The impact of NR Scheduling Timings on End-to-End Delay for Uplink Traffic

Natale Patriciello‡, Sandra Lagen‡, Lorenza Giupponi, Biljana Bojovic
Centre Tecnològic de Telecomunicacions de Catalunya (CTTC), Castelldefels, Barcelona, Spain
{npatriciello, slagen, lgiupponi, bbojovic}@cttc.cat
‡ These authors equally contributed to this work.

***Abstract*—One of the main design targets of New Radio (NR) is to support multiple applications, including low-latency data transmissions. To achieve that, multiple features have been introduced, among which dynamic scheduling timings (denoted by K0, K1, K2) is the one which determines the delay between the different paired control and data transmissions. For example, K2 is the delay (in unit of slots) between an uplink grant reception and the corresponding uplink data transmission. The End-to-End (E2E) latency would be highly impacted by these scheduling timings. Based on the common observation, lower scheduling timings would lead to low E2E latency. However, in this paper, especially for uplink traffic, we show that this is not always true, and there are cases in which increasing the scheduling timings provides better E2E delay performance. This is due to the interplay between traffic patterns, gNB-UE process, and the NR numerology. To show such impact of scheduling timings on E2E latency with different uplink traffic patterns and NR numerologies, we implement an E2E ns-3 based NR simulator.**

***Index Terms*—NR, 3GPP 5G, E2E delay, ns-3, processing delays, uplink scheduling, numerologies, traffic patterns.**

I. Introduction

The 3rd Generation Partnership Project (3GPP) is devoting significant efforts to define the fifth Generation (5G) New Radio (NR) access technology [1], which has a flexible, scalable, and forward-compatible Physical (PHY) layer to support a wide range of carrier frequencies, deployment options, and use cases. To achieve this flexibility, one of the key features of NR is a flexible Orthogonal Frequency Division Multiplexing (OFDM) system through support of multiple numerologies [2]. Each numerology in NR is characterized by a Sub-Carrier Spacing (SCS) and a Cyclic Prefix (CP) overhead [1]. Different numerologies can be used based on the carrier frequency as well as the specific requirements of the NR service being supported [3], [4], i.e., enhanced Mobile BroadBand (eMBB), massive Machine Type Communications (mMTC), and Ultra-Reliable and Low-Latency Communications (URLLC). The End-To-End (E2E) latency[1] is a key performance indicator for 5G future applications, specifically for URLLC applications. To account for that, in addition to variable Transmission Time Interval (TTI), which is possible due to different numerologies, NR also supports mini-slots, flexible Frequency Division Duplex (FDD) and Time Division Duplex (TDD) formats. In order to control the timings that govern the communications between next-Generation Node B (gNB) and User Equipment (UE), new scheduling/HARQ-ACK feedback timing parameters, i.e., K0, K1, K2 have been introduced in NR [5], [6]. The value of these timings parameters is generally derived by taking gNB/UE processing timings and/or E2E latency requirements into account.

In this paper, we provide an analysis on the inter-relations between UpLink (UL) traffic patterns (e.g., application type, Inter-Packet Arrival Time (IPAT), packet size), numerologies, scheduling timing and E2E latency. We will focus on the anomalies, which we reported in [7] with respect to the numerology and E2E latency relation. We observed that for DownLink (DL) flows, an increase in the SCS (numerology) leads to lower latency, which is a well-known phenomenon already seen in literature [8]. However, for UL flows, an increase in SCS (numerology) does not necessarily lead to lower E2E latency. This is due to the interaction between the IPAT, scheduling timings, processing delays, and the slot length (which is numerology-dependent). The main factors which influence the UL latency are summarized in the following [7]:

- An UL transmission generally starts with a control message exchange (i.e., Scheduling Request (SR)), which contributes to the E2E latency. It further continues through the exchange of Buffer Status Report (BSR), which is generally piggy-backed with data. As a result, any increase in the number of SRs required to complete the transmission would negatively affect the E2E latency;
- If an UE is assigned more resources than the requested resources (e.g., the default allocation is more than the request number of resources), then the UE's ability to make use of the assigned resources as much as possible would positively affect the E2E latency. This situation may occur, when new packets arrive between the UE's request (i.e., for grant) and the UE's transmission on the assigned resources.

In case of UL traffic, the value of K2 represents the slot offset between an UL grant and the corresponding UL transmission. In this paper, through simulation analysis, we show that with an appropriate K2, an optimum situation based on arrival traffic pattern may be achieved in terms of E2E

[1]Through this paper we use E2E latency and E2E delay interchangeably, to refer to the time taken for a packet to be transmitted across a network from source to destination. In case of uplink transmissions, the message exchange procedure used to request resources contributes to the E2E latency. Also, both for downlink and uplink, the scheduling/HARQ-ACK feedback timings as well as the processing delays, contribute to it.

latency. In particular, we show why an increase in K2 may lead to a lower E2E latency. The value of K2 is applied at UE side in between UL grant reception and UL data transmission, but it is determined and indicated from the gNB. However, in this context, we foresee the UE to be a proactive element for the derivation of the appropriate K2 timing and recommending it to the gNB. This is because the UE is aware of fundamental information that influences the latency changing factors, like the application IPAT, the packet size, and the flow multiplexing ability to fill the assigned Transport Block Size (TBS).

To perform such analysis, we use a network simulator, which we have built on top of the open-source ns-3 discrete-event network simulator [9], [10]. We modeled the NR technology with a high-fidelity full protocol stack [11], [12]. While common low-level simulators focus on link level simulations, our simulator offers more abstraction of the PHY layer and high-fidelity implementations from the Medium Access Control (MAC) to the Application layer. Thanks to these design choices, it is possible to extract E2E results, that span different domains, as reported in [8], [7].

The remainder of the paper is organized as follows. In Section II, we review scheduling timings and UL handshake procedure for UL grant-based access in NR. In Section III, we discuss the different factors that influence the UL E2E latency and introduce our proposal to address the interaction of processing delays, numerologies, and traffic patterns. In Section IV, we validate the factors and the proposal through the ns-3 NR simulator. Finally, Section V concludes the paper.

II. Scheduling Timings and UL Access

In this section, we review three important aspects that influence E2E delay for UL data: the scheduling timings, the processing delays, and the UL grant-based handshake procedure.

A. Scheduling timings and processing delays

In NR, the following timings are defined in terms of scheduling/HARQ-ACK feedback timing:

- K0: Delay (i.e., slot offset) between DL allocation (in Physical Downlink Control Channel (PDCCH)) and corresponding DL data (in Physical Downlink Shared Channel (PDSCH)) reception [6, Sect. 5.1.2.1].
- K1: Delay between DL data (PDSCH) reception and corresponding HARQ-ACK feedback transmission on UL (Physical Uplink Control Channel (PUCCH)) [13, Sect. 9.2.3].
- K2: Delay between UL grant reception in DL (PDCCH) and corresponding UL data (Physical Uplink Shared Channel (PUSCH)) transmission [6, Sect. 6.1.2.1].

These scheduling timings are determined at the gNB after taking the UE processing time into account. In case of UL flows, the most important parameter is K2, i.e., the number of slots assigned for the UE to decode the UL grant in PDCCH and prepare UL data to transmit in the indicated scheduling opportunity over PUSCH. The time-domain resource assignment, which conveys K2, mapping type, symbol start, and length of the UL scheduled transmission, is indicated in the UL grant to the UE [6, Sect. 6.1.2.1], [14, Sect. 7.3.1.1.2]. In particular, according to NR specifications, K2 may take any integer value from 0 to 32 (slots) [5, Sect. 6.3.2].

At the UE side, the following terminologies are used for the processing delays:

- N1: the number of OFDM symbols required for UE processing from the end of DL data (PDSCH) reception to the earliest possible start of the corresponding ACK/NACK transmission.
- N2: the number of OFDM symbols required for UE processing from the end of DL control (PDCCH) reception containing the UL grant to the earliest possible start of the corresponding UL data (PUSCH) transmission.

The specific values of N1 and N2 are detailed in [15], [16] for different configurations, numerologies, and UE capabilities. These values are communicated to the gNB as part of the UE capability, which are further used for derivation of the scheduling timings. For example, K2 *should refer to a time interval larger than or equal to* N2, to give time for the UE to prepare UL data.

In the ns-3 NR simulator, we implemented a flexible scheme for introducing these delays, which is easily extensible and may take future standard modifications into account. In particular, the scheduling timings and PHY/MAC processing delays are introduced through the following three parameters:

- L2L1processing: the time that the PHY/MAC layers at gNB need to encode control and/or data channels. From a simulator point of view, it is a delay between the control/data acquisition from the Radio Link Control (RLC) class by the MAC class and the moment at which the control/data is available to go over the air.
- decodeLatency: the time that the PHY layer needs to process/decode the incoming data. From a simulator point of view, it is a delay between the data acquisition from the air by the PHY class and the moment at which the data block is available to process at the MAC class. In case of UL, it applies at gNB side.
- K2: the time between UL grant reception in DL and UL data transmission. From a simulator point of view, it is a delay between the control acquisition from the air by the UE PHY class and the moment at which the control/data is available to go over the air.

B. UL grant-based handshake

NR allows both grant-based and grant-free access (also known as autonomous UL) schemes for UL [1]. The former is a dynamic scheduled-based access, which is similar to LTE DL/UL and NR DL. However, the latter is a contention-based scheme. In this paper, we focus on the UL grant-based access.

The UL grant-based scheme works as follows. Upon data arrival at UE RLC queues, the UE requests an UL grant by sending a SR to the gNB over PUCCH. Then, the gNB

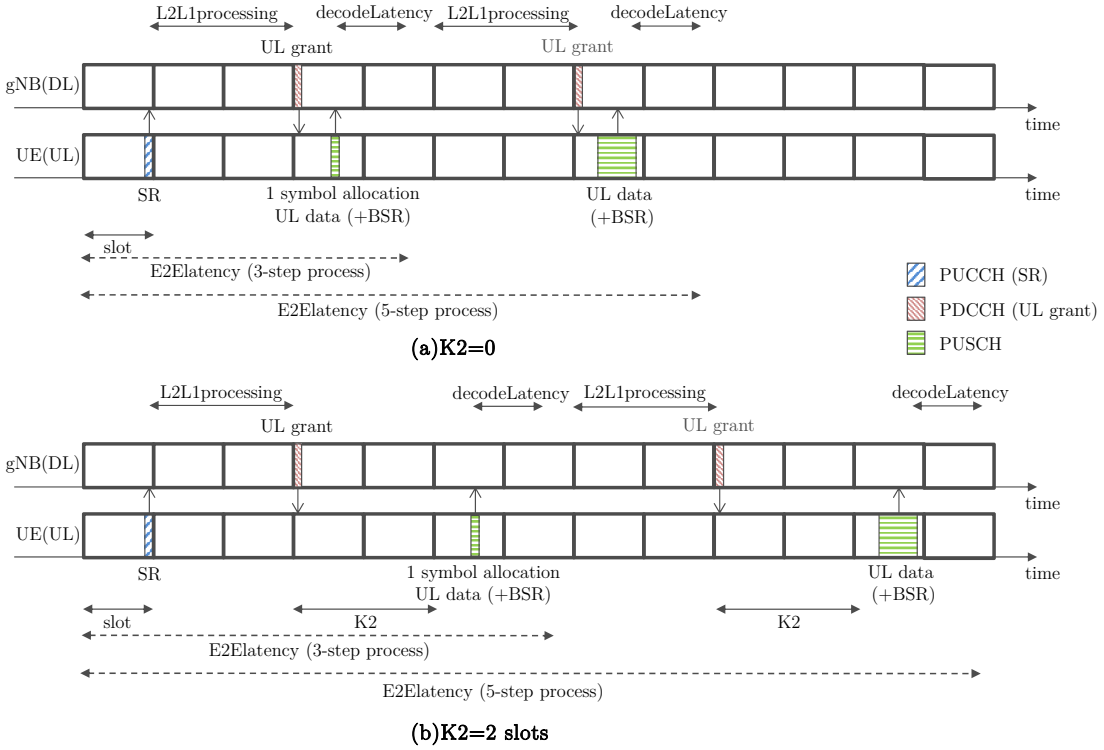

Fig. 1: UL grant-based access procedure, including scheduling timings and processing delays, as implemented in the ns-3 NR simulator. (a) K2=0, (b) K2=2 slots. E2E latency is also shown both for 3-step and 5-step processes.

sends the UL grant (Downlink Control Information (DCI) in PDCCH) to indicate the scheduling opportunity for the UE to transmit. Note that the first scheduling assignment may not be sufficient for the complete UL data transmission, since the gNB does not know the accurate requirement (e.g., buffer size) at the UE yet. In this regard, since this is implementation-specific, we assume that the first scheduling opportunity consists of the minimum number of symbols that permit at least a 4 bytes transmission. In the majority of cases, this value equals to 1 OFDM symbol. After receiving the UL grant, the UE performs the data transmission in the allocated resources over PUSCH, which may contain UL data and/or BSR. If a BSR is received, the gNB knows the UE RLC buffer status and can proceed with another UL grant to account for the remaining data. Note that, depending on the packet size and the TBS of first scheduling assignment, the UL packet transmission may end either after a 3-step process (SR $\rightarrow$ UL-grant $\rightarrow$ UL-data) or after a 5-step process (SR $\rightarrow$ UL-grant $\rightarrow$ UL-data + BSR $\rightarrow$ UL-grant $\rightarrow$ UL-data).

Before sending the UL grants, L2L1processing delay occurs at the gNB side. Also, upon reception of an UL grant by the UE, the UL transmission (e.g., data and/or BSR) is sent after K2 slots, where K2 is indicated in the UL grant. In Fig. 1, UL grant-based procedure is shown, including the scheduling timings and processing delays (i.e., L2L1processing, K2, and decodeLatency) involved in the procedure, as well as the resulting E2E delay is illustrated for 3-step and 5-step processes. Fig. 1 (a) and (b) are shown for K2=0 and K2=2 slots, respectively. We assume a NR-compliant TDD slot structure, for which the slot is composed of 14 symbols. As shown in the Fig. 1, we assume that PDCCH and PUCCH are sent in the 1st and 14th symbol, respectively. The remaining symbols in between are used for shared channels (e.g., PDSCH and/or PUSCH). The BSR is prepared shortly before the PHY transmission in the UL, reflecting the status of the RLC queue without including the current transmission.

III. K2 Timing Adjustment

As mentioned in the Section I, the traffic pattern (e.g., IPAT) and the scheduling timings would affect the E2E latency. For example,

- If IPAT is higher than the time required by the protocol to entirely transmit a single packet, the UE cannot take advantage of piggy-backing a non-zero BSR. Therefore, for the upcoming traffic, it needs to restart the SR procedure, which will increase the E2E latency.
- If IPAT is higher than the time between the grant and the transmission, the UE may not use all the excess resources allocated in the UL. For instance, if the next packet arrives shortly after the transmission of data following an oversized grant, it misses the opportunity to be transmitted along with the previous allocation (with a lower E2E latency).

In order to avoid such situations, the scheduling timing, i.e., the value of K2, can play a key role. For example,

- When a packet size (e.g., first packet) requires the 5-step process, i.e., SR → UL-grant → UL-data + BSR → UL-grant → UL-data, the value of K2 can be increased in such a way that either the first or the second UL-data transmission of the first packet carry the BSR for the second packet. Thus, only a 3-step process will be needed for the second packet, i.e., BSR → UL-grant → UL-data.
- When a packet size is small enough to complete the transmission with a 3-step process, i.e., SR → UL-grant → UL-data, the value of K2 can be increased in such a way that the exceeding resources allocated by default in the UL grant following the SR, can be used by the next packet. For example, compare Fig. 1.(b) with respect to Fig. 1.(a), for the case when the first packet arrives at slot 0 and the second packet at slot 4.

Therefore, an increment (or decrement) of K2 may lead to a better use of scheduling opportunities. Note that maybe only the UE knows the details of the traffic patterns for the UL data transmission. In such a case, it may be beneficial if the UE derives the appropriate scheduling timings, i.e., value of K2, based on its traffic conditions and the numerology used for the UL transmissions. Then, the UE can recommend the derived value of K2 to the gNB. However, this procedure of deriving K2 and communicating it to the gNB will incur additional delay. Therefore, this solution may not be beneficial for short-lived connection, where, the overhead due to exchanging the value of K2 can be higher than the data latency reduction obtained due to this solution, but rather it will be more beneficial for medium/long-lived connections. On the other hand, in case that the gNB can observe and/or predict the UL traffic patterns, then the gNB can directly adapt the K2 value, thus avoiding the additional delays. In any case, with this solution, the timing between the UL grant reception and the UL transmission, as well as the time required to entirely transmit a packet will adapt to the IPAT and slot length.

An analytic expression cannot be derived for K2 due to the inter-layer effects, multiple control message exchanges, data transmissions to complete a packet transmission, and different delays of the different packets. However, the UE (or gNB) could learn it dynamically. For example, for packet sizes that require the 5-step process, increasing K2 may lead to E2E delay reduction, if $(2\times\text{L2L1processing}+2\times\text{K2}+2)\times s_l$ $<$ IPAT, where, s_l is the slot length. In this case, the UE may suggest increasing K2 (or gNB may directly implement the update if it monitors the UL traffic pattern) to avoid sending SRs for every packet and enable piggy-backing BSR to UL MAC Packet Data Unit (PDU)s.

IV. Experimental Validation

To verify the effectiveness of changing K2 to reduce the E2E latency, we run an extensive simulation campaign using the ns-3 NR simulator. We implemented a NR specification-compliant UL SR process mechanism (as depicted in Fig. 1). Such modifications did require an in-depth redesign of the scheduler operation, since for UL, the scheduler at the gNB needs to work L2L1processing+K2 slots in advance, while for DL only L2L1processing slots.

The simulation setup is as follows. We consider a backbone connection between the Evolved Packet Core (EPC) and the remote node, modeled as 100 Gbps point-to-point link. The link between the gNB and the EPC that represents the core network is modeled through another point-to-point connection with a maximum rate of 10 Gbps. Regarding the Radio Access Network (RAN), we consider one gNB and one UE, to especifically analyze the impact of the scheduling timings. The UE with a UDP Constant Bit Rate (CBR) flow for UL transmission, i.e., that goes in the UL towards the remote node on the Internet. The UDP flow is characterized by IPAT and packet payload, whose product gives the UDP flow rate, which is varied through simulations. The gNB-UE radio link uses 28 GHz carrier frequency with 100 MHz channel bandwidth. The transmit power is 4 dBm (UE). The noise power spectral density and the noise figure are set to -174 dBm/Hz and 5 dB (gNB), respectively. The number of antennas at gNB and UE is 64 and 16, respectively. The UE is deployed at a random distance from the gNB, between 0 and 30 meters. The Urban Micro (UMi) propagation model with Line-of-Sight (LoS) condition is used. To obtain statistical significance, we repeat the same simulations using five different random seeds that affect the channel model.

We evaluate the E2E latency for each UDP packet from source to destination under different IPATs and K2 values. The K2 parameter is varied from 0 to 7 slots, which is enough to see the expected changes. The K2 value is fixed in each simulation, in other words, within a simulation, dynamic change of the K2 parameter is not implemented. For the processing delays, the L2L1processing is set to an LTE-compatible value of 2 slots. It is reasonable for L2L1processing to be numerology-dependent because the MAC scheduler at gNB works on a slot-basis. The decodeLatency is set to a fixed value (0.1 ms). This is because, for a device, the decoding time is mainly related to the CPU rate and the available energy to perform the task. It is independent of the numerology being used.

Due to the space constraint, in this paper, we only show the results for numerology μ=2; however, similar results were observed for μ=3 and μ=1. Assuming K2=0, for small packet sizes that fit in the minimum scheduling unit (e.g., a TBS lower than 850 bytes fits in 1 OFDM symbol with Modulation Coding Scheme (MCS) = 28, μ=2 and 100 MHz channel bandwidth), the minimum theoretical single packet delay (assuming no retransmission) will be 3 slots + 0.1 ms, i.e., 0.85 ms (for μ=2). Under the same setup, for packet sizes larger than 850 bytes that require the 5-step process, the single packet delay will be 6 slots + 0.1 ms, i.e., 1.6 ms. However, the 1.6 ms can be reduced to 0.85 ms in case if BSR is piggy-backed with the previous UL MAC PDU. Based on this analysis, for the simulations, we select two packet sizes, 500 bytes and 1000 bytes. Also, we select two packet arrival patterns, IPAT of 0.5 ms and another of 2 ms.

Figs. 2-5 show the E2E latency (in ms, in left y-axis) for

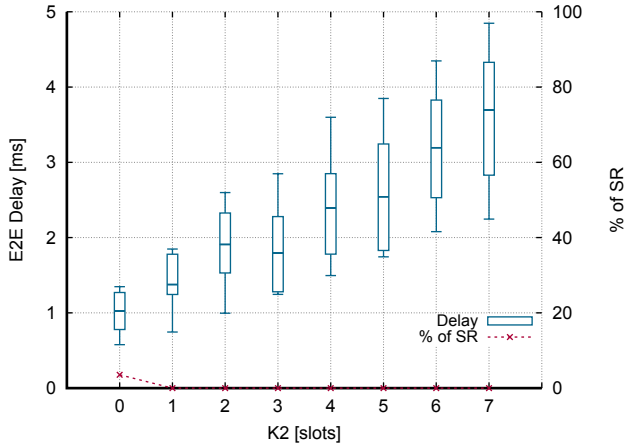

Fig. 2: E2E delay (left-axis) and percentage of SRs over the total number of packets (right-axis) vs. K2, for payload=500 bytes and IPAT=0.5 ms.

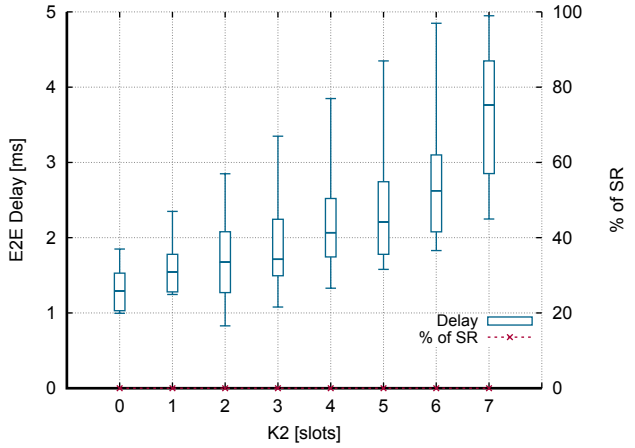

Fig. 3: E2E delay (left-axis) and percentage of SRs over the total number of packets (right-axis) vs. K2, for payload=1000 bytes and IPAT=0.5 ms.

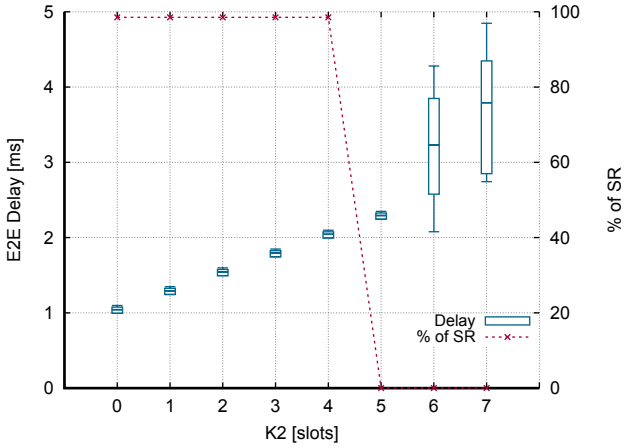

Fig. 4: E2E delay (left-axis) and percentage of SRs over the total number of packets (right-axis) vs. K2, for payload=500 bytes and IPAT=2 ms.

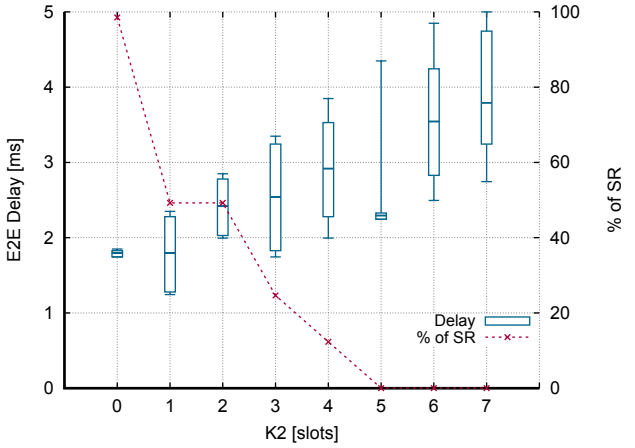

Fig. 5: E2E delay (left-axis) and percentage of SRs over the total number of packets (right-axis) vs. K2, for payload=1000 bytes and IPAT=2 ms.

the UL flow as a function of K2 (in slots), for different traffic patterns. Top whisker represents the maximum value, the box spans from the 20th to the 80th percentile, and the bottom whisker is the minimum value found in the dataset. The mean is painted as a horizontal straight line in the box. In addition, in the right y-axis of the figures, we depict the percentage of SRs over the total number of packets in the simulation. The following traffic patterns are used: IPAT = 0.5 ms and packet payload = 500 bytes (Fig. 2), IPAT = 0.5 ms and packet payload = 1000 bytes (Fig. 3), IPAT = 2 ms and packet payload = 500 bytes (Fig. 4), and IPAT = 2 ms and packet payload = 1000 bytes (Fig. 5).

In Fig. 2 (IPAT = 0.5 ms and payload = 500 bytes), we can see that increasing K2 from 2 to 3 reduces the mean E2E delay by 6%. We can observe that the number of SRs is negligible (when K2 $>$ 0, we have only one SR) and, therefore, the 6% latency reduction is entirely due to the better usage of the scheduling opportunity of 1 OFDM symbol. To further demonstrate our statement, we must prove that when the packet payload is larger than the TBS that fits in 1 OFDM symbol, this effect should disappear. As shown in Fig. 3 (IPAT = 0.5 ms and payload = 1000 bytes), although there is similar mean value for K2 = 2 and K2 = 3, but there is not latency reduction due to increasing the K2, which proves our statement.

Fig. 4 and 5 show the results with higher IPAT, i.e., 2 ms ($>$ 1.6 ms needed for the theoretical 5-step process with K2 = 0). In Fig. 4, the packet payload is 500 bytes, which is small enough to fit in the default TBS assigned after a SR. In this case, almost each packet requires 3-steps to complete the transmission. For K2 $<$ 5, almost every packet requires an SR. For K2 $\geq$ 5, the percentage of SR is nearly zero. It is because, starting from K2 = 5, all the packets are reported through BSRs during the transmission of the previous one. Although, the number of SRs are decreased, but each packet still needs three steps to complete the transmission. Therefore, there is an almost linear increase in the mean packet delay with respect to K2.

In Fig. 5 (IPAT = 2 ms and payload = 1000 bytes), we observe that setting K2 to 5 effectively reduces the E2E delay as well as number of SRs as compared to K2 = 2, 3, and 4. For example, there is a 22% E2E mean delay reduction with K2 = 5 as compared to K2 = 4 slots and of 5% with respect to K2 = 2. This is thanks to piggy-backing the BSR to the UL data transmission, which avoids sending the SR and the 5-step process in the forthcoming packets.

Let us finally recall that, device capabilities usually impose a minimum K2 of 2 slots to account for the UE processing times, especially at high numerologies. Therefore, the common belief that the lowest allowed K2 is the better one in terms of delay is shown not true in some situations.

V. Conclusions

In this paper, we analyzed the impact of traffic patterns, numerology, processing delays and scheduling delays on the E2E latency for uplink transmissions. To show that, an extensive simulation analysis is carried out on an E2E, full stack, and high fidelity NR simulator. Based on the analysis, a solution to achieve low E2E latency is proposed, in which, a dynamic update of the value of K2 (the timing in between UL grant reception and the corresponding UL data transmission) is suggested. In particular, it is shown that by introducing appropriate K2 timings in the gNB-UE process, the capacity of the allocated scheduling opportunities can be efficiently occupied and the number of scheduling requests can be reduced. It is also observed that the lowest permitted K2 timing value may not always lead to a lower E2E delay, instead, a longer one may help to improve the E2E performance. Based on that, mechanisms to dynamically update K2 are beneficial, where either the gNB monitors the UL traffic pattern and performs the update or the UE suggests an appropriate value based on its actual arrival traffic patterns and configured numerology.

VI. Acknowledgments

This work was partially funded by Spanish MINECO grant TEC2017-88373-R (5G-REFINE) and Generalitat de Catalunya grant 2017 SGR 1195. Also, it was supported by InterDigital Communications, Inc.

References

[1] 3GPP TS 38.300, *TSG RAN; NR; Overall description; Stage 2 (Release 15), v15.3.0*, Sept. 2018.

[2] A. A. Zaidi *et al.*, "Waveform and numerology to support 5G services and requirements," *IEEE Commun. Mag.*, vol. 54, pp. 90–98, Nov. 2016.

[3] A. Yazar and H. Arslan, "A flexibility metrix and optimization methods for mixed numerologies in 5G and beyond," *IEEE Access*, vol. 6, pp. 3755–3764, Jan. 2018.

[4] S. Lagen *et al.*, "Subband configuration optimization for multiplexing of numerologies in 5G TDD New Radio," *IEEE Int. Symp. Personal, Indoor and Mobile Radio Commun.*, Sept. 2018.

[5] 3GPP TS 38.331, *TSG RAN; NR; Radio Resource Control (RRC) protocol specification*, Release 15, v15.3.0, Sept. 2018.

[6] 3GPP TS 38.214, *TSG RAN; NR; Physical layer procedures for data*, Release 15, v15.3.0, Sept. 2018.

[7] N. Patriciello *et al.*, "5G New Radio numerologies and their impact on the end-to-end latency," *IEEE Int. Workshop on Computer-Aided Modeling, Analysis and Design of Commun. Links and Networks*, Sept. 2018.

[8] B. Bojovic, S. Lagen, and L. Giupponi, "Implementation and evaluation of frequency division multiplexing of numerologies for 5g new radio in ns-3," in *Proceedings of the 10th Workshop on ns-3*, pp. 37–44, ACM, 2018.

[9] T. R. Henderson, S. Roy, S. Floyd, and G. F. Riley, "ns-3 project goals," in *Proceeding from the 2006 workshop on ns-2: the IP network simulator*, p. 13, ACM, 2006.

[10] G. F. Riley and T. R. Henderson, "The ns-3 network simulator," in *Modeling and tools for network simulation*, pp. 15–34, Springer, 2010.

[11] N. Patriciello, S. Lagen, B. Bojovic, and L. Giupponi, "An E2E simulator for 5G NR networks," *Simulation Modelling Practice and Theory*, vol. 96, p. 101933, 2019.

[12] N. Patriciello, S. Lagen, L. Giupponi, and B. Bojovic, "An Improved MAC Layer for the 5G NR Ns-3 Module," in *Proceedings of the 2019 Workshop on Ns-3*, WNS3 2019, (New York, NY, USA), pp. 41–48, ACM, 2019.

[13] 3GPP TS 38.213, *TSG RAN; NR; Physical layer procedures for control*, Release 15, v15.2.0, June 2018.

[14] 3GPP TS 38.212, *TSG RAN; NR; Multiplexing and channel coding*, Release 15, v15.1.1, Apr. 2018.

[15] Qualcomm, 3GPP R1-1721515, 3GPP TSG RAN WG1 91 Meeting, *Summary of DL/UL scheduling and HARQ management*, Dec. 2017.

[16] Huawei, HiSilicon, 3GPP R1-1719401, 3GPP TSG RAN WG1 89 Meeting, *Remaining issues on HARQ*, Dec. 2017.