5G New Radio Numerologies and their Impact on the End-To-End Latency

Natale Patriciello, Sandra Lagen, Lorenza Giupponi, Biljana Bojovic
Centre Tecnològic de Telecomunicacions de Catalunya (CTTC/CERCA)
Av. Carl Friedrich Gauss 7, 08860 Castelldefels (Barcelona), Spain

***Abstract*—In this paper, we use a New Radio (NR) simulator, based on ns-3, to assess the impact of 5G NR numerologies on the end-to-end (E2E) latencies in a realistic and complex scenario, including TCP and UDP flows. As expected, we found that TCP goodput increases with the numerology, since a larger numerology allows reducing the round-trip-time. However, although counter-intuitive, simulation results exhibit that the E2E latency of uplink (UL) UDP flows may not be reduced with the numerology. In fact, it depends on two key factors and their relationship: the processing delays (fixed or numerology-dependent) and the inter-packet arrival time, which depends on the UDP flow rate and the packet size. We demonstrate how, in some cases, the latency is worsened by an increasing signaling exchange that grows with the numerology. In particular, this is due to a handshake mechanism in UL (scheduling request and UL grant) that is performed each time a data packet encounters empty RLC buffers. For some combination of flow rate, packet size, and processing delays that are not numerology-dependent, increasing the numerology may not reduce the E2E delay. Therefore, we conclude that the selection of the numerology in an NR system should be carefully made by taking into account the traffic patterns and the processing delays.**

***Index Terms*—New Radio, ns-3, numerologies, processing delays, TCP, UDP.**

I. Introduction

We are seeing incredible efforts by the partners of 3rd Generation Partnership Project (3GPP) to define the fifth Generation (5G) New Radio (NR) access technology [1]. The definition is expected to be flexible to be able to work in a wide range of bands and address many different use cases, to be able to reach its objectives [2]. In this regard, one of the key steps is the inclusion of a flexible Orthogonal Frequency Division Multiplexing (OFDM) system [3]. NR defines a set of numerologies, which specify a SubCarrier Spacing (SCS) and a cyclic prefix overhead, to handle a wide range of frequencies and deployment options [1]. Also, a base station (a.k.a. next-Generation Node B (gNB)) should provide access to different types of services, such as enhanced Mobile BroadBand (eMBB), massive Machine Type Communications (mMTC), and Ultra-Reliable and Low Latency Communications (URLLC). Theoretically, a latency-throughput trade-off appears at Physical (PHY) layer when attempting the selection of the proper numerology for gNB operation in full-buffer conditions: larger SCS is better to reduce latency (i.e., for URLLC traffic), while lower SCS is better for high PHY throughput performance (i.e., for eMBB traffic).

In this paper, we analyze a complex and realistic 5G future scenario, where traffic from multiple 5G applications is transmitted over both UDP and TCP transport protocols. We study the impact of processing and decoding delays in different numerologies, and how they affect the end-to-end performance. We prove that not always a higher SCS guarantees a lower delay, especially in the UpLink (UL) case. Our results indicate that increasing the numerology (in particular, in the first three numerologies) may trigger an increase of Scheduling Request (SR) messages, which in turn increases the latency. Similarly, another unexpected effect that we observe when increasing the numerology is with over-dimensioned Transport Block Size (TBS). With lower numerologies, the exceeding allocated resources can be filled by newly arrived packets, since slot times are larger and there is a higher probability that a packet arrives. On the other hand, for higher numerologies, the impact of over-dimensioned allocated resources is reduced, because slot times are smaller. Therefore, these resources can be left unused, increasing the end-to-end latency. These phenomena are due to a combination of the inter-packet arrival time and the processing delays, and changing the inter-packet arrival time (without changing the flow rate or the fixed delays) helps to return to more intuitive results.

We gathered the results with a novel network simulator of NR technology that we have developed as extension of the well-known simulator ns-3 (an open-source, free software, discrete-event network simulator popular in research and academia). Our development is a branch of the mmWave module of ns-3 designed by New York University (NYU) Wireless and the University of Padova [4]. The software includes a rewrite of the PHY and Medium Access Control (MAC) layers of the module done by CTTC [5] to simulate Long Term Evolution (LTE) networks, adapting them to the challenges of the mmWave communication (such as propagation, beamforming, antenna models). The features that we developed are already defined in [1]. For instance, we produced a 3GPP-compliant frame structure, as well as different numerologies (from 0 to 5) and the Frequency Division Multiplexing (FDM) of such numerologies. The interested reader can find more technical information about our features in [6]. While common low-level simulators focus on link layer simulations, our simulator offers more abstraction of the PHY layer and high fidelity implementations from the MAC to the Application layer. Besides, all segments of the network are adequately developed, and End-To-End (E2E) results can be evaluated.

The outline of the paper is organized as follows. In Section II we will briefly explain what is a 5G NR numerology.

	μ=0	μ=1	μ=2	μ=3	μ=4
SCS [kHz]	15	30	60	120	240
OFDM symbol length [us]	66.67	33.33	16.67	8.33	4.17
CP length [us]	∼4.8	∼2.4	∼1.2	∼0.6	∼0.3
Subframes in a frame	10	10	10	10	10
Slots in a subframe	1	2	4	8	16
Slot length [us]	1000	500	250	125	62.5
OFDM symbols in a slot	14	14	14	14	14
Subcarriers in a PRB	12	12	12	12	12
PRB width [MHz]	0.18	0.36	0.72	1.44	2.88

TABLE I: Numerologies in 5G NR.

Then, in Section III we explain the processing delays. In Section IV we present the simulation scenario and the fixed simulation parameters, as well as the various settings that we changed to provide the overall evaluation. In Section V we present and scrutinize all the obtained results, and finally in Section VI we conclude the paper.

II. 5G NR Numerologies

The NR access technology has a flexible OFDM system to allow operation in a wide range of bands, cover multiple deployment options, address different use cases, and operate under multiple spectrum access paradigms [1]. NR Rel-15 addresses the ranges up to 52.6 GHz and defines two frequency ranges: FR1 (sub 6 GHz, 0.45 - 6 GHz) and FR2 (mmWave, 24.25 - 52.6 GHz). Larger frequency bands will be considered in NR Rel-16, although not defined yet.

With flexibility in mind, NR includes multiple numerologies, being each defined by an SCS and a Cyclic Prefix (CP) [1]. The numerology μ can take values from 0 to 4 and specifies an SCS of $15\times2^{\mu}$ kHz and a slot length of $1/2^{\mu}$ ms. The supported SCSs are in the range from 15 to 240 kHz in NR Rel-15, where larger SCSs are used at higher carrier frequencies. Note however that, in NR Rel-15, not every numerology can be used for every physical channel and signals: μ=4 is not supported for data channels and μ=2 is not supported for synchronization signals [7]. Also, for data channels, only μ=0, 1, 2 is supported in FR1 and μ=2, 3 in FR2. Likely, in NR Rel-16, larger μ and support options will be included for mmWave bands. The number of subcarriers per Physical Resource Block (PRB) in NR is fixed to 12 so that the PRB width is equal to $180\times2^{\mu}$ kHz. The frame length is 10 ms, and a frame is composed of 10 subframes of 1 ms each, to maintain backward compatibility with LTE. Each subframe has a slot number equal to $1/2^{\mu}$, which depends on the numerology configuration, and a slot is composed of 14 OFDM symbols. Therefore, the OFDM symbol length (without CP) is $1/(14\times2^{\mu})$ ms.

Table I shows the parameters of different NR numerologies. μ=0 corresponds to the LTE system configuration, while μ>0 enables larger bandwidth and shorter Transmission Time Interval (TTI), which is useful for mmWave bands and delay-critical services. Compared to LTE, in NR the SCS and the OFDM symbol length can have different values depending on the numerology that is configured, thus reducing the TTI and the access delay [8].

In addition to the numerologies, to reduce the communication delay, NR includes mini-slots and self-contained slots [7]. The purpose of mini-slots is to reduce the latency by providing more flexibility for the transmission of the small amounts of data. Mini-slots are composed of 2 OFDM symbols up to the slot length - 1 in any band, and of 1 symbol, at least above 6 GHz. Slot formats have been defined in [9], and they can contain all DownLink (DL), all UL, or at least one DL part and at least one UL part. The self-contained slot concept involves data and control in the same slot, to provide latency reduction by reducing a delay in reception of signaling such as HARQ feedback or UL grant. For example, ACK/NACK is scheduled in the same slot as DL data, or the UL transmission follows, in the same slot, the UL grant.

III. Processing Delays

An aspect that, to the best of our knowledge, is not yet adequately investigated is the impact of PHY/MAC processing delay, i.e., the time that the PHY and MAC layers at devices need to encode/decode control/data channels, on network performances. In LTE, the PHY/MAC processing delay is 2 ms (equal to the time of two subframes), so that, for example, the minimum time between a UL grant reception and a UL data transmission is 4 ms, since it involves decoding of the UL grant plus encoding of the UL data. In NR, the PHY/MAC processing delays depend on multiple factors: (i) the operational numerology, (ii) the device capability (1-baseline, or 2-aggressive), (iii) scheduling operation (slot-based, or non-slot-based), (iv) carrier aggregation (CA) mode (CA, or non-CA), (v) DMRS (demodulation reference signals) configuration, and (vi) single or multiple numerology configuration for data and control channels. Even if we have already witnessed an air time of less than 1 ms [10], detailed information on processing times of industrial implementations are not publicly available. More details about the standard procedures and definitions are presented in [11], [12].

If we take a look at the User Equipment (UE) side, we have an agreed terminology about the processing delays that affect the scheduling and HARQ timing:

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

The specific values of N1 and N2 are detailed in [11], [12] for different configurations and numerologies.

In the ns-3 NR simulator, we implemented a flexible scheme for introducing these delays, that would be easily extendible, and that would take into account even future modifications to the standard. Therefore, the previously explained PHY/MAC processing delays are introduced through the following two parameters:

	# *Flows*	*Start time* (s)	*App. Rate (Mb/s)*	*Segment Size (B)*	*RAN Dir.*	*TCP ACKs Dir.*
Video (UDP)	4	2	10	1400	UL	X
Sensor (UDP)	6	2	1.6	500	UL	X
Smartphone Upload (TCP)	25	[25 , 75]	X	1440 (ACK 40)	UL	DL
Smartphone Download (TCP)	125	[5 , 95]	X	1440 (ACK 40)	DL	UL

TABLE II: Application settings, if a setting does not apply it is marked with an "X"

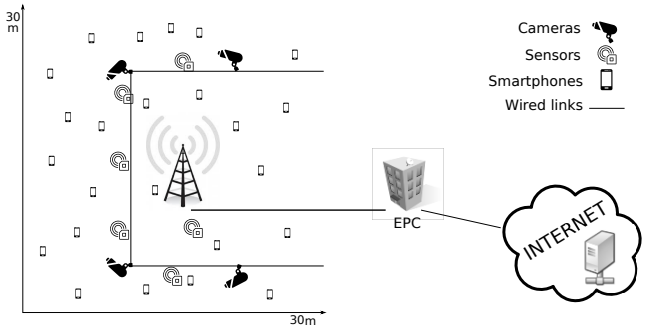

Fig. 1: Reference scenario

Parameter	*Value*
Channel Model	3GPP
Channel Condition	Line-Of-Sight
Channel bandwidth	100 MHz
Channel central freq.	28 GHz
Scenario	Urban (UMa)
Shadowing	false
Beam Angle Step	10 degrees
Beamforming Method	Beam Search
Modulation Coding Scheme	Adaptive
Ctrl/Data encode latency	2 slots
Radio Scheduler	Round-Robin

TABLE III: Relevant simulation parameters

- *decodingLatency*: the time that the PHY layer needs to process 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 is available to process in the MAC class. It applies both at gNB PHY and UE PHY.
- *gNBLatency*: 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 RLC class by the MAC class and the moment at which the control/data is available to go over the air.

According to NR specifications, these parameters are numerology- and device-dependent. For the parameter *gNBLatency* we set an LTE-compatible value of 2 slots: it is reasonable for *gNBLatency* to be numerology-dependent because the MAC scheduler at gNB works on a slot-basis. We vary the *decodingLatency* to analyze its E2E impact, setting slot-dependent values as well as fixed values. These settings make sense for a device, independently of its operational numerology, because the decoding time is related to the Central Processing Unit (CPU) rate, as well as the available energy to perform the task.

IV. Simulation Settings and Parameters

A. Reference Scenario

To model a real-world scenario, we base our simulation on the setup of Figure 1. At a high level, we have a backbone connection between Evolved Packet Core (EPC) to remote nodes modeled as 100 Gb/s point-to-point link. The link between the gNB and the EPC that represents the Core Network (CN) is made with another point-to-point connection with a maximum rate of 10 Gb/s, without propagation delay.

Regarding the Radio Area Network (RAN), we consider the use case of a next-generation school, served by a single gNB, in which different but connected objects share the connectivity. We have twenty-five smartphones, six sensors, four IP cameras distributed over a circular area of 30 m of diameter. The position of each UE in the reference scenario depicted in Figure 1 is indicative because in the simulations we have located the UEs in random positions to gather more statistical significance in the results.

For the traffic types, the video and sensor nodes have one UDP flow each, that goes in the UL towards a remote node on the Internet. These flows are fixed-rate flows: we have a continuous transmission of 10 Mb/s for the video nodes, to simulate a 720p24 HD video, and the sensors transmit a payload of 500 bytes each 2.5 ms, that gives a rate of 1.6 Mb/s. Table II summarizes the UDP flow characteristics. For the smartphones, we use TCP as the transmission protocol, with a state-of-the-art implementation [13], [14]. Each UE has to download five times a 5 MB file (so the downloads count as five different flows) and to upload one file of 15 MB. These flows start at different times: the upload can start at a random time between the 25th and the 75th simulation seconds, while each download can start between the 5th and the 95th simulation seconds. Table II summarizes the details.

B. Simulations campaign

We compare NR numerologies, from 0 to 4, and analyze the TCP goodput (the average rate at which the receiver application gets the data) and the UDP one-way delay (the average latency of each UDP packet from source to destination), as well as the average per-second rate. Other relevant parameters for the simulations are reported in Table III. For each μ, we have performed multiple sets of simulations in the ns-3 network simulator, with different decoding latencies, represented by the parameter *decodingLatency*.

We consider four values for the decoding latency setting: 1) the ideal condition, in which the signal takes no time to reach the MAC layer (0 ms case); 2) a fixed value of 0.1 ms, representing high-speed decoding; 3) a fixed value of 0.5 ms, as in literature [15]; 4) a slot-dependent latency value, of twice the slot length (which varies accordingly with the numerology). Inside a single simulation, we average the flow

performance of each class (video, sensor, TCP download, TCP upload) by using a geometric mean.

To obtain statistical significance, we repeat the same simulations using five different random seeds. In this way node positions, flow start times, and many other factors result randomized. Then, we use the geometric mean to average the result of the same traffic class with different seeds.

V. Simulation Results

In this section we evaluate the E2E goodput (TCP) and latency (UDP) for different numerologies and processing delays. In the first subsection, we will show the results gathered by using the segment size defined in the previous section. In the second subsection, we show that with different packet sizes (but same rate) we can impact end-to-end UL latency results.

A. Typical packet sizes

TCP Upload Goodput. We start the analysis by looking at the goodput for the TCP upload flows, in Figure 2. For the numerology μ=0, the performances are almost the same for all the delay configurations, with values close to the 20 Mb/s mark. When we increase the numerology, from 0 to 1, we see a remarkable increase in goodput, almost reaching the 80 Mb/s mark. Here we start to see some differences between the ideal case and the others. Increasing to μ=2, we gain additional 40 Mb/s in each case. Numerologies, 3 and 4 offer an additional increase of about 20 Mb/s. The different configurations achieve, more or less, the same performance. We observe that normally best results are obtained for the case of 0 ms processing delay. Another important thing is that the processing delay dependant on the slot length offers the worst performance in the lowest numerology (0 and 1), but starts to recover (and eventually in the last numerology outperforms the others) with the reduction of the slot time itself, due to the increasing numerology.

TCP Download Goodput. In Figure 3 we can analyze the performance of the TCP Download flows. The trend while increasing the numerology follows what we have seen in the TCP Upload, with an increasing goodput each time the numerology is increased, and the slot-dependent delay that is gaining at the end, with the smallest slot length. In absolute values, the downloads have a slightly higher performance compared to uploads. In particular, comparing the upload and the download goodput in the same numerology, it is easy to see that *the download goodput is almost 10 Mb/s higher than the upload goodput*. This difference is due to the absence, in the DL, of the SR/UL Grant control messaging. When the data arrives in the buffers of the gNB, it will take scheduling decision almost immediately. When the data is waiting in the UE's buffer, instead, the permission to transmit is not immediate, but has to be granted by the gNB, involving a signaling exchange that, albeit slightly, increases the round trip time and therefore reduces the TCP goodput.

TCP remarks. In general, when using a low SCS, the PRB width is lower, so that there are more PRBs available and

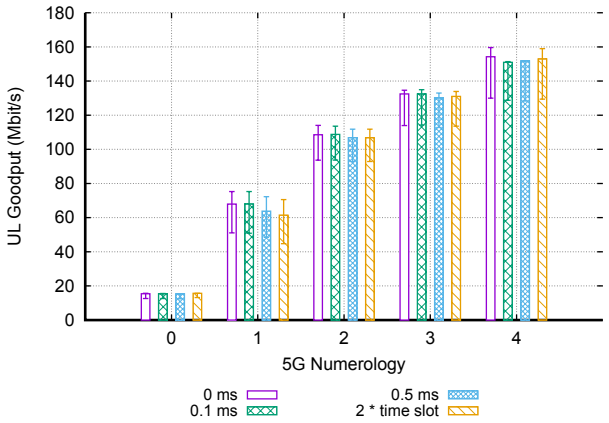

Fig. 2: TCP Upload goodput. Top whisker represents the maximum value, the bottom whisker represents the minimum value, and the box is at the 80th percentile.

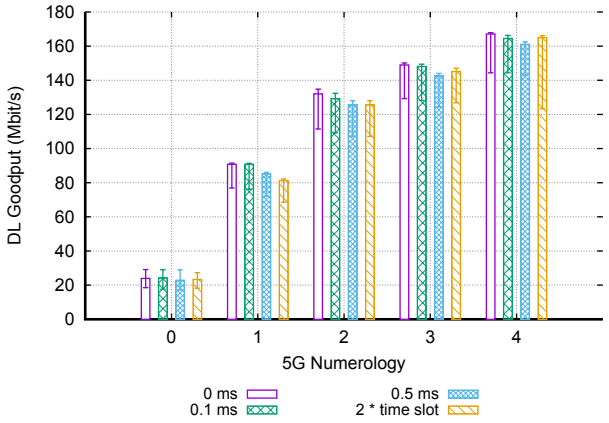

Fig. 3: TCP Download goodput. Top whisker represents the maximum value, the bottom whisker represents the minimum value, and the box is at the 80th percentile.

the bandwidth is more efficiently utilized, which results in a larger achieved goodput. However, when using TCP, and especially variants that derive from TCP NewReno, we can observe an opposite trend. In particular, the rate increase is directly proportional to the reduced E2E round trip time, and so it benefits from higher numerologies. Therefore, a flow that has a fixed amount of data to transmit will transmit it in a shorter time at higher numerologies (because of the reduced latency), which results in a higher goodput as well.

Sensor UDP Delay. In Figure 4 we can see the latency performance of the sensor flows. In the first three numerologies, the worst performance is achieved by the delay configuration that is tied to the slot length. The explanation naturally follows if we keep in consideration that, in these numerologies, the slot length is much more than the fixed values we are considering. Instead, when the slot length is reduced, the performance starts to equal the fixed delays (the perfect example is represented by the equality, for μ=2, of the last two cases: in fact, the slot length is equal to 0.25 ms, exactly half of the fixed delay of 0.5 ms). The best performance is offered by the ideal case of 0 ms decoding latency. In absolute values, increasing the decoding latency from 0 ms to 0.1 ms adds approximately 0.1 ms to the latency performance. The linear increase also

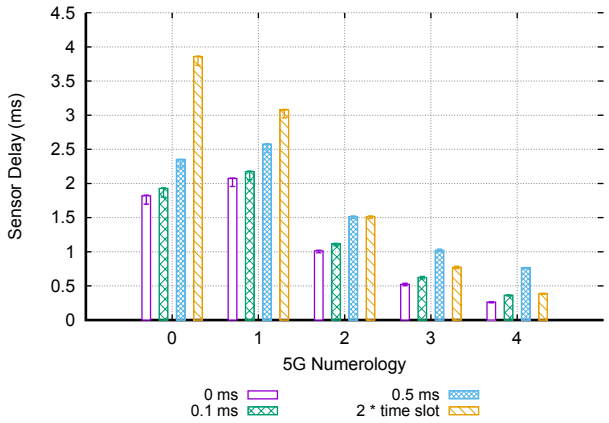

Fig. 4: Sensor delay. Top whisker represents the maximum value, the bottom whisker represents the minimum value, and the box is at the 80th percentile.

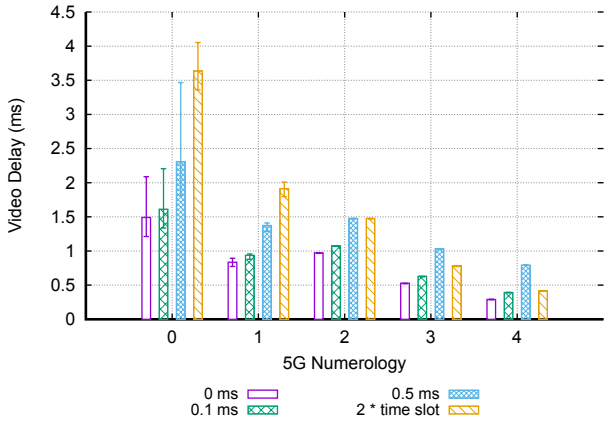

Fig. 5: Video delay. Top whisker represents the maximum value, the bottom whisker represents the minimum value, and the box is at the 80th percentile.

applies when passing from 0.1 ms to 0.5 ms: that difference is added, almost without change, in the end-to-end delay value. These observations allow us to conclude that *the analyzed fixed delays in the decoding impact the overall latency linearly*, without affecting other phases.

Video UDP Delay. For the latency performance of the Video flows, we can refer to Figure 5. Here we observe a similar trend to that shown by the sensor delay, but with more variance in the first two numerologies. If we compare the video latency values with the sensor values, we can see that in some cases a video packet experiences a lower delay than a sensor packet, regardless of the decoding time. For instance, let us consider the slot-tied case, on the right side of each numerology. In the sensor case, for μ=1 we have a value of 3.08 ms, while for the video case the latency is only 2.0 ms. For μ=2, the video delay is 1.47 ms, and sensor latency is 1.52. It can seem counter-intuitive that a low rate flow (only 1.6 Mb/s injected into the network) experiences higher latency than a higher rate flow at 10 Mb/s. The reason for that is that, before doing an UL transmission, it is necessary to have the UL Grant from the gNB. A grant comes from an explicit SR, or following a Buffer Status Report (BSR) message sent along user data in a previously granted space. If a data packet meets an empty Radio Link Control (RLC) buffer, the UE is forced to send the SR message to get an UL grant from the gNB MAC scheduler. On the other side, if the RLC buffer already contains data at the time the packet arrives, it is very likely that the UE sent earlier the SR, and all the upcoming data (until the buffer will be emptied) will be sent in grants that come automatically after the BSRs. This effect is also emphasized by the fact that, a continuously filled RLC buffer can exploit granted, but unused, uplink TBS space.

In light of the previous explanation, we can demonstrate why the lighter sensor flows experience higher delay than the video flows. A sensor flow, to achieve a rate of 1.6 Mb/s, transmits 500 B of data for 400 times in a second. The video flow, instead, has to achieve the rate of 10 Mb/s with packets of 1400 B, and therefore it sends a packet almost every millisecond (in one second the camera should send almost 893 packets). As a result, we deduce that the increased latency is due to the SR control message, generated by the lighter flow each time a packet is received in the RLC buffer. The RAN is so fast that the time in between two consecutive sensor packets is enough to empty the RLC buffer, thus requiring, that an extra control message is sent without piggy-backing it with data, if new data arrives. To remark why this is so important, we must say again that the UE sends the SR without data, and it must wait until the UL Grant arrives, before transmiting. In case of the video flows, the arrival rate in the RLC buffer is higher, and therefore there is a higher probability that the BSR sent with the data would trigger others UL grants, without incurring in the penalization of sending SR without any data piggy-backed. *In this way, the flow of data is never interrupted, because bits and BSR sent in this slot will automatically create an UL Grant for new data in one of the following slots.* In some cases, the newly arrived data can be aggregated an transmitted in unused over-dimensioned TBS allocated resources, further reducing the latency experienced by some packets.

UDP Remarks. Looking at the fixed-value case, with decoding latency of 0 ms - 0.1 ms - 0.5 ms, we can see that there are two strange cases in which an increased numerology corresponds to an increase in latency. The increase happens in the sensor flows when passing from μ=0 to μ=1 and in the video flow when passing from μ=1 to μ=2. How is it possible that half the slot time corresponds to more latency experienced by a single packet? The reason lies, again, in the SR mechanism. The data arrival rate in the RLC buffers, together with the transmission and the processing time, determines if the UL flow needs a SR, or it can rely on the BSR, to continue the UL transmission. The data arrival rate is fixed in all the experiments, while the transmission and processing time change with the numerology. The two components generate a synergy for which it is necessary to send a SR to restart the data flow, increasing the latency. This generates the unfortunate event in which for lower numerologies the number of SR is lower than in the higher numerologies. Even if the slot time is lower, the overhead for the increased number of

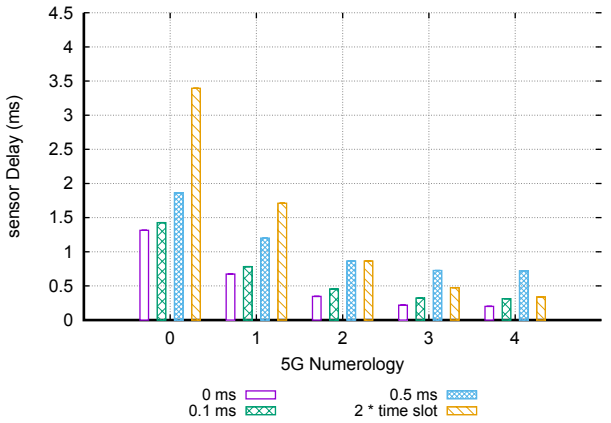

Fig. 6: Sensor delay with 50 B packets. The box represent the geometric average of the scenario sensor flows.

SR is reflected in the latency value plotted in Figure 4 (μ=1) and Figure 5 (μ=2). The overall increase does not appear for very high numerologies (μ=3 and μ=4) because the slot time starts to have very small values (less than 0.2 ms) and a single packet (of 500 or 1400 B) in the middle of many concurrent flows requires more than one slot to be sent entirely, therefore continuing to get grants through the mean of BSR.

B. Smaller Packets

When we reduce the packet sizes, but we maintain the same generation rate (as in Table II), we increase the packet arrival rate in the RLC buffers. We chose a packet size of 50 B, enough to increment the packet arrival rate to 4000 Packets Per Second (pps) in the case of the sensor flow, and 25000 pps in the case of the video flows. As we can see in Figure 6, in which we plot the new values of the sensor delays, increasing the numerology always corresponds to a decreased latency. Here, the number of SR does not increase, because there will always be data in the buffers, and therefore the flows will continue to transmit data through grants given through BSR. In addition, packets that arrive after the reception of the grant can be aggregated and transmitted in the over-dimensioned allocation in terms of TBS, if it is the case. The latency trend corresponds to what we would expect by reducing the slot time, and even if not depicted for space constraints, holds for the video flows as well. Therefore, it is important to remember that for UL flows, especially in the first three numerologies, the inter-packet arrival rate in the RLC buffers has a dominant impact on the resulting latency.

VI. Conclusions

In this paper, we have analyzed a complex and realistic 5G future scenario, in which the traffic from different 5G applications was competing for resources on the NR RAN and in the Core Network. We have studied the impact of processing and decoding delays in different numerologies and shown that not always an increase in numerology corresponds to a decrease in latency. We have shown that in the uplink case, the inter-packet arrival rate plays an essential role in the overall latency. The reason is that, for some combination of inter-packet arrival rate and fixed decoding delays, increasing the numerology also increases the overhead of scheduling messages, which impacts the overall latency. Besides, there are cases in which, with a higher numerology, the over-dimensioned TBS cannot be filled due to the absence of data packets in the RLC layer. Thus, we argue that the selection of the numerology in 5G systems should take into account the traffic pattern, to foster piggy-backed Buffer Status Report messages into uplink data, and so avoid as much as possible the latency-killing increase of Scheduling Request messages. As future work, we plan to investigate the differences in throughput and delay between UL grant-free and grant-based access when changing the numerology, for different application loads. In addition, another interesting problem is the automatic selection of the appropriate numerology for each UE.

VII. Acknowledgments

This work was 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 TR 38.912, *TSG RAN; Study on New Radio (NR) access technology*, Release 14, v14.0.0, Mar. 2017.

[2] F. Boccardi, R. W. Heath, A. Lozano, T. L. Marzetta, and P. Popovski, "Five Disruptive Technology Directions for 5G," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 74–80, 2014.

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

[4] M. Mezzavilla, M. Zhang, M. Polese, R. Ford, S. Dutta, S. Rangan, and M. Zorzi, "End-to-End Simulation of 5G mmWave Networks," *IEEE Communications Surveys & Tutorials*, pp. 1–1, 2018.

[5] N. Baldo, M. Miozzo, M. Requena-Esteso, and J. Nin-Guerrero, "An Open Source Product-oriented LTE Network Simulator Based on Ns-3," in *Proceedings of the 14th ACM International Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems*, MSWiM '11, (New York, NY, USA), pp. 293–298, ACM, 2011.

[6] 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 2018 Workshop on ns-3, Surathkal, India, June '18*, WNS3 2018, pp. 37–44, ACM, 2018.

[7] 3GPP TS 38.300, *NR; Overall description; Stage-2*, Release 15, v15.1.0, Apr. 2018.

[8] S. Lagen, B. Bojovic, S. Goyal, L. Giupponi, and J. Mangues-Bafalluy, "Subband Configuration Optimization for Multiplexing of Numerologies in 5G TDD New Radio," *IEEE Int. Symp. on Personal, Indoor and Mobile Radio Commun.*, Sep. 2018.

[9] 3GPP TR 38.211, *TSG RAN; NR; Physical channels and modulation*, Release 15, v15.1.0, Apr. 2018.

[10] J. Pilz, M. Mehlhose, T. Wirth, D. Wieruch, B. Holfeld, and T. Haustein, "A Tactile Internet demonstration: 1ms ultra low delay for wireless communications towards 5G," in *2016 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pp. 862–863, April 2016.

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

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

[13] M. Casoni and N. Patriciello, "Next-generation TCP for ns-3 simulator," *Simulation Modelling Practice and Theory*, vol. 66, pp. 81 – 93, 2016.

[14] N. Patriciello, "A SACK-based Conservative Loss Recovery Algorithm for Ns-3 TCP: A Linux-inspired Proposal," in *Proceedings of the Workshop on Ns-3, Porto, Portugal, June '17*, WNS3 2017, pp. 1–8, ACM, 2017.

[15] T. Wirth, M. Mehlhose, J. Pilz, B. Holfeld, and D. Wieruch, "5G New Radio and Ultra Low Latency Applications: A PHY implementation perspective," in *2016 50th Asilomar Conference on Signals, Systems and Computers*, pp. 1409–1413, Nov 2016.