An E2E Simulator for 5G NR Networks

Natale Patriciello, Sandra Lagen, Biljana Bojovic, Lorenza Giupponi

Centre Tecnològic de Telecomunicacions de Catalunya (CTTC/CERCA)
Avinguda Carl Friedrich Gauss, 7
08860 Castelldefels, Barcelona, Spain
{*npatriciello, slagen, bbojovic, lgiupponi*}*@cttc.cat*

Abstract

As the specification of the new 5G NR standard proceeds inside 3GPP, the availability of a versatile, full-stack, End-To-End (E2E), and open source simulator becomes a necessity to extract insights from the recently approved 3GPP specifications. This paper presents an extension to ns-3, a well-known discrete-event network simulator, to support the NR Radio Access Network. The present work describes the design and implementation choices at the MAC and PHY layers, and it discusses a technical solution for managing different bandwidth parts. Finally, we present calibration results, according to 3GPP procedures, and we show how to get E2E performance indicators in a realistic deployment scenario, with special emphasis on the E2E latency.

Keywords: ns-3, NR, network simulator, E2E evaluation, calibration.

1. 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 flexible, scalable, and forward-compatible Physical (PHY) and Medium Access Control (MAC) layers to support a wide range of center carrier frequencies, deployment options, and variety of use cases. To account for that, and as

Email addresses: `natale.patriciello@cttc.cat` (Natale Patriciello), `sandra.lagen@cttc.cat` (Sandra Lagen), `biljana.bojovic@cttc.cat` (Biljana Bojovic), `lorenza.giupponi@cttc.cat` (Lorenza Giupponi)

compared to Long Term Evolution (LTE), NR includes new features, such as a flexible frame structure by means of multiple numerologies support, dynamic Time Division Duplex (TDD), support for new millimeter-wave (mmWave) frequency bands, beam management-related operations, support for wide channel bandwidth operations and frequency-division multiplexing of multiple bandwidth parts, symbol-level scheduling through mini-slots and variable Transmission Time Interval (TTI)s, and new channel coding schemes. The new NR features span over all the protocol stack, also introducing a new layer above Packet Data Convergence Protocol (PDCP), called Service Data Adaptation Protocol (SDAP), standalone and non-standalone architectures, and a mandatory split of control and user planes in the core network.

Research institutions or Small and Medium Enterprise (SME)s that cannot develop sophisticated simulation tools capable of simulating 5G and beyond networks due to the cost, time effort, and required human resources, are at risk of being cut out from the early stages of the development process. Some of them rely on analytic simulation methods. However, the assumptions and simplifications in the sender and receiver nodes, as well as in other network segments and layers, limit the generality of the extracted results. Moreover, it is tough to represent external network dynamics (like the burstiness of data traffic) or to assess the interaction with the core network and the mobility of the users, without a solid full-stack E2E simulation model.

As researchers, we are not only interested in the low-level characterization of the previously mentioned NR features but also we want to have an overall view of the system, which starts from the application level to the PHY layer and includes an E2E performance evaluation from the User Equipment (UE) to the remote host. Our objective is to properly evaluate the performance of a sophisticated and flexible technology, NR, and to be able to conduct interoperability studies with other technologies. As a result, in this work, we present an NR network simulator that has been built as a pluggable module to ns-3[1]. The simulator

[1] www.nsnam.org

models the NR technology with a high-fidelity full protocol stack, and it has been calibrated according to 3GPP procedures. In particular, our simulator offers an abstraction of the PHY layer and high-fidelity implementations from the MAC to the application layer. It can be used to evaluate cross-layer and E2E performance, as well as a platform to assess the coexistence of NR with other technologies. As an example of the capabilities offered by the simulator, we already used it to evaluate the impact of the processing and decoding times in the E2E delay in [2], for different NR numerologies.

We believe that the open distribution of this simulator, under the terms of the GPLv2 license, represents an unprecedented contribution to the community and facilitates innovation in the area of 5G. The GPLv2 adoption is necessary because our simulator is derived from ns-3, which is GPLv2-licensed. Even if in a lot of GPL-covered software there is an explicit clause that permits the user to choose any later version of the license, the ns-3 simulator explicitly disabled this option, thus making our simulator effectively released under the sole GPLv2 license (as much as the Linux kernel). People used to GPLv3 software have to keep in mind the following main differences between the two GPL versions that apply to our software:

- GPLv2 code cannot be combined with a range of software licenses that are now compatible with GPLv3 (the most prominent case is Apache v2, the license of the OpenAir Interface software);
- The GPLv2 license does not contain an explicit patent license.

There is a vast literature on the GPLv3 versus GPLv2 license, so for space constraints, we will not expand on all the specific differences. The interested reader is referred to the GPL FAQ[2] or the many resources available on the Web[3]. We would like though to remind that, in the pure open spirit, we are using a very widely used license to foster the development and research around

[2] https://www.gnu.org/licenses/gpl-faq.html

[3] https://en.wikipedia.org/wiki/GNU_General_Public_License

5G concepts.

The objective of this paper is to give the reader a complete overview of the NR simulator, including its supported features and descriptions regarding the additions and modifications with respect to the original ns-3 modules from which it was generated, i.e., the LTE [3] and the mmWave [4] ns-3 simulation models. We have designed the NR model by following as much as possible the latest 3GPP specifications, which we reference through the document when appropriate. Besides, we present some examples of usage and simulation results. We start in Section 2 by giving a brief overview of the related work in the 5G simulation and emulation domain, as well as positioning our contribution with respect to the scientific community. Section 3 gives an overview of the NR simulator components, before entering into the details of the PHY layer in Section 4 and the MAC layer in Section 5. In Section 6, our innovative work for managing different bandwidth parts is described. Then, we present the calibration procedure and an example of usage in a realistic 5G scenario in Section 7. Finally, Section 8 discusses our roadmap and future plans, and Section 9 concludes the work.

2. Scientific Contribution

In this section, we briefly analyze the existing available simulation tools, and then we highlight the novelty with respect to other works, and the contribution to the scientific community provided by our simulator.

2.1. Related work

A key challenge to perform new technology evaluations is that, despite the large body of results presented in the literature and produced by 3GPP evaluation working groups, the simulators are not publicly available. Usually, the obtained results are not reproducible, and system performance metrics are presented without much detail revealed about the underlying models and assumptions. Normally, simulators used by companies in 3GPP are required to pass through a calibration procedure, but they are private, and consequently not available to the research community. There are private commercial simulators that are available after paying an annual license fee for using them. Often, if not in all cases, the license is very restrictive and does not allow modifications or inspection of the source code, which is a clear limit for the research and the potential innovation. Our simulator, in turn, is GPLv2-licensed and guarantees the freedoms of the free software (free as in speech, not free as in beer) movement. We do not advocate for any political positions in this paper, but at the same time, we are interested in fostering the reproducibility of results, the collaborative development, and the support to the open innovation. Therefore, in the short review that follows, we only focus on software that is openly available and that guarantees, at least for academic purposes, the same freedom as we are guaranteeing.

The OpenAirInterface [5] is an open source platform for the simulation of wireless networks. Since April 2018, it supports the NR specifications. It can be used in a real testbed, but it misses a more comprehensive simulation setup that includes configuration and tracing of variables, as well as integration with other technologies to conduct interoperability experiments. Differently, our simulator is covered by the ns-3 umbrella, which includes models for multiple technologies like LTE and WiFi, among others, and therefore it offers the option of evaluating multi-Radio Access Technology (RAT) coexistence scenarios.

Concerning system-level simulators, an interesting software is the Vienna Simulator [6]. The research group from TU Wien developed a MATLAB tool that allows researchers to perform link- and system-level simulations for LTE and NR. In combination with several propagation models, the Vienna simulator allows simulating the network performance based on signal strength and accumulated interference. Generally speaking, the simulator is of interest to people working at the PHY and MAC layers. Differently, network simulators need to abstract the PHY layer through look-up tables to reduce the computational time and focus more on the MAC and higher layers.

The last category is the domain of the network simulation, in which our simulator is classified. Besides ns-3, from which we derive for the reasons that we explain in the next subsection, there is OMNeT++ [7]. OMNeT++ has support for LTE and LTE-Advanced features but lacks support for NR.

2.2. Contribution

The NR module starts as a fork of the ns-3 mmWave simulation tool developed by New York University (NYU) and University of Padova [4]. The mmWave simulation model imports fundamental LTE features from the ns-3 LTE module (LENA) [3], which has been entirely designed and developed at Centre Tecnològic de Telecomunicacions Catalunya (CTTC). Our contributions entail a comprehensive and intensive work to align the mmWave module to the latest NR standard published by 3GPP. We have chosen as a base the popular ns-3 framework, as well as the mmWave simulation tool, for many reasons.

On the one hand, ns-3 is an open-source discrete-event network simulator, and thus we inherit the capability of tracing internal events, a flexible configuration system, and a variety of modules to simulate other technologies, such as Ethernet, LTE, or WiFi, and multi-technology scenarios. As such, we can model different segments of the same network and, when needed, inter-connect different systems. The ns-3 simulator is widely adopted, recognized in research and academia, well maintained by an active community, and receives typically also support from the Google Summer of Code, as a flagship open source project. Thanks to ns-3, we also have the possibility of running our model in real time, therefore leaving simulation and entering the emulation domain with real equipment. We discussed the initial development of our simulator in [8].

On the other hand, we have chosen the mmWave simulation tool as our starting point because it includes already multiple features that are of interest for NR, mainly to access the mmWave spectrum above 6 GHz. In particular, the researchers from NYU and the University of Padova have done a solid job in modeling the aspects of beamforming, antenna gain, and propagation channel models, which makes the resulting work a milestone in the history of ns-3. However, the implementation of the mmWave module started in a moment in time when NR 3GPP specifications were not available, and the general vision of the technology was not as stable as it is today. As a result, many implemented aspects were not standard compliant and needed a revision, like the frame structure. Other things, such as the modeling of the channel and the beam management representation, were, on the contrary, entirely in line with the 3GPP standard, and therefore we have not modified them.

The mmWave module derives from the ns-3 LTE module (LENA) [3], so that both the mmWave and NR modules are also highly influenced by the previous design of the LTE module. In particular, both modules reuse from LTE all the higher protocol (Radio Link Control (RLC), PDCP, Radio Resource Control (RRC), Non-Access Stratum (NAS)), as well as the Evolved Packet Core (EPC). However, we have done work to upstream the NR module to the ns-3 framework. As such, the NR module will further benefit from additions that future contributors will make to the ns-3 simulator, and it can reuse exciting features such as the Direct Code Execution [9]: users can perform simulations with realistic TCP/IP implementations and existing applications. In addition, in the spirit of ns-3, and inherited from the LTE module, the NR model abstracts some low-level details, interpolating the values from a static set of look-up tables, to reduce computational aspects and facilitate simulating wider and complex scenarios with many base stations and users.

The rest of the paper is dedicated to explaining the design and implementation of our Non-Standalone (NSA) NR simulator, which includes 4G EPC and 5G Radio Access Network (RAN). The main NR features that we have added and modified to the mmWave tool are:

- flexible and automatic configuration of the NR frame structure through multiple numerologies;
- Orthogonal Frequency-Division Multiple Access (OFDMA)-based access with variable TTIs;
- restructuring and redesign of the MAC layer, including new flexible MAC schedulers that simultaneously consider time- and frequency-domain re-

sources (resource blocks and symbols) both for Time-Division Multiple Access (TDMA) and OFDMA-based access schemes with variable TTI;

- UpLink (UL) grant-based access scheme with scheduling request and 3GPP-compliant buffer status reporting;
- NR-compliant processing timings;
- new Bandwidth Part (BWP) managers and the architecture to support operation through multiple BWPs.

3. NR module overview

We designed the NR module to be able to perform E2E simulations of 3GPP-oriented cellular networks. The E2E overview of a typical simulation with the NR model is drawn in Figure 1. On one side, we have a remote host (depicted as a single node in the figure, for simplicity, but there can be multiple nodes) that connects to an Service GateWay (SGW)/Packet data network GateWay (PGW), through a link. Such a connection can be of any technology that is currently available in ns-3. It is currently implemented through a single link, but it can be replaced by an entire subnetwork with many nodes and routing rules. Inside the SGW/PGW, the EpcSgwPgwApp encapsulates the packet using the GPRS Tunneling Protocol (GTP) protocol. Through an IP connection, which represents the backhaul of the NR network (again, represented with a single link in the figure, but the topology can vary), the GTP packet is received by the next-Generation Node B (gNB). There, after decapsulating the payload, the packet is transmitted over the RAN through the entry point represented by the class NRGnbNetDevice. The packet, if received correctly at the UE, is passed to higher layers by the class NRUeNetDevice. The path crossed by packets in the UL case is the same as the one described above but on the contrary direction. We will detail our modifications to support NR in the PHY classes in Section 4.

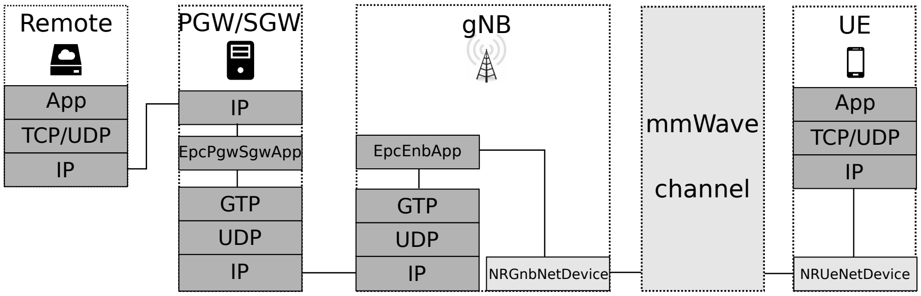


Figure 1: End-to-end class overview. In dark gray, we represent the existing, and unmodified, ns-3 and LENA components. In light gray, we represent the mmWave/NR components.

Concerning the RAN, we detail what is happening between NRGnbNetDevice and NRUeNetDevice in Figure 2. The NRGnbMac and NRUeMac MAC classes implement the LTE module Service Access Point (SAP) provider and user interfaces, enabling the communication with the LTE RLC layer. The module supports RLC Transparent Mode (TM), Saturation Mode (SM), Unacknowledged Mode (UM), and Acknowledged Mode (AM) modes. The MAC layer contains the scheduler (NRMacScheduler and derived classes). Every scheduler also implements a SAP for LTE RRC layer configuration (LteEnbRrc). The NRPhy classes are used to perform the directional communication for both DownLink (DL) and UL, to transmit/receive the data and control channels. Each NRPhy class writes into an instance of MmWaveSpectrumPhy class, which is shared between the UL and DL parts. We did not modify the internal of MmWaveSpectrumPhy and, as for the original design of mmWave, it contains many PHY-layer models: interference calculation, Signal-to-Interference-plus-Noise Ratio (SINR) calculation, the Mutual Information (MI)-based error model (to compute the packet error probability), as well as the Hybrid ARQ PHY-layer entity to perform soft combining. We will detail our modifications to support NR in the MAC classes in Section 5.

Interesting blocks in Figure 2 are the NRGnbBwpM and NRUeBwpM layers. 3GPP does not explicitly define them, and as such, they are virtual layers, but they help construct a fundamental feature of our simulator: the

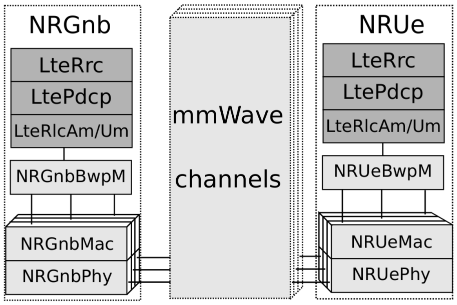


Figure 2: RAN class overview. In dark gray, we represent the existing, and unmodified, LENA components. In light gray, we represent the mmWave/NR components.

multiplexing of different BWPs. NR has included the definition of BWP for energy-saving purposes, as well as to multiplex a variety of services with different Quality of Service (QoS) requirements [1, Sect. 6.10]. In our simulator, it is possible to divide the entire bandwidth into different BWPs. Each BWP can have its own PHY and MAC configuration (e.g., a specific numerology, scheduler rationale, and so on). We added the possibility for any node to transmit and receive flows in different BWPs, by either assigning each bearer to a specific BWP or distributing the data flow among different BWPs, according to the rules of the BWP manager. The introduction of a proxy layer to multiplex and demultiplex the data was necessary to glue everything together, and this is the purpose of these two new classes (NRGnbBwpM and NRUeBwpM). Everything briefly explained here will be analyzed in Section 6.

4. Physical Layer

One of the fundamental features that the NR simulator must support is the flexible frame structure defined in NR specifications [1].

LTE considers three types of frame structures: Type 1 for Frequency Division Duplex (FDD), Type 2 for TDD, and Type 3 for use in the unlicensed spectrum (Licensed-Assisted Access). In general, each of these frames contains ten subframes, of 1 ms each. The ns-3 LTE model supports FDD and focuses on Type 1 frame structure. In that model, the subframe is, as per [10], organized into two slots of 0.5 ms length. The model considers the implementation of the normal Cyclic Prefix (CP), which results in seven Orthogonal Frequency Division Multiplexing (OFDM) symbols per slot. The MAC scheduler of the ns-3 LTE model, however, only allocates resources within the subframe granularity in the Physical Downlink Shared Channel (PDSCH) and Physical Uplink Shared Channel (PUSCH).

Differently, NR introduces the concept of numerology, defined by a SubCarrier Spacing (SCS) and a CP [11, 1], and defines a set of numerologies to be supported. Currently, six frame structures are supported in the standard: 1) $\mu = 0$ (SCS=15 KHz) and normal CP, 2) $\mu = 1$ (SCS=30 KHz) and normal CP, 3) $\mu = 2$ (SCS=30 KHz) and normal CP, 4) $\mu = 2$ (SCS=30 KHz) and extended CP, 5) $\mu = 3$ (SCS=60 KHz) and normal CP, and 6) $\mu = 4$ (SCS=120 KHz) and normal CP [12, Sect. 4.3.2].

The NR frame length is 10 ms, and an NR frame is composed of ten subframes of 1 ms each, to maintain backward compatibility with LTE. However, differently from LTE, each subframe is split in the time domain into a variable number of slots that depends on the configured numerology. In particular, as the SCS increases, the number of slots per subframe increases and the slot length reduces. The number of OFDM symbols per slot depends on the CP length: 14 OFDM symbols per slot for normal CP, and 12 OFDM symbols per slot for extended CP. In the frequency domain, the number of subcarriers per Physical Resource Block (PRB) is fixed to 12 (as in LTE). Actually, for $\mu = 0$, the NR frame structure is equal to that of LTE, although the concept of *slot* has changed. Thus, according to NR specifications and differently from LTE, the NR frame structure is flexible and allows different OFDM symbol lengths, slot lengths, and the number of slots per subframe. In addition, more flexibility is added to the NR MAC scheduler, which works on a numerology-dependent slot-basis (instead of a fixed 1 ms subframe-basis, as in LTE).

The ns-3 NR model implements TDD and supports the different NR numerologies, as shown in Table 1 for normal CP. As it can be observed, differently from LTE, the SCS, the slot length, the OFDM symbol length, and the

	$\mu = 0$	$\mu = 1$	$\mu = 2$	$\mu = 3$	$\mu = 4$
SCS [kHz]	15	30	60	120	240
OFDM symbol length [us]	66.67	33.33	16.67	8.33	4.17
Cyclyc prefix [us]	∼4.8	∼2.4	∼1.2	∼0.6	∼0.3
Number of subframes in frame	10	10	10	10	10
Number of slots in subframe	1	2	4	8	16
Slot length [us]	1000	500	250	125	62.5
Number of OFDM symbols in slot	14	14	14	14	14
Number of subcarriers in a PRB	12	12	12	12	12
PRB width [MHz]	0.18	0.36	0.72	1.44	2.88

Table 1: NR numerologies

CP length have different values depending on the numerology that is configured. Also, as the SCS varies with the numerology, the number of PRBs within the system bandwidth is numerology-dependent, as illustrated in Table 1.

To configure the numerologies in the NR module in a user-friendly way, we have a single numerology attribute value to be set in the class NrPhyMacCommon, which refers to μ. Then, differently to the original code, we derive the PHY layer parameters as follows:

- The number of slots per subframe (n) is: $n = 2^{\mu}$;
- The slot period (t_s) is set to: $t_s = \frac{1}{n}$ ms;
- The number of OFDM symbols per slot is fixed to 14, according to normal CP;
- The OFDM symbol period (t_{os}) is set to: $t_{os} = \frac{t_s}{14}$ ms;
- The number of subcarriers per resource block is fixed to 12, as per 3GPP specifications;
- The subcarrier spacing (SCS) is: $SCS = 2^{\mu} * 15 * 1000$ Hz;
- For a total bandwidth of BW, the number of resource blocks (N) is then: $N = \lfloor \frac{BW}{SCS*12} \rfloor$, being $\lfloor . \rfloor$ the floor function.

Our implementation currently supports the numerologies[4] shown in Table 1. It also supports $\mu = 5$, which might be used in future NR releases for operation at high carrier frequencies. $\mu = 5$ is defined by SCS of 480 kHz, OFDM symbol length of 2.08 us, CP of 0.15 us, slot length of 31.25 us, and contains 32 slots in a single subframe.

In Figure 3, we illustrate the implemented NR frame structure in time- and frequency- domain when configured for $\mu = 3$ (i.e., SCS=120 kHz) and a total channel bandwidth of 400 MHz. To simplify modeling in the simulator, the CP is included jointly with the OFDM symbol length. For example, for $\mu = 3$, the OFDM symbol length including CP is 8.92 us, which accounts for the real OFDM symbol length 1/SCS=8.33 us plus a CP of 0.59 us.

To support a realistic NR simulation, we properly model (as per the standard) the numerology-dependent slot and OFDM symbol granularity [8]. Differently from the original mmWave and LTE module (which only disposed of frame/subframe/symbol granularities) we have introduced the *slot* granularity. Therefore, our code is able to support the NR frame structure in the time domain, and the scheduling operation per slot. Accordingly, we have adapted other parts of the simulator to the new NR frame structure: the PHY transmission/reception functionality, the MAC scheduling and the resource allocation information, the processing delays, and the interaction of the PHY layer with the MAC layer, as detailed in the following.

Firstly, the transmission and the reception have been updated to be performed on a slot basis. The corresponding functions are StartSlot() and EndSlot(), respectively. These functions are executed at both gNB and UE every 14 OFDM symbols, so that the time periodicity depends on the configured numerology. According to the NR definition [11], one TTI corresponds

[4] In NR Rel-15, not every numerology can be used for every physical channel and signal: μ=4 is not supported for data channels, and μ=2 is not supported for synchronization signals [1]. Also, for data channels, only μ=$0, 1, 2$ are supported in frequency range 1 (sub 6 GHz, 0.45 - 6 GHz) and μ=$2, 3$ in frequency range 2 (mmWave, 24.25 - 52.6 GHz).

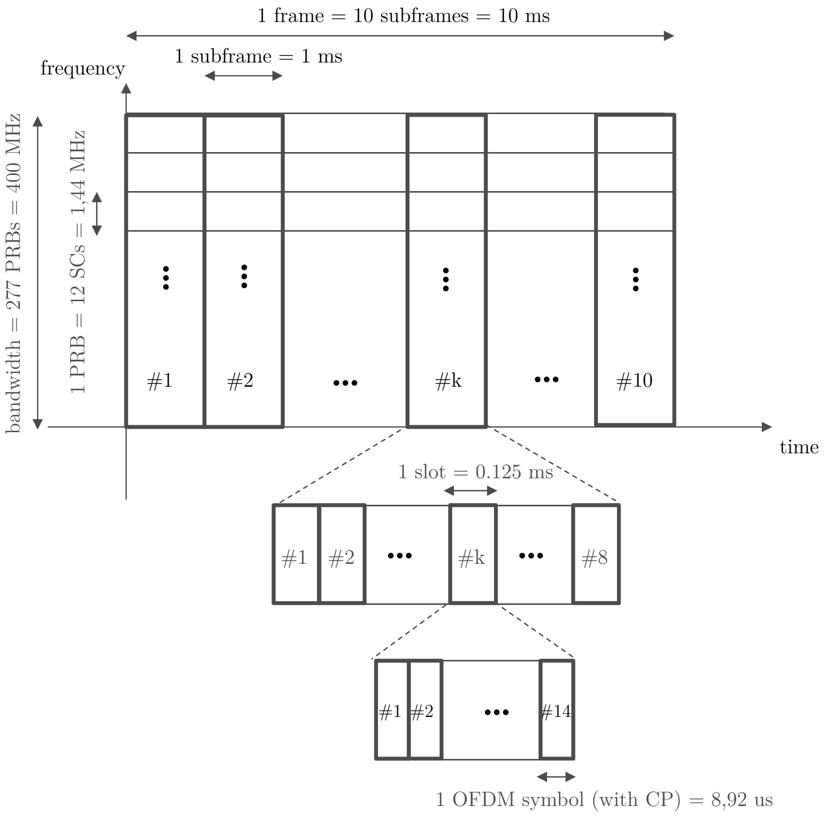

1 frame = 10 subframes = 10 ms
1 subframe = 1 ms
frequency
bandwidth = 277 PRBs = 400 MHz
1 PRB = 12 SCs = 1,44 MHz
#1
#2
#k
#10
time
1 slot = 0.125 ms
#1
#2
#k
#8
#1
#2
#14
1 OFDM symbol (with CP) = 8,92 us

Figure 3: NR frame structure in time- and frequency- domain for $\mu = 3$ and a total channel bandwidth of 400 MHz.

to many consecutive OFDM symbols in the time domain in one transmission direction, and different TTI durations can be defined when using a different number of OFDM symbols (e.g., corresponding to a mini-slot, one slot, or several slots in one transmit direction). Thus, the TTI is in general of variable length, regardless of the numerology. Transmission and reception of TTIs of variable length, as per NR, are handled by StartVarTti() and EndVarTti() functions, respectively.

Secondly, to support operation per slot, we introduced SlotAllocInfo that stores the resource allocation information at a slot level. A single SlotAllocInfo includes a list of VarTtiAllocInfo elements where each element contains the scheduling information per TTI, whose duration is no longer than that of one slot. For example, VarTtiAllocInfo objects are populated at the UE after reception of Downlink Control Information (DCI) messages. For each VarTtiAllocInfo, the MAC scheduler specifies the assigned PRBs and OFDM symbols, along with the information whether the allocation is DL or UL, and whether it is control or data. More details regarding the scheduling process and the scheduling structures will be provided in Section 5.

Currently, as per the flexible slot structure in NR [1], the first and the last OFDM symbols of the slot are reserved for DL control and UL control, respectively, while the OFDM symbols in between can be dynamically allocated to DL or UL data. This enables dynamic TDD and also a flexible and configurable slot structure to allow fast DL-UL switch for bidirectional transmissions [12, Sect. 4.3.2], [13].

Thirdly, regarding the processing delays, we have configured the MAC-to-PHY processing delay at the gNB to be numerology-dependent. Such delay is defined as a specific number of slots and configured by default to 2 slots. Differently, the transport block decoding time at the UE is by default fixed and equals to 100 us, but it could also be easily configured to be numerology-dependent.

Finally, the slot inclusion at PHY requires also interaction with the MAC layer, because the MAC scheduling is performed per slot. Accordingly, a slot indication to the MAC layer has been included to trigger the scheduler at the beginning of each slot, to allocate a future slot. We enter into the details of the new NR schedulers in Section 5.

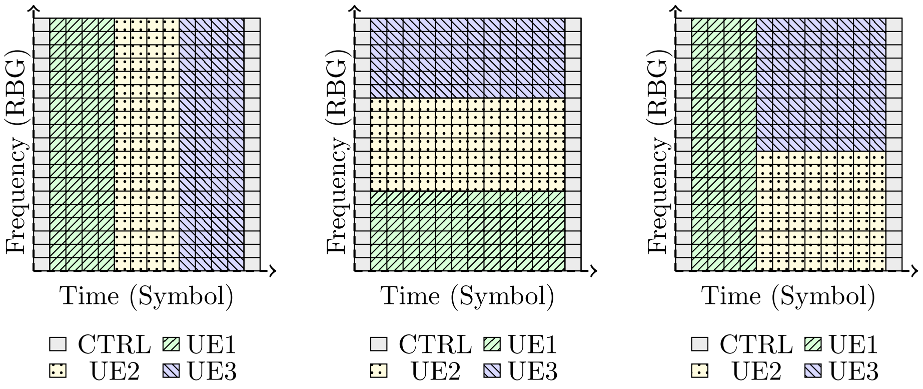
5. MAC Layer

We have implemented the MAC layer in the classes NRGnbMac and NRUeMac. They interact directly with the PHY layer through a set of SAP APIs, and indirectly with the RLC layer. The messages exchanged through the API between RLC and MAC are captured and adequately routed by the bandwidth part manager (see more details about this in Section 6). As an example, the RLC sends to the MAC many Buffer Status Report (BSR) messages (one per bearer) to inform the scheduler of the quantity of data that is currently stored in the RLC buffers. The scheduler, based on such information, takes then scheduling decisions. We have completely transformed the multiple access schemes, the UL scheduling schemes, the scheduler timings, and the scheduler implementation part inside the MAC layer, as detailed in next subsections.

5.1. Multiple Access Schemes

We support OFDMA with variable TTI and TDMA with variable TTI schemes. In the case of OFDMA, we adapted the code to be able to assign a variable number of OFDM symbols in time and Resource Block Group (RBG)s in frequency inside a slot. Visually, a TDMA-based scheme looks as depicted in Figure 4a. Three UEs are scheduled, each one during a period of time that spans four OFDM symbols and with data in all the RBGs. A pure-OFDMA scheme allocates data of different UEs on different RBGs, but using all the available OFDM symbols, as shown in Figure 4b. The new OFDMA-based scheme with variable TTI, instead, is the most flexible way to assign resources. It can allocate different RBGs and a variable number of OFDM symbols. An example is reported in Figure 4c: UE1 is allocated in a TDMA fashion in the first part of the slot, while UE2 and UE3 are scheduled in the rest of the OFDM symbols, each one with a different set of RBGs. Notice that while TDMA was available in the ns-3 mmWave module, and the pure-OFDMA was the original access supported in the ns-3 LTE module, we have extended the MAC layer to support both these schemes, plus the new OFDMA-based scheme with variable TTI.

In the NR simulator, these multiple access schemes, as well as the scheduler policies for them, can be freely chosen (as we will explain later). However, it is worth noting that there are physical limitations when applying them to different spectrum regions. For instance, in the higher spectrum region (e.g., mmWave bands) it would be more difficult to use the pure-OFDMA scheme



(a) Pure TDMA scheme.

(b) Pure OFDMA scheme.

(c) OFDMA with variable TTI scheme.

Figure 4: Possible allocation schemes for a slot.

due to incompatibility with the radio-frequency architectures that are based on single-beam capability [14].

5.2. Scheduling Schemes

NR, like LTE, uses dynamic scheduled-based access for DL, based on which the gNB makes the scheduling decisions. Each UE monitors the Physical Downlink Control Channel (PDCCH), and upon the detection of a valid DCI, follows the given scheduling decision and receives its DL data. In the case of UL, NR considers UL grant-based and UL grant-free access (also known as autonomous UL) schemes [1]. The former is the conventional dynamic scheduled-based access, as per LTE DL/UL and NR DL, based on which the gNB makes the scheduling decisions in both UL and DL. Each UE monitors the PDCCH and, upon the detection of a valid DCI, follows the given scheduling decision and transmits its UL data. The latter is a contention-based scheme. At the time of writing, we have implemented only the UL grant-based access, as per NR specifications, but the UL grant-free implementation is in our future roadmap. As a result, the NR module supports dynamic scheduled-based accesses both for DL and UL.

The design that we followed aims to adopt different scheduling policies (round-robin, proportional fair, etc.) to a TDMA with variable TTI, a pure-OFDMA, or an OFDMA with variable TTI multiple access schemes. Also, we aim to reduce to the minimum the amount of duplicated code, while respecting the FemtoForum specification for LTE MAC Scheduler Interface. To do so, we considered that the primary output of a scheduler is a list of DCIs for a specific slot, each of which specifies (among other values) three crucial parameters. The first is the starting symbol, the second is the duration (number of OFDM symbols), and the last one is the RBG bitmap, in which a value of 1 in the position m represents a transmission in the RBG number m. This is compliant with DL and UL resource allocation Type 0 in NR [15, Sect. 5.1.2.2 and Sect. 6.1.2.2], as far as frequency-domain is concerned, and follows the standard time-domain resource allocation that includes Start and Length Indicator Value (SLIV) for both DL and UL [15, Sect. 5.1.2.1 and Sect. 6.1.2.1].

Scheduler Timings: We consider that the scheduler works "ahead“ of time: at time t, when the PHY is transmitting slot x over the air, the MAC is working to allocate slot $x + d$, where d is a configurable delay, defined as a function of the number of slots. It represents the operational latency and, in the simulator, it is configured through the attribute L1L2CtrlLatency and L1L2DataLatency of the class NrPhyMacCommon). For the DL DCIs, this is the only delay to consider: when the slot $x+d$ is over the air, the DL DCIs are transmitted in the first symbol and will apply for the same slot. However, for the UL case, we must consider an additional delay which represents the time needed by the UE to decode the DCI and to prepare the UL data to transmit. The standard refers to this further delay as K2 [15, Sect. 6.1.2.1], which is measured in number of slots and can take any integer value from 0 to 32 slots. We model it through the attribute UlSchedDelay of the class NrPhyMacCommon. To keep it in consideration, if the PHY is transmitting over the air the slot x, the MAC will work on the UL part of the slot $x + d+$ K2. These DCIs containing the UL grant are transmitted over the air in slot $x + d$, and the UE has K2 slots of time for preparing its UL data. Figure 5 illustrates the scheduler operation

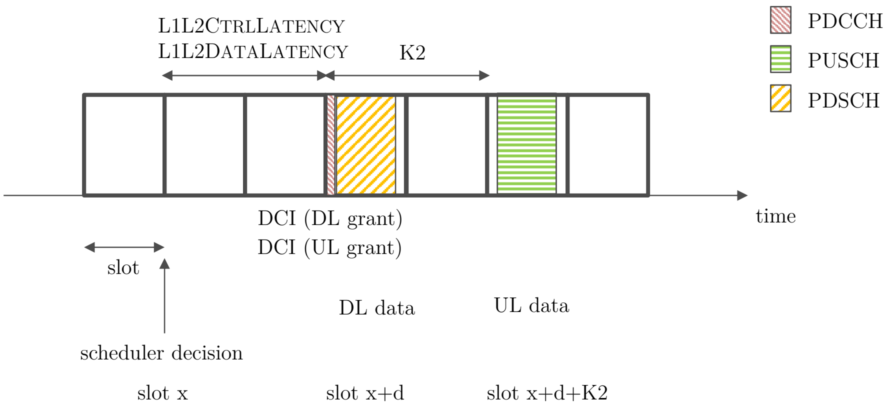
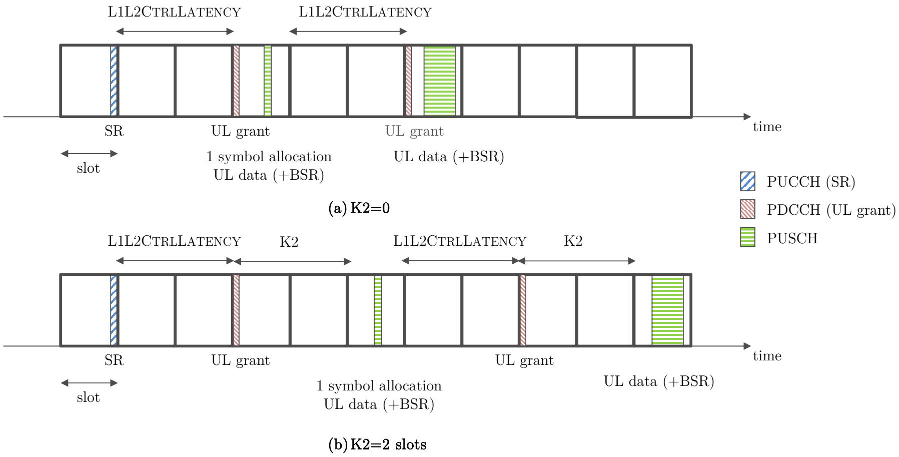


Figure 5: Scheduler timings in the ns-3 NR simulator, for K2=2 slots and L1L2DataLatency=L1L2CtrlLatency=2 slots.

and the DL/UL transmissions by taking into account these timings, for K2=2 slots and L1L2DataLatency =L1L2CtrlLatency=2 slots.

UL handshake: We have improved the dynamic scheduled-based access for UL (i.e., the UL grant-based scheme) as follows. Upon data arrival at the UE RLC queues, the UE sends an Scheduling Request (SR) to the gNB through the Physical Uplink Control Channel (PUCCH) to request an UL grant from its gNB. Then, the gNB sends the UL grant (DCI in PDCCH) to indicate the scheduling opportunity for the UE to transmit. Note that the first scheduling assignment is blind since the gNB does not know the 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 amount of OFDM symbols that permits at least a 4 bytes transmission. In the majority of cases, this value equals to 1 OFDM symbol. Next, the UE, after receiving the UL grant, performs the data transmission in the PUSCH, which may contain UL data and/or BSR. After that, 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 the main difference in the NR module with respect to mmWave and LTE ns-3 modules is that we have introduced the SR in the PUCCH and the BSR can only be sent in conjunction with the MAC Packet Data Unit (PDU) (since,



(a) K2=0

(b) K2=2 slots

Figure 6: UL handshake procedure for NR UL grant-based access, including NR timings and processing delays, as implemented in the ns-3 NR simulator. (a) K2=0, (b) K2=2 slots.

according to NR specifications, the BSR is part of the MAC header), while in previous ns-3 modules the BSR was sent periodically and ideally.

Before sending the UL grants, L1L2CtrlLatency delay has to be considered at the gNB side. Also, upon reception of an UL grant, the UE should send UL data and/or BSR after K2 slots, being K2 indicated in the UL grant. So, these two parameters influence the UL handshake. In Figure 6, we show the UL handshake, including also the timings and processing delays that influence it (i.e., L1L2CtrlLatency and K2) for K2=0 (top) and K2=2 slots (bottom). Recall that we have a TDD slot structure with 14 OFDM symbols per slot, in which PDCCH is sent in the 1st symbol, PUCCH in the 14th OFDM symbol, and the OFDM symbols in between are devoted to shared channels that may contain data (PDSCH and/or PUSCH). Our implementation in the ns-3 NR simulator follows exactly the handshake and timings that are illustrated in Figure 6. The BSR is prepared shortly before the PHY transmission in the UL, reflecting the status of the RLC queue without including the current transmission.

5.3. Scheduler Policies and Implementation at the gNB

The core class of the NR schedulers design is NrMacSchedulerNs3. This class defines the core scheduling process and splits the scheduling logic into logical blocks. The FemtoForum API splits the UL and the DL scheduling. In the following, we will consider only the DL, but the description also applies to the UL case. The differences lie in the variable and function naming, as well as the delays involved, as explained before. As a starting point, we prepare a list of active UE and their requirements, organized based on the concrete beams they belong to.

We start with the scheduler implementation details for OFDMA-based schemes. The first step of the procedure consists of distributing OFDM symbols among multiple beams. We need this block for the OFDMA-based schemes because we chose to support single-beam capability only. At high frequencies, the beam is shaped after digital-to-analog conversion due to limitations in the implementation phase. Therefore, with analog beamforming, there is the constraint that a receive or transmit beam can only be formed in a single direction at any given time instant, meaning that if we want to transmit towards two UEs with different beams, we must do so in different time instants. We provide two different ways to assign symbols to the beams: in a load-based or round-robin fashion. We consider as the beam load the sum of the bytes queued in the RLC layer of the UEs that belong to that beam. The round-robin assignment merely assigns the same number of OFDM symbols to all beams.

After the symbols/beam selection in OFDMA schedulers, it is necessary to distribute the available RBGs in the time/frequency domain among active UEs in each beam. This step depends on the specific scheduling algorithm that the user has chosen. The RBGs can be distributed following a round-robin, proportional fair, or maximum rate algorithm. The resources to be allocated are groups of RBGs spanned over one, or more, symbols.

Finally, the last step consists in the creation of the corresponding DCI, based on the number of assigned resources made in the previous block. The assigned RBGs should be grouped to create a single block for each UE. Then, the RBG bitmap is created[5], so that DCIs for different UEs do not overlap. The bitmap will be an input, later on, for the PHY layer. At the transmission or reception time, the PHY translates the bitmap into a vector of enabled PRB. As the standard indicates in [15, Sect. 5.1.2.2 and 6.1.2.2], each RBG is grouping 2, 4, 8, or 16 PRB depending on the BWP size. Then, the transmitter distributes the power, and the receiver decodes, only among the active PRBs.

The design also takes into consideration HARQ retransmissions. They have a higher priority in the scheduling policies. When a NACK is received, the scheduler takes the old DCI and tries to put it in the current slot for retransmission. If that is not possible, then it will be queued for the next slot. It is important to remark that the simulator only supports a round-robin policy to select the HARQ process to retransmit.

The user can select different schedulers and different assignment modes by swapping class name during the configuration phase. The available OFDMA schedulers are NrMacSchedulerOfdmaRR (round-robin), NrMacSchedulerOfdmaPF (proportional fair), and NrMacSchedulerOfdmaMR (maximum rate). Our OFDMA schedulers are all using the variable TTI strategy, so they are allowed to create TTIs of different length. The configuration into pure OFDMA schedulers is straightforward.

For TDMA-based schedulers, the first step (symbols/beam selection) is not performed, as entire symbols are assigned to the UEs and the PHY layer is then perfectly capable of switching the beam in time (under the single-beam capability assumption explained before). Therefore, the assignment phase, in which the scheduler decides how many OFDM symbols are assigned to each UE, is directly executed. We support round-robin (NrMacSchedulerTdmaRR), proportional fair (NrMacSchedulerTdmaPF), and maximum rate (NrMacSchedulerTdmaMR)

[5] The ns-3 NR module follows NR resource allocation Type 0, as per [15], in which the resource allocation is specified through a bitmap. It provides more flexibility to the scheduler operation, as compared to NR resource allocation Type 1 that specifies the PRB start and number of PRB allocated.

schedulers.

These classes, no matter the access mode, follow the same principles:

- *Round-robin*: The scheduler evenly distributes the available RBGs among UEs associated with that beam (OFDMA), while for TDMA evenly distributes the available symbols.

- *Proportional fair*: In the OFDMA mode, the scheduler evenly distributes the available RBGs among UEs according to a metric that considers the actual rate, based on the Channel Quality Indicator (CQI)) elevated to α and the average rate that has been provided in the previous slots to the different UEs. By changing the α parameter the metric also changes. For $\alpha = 0$, the scheduler selects the UE with the lowest average rate. For $\alpha = 1$, the scheduler selects the UE with the largest ratio between actual rate and average rate. For TDMA, the resources to distribute are entire symbols.

- *Maximum rate*: The scheduler evenly distributes the available RBGs (or the available symbols in case of TDMA) among UEs according to a maximum rate metric that considers the actual rate (based on the CQI) of the different UEs.

In the UL, we currently support only TDMA. This means that, even for OFDMA schedulers, such a phase is treated as it was in the TDMA schedulers.

6. Bandwidth Part Manager

An additional level of flexibility in the NR system can be achieved by implementing the multiplexing of numerologies in the frequency domain. As an example, Ultra-Reliable and Low-Latency Communications (URLLC) traffic requires a short slot length to meet strict latency requirements, while enhanced Mobile BroadBand (eMBB) use case in general aims at increasing throughput, which is achieved with a large slot length [16]. Therefore, among the set of

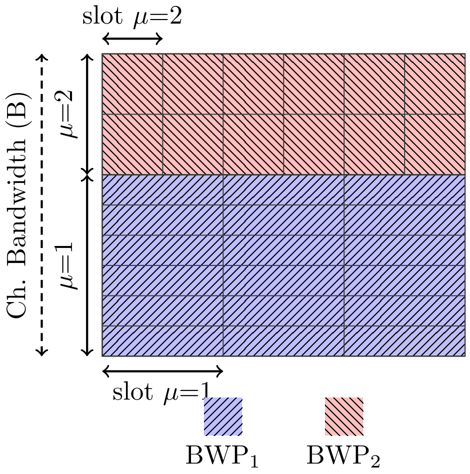


Figure 7: An example of BWPs configuration.

supported numerologies for a specific operational band and deployment configuration, URLLC can be served with the numerology that has the shortest slot length, and eMBB with the numerology associated to the largest slot length [17]. That is, the numerology for URLLC is recommended to be larger than the numerology for eMBB, $\mu_u > \mu_e$, where μ_u is numerology used for URLLC and μ_e for eMBB. Hence, the main objective of Frequency Division Multiplexing (FDM) of numerologies is to address the trade-off between latency and throughput for different types of traffic by physically dividing the bandwidth in two or more BWPs. In Figure 7, we illustrate an example of FDM of numerologies. The channel is split into two BWP that accommodate the two numerologies (μ_u and μ_e) multiplexed in frequency domain. The total bandwidth B is then divided into two parts of bandwidth B_u for URLLC and B_e for eMBB, so that $B_u + B_e \leq B$. The number of PRBs for URLLC is N_u and N_e for eMBB. Note, that the PRB width varies with the numerology.

6.1. FDM Model

In the ns-3 NR simulator, the user can configure FDM bands statically before the simulation starts. This is a critical design assumption based on two main reasons. First, the NR module relies on the channel and the propagation loss model defined in the mmWave module, which is not able to allow runtime modifications of the physical configuration parameters related to time/frequency configuration (such as the system bandwidth, the central carrier frequency, and the symbol length). Thus, until the current mmWave/NR module channel model is not modified to allow these runtime configuration changes, it will not be possible to perform semi-static reconfiguration of BWPs. The second reason is that in the simulator the RRC messaging to configure the default bandwidth part, as well as the bandwidth part reconfiguration, are not implemented yet.

Regarding the data plane, there is a similarity among the concept of the component carriers (CCs) in LTE and the BWPs in NR. While the objective is different, since LTE aims to aggregate narrower bandwidth to achieve a wider capacity, while NR BWP intend to subdivide the bandwidth to use it for multiple and different purposes, the idea of having aggregated PHY layer instances remains the same. The main difference, at the implementation level, is that in LTE, the various carriers have the same OFDM symbol, subframe, and frame boundary, while in NR only the subframe and frame boundaries of different BWPs are aligned, the slot and OFDM symbol may not be. Everything else can be different, and therefore we can have contiguous BWPs with different PHY parameters.

Following the previous discussion, it naturally comes that one possible way to implement the FDM is to start from the Carrier Aggregation (CA) feature of ns-3 LTE [18]. This is exactly what we have done. In this line, the models of both NrGnbDevice and NrUeDevice have been extended to support the installation of instances of MAC and PHY per carriers, following the design (and inheriting the architecture) of the ns-3 LTE CA feature.

In the current implementation, we support the transmission of the scheduling information through a dedicated control channel in each BWP, and the MAC scheduling and HARQ processes are performed per BWP. Finally, according to our model, the multiplexing of the data flows based on the type of traffic is performed by a new layer, which is implemented by an entity called Bwp-

Manager. Its role is similar to that of CC manager in the LTE module, and BwpManager can use 5G QoS identifiers (as defined in [19]) to determine on which BWP to allocate the packets of a radio bearer and to establish priorities among radio bearers.

6.2. Implementation of FDM of Numerologies

The main challenges to implement FDM of numerologies lie in the modifications of the NR SAP interfaces between the layers of the NR stack to include the new BWP layer according to the FDM design. The basic block is a CC, which we will use as a synonym of BWP. A CC consists of two instances, one for the MAC and the other for the PHY layer. Moreover, for each PHY there is a separate channel model, that is shared among the attached UEs. Each CC/BWP can be configured independently with different parameters (for instance, each PHY instance can be configured with a different transmit power) but the configuration should match with that of the attached UEs. In other words, the BWPs configuration for the gNB and the UEs should be the same. Practically speaking, when the final user configures one or more BWPs for one gNB, the code is automatically creating CCs and the configuration for the gNB and all the attached UEs.

Then, we have a BWP Manager entity that is responsible for routing the traffic and the signaling messages to the correct BWP, based on their QoS requirements. Currently, this class is located above all the instances of MAC and PHY in UEs' and gNBs' stack. Each message that arrives at each CC is automatically redirected to the BWP Manager class, NRGnbBwpM. The current implementation of NRGnbBwpM supports all LTE EPS bearer QoS types, and the assignment of the corresponding BWP is based on the static configuration provided by the user that maps a bearer to a specific bandwidth part. So, the routing job is done through a lookup table, in which each message is mapped into an identifier, and then passed to the CC with that identifier. Similarly to the changes in gNB device, we have also extended the NR UE model to support the CCs and UE CC manager, to be able to route the UL traffic properly. A visual representation of this class hierarchy is depicted in Figure 2, presented when we explained the overall architecture of the simulator.

The mmWave module channel models (MmWaveBeamforming, MmWaveChannelMatrix, MmWaveChannelRaytracing, MmWave3gppChannel) depend on the various PHY and MAC configuration parameters specified through a single instance of the class MmWavePhyMacCommon. Hence, it is necessary to install as many MmWave3gppChannel channel model instances as BWPs to configure. However, an important limitation of this design is that all gNBs and UEs in the simulation have the same BWP configuration. Also, the model does not consider interference between BWPs/CCs, so that appropriate band guards are to be left between contiguous BWPs/CCs.

Note that the architecture of the current implementation can be used either as a way to implement RAN slicing with a dedicated resource model [20], by allocating different flows to orthogonal BWPs, or can be reused, with slight modifications, as a way to implement Carrier Aggregation [18]. Those are different use cases, with the same implementation blocks, for which the only difference lies in the logic of the BWP Managers at gNB and UE sides.

7. Use Cases

In this section, we report two important simulation outcomes that we have obtained using the simulator. First, we discuss the validation of the models, following 3GPP calibration procedures. This ensures that the simulator, besides being properly tested with both system and unit tests, provides expected results as compared to those achieved by similar proprietary simulators. Second, we evaluate the E2E performance, as a function of different numerologies, in the context of a complex realistic 5G scenario, in order to show the potentiality of the simulator.

7.1. Calibration

The accuracy of the simulation results is very important when the simulator is used as a base to take design decisions or when it is needed to evaluate the

Parameter	Value
Carrier freq.	30 GHz
Bandwidth	40 MHz
SCS	60 kHz ($\mu = 2$)
Channel	Indoor TR 38.900
BS Tx Power	23 dBm
BS Antenna	M=4, N=8, 1 sector, height=3 m, vertical polarization
UE Antenna	M=2, N=4, 1 panel, height=1.5 m, vertical polarization
BS noise figure	7 dB
UE noise figure	10 dB
UE speed	3 km/h
Scheduler	TDMA PF
Traffic model	Full Buffer

Table 2: Simulation parameters for calibration experiments

effectiveness of a proposal. 3GPP holds calibration campaigns to align both link-level and system-level simulation results of different simulators. To align our simulator, we have followed the Indoor Hotspot (InH) system-level calibration for multi-antenna systems, as per [21, Annex A.2]. Details of the evaluation assumptions for Phase 1 NR MIMO system-level calibration are provided in [22], with further clarifications in [23], and are summarized in Table 2.

As reference curves, we use the results provided by the companies in [24]. We consider the Cumulative Distribution Function (CDF) of the wideband SINR with beamforming, and the CDF of the wideband SNR with step b (i.e., with analog TX/RX beamforming, using a single digital TX/RX port). For each case, we depict as reference the average of the companies contributing to 3GPP, as well as the company that gets the minimum and the maximum of the average wideband SNR/SINR, so that an optimal region for calibration is defined. As a scenario, we are using the standard 3GPP calibration deployment, composed by 12 gNBs, deployed in two rows of six gNBs each, equally spaced by 20 meters. Horizontally, between every gNBs there are 20 m. Then, we locate 120 UEs (100% indoor) that are randomly dropped in a 50 m × 120 m area.

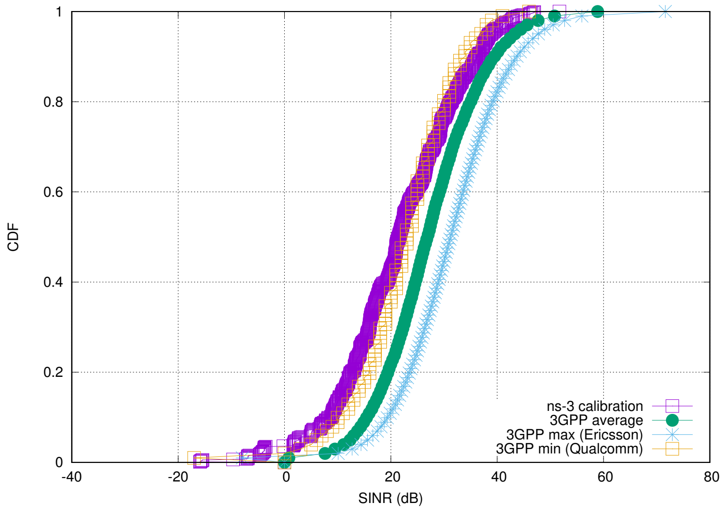


Figure 8: SINR for the case OfficeOpen with shadowing disabled.

In Figure 8 and Figure 9 we show the SINR and SNR, respectively, for the InH Office-Open propagation model, with shadowing disabled. In Figure 10 and Figure 11, we depict the SINR and SNR, respectively, of the InH Shopping-Mall propagation model, with shadowing enabled. In both cases, we can observe that the SINR is close to the lowest 3GPP reference curve. Regarding SNR, it lies entirely within the calibration region, with a perfect match with the average 3GPP SNR in the first configuration setup.

7.2. E2E Latency and NR Numerologies

In this subsection, we analyze a complex and realistic 5G future scenario, where traffic from multiple 5G applications is transmitted, using both UDP and TCP transport protocols. We study the impact of processing and decoding delays for differently configured numerologies and analyze how they affect the E2E performance.

To model a real-world scenario, we base our simulation on the setup shown in Figure 12. At a high level, we have a backbone connection between the 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

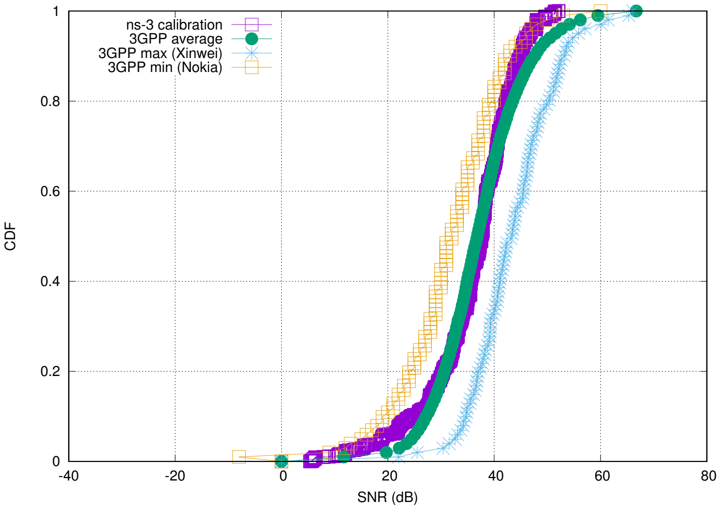


Figure 9: SNR for the case OfficeOpen with shadowing disabled.

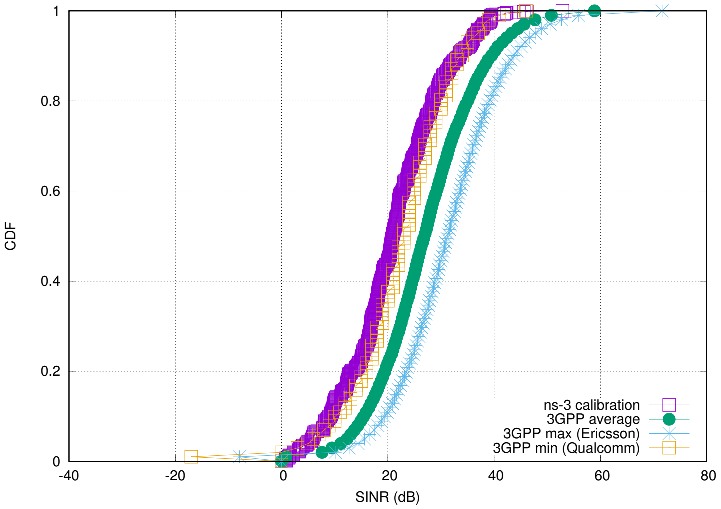


Figure 10: SINR for the case ShoppingMall with shadowing enabled.

another point-to-point connection with a maximum rate of 10 Gb/s, without propagation delay.

Regarding the 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

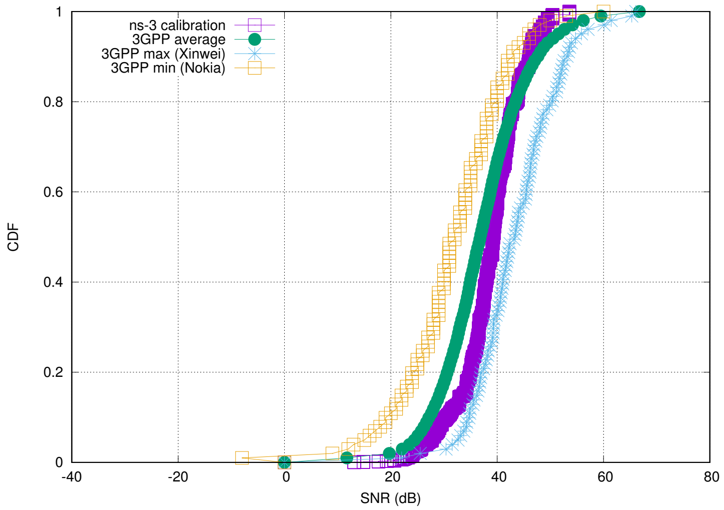


Figure 11: SNR for the case ShoppingMall with shadowing enabled.

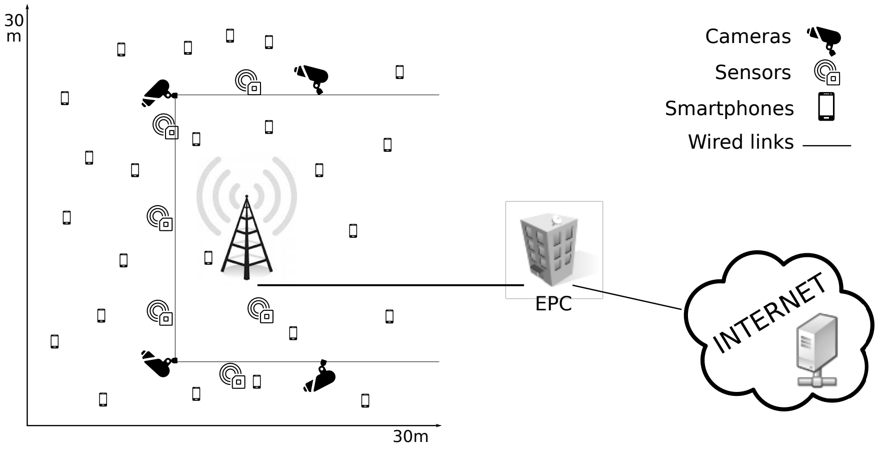


Figure 12: Reference scenario for analyzing E2E latency and NR numerologies.

distributed over a circular area of 30 m of diameter. The position of each UE in the reference scenario depicted in Figure 12 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, each of the video and sensor nodes has one UL UDP flow towards a remote node on the Internet. These flows are fixed-rate flows:

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

Table 3: Application settings, if a setting does not apply it is marked with an "X"

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 4: Relevant simulation parameters

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 3 summarizes the UDP flow characteristics. For the smartphones, we use TCP as the transmission protocol, with the state-of-the-art ns-3 implementation [25, 26]. 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 3 summarizes the details.

Simulations campaign. We compare NR numerologies, from $\mu = 0$ to $\mu = 4$, 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). Other relevant parameters for the simulations are reported in Table 4. For each μ, we have performed multiple sets of simulations in the ns-3 network simulator to obtain data statistics. We have done the experiments using 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 [27]; and 4) a slot-dependent latency value that is equal to two times the slot length (so that it 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 repeated 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.

Sensor and Video UDP Delay. In Figure 13 we can see the latency performance of the sensor flows. In the first two 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 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 observa-

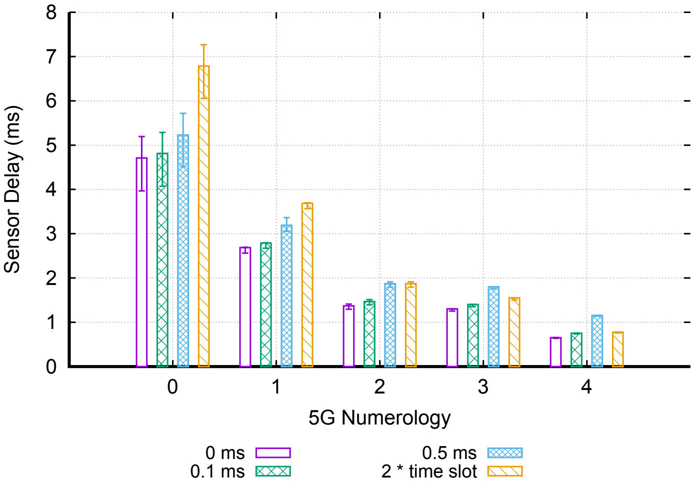


Figure 13: Sensor delay. Top whisker represents the maximum value, the bottom whisker represents the minimum value, and the box is at the 80th percentile.

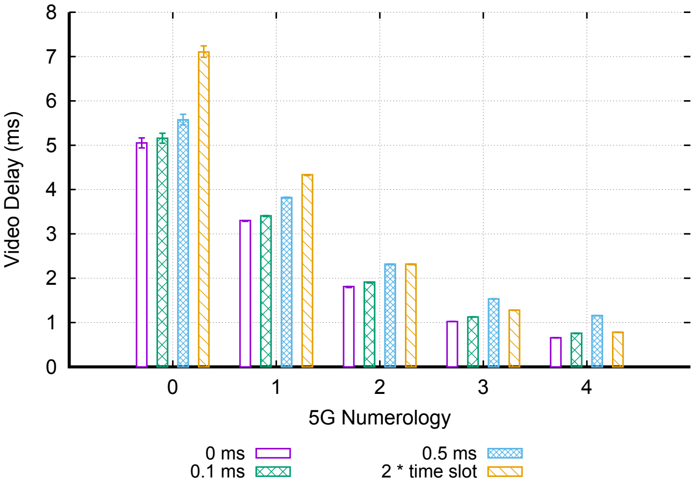


Figure 14: Video delay. Top whisker represents the maximum value, the bottom whisker represents the minimum value, and the box is at the 80th percentile.

tions allow us to conclude that *the analyzed fixed delays in the decoding impact the overall latency linearly*, without affecting other phases.

For the latency performance of the Video flows, we can refer to Figure 14.

Here we observe a similar trend to that shown by the sensor delay, but with slightly higher values.

UDP Remarks. Looking at the fixed-value case, with decoding latency of 0 ms, 0.1 ms, and 0.5 ms, we can see that there is a strange case in which an increased numerology corresponds to an increase in latency (or at least, not in a latency reduction). The increase happens in the sensor flows when passing from μ=2 to μ=3. How is it possible that half the slot time corresponds to more latency experienced by a single packet? The reason lies in the SR mechanism. 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 BSR message sent along with user data in a previously granted space. If a data packet meets an empty 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.

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 re-start the data flow. This generates the unfortunate event in which for the sensors, in numerology 2 the number of SR is lower than in the numerology 3. Even if the slot time is lower, the overhead for the increased number of SR is reflected in the latency value plotted in Figure 13 (μ=2). We do not see the same effect for video since the phenomenon is correlated to the inter-packet arrival time in the RLC buffers, that are different between sensor and video flows.

TCP Upload Goodput. We start the analysis by looking at the goodput for the TCP upload flows, in Figure 15. With higher numerologies, the slot

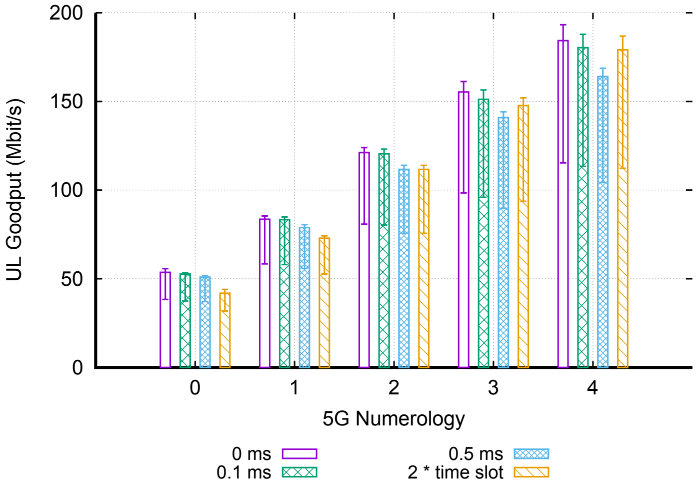


Figure 15: TCP Upload goodput. Top whisker represents the maximum value, the bottom whisker represents the minimum value, and the box is at the 80th percentile.

time reduces, allowing to have smaller Round-Trip Time (RTT). Therefore, the goodput increases, reaching almost 200 Mb/s for the numerology 4. 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 16 we can analyze the performance of the TCP Download flows. 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 decisions 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,

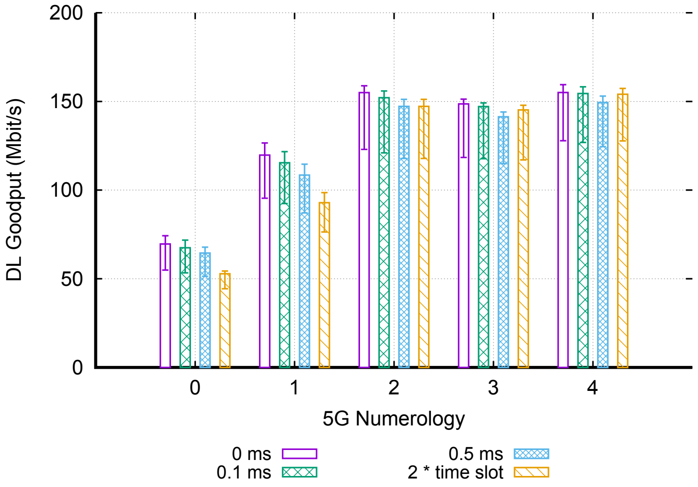


Figure 16: TCP Download goodput. Top whisker represents the maximum value, the bottom whisker represents the minimum value, and the box is at the 80th percentile.

involving a signaling exchange that, albeit slightly, increases the round trip time and therefore reduces the TCP goodput. The trend of the goodput while increasing the numerology follows what we have seen in the TCP Upload, but stops after numerology 2. This is because our scheduler prioritizes the UL flows by assigning more usable symbols. So, at one point, the download flows are reaching their cap assigned by scheduler.

8. Future Work

In this section, we describe our roadmap and future development plans regarding different layers of the protocol stack, according to NR specifications.

8.1. PHY

8.1.1. Mini-slots

NR defines mini-slots composed of 2 OFDM symbols up to the slot length - 1 in any band [28], [29, Sect. 8.1], and of 1 symbol, at least above 6 GHz. Although we already support flexible transmission duration thanks to the variable TTI concept, true mini-slots also include the PDCCH. In the current implementation, we transmit PDCCH only in the first symbol of a slot regardless of the TTI number in such slot. So, this will be extended by including the PDCCH inside the first symbol of the TTI allocation.

8.1.2. PHY layer abstraction

The current version of the simulator imports the PHY layer abstraction of LTE. One of the important changes of NR with respect to LTE (which used Turbo Codes) is the adoption of Polar Codes for control channels and low Density Parity Check (LDPC) coding for data channels. Turbo Codes and LDPC are shown to have similar performances for large packet sizes, however, the differences lie in the implementation complexity and when packet sizes are small. In this regard, we are currently working to include the specific NR PHY layer abstraction through a proper Link to System Mapping (L2SM) and error model of NR. We are following the same approach like that of the ns-3 LTE module, in which the multi-carrier compression metrics are combined with link-level performance curves matching, to obtain lookup tables of Block Error Rate (BLER) versus SINR. The main objective is to predict the Transport Block Error Rate (TBLER) at MAC layer [30]. In addition to the error model that has to be abstracted from a link-level simulator, the Code Block (CB) sizes and the number of CBs that map to a Transport Block (TB) need to be updated as per [31, Sect. 5.2.2].

8.1.3. 256-QAM

NR supports 256-QAM modulation, which is still not available in the NR module. For that, we need to update the Modulation Coding Scheme (MCS)s as per [15, Table 5.2.2.1-2], include the Mutual Information (MI) mapping for 256-QAM, for which the mapping in [32] can be used, and add new lookup tables of BLER versus SINR for these new MCSs. We are working on this, which goes in parallel with the PHY layer abstraction, and will be released soon.

8.1.4. Spatial user multiplexing and MIMO

Currently, we only support single-beam capability at a time, multiplexing of UEs in frequency/time domain, and single stream per UE (i.e., we have beamforming gain supported, but not spatial multiplexing gain). To support transmission/reception to/from multiple UEs simultaneously at the same time/frequency resources, we should extend the interference model and add power distribution per UE. To support the transmission of multiple streams per UE, the PHY abstraction model could be extended to support multiple streams per UE, for which also the precoders would need to be updated and redesigned.

8.1.5. Frequency Division Duplexing

Currently the simulator supports dynamic TDD, which is the interesting choice for higher frequency ranges and is the approach currently mainly considered for deployment due to regulatory requirements, especially in the United States. However, support for FDD will be needed to simulate future deployments, as the regulation will progress.

8.2. MAC

8.2.1. UL grant-free scheme

NR has introduced a contention-based (non-scheduled) access scheme for URLLC, named UL grant-free or autonomous UL [21, Sect. 8.1.2.1] [33]. In UL grant-free, the UE is allowed, upon activation, to transmit UL data on resources devoted to contention-based access without a UL grant. Therefore, the UL grant-free scheme eliminates the handshake of SR, BSR, and UL grant, as compared to the UL grant-based access shown in Figure 6. The resource allocation for UL grant-free in NR is as follows: time-domain resources for UL grant-free are configured by RRC signaling, and then the activation/deactivation is done through the DCI in PDCCH, which indicates to the UE the RBs and MCS to use if UE wants to access such resources. Its main problem is that, as it is non-scheduled, collisions may arise. Accordingly, to introduce this model in the ns-3 NR simulator we would need to (i) define resources for UL grant-free access, to be configured by RRC, and (ii) to include an error model for the collisions therein.

8.2.2. Bearer prioritization

Currently, the scheduler is not prioritizing flows based on their bearer. In other words, it does not take into account guaranteed bit rate or latency deadlines when making decisions. A crucial step will be to insert a general policy to ensure such constraints to the active flows.

8.2.3. Punctured scheduling

To efficiently multiplex eMBB and URLLC traffics, NR has defined procedures to enable punctured scheduling in DL [34]. This is useful in case of a sudden need for resources for URLLC traffic that has strict latency requirements. URLLC latency targets can be met by puncturing the resources already allocated to eMBB traffic and indicating so to those UEs through an indicator of preemption. To support punctured scheduling in the NR simulator we should: (i) allow the scheduler to work symbol-by-symbol, instead of in a per-slot basis, when there is a new URLLC packet arrival in the middle of the slot, (ii) include an indication of preemption in DL control channels to indicate preemption of eMBB data, and (iii) redesign procedures for eMBB flows to ignore scheduling assignments and avoid decoding, as well as the subsequent HARQ-ACK feedback generation.

8.3. Upper layers extensions

Currently RRC, RLC, PDCP layers are relying on LTE implementation. Different simplifications have been proposed for RLC and PDCP, in order to facilitate the targeted latency reduction that NR should support. RRC needs also different extensions already in its original implementation, since it was mainly designed for operation in CONNECTED mode. As a result of that, updates and extensions will be considered for inclusion.

8.4. SDAP

In the new QoS framework for the 5G network, there is a new layer above PDCP. Its name is SDAP, and its role is to map distinct QoS flows into data radio bearers. In the simulator, we miss an SDAP entity that receives Service Data Unit (SDU) from upper layers and sends SDAP PDU to its peer SDAP entity via lower layers. These entities should be able to mark the QoS flows appropriately, and it should be possible to configure the mapping between flow and data radio bearer, through a static or a dynamic configuration via RRC.

8.5. Core Network

In 5G, the core network will support the separation of control and user plane, and each service will be provided as a network function, creating an overall service-based architecture. We are incorporating changes from ns-3 LTE module, such as the EPC functional split between PGW and SGW that has been recently included in LTE ns-3, but we should prepare a compatible API to offer functionalities to external users.

8.6. Operation in unlicensed bands

Differently from LTE, NR includes native support for operation in unlicensed bands. A work item is currently ongoing to define the operation in the band below 7 GHz, and for Release 17 extensions are expected for mmWave bands. The operation will include Listen Before Talk functionalities, able to facilitate coexistence in the same band with technologies like WiFi and WiGig. Thanks to the multi-technology characteristic of ns-3, we will be able to propose such an extension in our module.

9. Conclusion

In this paper, we have presented a complete overview of a novel full-stack NR simulator. In particular, we have discussed the additions and modifications that we have done to the mmWave simulation tool developed on top of ns-3.

The objective was to support an end-to-end simulation of NR networks, through an up-to-date and standard-compliant platform. The work has been validated and calibrated in different indoor scenarios, as compared to other proprietary simulators that follow similar purposes inside 3GPP, and by following 3GPP recommendations. In addition, as an example of the potentiality of the tool, we have created a complex end-to-end simulation campaign to assess the impact of different NR numerologies on the overall E2E latency of different devices. We have studied parameter sensitivities and shown that when considering a full protocol stack and high fidelity models, unexpected behaviors are observed, which hardly could be highlighted with other types of simulators. We have concluded the discussion with the presentation of the roadmap that we are currently following, and wish to follow, in line with current and future 3GPP developments.

Acknowledgment

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

[2] 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 (2018).

[3] N. Baldo, M. Miozzo, M. Requena-Esteso, 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, ACM, New York, NY, USA, 2011, pp. 293–298.

[4] M. Mezzavilla, M. Zhang, M. Polese, R. Ford, S. Dutta, S. Rangan, M. Zorzi, End-to-End Simulation of 5G mmWave Networks, IEEE Communications Surveys Tutorials 20 (2018) 2237–2263.

[5] N. Nikaein, M. K. Marina, S. Manickam, A. Dawson, R. Knopp, C. Bonnet, OpenAirInterface: A flexible platform for 5G research, ACM SIGCOMM Computer Communication Review 44 (2014) 33–38.

[6] M. K. Müller, F. Ademaj, T. Dittrich, A. Fastenbauer, B. Ramos Elbal, A. Nabavi, L. Nagel, S. Schwarz, M. Rupp, Flexible multi-node simulation of cellular mobile communications: the Vienna 5G System Level Simulator, EURASIP Journal on Wireless Communications and Networking 2018 (2018) 227.

[7] A. Virdis, G. Stea, G. Nardini, Simulating lte/lte-advanced networks with simulte, in: Simulation and Modeling Methodologies, Technologies and Applications, Springer, 2015, pp. 83–105.

[8] B. Bojovic, S. Lagen, L. Giupponi, Implementation and evaluation of frequency division multiplexing of numerologies for 5G new radio in ns-3, submitted to ns-3 Workshop 2018 (2018).

[9] H. Tazaki, F. Uarbani, E. Mancini, M. Lacage, D. Camara, T. Turletti, W. Dabbous, Direct Code Execution: Revisiting Library OS Architecture for Reproducible Network Experiments, in: Proceedings of the Ninth ACM Conference on Emerging Networking Experiments and Technologies, CoNEXT '13, ACM, New York, NY, USA, 2013, pp. 217–228.

[10] Evolved Universal Terrestrial Radio Access (E-UTRA); Physical channels and modulation, 3GPP TS 36.211, Release 8, v15.5.0, Mar. 2019.

[11] 3GPP TSG RAN, Study on New Radio (NR) Access Technology (Release 14), 3GPP TR 38.912 V14.0.0, 2017.

[12] TSG RAN; NR; Physical channels and modulation, 3GPP TS 38.211, Release 15, v15.5.0, Mar. 2019.

[13] The 5G unified air interface, Qualcomm Technologies, Inc. (2015).

[14] J. G. Andrews, et al., Modeling and analyzing millimeter wave cellular systems, IEEE Trans. Commun. 65 (2017).

[15] TSG RAN; NR; Physical layer procedures for data, 3GPP TS 38.214, Release 15, v15.2.0, June 2018.

[16] A. A. Zaidi, et al., Waveform and numerology to support 5G services and requirements, IEEE Commun. Mag. 54 (2016) 90–98.

[17] Vivo 3GPP R1-1707238, Discussion on NR resource allocation, 3GPP TSG RAN WG1 88bis Meeting, 2017.

[18] B. Bojovic, D. A. Melchiorre, M. Miozzo, L. Giupponi, N. Baldo, Towards lte-advanced and lte-a pro network simulations: Implementing carrier aggregation in lte module of ns-3, in: Proceedings of the Workshop on ns-3, WNS3 '17, ACM, New York, NY, USA, 2017, pp. 63–70.

[19] 3GPP TSG SSA, System Architecture for the 5G System; Stage 2 (Release 15), 3GPP TS 23.501 V15.0.0, 2017.

[20] A. Ksentini, N. Nikaein, Towards enforcing network slicing on RAN: flexibility and resources abstraction, IEEE Commun. Mag. 55 (2017) 102–108.

[21] TSG RAN; NR; Study on new radio access technology Physical layer aspects, 3GPP TR 38.802, Release 14, v14.2.0, Sept. 2017.

[22] Evaluation assumptions for Phase 1 NR MIMO system level calibration, ZTE, ZTE Microelectronics, 3GPP R1-1703534, 3GPP TSG RAN WG1 88 Meeting, Feb. 2017.

[23] Further clarification on assumptions for Phase 1 NR MIMO calibration, ZTE, ZTE Microelectronics, 3GPP R1-1700144, 3GPP TSG RAN WG1 1 Meeting, Jan. 2017.

[24] Calibration results for Phase 1 NR MIMO system level calibration, ZTE, 3GPP R1-1709828, 3GPP TSG RAN WG1 89 Meeting, May 2017.

[25] M. Casoni, N. Patriciello, Next-generation TCP for ns-3 simulator, Simulation Modelling Practice and Theory 66 (2016) 81 – 93.

[26] 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, ACM, New York, NY, USA, 2017, pp. 1–8.

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

[28] Unified design for slot and mini-slot, Huawei, HiSilicon, 3GPP R1-1708121, 3GPP TSG RAN WG1 89 Meeting, May 2017.

[29] Study on New Radio (NR) access technology (Release 14), V14.1.0, 3GPP TR 38.912, Aug. 2017.

[30] M. Mezzavilla, et al., A lightweight and accurate link abstraction model for system-level simulation of LTE networks in ns-3, Proceedings of the 15th ACM Int. Cont. on Modeling, Analysis and Simulation of Wireless and Mobile Systems (2012).

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

[32] J. Yang, et al., An effective SINR mapping models for 256QAM in LTE-Advanced system, IEEE Int. Symp. on Personal, Indoor and Mobile Radio Commun. (2014).

[33] B. Singh, et al., Contention-based access for ultra-reliable low latency uplink transmissions, IEEE Wireless Commun. Lett. 7 (2018) 182–185.

[34] K. Pedersen, et al., Agile 5G scheduler for improved E2E performance and flexibility for different network implementations, IEEE Commun. Mag. 56 (2018).