On Radio Access Network Uplink Latency and Jitter : Measurements and Analysis

Flavien Ronteix–Jacquet
IMT Atlantique (Rennes, France), Orange Labs Networks (Lannion, France)
Email: flavien.ronteix-jacquet@imt-atlantique.fr

***Abstract*—This work studies the latency experienced by a TCP connection in an operational LTE RAN. After exhibiting the well-known downlink bufferbloat phenomenon, our experiments shed some light on the less studied RAN uplink jitter. We explain this uplink jitter by the uplink grant-based access method. These results are reproduced in a lab environment based on the OpenAirInterface software RAN, and demonstrate the importance of RAN configuration and limitations in the current LTE standard. We conclude on open issues in the 5G grant allocation process and the current grant-free access methods.**

***Index Terms*—Cellular Networks, Latency, Jitter, Uplink, Measurements.**

My thesis

My thesis is about low latency in cellular Radio Access Networks (RAN) to unleash transmission performances. I am working on it since 21 months and I plan a graduation on October 2022. It is conducted under the control of a doctoral advisor, *Xavier Lagrange* (IMT Atlantique, France) and two supervisors, *Alexandre Ferrieux* and *Isabelle Hamchaoui* (Orange Labs Networks, France).

I. Introduction

LTE cellular technology deployment since 2014 has enabled the current era of high speed mobile communications. However, they are impeded by significant end-to-end latency [1], higher than fiber access for example. 5G aims to revolutionize cellular networks with new capacities such as higher throughput and, above all, lower latency [2]. It should be noted that in both LTE and 5G mobile networks, the main contribution to end-to-end latency is clearly the access segment, for physical reasons. Indeed, the air interface capacity, shared between several mobile users, is physically limited and variable over time due to fading, interference, etc. - depending on user location. This noise can generate a lot of losses, thus error correction and retransmission mechanisms are required (Hybrid-ARQ (HARQ)). These elements make the access network a natural bottleneck for all communications in a cell.

Unsurprisingly, LTE and 5G latency reduction is a popular research subject that has generated extensive literature in the last decade [3]. However, most of these studies focused on the downlink contribution to latency, as this direction conveys the bulk of the traffic in a typical download scenario. The uplink direction was seldom studied, as its effect on overall performance was deemed negligible. In this article, we show that this is not always true: the uplink segment may constitute an important part of latency in specific configurations or bad radio conditions [4], particularly when edge computing is considered.

Specific mechanisms have been designed by 3GPP to overcome this issue via dedicated bearers, some adapted to low latency requirements [5]. Indeed, bearers are used to carry packets of a given user, one for each level of quality of service; one of them at least, the default bearer, is in Best Effort (BE) mode. Radio resources are allocated to bearers by the Base Station (BS) on both the forward and return paths depending on the bearer quality level. Several allocation schemes are possible [6], from low-delay resource-intensive to scalable BE with no delay commitment. In this work, we focus on this best effort mono-bearer configuration since it is nearly the only one encountered in operational networks

Our first contribution reveals this uplink bottleneck via experiments in an operational Radio Access Network (RAN). The second contribution exhibits the bursty pattern of uplink traffic, due to grant-based access methods, that are typical of mobile networks. Finally we conclude on the implication of such access methods on applications and services which relies on uplink transmission for a good service.

II. Experimental setups

A. Experiment on a production network

We first investigate this issue in real life, using the mobile network of an Orange affiliate. For this purpose, we simulate an end-to-end scenario with both a controlled user and server. To ease clock synchronisation, we co-locate the UE and the server on the same PC with 2 interfaces, as presented in Figure 1. The User Equipment (UE) part is connected to the mobile network via a LTE modem and the server part is connected to the network through an Ethernet interface. Our server is based on linux kernel 4 TCP stack with *Cubic nohystart* as Congestion Control Algorithm (CCA).

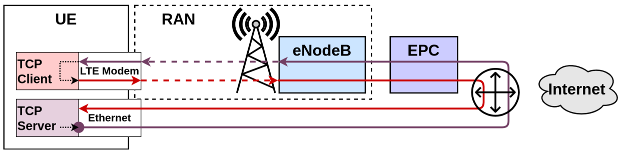
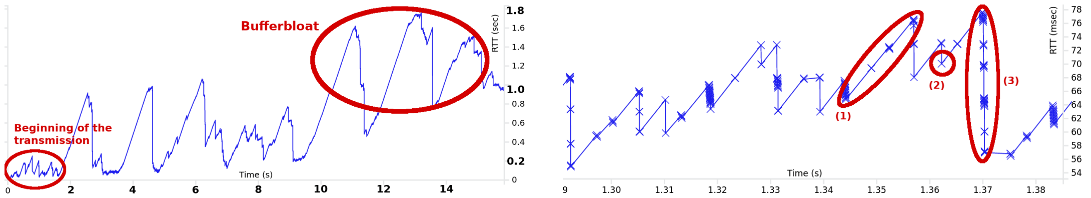


Figure 1: Experiment architecture on an operational network.

The UE is in a fixed location and in good radio condition, *i.e.* the SNR is high and constant. Hence, the air interface

ISBN 978-3-903176-43-0 © 2021 IFIP



(a) RTT during all the transmission

(b) Zoom on RTT when downlink queuing delay is low

Figure 2: RTT experienced by TCP packets in a loaded cell $\rho = 0.7$

capacity is high and does not vary. We chose this ideal configuration to minimize radio layer retransmissions, left for further study.

In our operational network, delays introduced by the backhaul and core networks are negligible with respect to the RAN's contribution. As for any customer, the BE Data Radio Bearer (DRB) is used. The load of the cell ρ is varying along with the hours of day. Thanks to this, we can measure TCP performances under a wide range of loads. The observations are performed from two synchronized capture points: one in the terminal at LTE interface (for UE side), and one at the Ethernet interface (for server side). We then visualize (*e.g Figure 3.*) the two captures on the same time axis: the delay for a given packet appears as the horizontal space between the two curves.

B. *Experiment on a lab testbed*

We reproduce the previous experiment in a controlled RAN for analysis purpose. In a pursuit of realism, we use COTS UE, a software Base Station and a real air interface. The software Base Station is OpenAirInterface (OAI)[1] Evolved Node B (eNB), using a USRP B210 as Software Defined Radio (SDR) and the core network is the OAI Evolved Packet Core (EPC).

III. Results

In our experiments, we exhibit two salient phenomena: Downlink bufferbloat and high uplink jitter.

A. *Downlink bufferbloat*

The cell load $\rho = 0.7$ of maximum capacity with indications of congestion. At the beginning of the transmission, the RTT is around 60ms (Figure 2a and Figure 2b), more than baseline RTT of 30ms. As the scheduling is made between users, when the number of users increases, the number of opportunities to transmit decreases, and the delay increases.

After 1.8sec, a first delay spike occurs, then the RTT keeps increasing and remains always higher than 200ms. A high steady delay is the hallmark of the so-called bufferbloat phenomenon [1]. To locate more precisely the bloated buffer, we note that the difference between cumulative volumes in Figure 3-a indicates an uplink latency of 20 to 70ms, much smaller than the total end-to-end delay. The bufferbloat is then located in the downlink transmission queue in the BS. The bottleneck buffer size is large, above 3MB, which can stores 1 second of packets (since there is no losses by buffer overflow and the bytes-in-flight is equal to 3MB). Delay spikes are the consequence of a lack of downlink transmission opportunity coupled with a slow adaptation of CCA to a decrease in radio capacity. Figure 2a is a good argument to focus latency research work on downlink bufferbloat.

B. *High uplink jitter*

However, the zoom on production network's RTT presented in Figure 2b reveals a large uplink jitter. Looking at Figure 3-b, we observe that while uplink traffic (made only of TCP ACKs) emitted by the UE (purple curve) is almost linear, the same uplink traffic received by the BS (red curve) exhibits large steps. The consequence is an extra delay for the first packets of each burst and a high variation of inter-arrival time between packet (= jitter). As it turns out, 1) an uplink queue builds up in the UE for lack of uplink grants; 2) some isolated packets are transmitted in uplink even if more are in the buffer; 3) sometimes, almost all awaiting packets are transmitted in a burst of a few milliseconds.

To get rid of any influence from the TCP stack, we reproduced a similar uplink traffic profile (low datarate with small packets) in User Datagram Protocol (UDP). We found the same behaviour as in TCP Figure (3-c).

Finally, to avoid any bias from uplink access contention, we study the empty cell ($\rho < 0.1$) case. There again, the uplink traffic profile in red (Figure 3-b) is discontinuous with burst and partial buffer emptying, exactly as in a loaded cell. The same applies to UDP traffic (Figure 3-d).

IV. Discussion

We interpret the uplink transmission profile as a direct consequence of grant-based access mechanisms: Scheduling Request (SR), Buffer Status Report (BSR) and scheduling algorithm.

Packets are stored in the Radio Link Control (RLC) transmission buffer of the BS in downlink and of UE in uplink before transmission. In downlink, the BS empties this buffer in a First-In First-Out (FIFO) style according to its scheduling algorithm (e.g. Round-Robin (RR) or Proportionnal-Fair (PF)), taking into account the data queue occupancy, possible retransmissions, requested quality and the radio channel quality. The BS allocation is then based on exact, real time information.

On the contrary, the uplink allocation is indirect and delayed, as UEs RLC queue occupancy is not readily available to the BS. Indeed, it is signaled to the BS by access requests

[1] https://www.openairinterface.org/

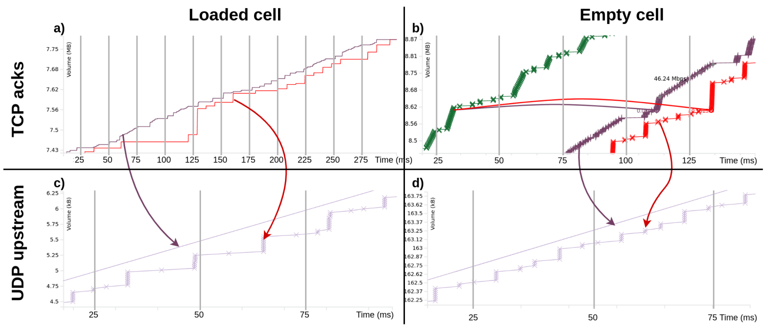


Figure 3: Time-sequence plot for UDP and TCP transmissions in a loaded cell ($\rho = 0.7$) and empty cell ($rho < 0.1$)

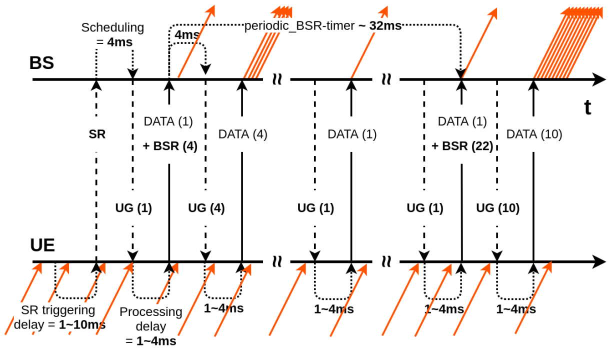


Figure 4: Uplink grant-based access procedure

emitted by the UE: UE can request radio resources with the (cheap and frequent) SR signal or quantifies its needs with (more expensive and rarer) BSR message which reports the RLC transmission buffer size. As presented in Figure 4, when a packet arrives in the UE transmission buffer, a SR is triggered. Without further information, the BS estimates the UE's uplink needs, and grants resources accordingly. Less frequently, the UE can transmit a BSR which gives a snapshot of its uplink buffer occupancy. The BS uses this BSR to update its estimation of UE uplink needs, a more precise allocation is done and grants are generated accordingly. The crucial feature is the timing of BSR generation, as too-low frequency of BSR undermines the whole mechanism.

The main triggering mechanism is a timer set by configuration. Between 2 BSRs, needs of transmission are estimated according to the last received BSR and the total amount of grant sent (method confirmed by the study of OAI and srsLTE code). Between that time, SR are triggered at a given frequency since the transmission buffer is not empty. The periodic BSR timer expiration allows a BSR and an estimation update at the BS. 8ms (=$T_{Sched}+T_{UG}$) after a BSR we observe a new burst of data. With these elements, we conclude on a BSR periodic timer of 16ms and SR alignment timer of 4ms in this Radio Access Technology (RAT) configurations.

We confirmed this analyze on a testbed with SR alignment timer, BSR periodic timer and scheduling buffer estimation method as parameters and got a better transmission pattern for transport protocols.

V. Conclusion and Future work

In these experiments we observe high uplink jitter in a 4G operational network and reproduce it in a lab setup. This bursty transmission pattern has an impact on TCP performance [7] since it falsely interprets RTT variations as congestion, leading to false network state estimation. This work on latency could be extended to 5G [8] since the access method does not change, even if 5G achieves lower latency thanks to spectrum flexibility, ultra-lean design and more efficient HARQ. Alternatives to the grant based mechanism have been proposed for low latency access, including Semi-Persistent Scheduling (SPS), pre-scheduling and contention-based. They effectively reduce latency at the expense of resource waste, making them out of reach for best effort traffic. An interesting development could be a grant-based access procedure addressing this uplink jitter issue at a lower cost than grant-free methods.

Acknowledgments

I would like to thank Isabelle Hamchaoui and Alexandre Ferrieux for their careful proofreading and their technical support.

References

[1] H. Jiang, Z. Liu, Y. Wang, K. Lee, and I. Rhee, "Understanding bufferbloat in cellular networks," in *Proceedings of the 2012 ACM SIGCOMM workshop on Cellular networks: operations, challenges, and future design*, 2012, pp. 1–6.

[2] E. Dahlman and S. Parkvall, "Nr - the new 5g radio-access technology," in *2018 IEEE 87th Vehicular Technology Conference (VTC Spring)*, June 2018, p. 1–6.

[3] B. Briscoe, A. Brunstrom, A. Petlund, D. Hayes, D. Ros, I. Tsang, S. Gjessing, G. Fairhurst, C. Griwodz, and M. Welzl, "Reducing internet latency: A survey of techniques and their merits," *IEEE Communications Surveys Tutorials*, vol. 18, no. 3, pp. 2149–2196, thirdquarter 2016, 6967689.

[4] M. Sahu, S. Damle, and A. A. Kherani, "End-to-end uplink delay jitter in lte systems," *Wireless Networks*, vol. 27, no. 3, p. 1783–1800, 2021.

[5] 3GPP, "System architecture for the 5G System (5GS)," 3rd Generation Partnership Project (3GPP), Technical Specification (TS) 23.501, 06 2021, version 16.9.0.

[6] K. Pedersen, G. Pocovi, J. Steiner, and A. Maeder, "Agile 5g scheduler for improved e2e performance and flexibility for different network implementations," *IEEE Communications Magazine*, vol. 56, no. 3, p. 210–217, March 2018.

[7] I. Johansson, "Congestion control for 4g and 5g access," Internet Engineering Task Force, Internet-Draft draft-johansson-cc-for-4g-5g-02, Jul. 2016, work in Progress. [Online]. Available: https://datatracker.ietf.org/doc/html/draft-johansson-cc-for-4g-5g-02

[8] R. Poorzare and A. Calveras, "Challenges on the way of implementing tcp over 5g networks," *IEEE Access*, 2020.