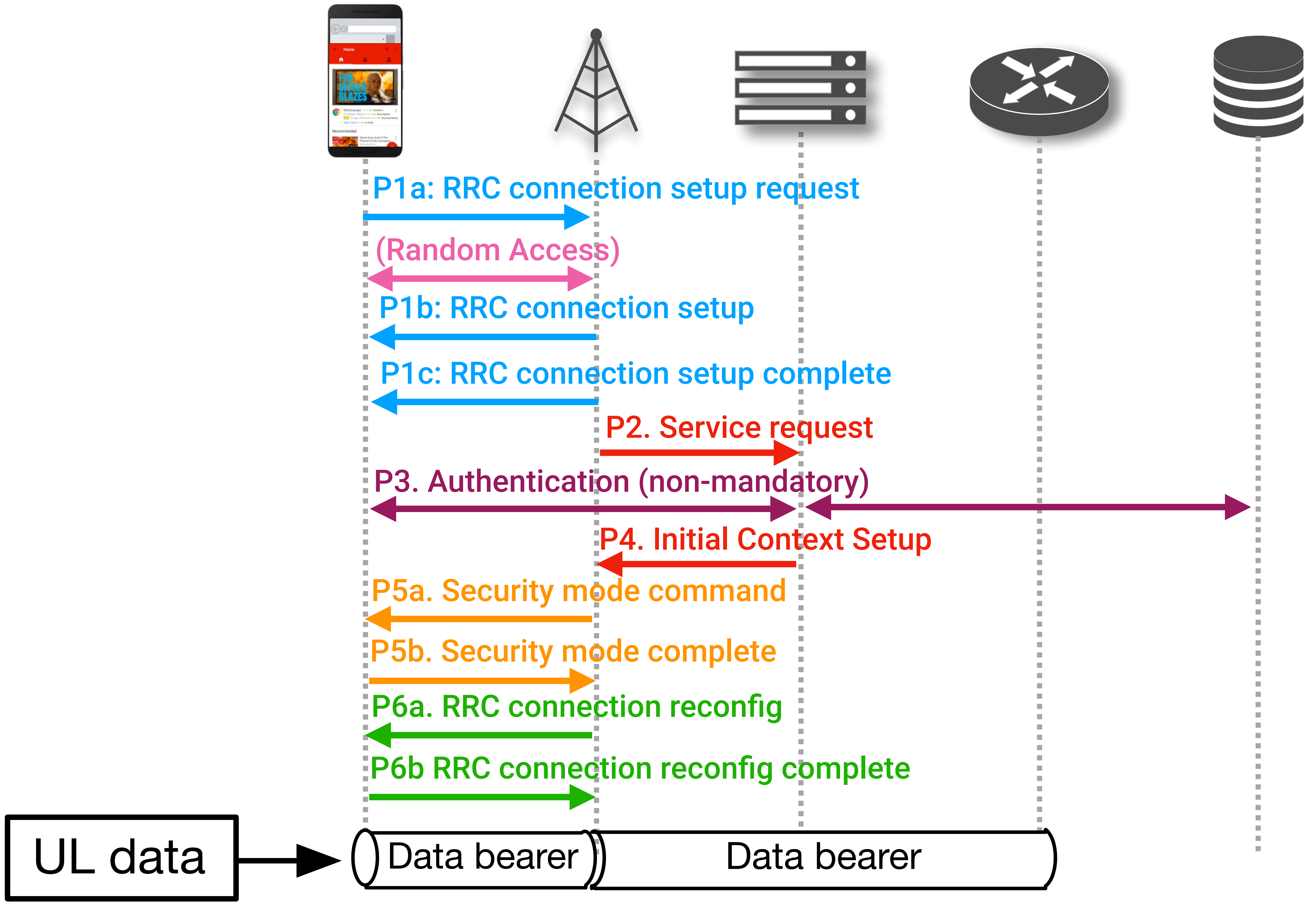
Timing breakdown of control plane operations



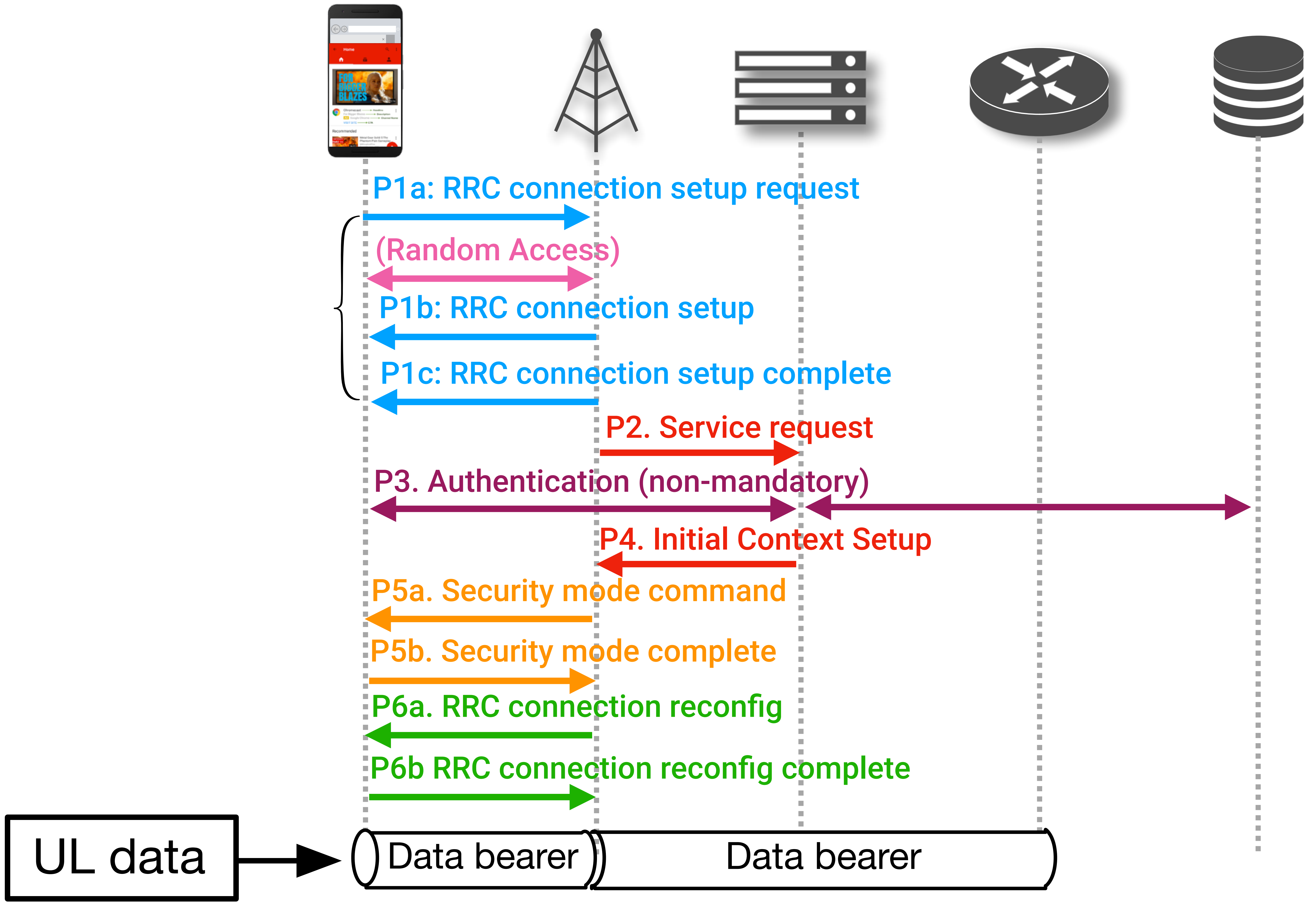
Timing breakdown of control plane operations



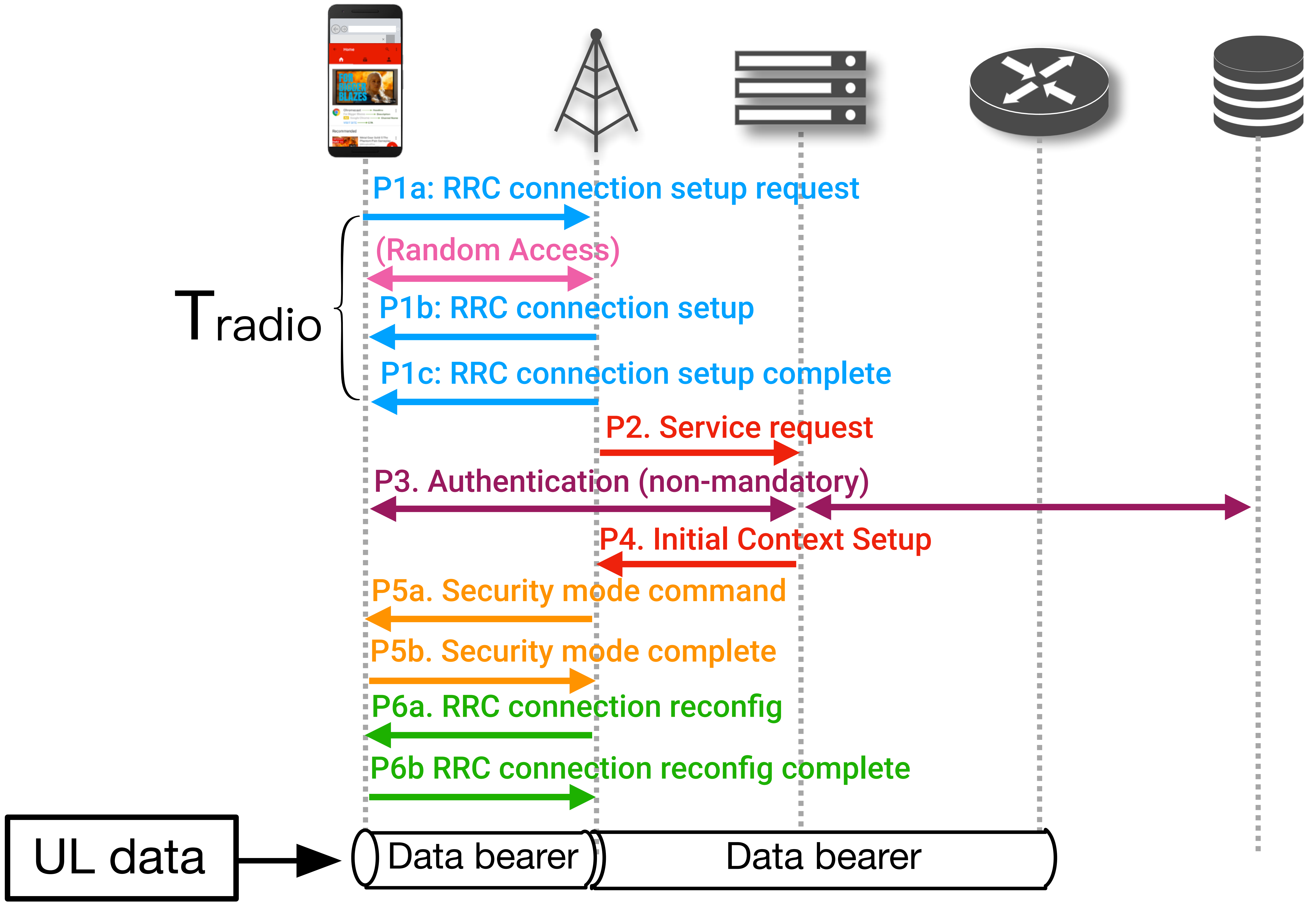
Timing breakdown of control plane operations



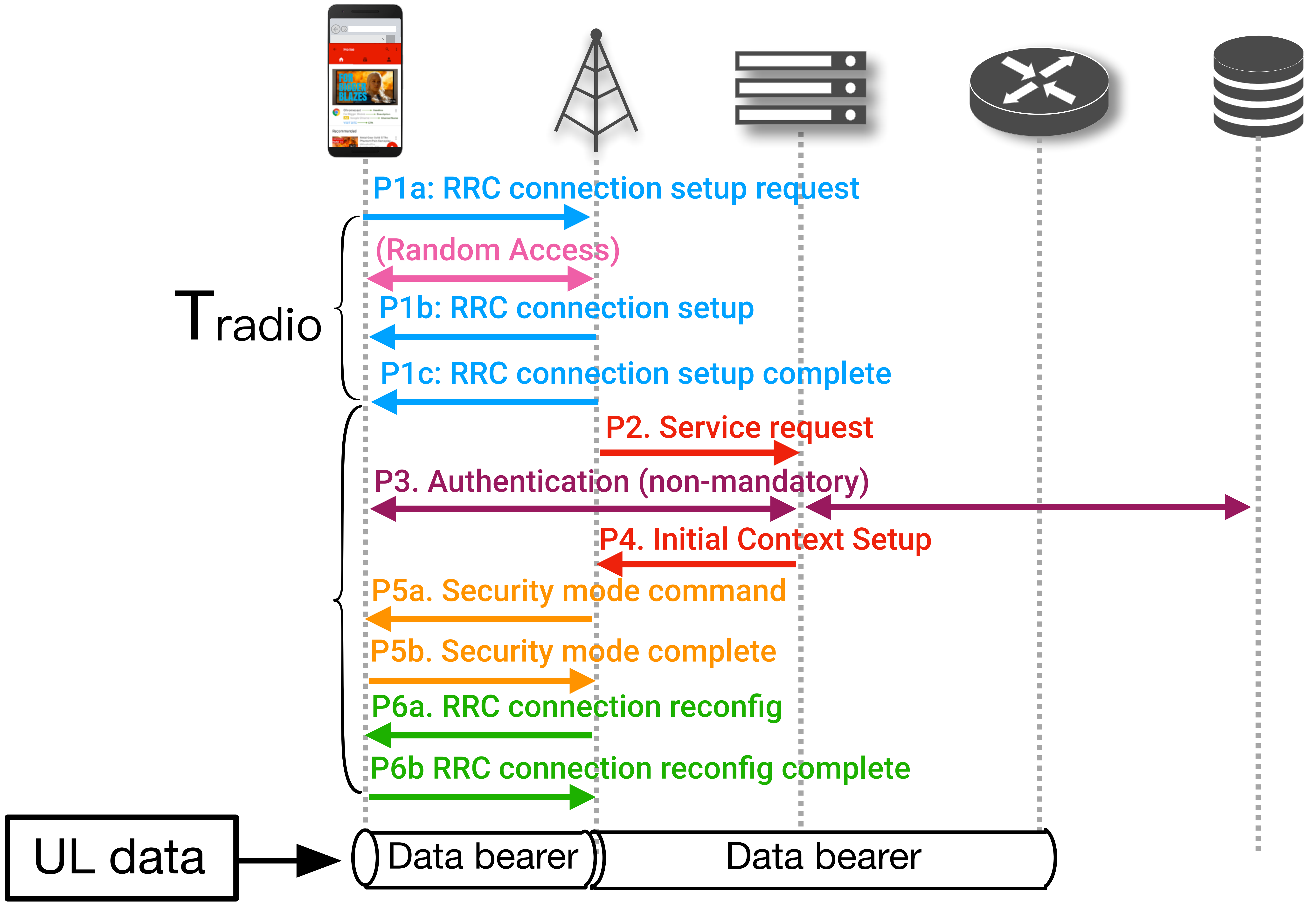
Timing breakdown of control plane operations



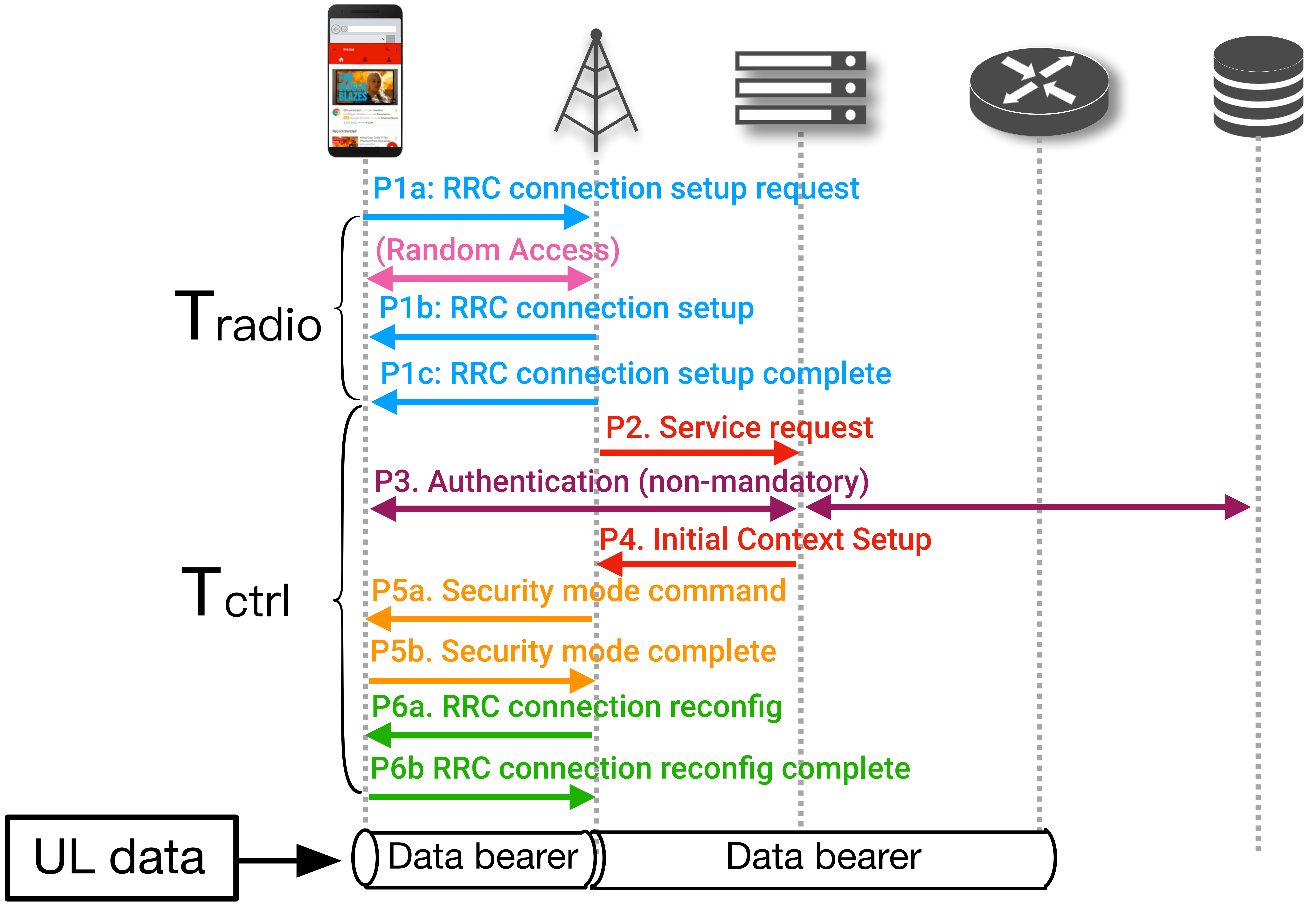
Timing breakdown of control plane operations



Timing breakdown of control plane operations



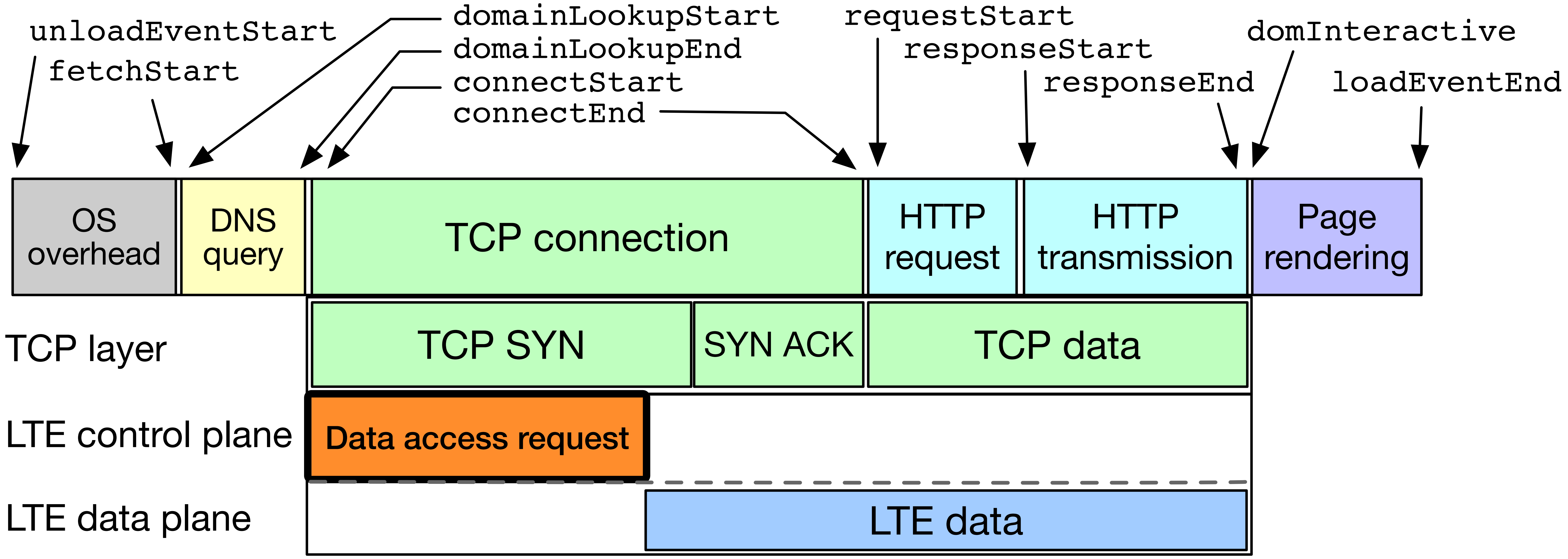
Timing breakdown of control plane operations



Learning latency: latency data sensing

Three-tiered timing data collection:

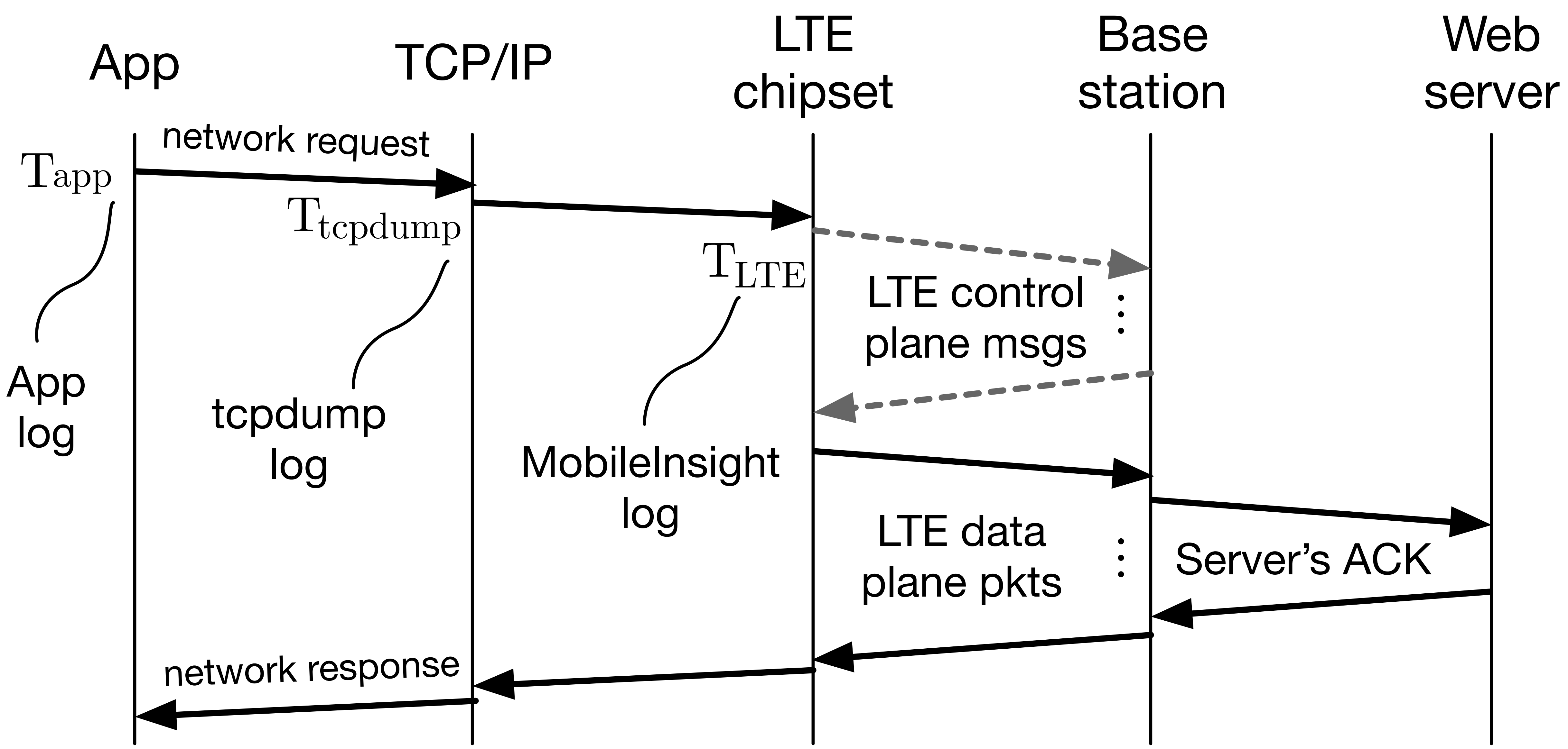
- App-specific semantic timing (e.g. Navigation Timing API, IM timing model)
- TCP/IP stack timing (from TCPDUMP)
- LTE stack timing (from MobileInsight)



Challenge: timestamp alignment

How to align timestamps at these layers?

- Domain-specific event tracing and mapping
- Machine-learning assisted



Pinpoint latency bottleneck in LTE: An example

Pinpoint latency bottleneck in LTE: An example

Run a small webpage (4 KB) in Chrome on Android

- User is static, under good 4G LTE signal (-95 dBm), T-Mobile

Pinpoint latency bottleneck in LTE: An example

Run a small webpage (4 KB) in Chrome on Android

- User is static, under good 4G LTE signal (-95 dBm), T-Mobile

Total Latency: 473 msec

- Clicking URL → page loading complete, Steps (a)—(f)

Pinpoint latency bottleneck in LTE: An example

Run a small webpage (4 KB) in Chrome on Android

- User is static, under good 4G LTE signal (-95 dBm), T-Mobile

Total Latency: 473 msec

- Clicking URL → page loading complete, Steps (a)—(f)

Pinpointing the latency bottleneck

- How to breakdown?

Control-plane latency breakdown: local analysis I

Control-plane latency breakdown: local analysis I

Major component from Navigation Timing API: DNS lookup, 250 ms out of 473 ms

Control-plane latency breakdown: local analysis I

Major component from Navigation Timing API: DNS lookup, 250 ms out of 473 ms

Is the DNS server slow to handle connection?

Control-plane latency breakdown: local analysis I

Major component from Navigation Timing API: DNS lookup, 250 ms out of 473 ms

Is the DNS server slow to handle connection?

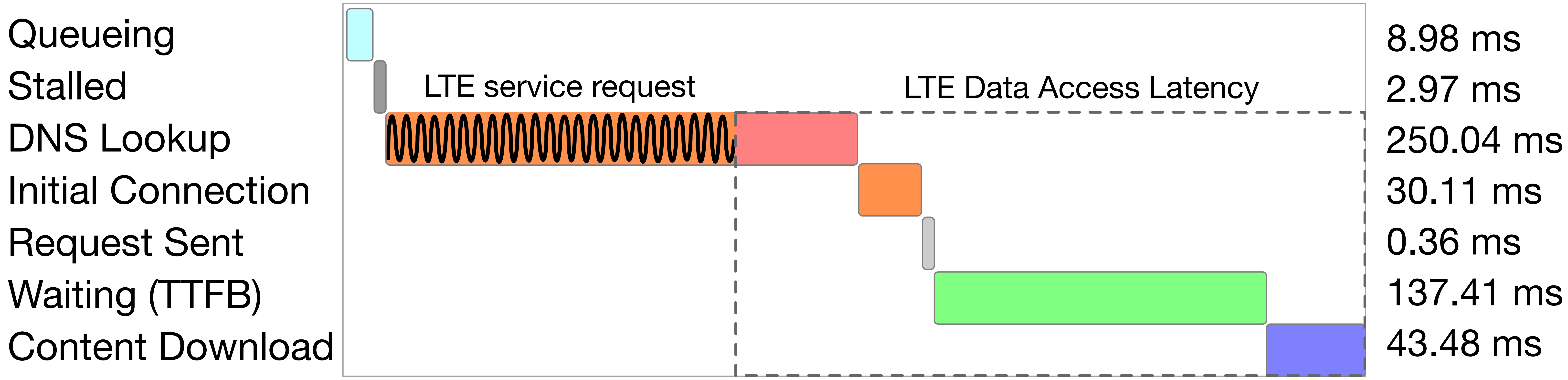
Further breakdown: *LTE service request* takes 172 ms before the DNS setup

Control-plane latency breakdown: local analysis I

Major component from Navigation Timing API: DNS lookup, 250 ms out of 473 ms

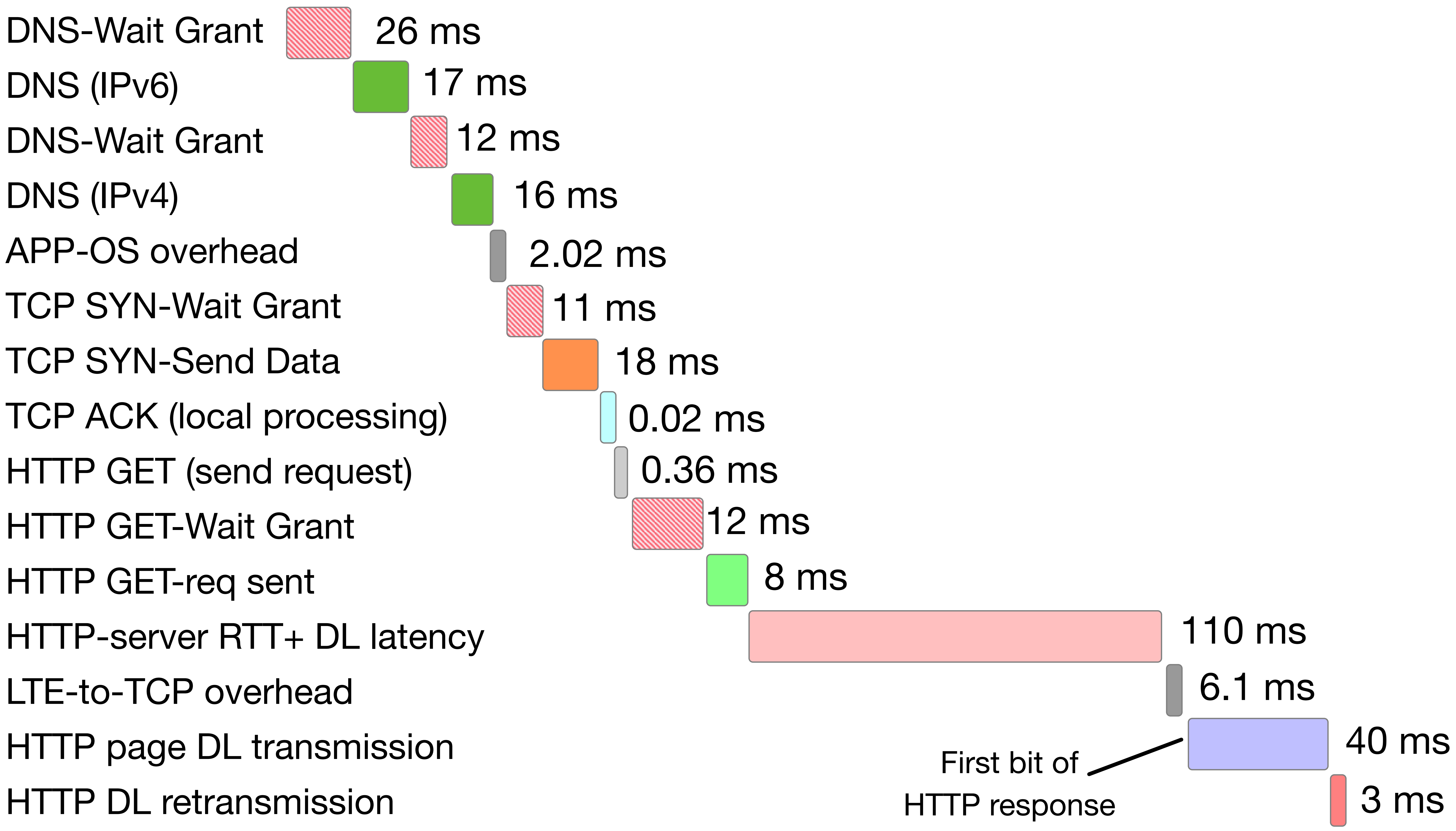
Is the DNS server slow to handle connection?

Further breakdown: *LTE service request* takes 172 ms before the DNS setup



Data-plane latency breakdown: local analysis II

Further zoom in and breakdown the remaining LTE data access latency (291 ms):

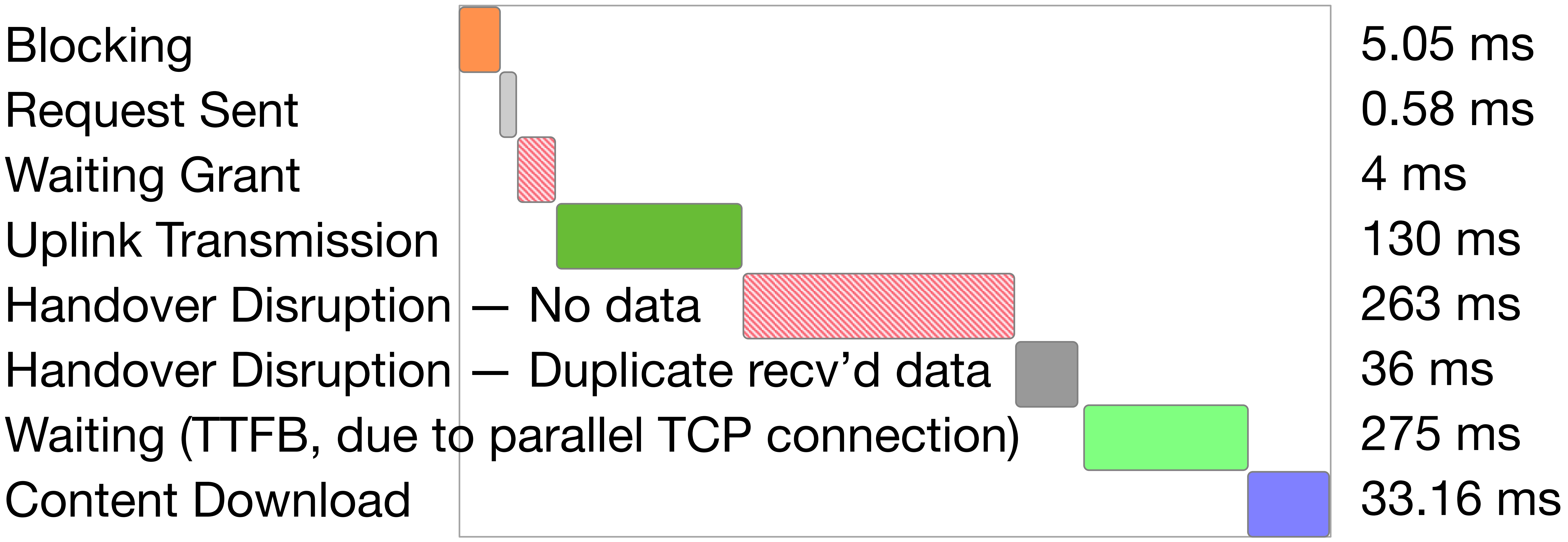


Latency mapping for failures: local analysis III

Example: data plane suspension due to radio reconnection and head-of-line blocking during handover

Latency mapping for failures: local analysis III

Example: data plane suspension due to radio reconnection and head-of-line blocking during handover



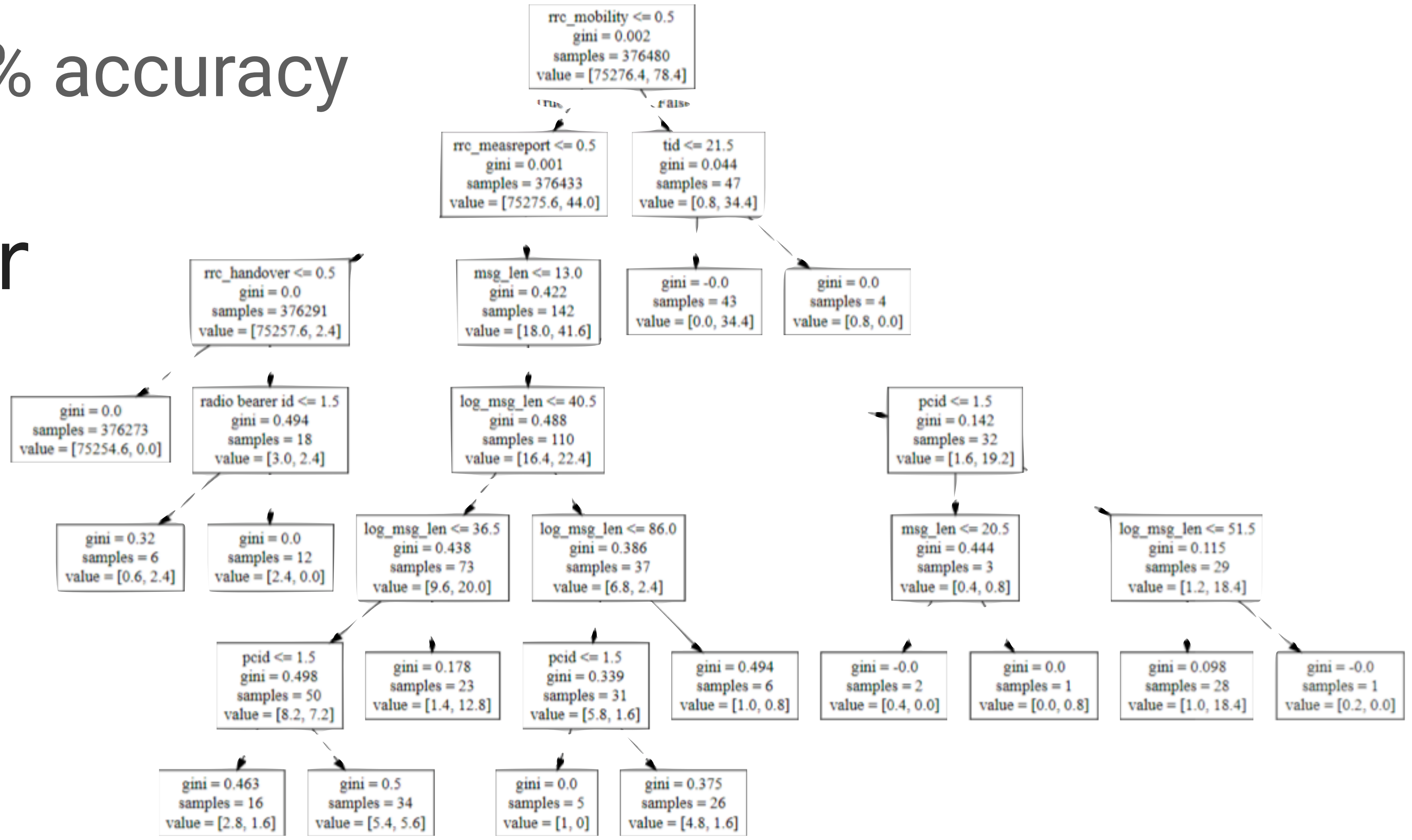
Machine learning scheme

We leverage domain-specific knowledge for ML-based predictions

Control plane: predict handover using a decision tree classifier

- Features from 3GPP standards
- Predicts handover 100ms before it occurs with >99% accuracy

Data plane: predict NACK/ACK flip at MAC layer



Synthesizer: global crowdsourcing analysis

Synthesizer: global crowdsourcing analysis

Four US carriers + Google Project Fi

Synthesizer: global crowdsourcing analysis

Four US carriers + Google Project Fi

23 phone models, 95,057 data sessions

Synthesizer: global crowdsourcing analysis

Four US carriers + Google Project Fi

23 phone models, 95,057 data sessions

Overall latency: 77 — 2956 ms in 500K samples

Synthesizer: global crowdsourcing analysis

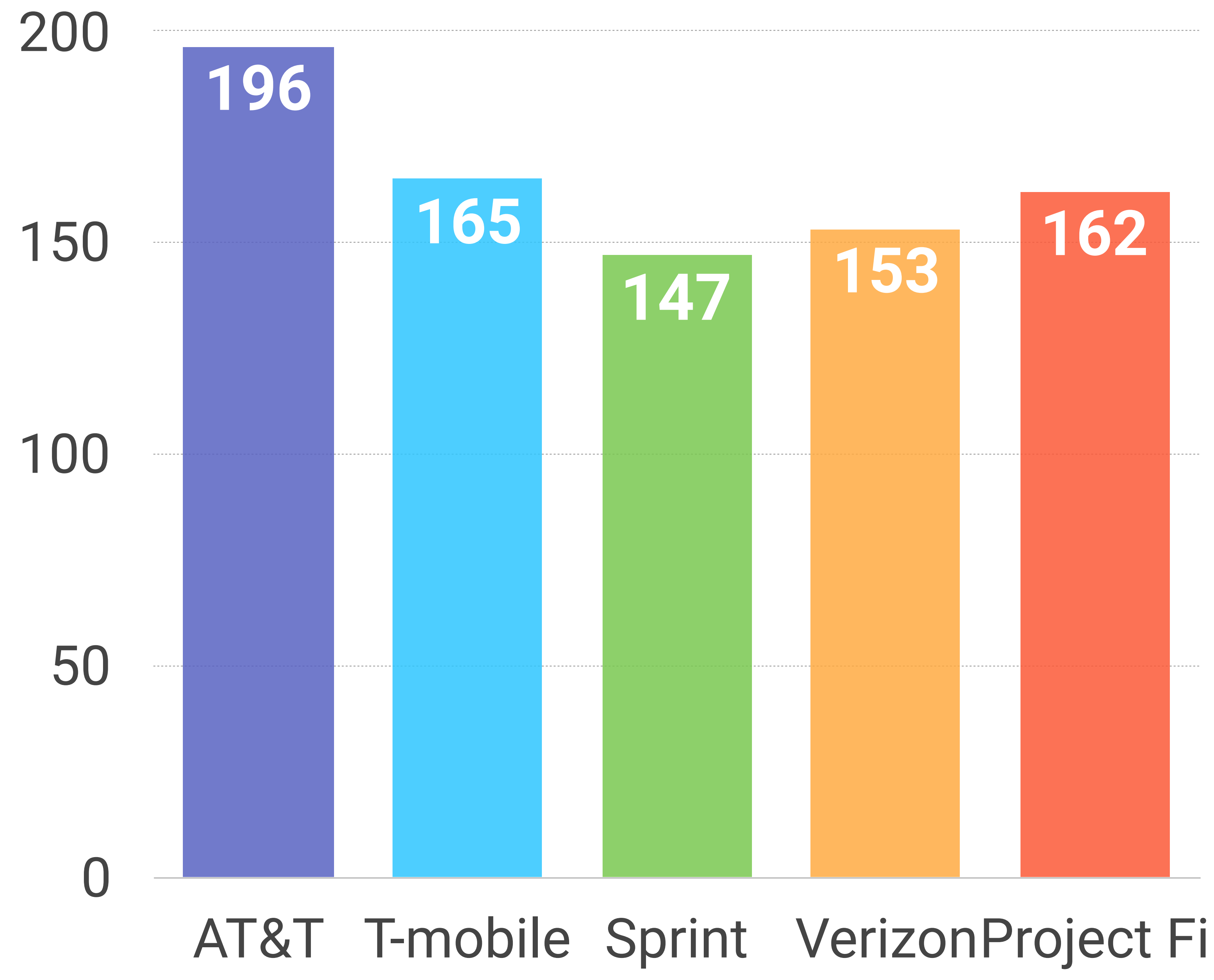
Four US carriers + Google Project Fi

23 phone models, 95,057 data sessions

Overall latency: 77 — 2956 ms in 500K samples

- Varies among different mobile carriers

Average Latency by LTE Data Access Setup (no mobility)



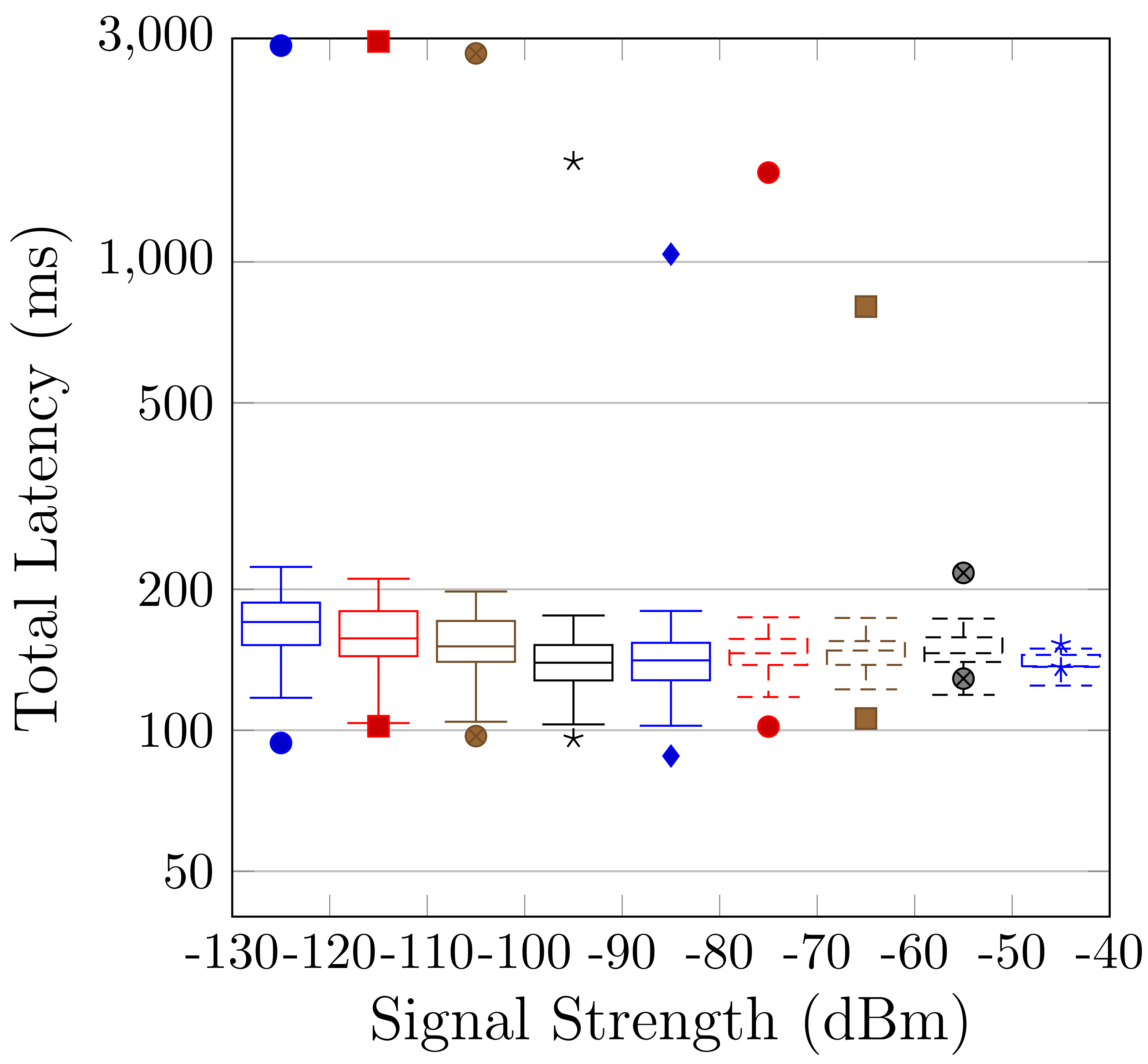
Synthesizer: global crowdsourcing analysis

Four US carriers + Google Project Fi

23 phone models, 95,057 data sessions

Overall latency: 77 — 2956 ms in 500K samples

- Varies among different mobile carriers
- Insensitive to varying radio link quality



LTE data access latency: how frequent?

Frequent data access setup operations

- every 58.8 sec (median); 133.6 sec (average)
- cause: frequently entering power-saving mode

Short-lived Radio connectivity lifetime

- every 10.8 sec (median); 17.3 sec (average)
- cause: inactivity timer (regulated by standards)

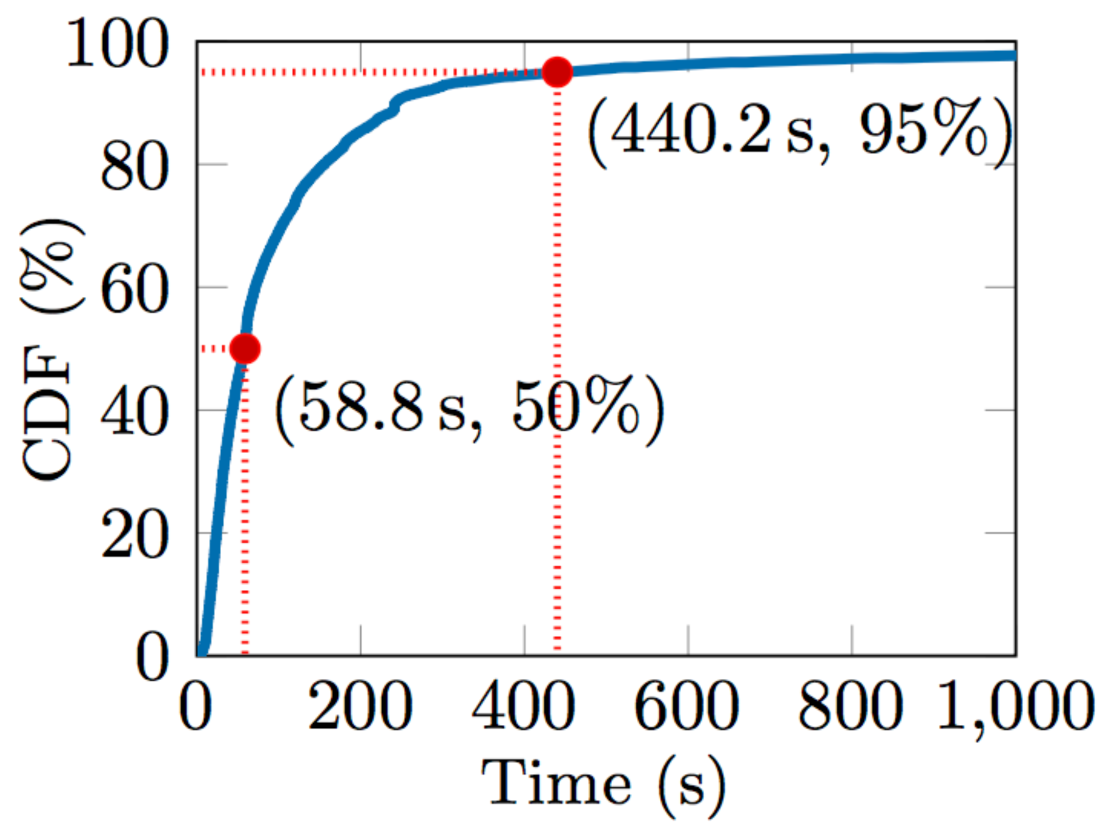
LTE data access latency: how frequent?

Frequent data access setup operations

- every 58.8 sec (median); 133.6 sec (average)
- cause: frequently entering power-saving mode

Short-lived Radio connectivity lifetime

- every 10.8 sec (median); 17.3 sec (average)
- cause: inactivity timer (regulated by standards)



(a) CDF for consecutive request interval

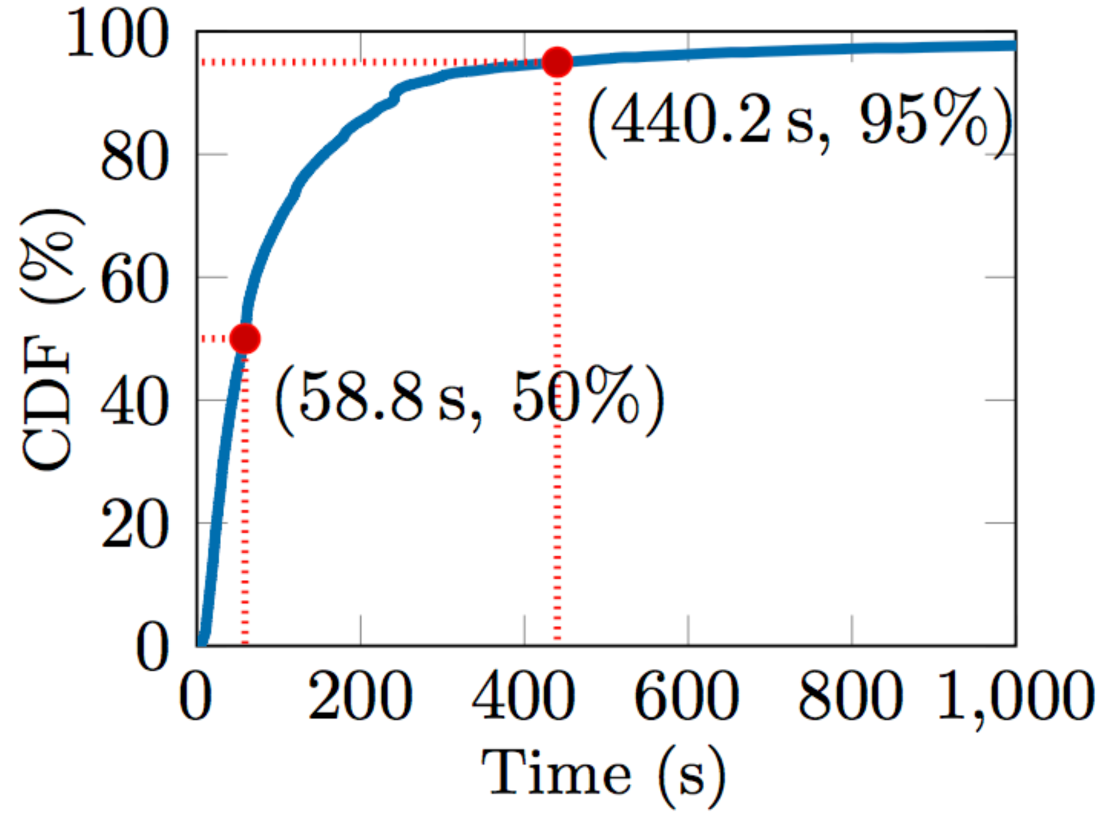
LTE data access latency: how frequent?

Frequent data access setup operations

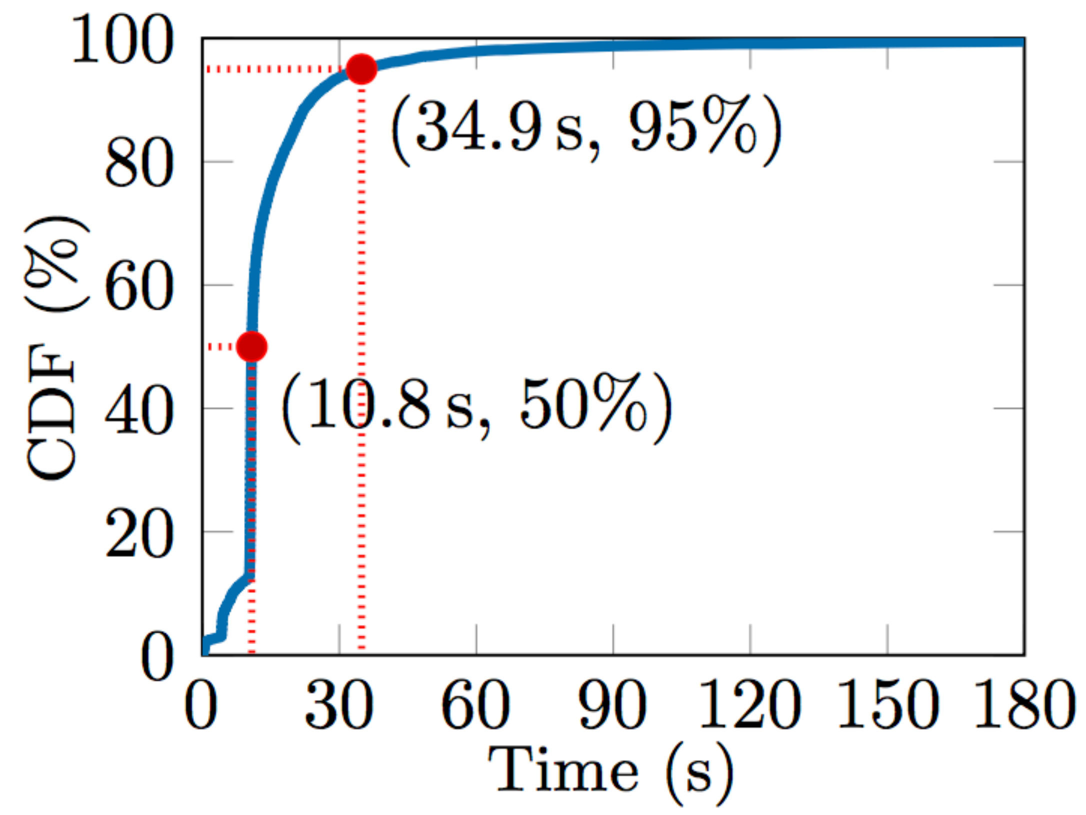
- every 58.8 sec (median); 133.6 sec (average)
- cause: frequently entering power-saving mode

Short-lived Radio connectivity lifetime

- every 10.8 sec (median); 17.3 sec (average)
- cause: inactivity timer (regulated by standards)



(a) CDF for consecutive request interval



(b) Radio connectivity lifetime

Overall latency and breakdown for major carriers

Findings Summary

Tradio: *Radio connectivity setup*

- It contributes 67.5 –1665.0 ms of the overall LTE access latency.
- On average, it contributes 39.7%, 44.0%, 61.9%, 64.2% and 43.7% of total latency in T-Mobile, AT&T, Verizon, Sprint and Project-Fi, respectively.

Tctrl: *Connectivity state transfer*

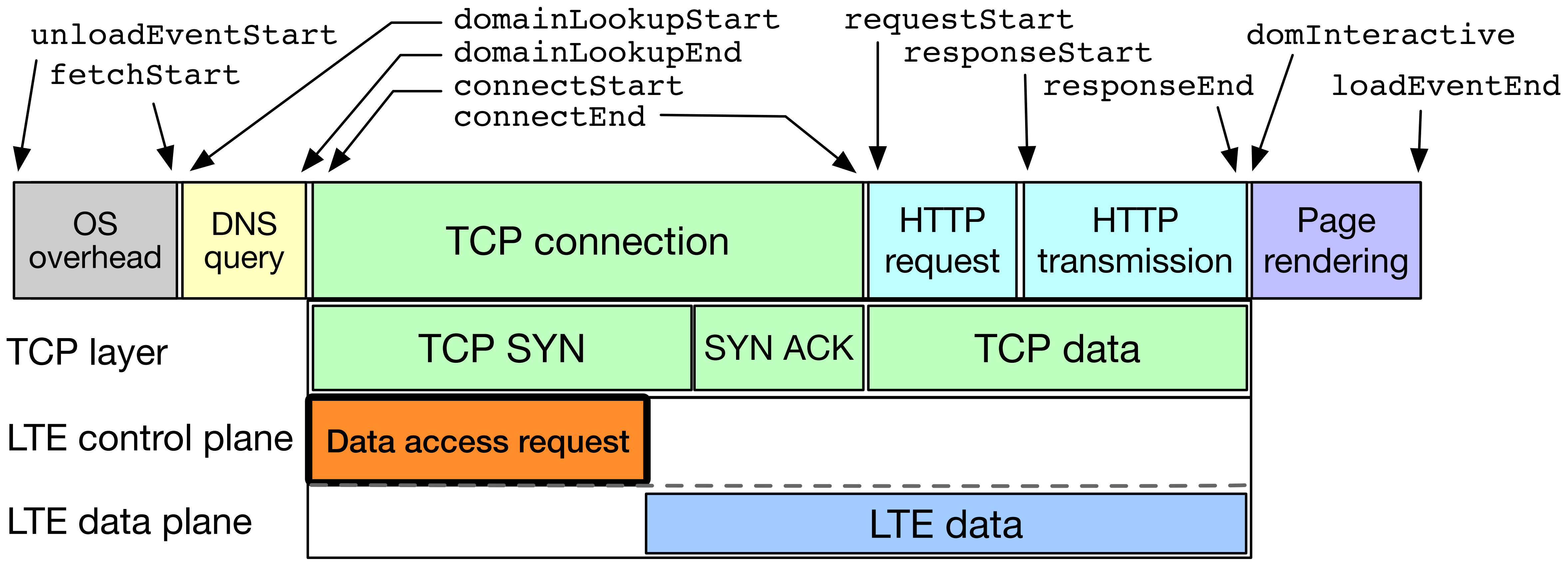
- It contributes 28.75 ms to 2286.25ms of the overall LTE access latency.
- On average, it contributes 60.3%, 56.0%, 38.1%, 35.8% and 56.3% of total latency in T-Mobile, AT&T, Verizon, Sprint and Project-Fi, respectively.

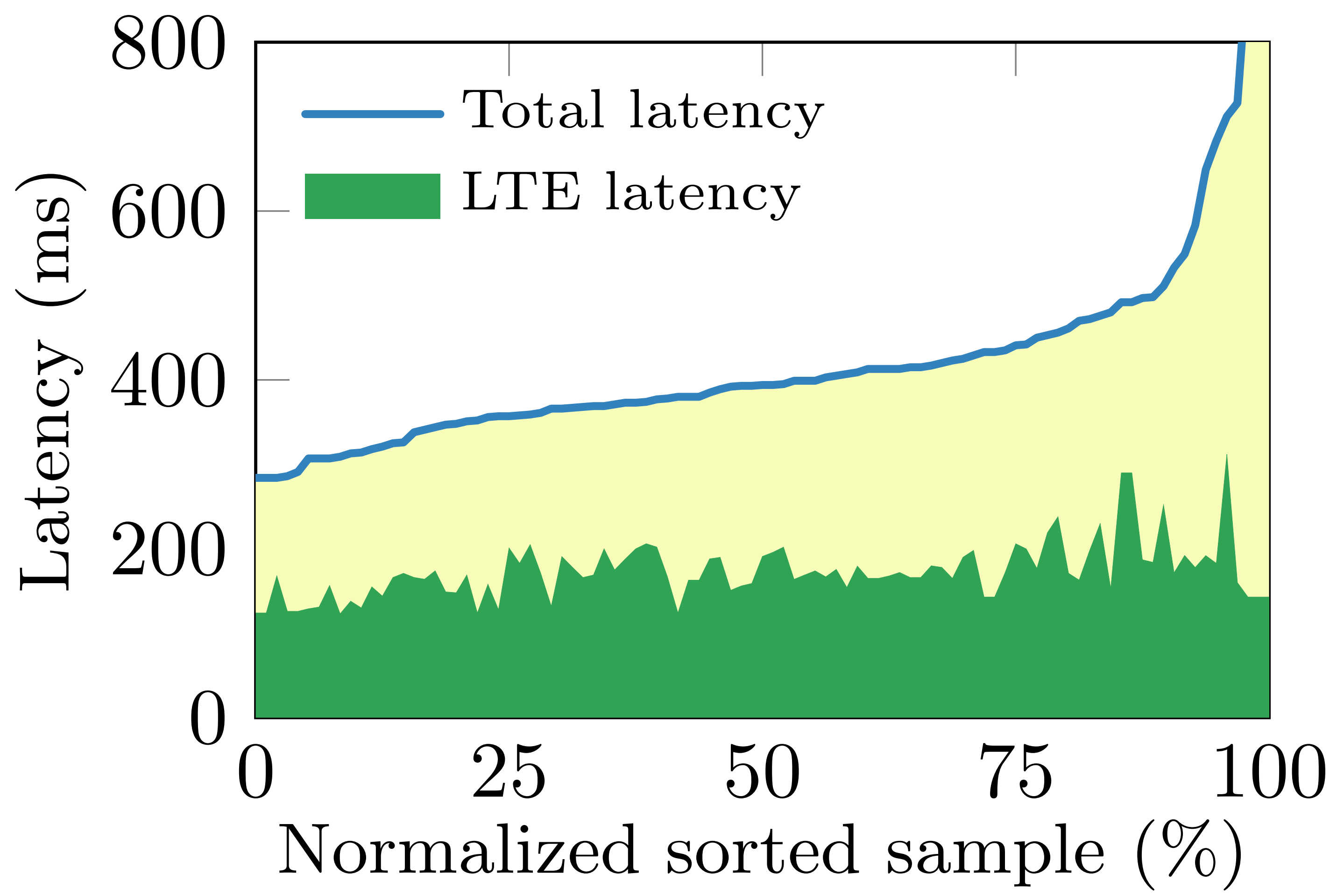
Impact on mobile Web app: Chrome

Average page loading time for tested webpage: 411 ms

- LTE data access setup: 174 ms
- **42.3%** total latency perceived

Similar results for Safari latency on iOS

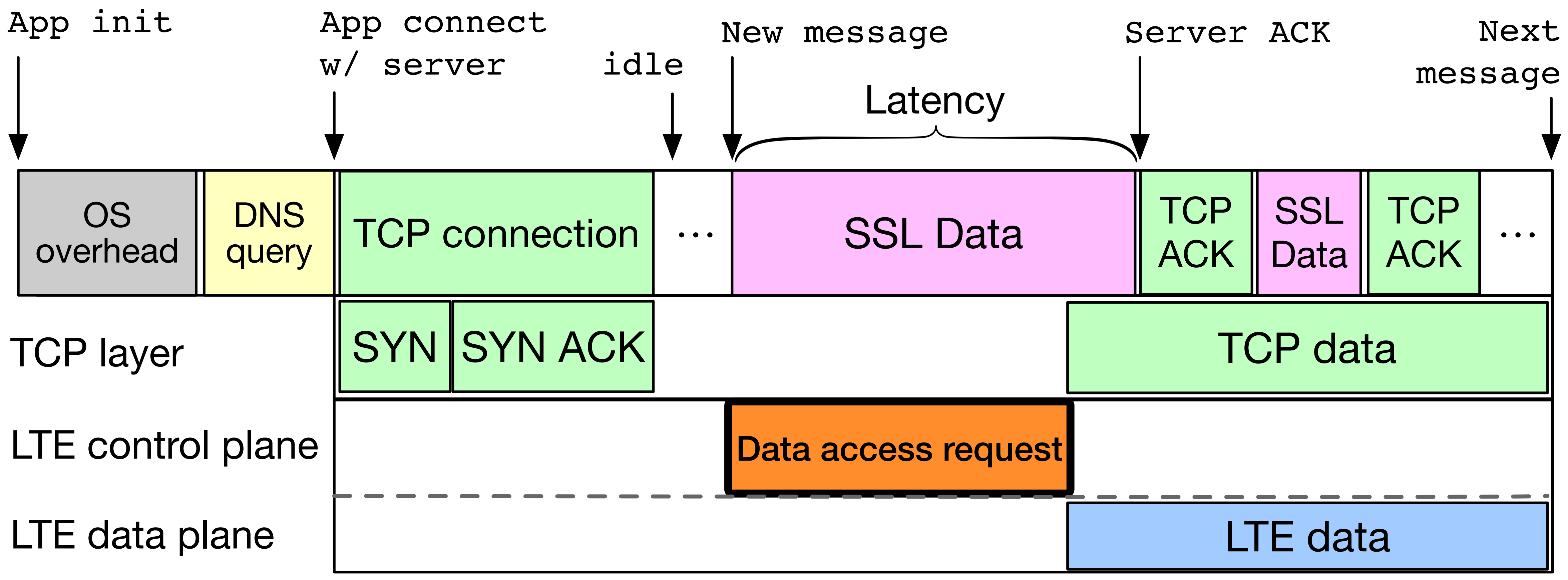


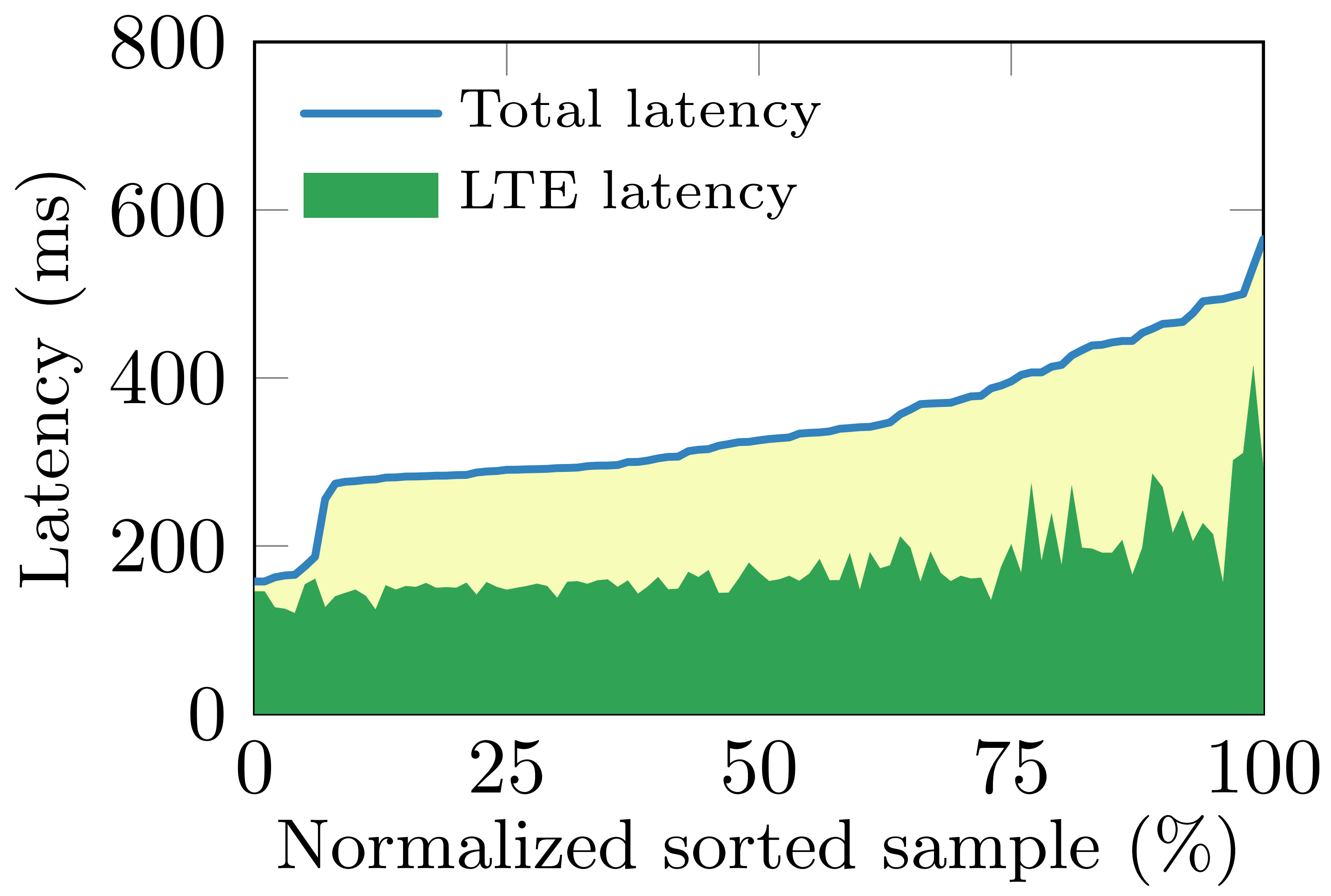


Impact on instant-messaging: WhatsApp

Average time first data packet being ACKed: 341 ms

- LTE data access setup: 175 ms
- **51.4%** total latency perceived





Discussion: reducing LTE latency

Data plane walk-arounds

- Mask the data setup latency by waking device in connected mode in advance

Control plane acceleration

- Speed up connectivity state transfer between the base station and the mobility controller (e.g. DPCM [ACM MobiCom’17])
- Handover prediction

Other issues

- Extending to other network metrics (e.g. loss, throughput, …)
- Theoretical bounds
- Privacy issues

Conclusion: ML-based analysis for next-gen mobile networks

Mobile networks are successful and will continue to prosper (5G, self driving, ...)

Mobile network analysis: paradigm shift to **device-centric**, **ML-based** scheme

- Device-centric: unveil the tightly-guided operation issues over 4G/5G mobile networks
- Two-tiered approach: a more open solution approach for the research community

Q & A

Backup