5G Radio Network Design for Ultra-Reliable Low-Latency Communication

Joachim Sachs, Gustav Wikström, Torsten Dudda, Robert Baldemair, and Kittipong Kittichokechai

Abstract

5G is currently being standardized and addresses, among other things, new URLLC services. These are characterized by the need to support reliable communication, where successful data transmission can be guaranteed within low latency bounds, like 1 ms, at a low failure rate. This article describes the functionality of both the NR and LTE radio interfaces to provide URLLC services. Achievable latency bounds are evaluated, and the expected spectral efficiency is demonstrated. It is shown that both NR and LTE can fulfill the ITU 5G requirements on URLLC; however, this comes at the cost of reduced spectral efficiency compared to mobile broadband services without latency or reliability constraints. Still, the impact on the overall network performance is expected to be moderate.

Introduction

After more than five years of research activities, the next generation mobile communication system, fifth generation (5G), has entered the standardization phase. One difference of 5G compared to earlier mobile communication systems is that it is expected to address a much broader range of applications and use cases. Earlier systems primarily focused on human-centric communication for mobile broadband (MBB) services, like telephony, multimedia, and mobile Internet. In addition to those, the International Telecommunication Union (ITU) has defined new requirements on 5G that enable machine-type communication in a variety of industrial and societal fields. One segment is the so-called *massive machine-type communication* for providing efficient wireless connectivity for connected sensor devices of the Internet of Things [1]. Another segment is critical communication services via *ultra-reliable low-latency communication* (URLLC) for, for example, industrial automation and control, real-time operation of the smart electrical power grid, and remote control of real-time operations. In this article we present the 5G design for URLLC and assess the achievable performance.

For the 5G specification, ITU has recently defined and finalized the required capabilities [2]. The Third Generation Partnership Project (3GPP) has — after an initial feasibility study and definition of performance requirements in Release-14 — started the specification for 5G in 3GPP Release-15 in early 2017. The standardization work has been divided into two phases: the fundamental 5G properties will be standardized in Release-15 in mid-2018, and further enhancements will be standardized in Release-16 in autumn 2019. Before the end of 2019, a full 5G specification will be submitted from 3GPP to ITU that will fulfill all the requirements defined by ITU in [2]. 5G is being standardized in the form of two radio technology components: a novel radio interface technology denoted as New Radio (NR), and Long Term Evolution (LTE). The evolution of LTE will comprise new LTE capabilities that achieve the 5G requirements in a backward-compatible manner, which means that new devices can make use of new capabilities, while the evolved LTE system will continue to serve pre-5G LTE devices within the same system. NR will initially focus on new deployments, for example, in new spectrum allocations. While LTE is only specified for deployments in spectrum below 6 GHz, NR is also designed for deployments in spectrum above 6 GHz [3]. The development of a 5G radio access network is complemented by a 5G system architecture evolution. The system design is changed toward a programmable network design, where network functionality is software-based (virtualized network functions) and is executed on computing nodes within the network. In addition, separation of control and user plane functions is ongoing, which allows for separate scaling of the user plane and control plane. The 5G system architecture allows for a service-aware realization of functionality, where, for example, functions can be either located at the network edge at the radio access (e.g., to enable low latency) or centralized in a large data center (e.g., to enable cost-efficient scaling). In this work we focus only on the radio access developments of 5G for URLLC.

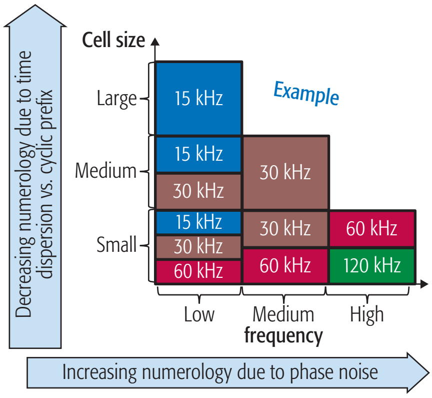
This article is structured in the following way. First, an overview of URLLC services and the corresponding service requirements are presented. Next, the 5G design for achieving low communication latency and increased reliability is described; an analysis of achievable latencies and spectral efficiency is provided. Finally, the 5G URLLC design and evaluation are summarized.

URLLC Services and Requirements

URLLC services have been identified in a variety of fields [4, 5]. A commonality among those services is low latency combined with high reliability. Reliability is specified by the failure probability of packets that are not successfully delivered to the receiver within the latency bound, as they are either erroneous or lost, or arrive too late. In other words, reliability guarantees that messages are successfully delivered, and latency bounds can be met. This has introduced a new performance metric for 5G, since in earlier networks latency

Digital Object Identifier:
10.1109/MNET.2018.1700232

The authors are with Ericsson Research.



OFDM numerologies for NR and LTE

	Sub-carrier spacing	Symbol duration	Cyclic prefix	Slot (14 symbols)	Mini-slot (7 symbols)		
					(7 symbols)	(4 symbols)	(2 symbols)
NR and LTE	15 kHz	66.67 µs	4.76 µs	1000 µs	500 µs	286 µs	143 µs
NR	30 kHz	33.33 µs	2.38 µs	500 µs	250 µs	143 µs	71 µs
NR	60 kHz	16.67 µs	1.19 µs	250 µs	125 µs	71 µs	36 µs
NR	120 kHz	8.33 µs	0.59 µs	125 µs	63 µs	36 µs	18 µs

Note: 60 kHz sub-carrier spacing is optional in Rel-15

FIGURE 1. Examples of feasible OFDM numerology options for different spectrum ranges and deployments.

was predominantly considered in terms of lowest achievable latency or average latency. Some examples of URLLC services are automation of the smart grid energy distribution, industrial process automation, factory automation, automated intelligent transport systems, remote control of machinery, and tactile Internet services. Most URLLC services introduce real-time control applications. In a smart grid this can be the automation of the energy distribution, including detection and restoration of faults; in factory automation it can be the real-time control of the manufacturing robots; and in intelligent transport systems (ITS) it can be the real-time maneuver coordination among autonomous vehicles and with the transportation infrastructure.

In order to address this wide field of URLLC services in a generic way, two representative 5G requirements have been defined:

- ITU and 3GPP [6] require 5G to be capable of successfully transmitting a 32-byte message over the 5G radio interface within 1 ms with a 1 – 10^{-5} success probability.
- 3GPP further requires 5G to be able to achieve a latency over the 5G radio interface of 0.5 ms that can be provided on average for multiple data transmissions (the fulfillment of this requirement is not needed for the 5G evaluation at ITU).

It should be noted that the service requirements for critical communication services are typically defined end to end. However, the URLLC requirements specified in 3GPP and ITU apply only to the one-way latency over the 5G radio network, which constitutes only a fraction of the end-to-end latency budget. An additional latency budget would need to be reserved for the other parts of the communication parts, like the core network and an external network. 5G radio functionality for URLLC should be complemented with low-latency core network design, which could be optimized by, for example, local hosting of application functionality [3, 7].

5G Design for Low-Latency Transmission

Waveform and Radio Slot Structure

Both LTE and NR use orthogonal frequency-division multiplexing (OFDM) as the waveform. OFDM, as an orthogonal modulation scheme, minimizes interference of other transmissions, which is a valuable property for highly reliable communication. A cyclic prefix (CP) is used to cope with time-dispersive channel propagation; only delay-spread of the signal that exceeds the CP length introduces inter-symbol interference. In uplink (UL), LTE uses discrete Fourier transform (DFT) precoding in order to reduce the peak-to-average-power ratio at the transmitter; DFT precoding is also available for NR uplink.

It should be noted that the service requirements for critical communication services are typically defined end to end. However, the URLLC requirements specified in 3GPP and ITU apply only to one-way latency over the 5G radio network, which constitutes only a fraction of the end-to-end latency budget.

One difference between NR and LTE is that LTE uses a fixed numerology of 15 kHz sub-carrier spacing (SCS), whereas NR Release-15 has a scalable numerology with sub-carrier spacings of 15, 30, and 60 kHz below 6 GHz, and 60 and 120 kHz above 6 GHz, as listed in Fig. 1.[1] At higher SCS, the symbol duration decreases, and hence also the length of a slot. The slot is the basic frame structure at which most physical channels and signals repeat; however, slots can be complemented by mini-slot-based transmissions (referred to as Type B scheduling in NR) to provide shorter and more agile transmission units than slots. In LTE and NR a slot comprises 14 OFDM symbols,[2] which leads to a slot length of 1 ms at 15 kHz SCS. By using higher numerologies in NR, the slot duration decreases, which is beneficial for lower latencies. The intention of NR is to support a mix of numerologies on the same carrier. Different numerologies are not orthogonal and interfere with each other; however, by leaving a guard band, this interference is reduced to acceptable levels. By applying, for example, windowing on the OFDM signal, it becomes more confined in the frequency domain, so a smaller guard band suffices. While Release-15 provides the framework to enable mixed numerology in a future NR release, a Release-15 user equipment (UE) is not expected to receive multiple data numerologies simultaneously, and no radio requirements w.r.t. mixing of numerologies will be defined for base stations (BSs). A more profound waveform analysis can be found in [8].

[1] The synchronization block SSB uses SCS of 15 or 30 kHz below 6 GHz and 120 or 240 kHz above 6 GHz.

[2] Strictly speaking, the correspondence to a slot in NR is called a subframe in LTE.

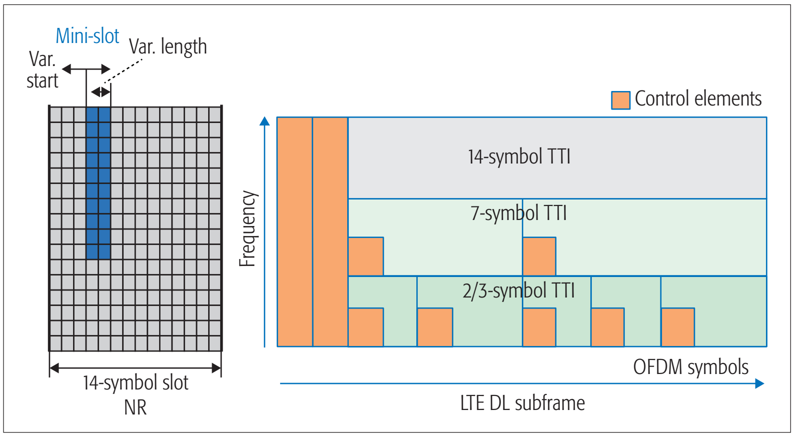


FIGURE 2. Slot structures for NR DL/UL and LTE DL.

In network deployments there are practical restrictions on which numerology is suitable, as depicted in Fig. 1. At higher frequencies (e.g., in millimeter-wave spectrum above 20 GHz), phase noise increases, and numerologies with larger sub-carrier spacing provide better robustness. Since the cyclic prefix is scaled together with the symbol duration, a high numerology has a shorter cyclic prefix and is less suitable for radio environments with large delay spread in the radio channel, as is typical for larger cell sizes. Accordingly, higher numerologies are better suited for smaller cell sizes. NR specifies for 60 kHz sub-carrier spacing an extended cyclic prefix that enables the usage of 60 kHz in larger cells at the cost of increased overhead.

A slot consists of 14 OFDM symbols and is transmitted within a transmission time interval (TTI) (Fig. 2). Different numerologies lead to different slot lengths, ranging from 1 ms at 15 kHz sub-carrier spacing to 125 μs at 120 kHz sub-carrier spacing, enabling shorter TTIs. Beyond numerology scaling in NR, the concept of non-slot-based transmission has been introduced for NR, which is also referred to as mini-slots and corresponds to the short TTI (sTTI) concept that is being standardized for LTE (Fig. 2). A mini-slot in NR can start at any OFDM symbol and can have a variable length; mini-slot lengths of 2, 4, or 7 symbols have been defined in the standard so far. This provides fast transmission opportunities for, for example, URLLC traffic that is not restricted by slot boundaries. Thus, mini-slots provide a viable solution to low-latency transmissions irrespective of sub-carrier spacing, while the usage of wider sub-carrier spacing for low latency is limited to small cells as described above. Similarly, LTE sTTI enables fast transmission opportunities of 2–3- or 7-symbol duration, which can be embedded into the existing LTE frame structure around existing control channels and reference symbols. These short TTIs are independently scheduled with new in-band control elements, thereby allowing lower scheduling latency at the cost of increased overhead.

The LTE enhancement for URLLC is focused on frequency-division duplex (FDD) spectrum allocations. For NR both time-division duplex (TDD) and FDD are addressed for URLLC. NR will support both static and dynamic uplink-downlink (UL/DL) TDD configuration options. For URLLC, the focus is on what latency can be provided with high reliability (e.g., 1–10^{-5}). This means that the TDD latency is largely determined by the worst case timing, for example, a DL packet arrives exactly at the beginning of a UL allocation or vice versa. A suitable TDD configuration for URLLC is when UL (mini-)slots alternate with DL (mini-)slots. Mini-slots shorter than 7 symbols are not considered for TDD, as the UL-DL switching overhead would become significant for too short slot lengths.

Resource Assignments and Channel Access

By reducing the transmission duration and interval, both the time over the air and the delay waiting for a transmit opportunity is reduced. Short transmissions that can be scheduled at short periodicity yield the lowest latency. Mini-slots in NR and short TTIs in LTE, as illustrated in Fig. 2, are therefore key design choices. In addition, for TDD systems a short UL-DL switching period is necessary to achieve low delays. Also, short turn-around times enable more retransmissions within a latency constraint, which can be converted into spectral efficiency. A shorter turn-around is enabled by faster data processing in the network and on the UE side.

For the UL, a significant part of delay comes from the exchange of a scheduling request and a UL transmission grant between the UE and the BS. Also, from a reliability aspect this signaling bears the risk that both messages need to be correctly decoded in order for the UL transmission to start. As a remedy for both the delay and robustness issues, a periodic grant can be configured for the UE. LTE as well as NR specify a semi-persistent scheduling (SPS) framework [9–11], as illustrated in Fig. 3, in which the UE is given a periodic grant that it uses only when it has UL data to transmit. Necessarily, these URLLC resources are tied up for the UE, but by assigning overlapping grants to multiple UEs the resource waste is reduced, and at lower rates the impact on reliability due to collisions can be manageable. This method of periodic grants reduces the latency by one feedback round-trip time and thereby enables UL URLLC with the strictest requirements.

A resource- and latency-efficient scheduling solution is to multiplex data with different TTI lengths (i.e., mini-slots and slots) and — in the case of resource limitations — let the high-priority data use resources from lower-priority data. This type of multiplexing is also referred to as preemption. For example, in NR DL, a mini-slot carrying high-priority or delay-sensitive data can preempt an already ongoing slot-based transmission on the first available OFDM symbols without waiting until the next free transmission resource. This operation enables ultra-low latency for mini-slot-based transmission, especially in the scenario where a long slot-based transmission has been scheduled. A similar concept is also considered for the UL, and in general for LTE. At the cost of degrading the longer transmission, no additional resources need to be reserved in advance for the URLLC device. The damaged longer transmission is then swiftly repaired with a transmission containing a subset of the code block groups (CBGs) in a later TTI, after providing the essential information to clean the contaminated soft values in the receive buffer from the preempted data.

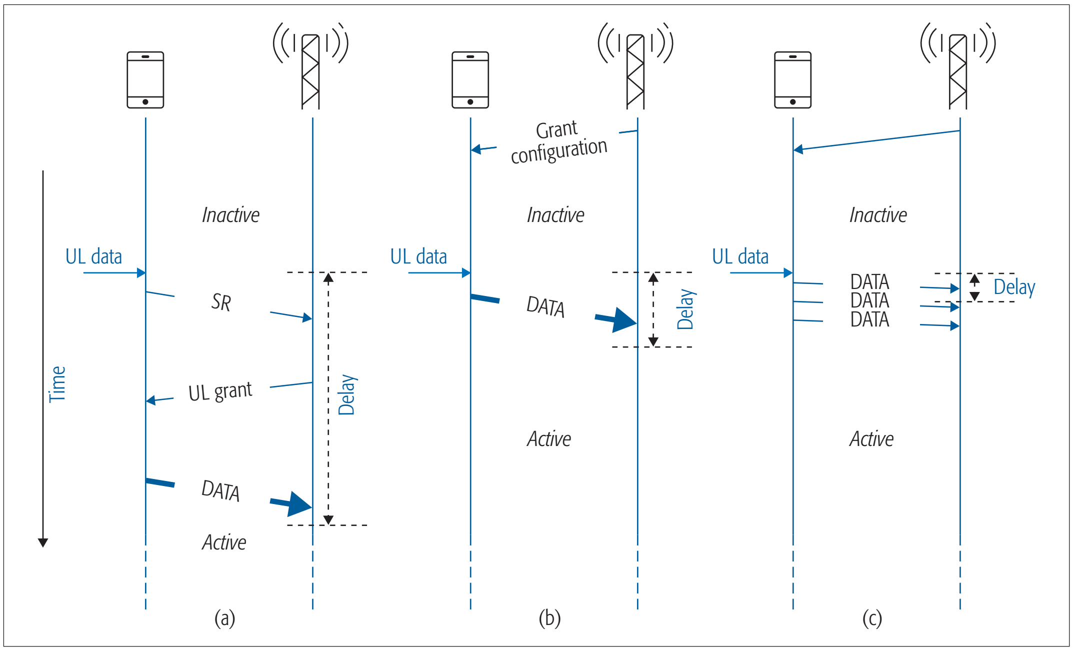


FIGURE 3. UL scheduling a) with scheduling request (SR) access; b) with SPS UL access; c) combined with shorter transmissions.

Achievable 5G URLLC Latencies

We now want to understand what latencies are achievable with different URLLC configurations. Since URLLC is about latencies that can be guaranteed with high reliability (e.g., $1 - 10^{-5}$), we are interested in the worst case latencies (e.g., with maximum slot alignment times) for each URLLC configuration. Both FDD and TDD carrier configurations are considered. For FDD both LTE evolution as well as NR are investigated; for TDD only NR is investigated, since LTE TDD is not considered for URLLC enhancements. Assumptions on timings are made according to the discussions in 3GPP.

The following questions are being investigated:

- How does the NR numerology impact latencies?
- How does the latency change when slots of 14 OFDM symbols are complemented by shorter transmissions of 7, 4, or 2 symbols?
- What benefits does SPS UL transmission provide over scheduling request (SR)-based UL transmission?

The latency for a DL packet transmission with k retransmissions can be expressed as

$$T_{tot} = T_{align} + T_{tx} + 2T_{proc} + k \cdot (2T_{turn} + 2T_{tx}).$$

The worst case alignment delay T_{align} corresponds to one (mini-)slot duration for FDD and two (mini-)slot durations for TDD, where we assume alternating UL-DL (mini-)slots in TDD as described earlier. The transmission time T_{tx} equals the (mini-)slot duration. We further assume one (mini-)slot duration as processing time for each transmitter and receiver, which represents the layer 1 and layer 2 processing for the packet transmission. For each retransmission round, two transmission times T_{tx} are added, plus two turn-around times T_{turn}. For NR the turn-around times in the device and the BS are set to allow for a processing time between 3 symbol durations at low numerologies (15–30 kHz) and 9 symbols at high numerology (120 kHz), rounded up to entire (mini-)slots. For LTE, the following turn-around times are assumed: 4 TTIs for 2/3-symbol sTTIs, 4 TTIs for 7-symbol sTTIs, and 3 TTIs for 14-symbol TTIs. For LTE a 2-symbol physical DL control channel (PDCCH) is assumed, and 3-symbol length is assumed for the 2/3-symbol mixed configuration (Fig. 2).

For the UL, a significant part of delay comes from the exchange of a scheduling request and a UL transmission grant between the user equipment and the base station. Also, from a reliability aspect this signaling bears the risk that both messages need to be correctly decoded in order for the UL transmission to start.

For UL data two cases are considered: SR-based with one SR resource per TTI (requiring one additional round-trip for UL resources) and SPS-based with one configured UL resource per TTI for the UE. Table 1 lists the worst case one-way radio access network (RAN) latencies without retransmission in DL and UL, as well as the additional latency for each hybrid automatic repeat request (HARQ) retransmission (retx). It can be seen that some configurations (those marked in red) do not achieve a RAN latency bound down to 1 ms. Furthermore, the following observations can be made:

- FDD can provide significantly lower latency bounds than TDD, as the latency is not restricted by the UL-DL switching configuration.
- Higher numerologies decrease latency, as expected.
- The usage of mini-slots significantly decreases the guaranteed worst case latencies. The relative latency improvement of mini-slots decreases at higher sub-carrier spacings.
- SPS can significantly reduce UL latencies, so UL latencies become similar to DL latencies.

FDD		Downlink (ms)	Uplink (ms)		Retx delay (ms)
			SPS-based	SR-based	
NR	30 kHz, 14 s mini-slot	1.7	1.7	3.2	1 .5
	30 kHz, 7 s mini-slot	0.86	0.86	1.6	0.75
	30 kHz, 4 s mini-slot	0.54	0.54	0.96	0.43
	30 kHz, 2 s mini-slot	0.39	0.39	0.75	0.36
	120 kHz, 14 s slot	0.46	0.46	0.83	0.38
	120 kHz, 7 s mini-slot	0.33	0.33	0.64	0.31
LTE	15 kHz, 14 s TTI	4.0	4.0	10	6.0
	15 kHz, 7 s sTTI	2.0	2.0	6.0	4.0
	15 kHz, 2 s sTTI	1.0	0.86	2.3	1.4
TDD (DUDU pattern)		Downlink (ms)	Uplink (ms)		Retx delay (ms)
			SPS-based	SR-based	
NR	30 kHz, 14 s slot	2.2	2.2	4.1	2.0
	30 kHz, 7 s slot	1.1	1.1	2.1	1.0
	30 kHz, 4 s mini-slot	0.68	0.68	1.3	0.57
	120 kHz, 14 s slot	0.58	0.58	1.1	0.5
	120 kHz, 7 s mini-slot	0.39	0.39	0.64	0.25

TABLE 1. Worst case RAN transmission latencies for different 5G URLLC configurations (note that average latencies can be lower).

- LTE URLLC configurations for FDD can provide low latencies that are on the order of NR URLLC for TDD configurations at numerologies of 30 kHz.
- The turn-around times in NR are lower than for LTE, so delays for retransmissions are lower.

5G Design for Ultra-Reliable Communication

Robust Transmission Modes

As described earlier, the general URLLC reliability requirement for one transmission of a packet is 1 – 10^{-5} for 32 bytes with a user plane latency of 1 ms. For the physical layer reliability, this corresponds to a maximum block error rate (BLER) of 10^{-5} (i.e., 0.001 percent) which needs to be achieved at a certain channel quality depending on the deployment in which the URLLC service is operated. For LTE, besides fulfilling the requirement above, the work item description on URLLC [12] also considers use cases with requirements that are less stringent in terms of combinations of reliability and latency.

The transmission of a packet can comprise several steps, which can include control signaling and data transmission, for example, for assigning resources or HARQ feedback for retransmissions. This implies that each individual part of the transmission chain should be reliable enough such that the overall reliability for the entire transmission sequence of the packet is achieved. In the following, we discuss some techniques applicable to both NR and LTE, and aspects of the new NR design that enable ultra-high-reliability transmission.

To achieve ultra-reliable transmissions over a fading radio channel, significant signal-to-noise ratio (SNR) margins are needed. Diversity is a key element for providing ultra-reliable transmission for both NR and LTE while also keeping the fading margins at reasonable levels [1, 13]. Diversity can be exploited, for example, in the time, frequency, and spatial domains. In the time domain, diversity can be achieved by repetitions or feedback-based retransmission when the radio channel has changed its fading (i.e., after the channel coherence time); retransmission within the coherence time still provides repetition or coding gains. It is often not possible to exploit time diversity for ultra-reliable low-latency services if the latency requirements are shorter than the channel coherence time. Diversity in the frequency domain can be exploited within the physical bounds of the channel and available bandwidth by using techniques such as distributed resource mapping or frequency hopping. In the spatial domain, multiple antenna configurations at the transmitter and receiver determine the diversity order. For example, the ITU evaluation configuration [14] suggests that for certain urban macro environments, up to eight transmitter/receiver UE antennas can be considered. With multiple transmit antennas, an open-loop transmit diversity technique such as precoder cycling, which cycles through a set of precoding matrices over the bandwidth of the transmission, can provide spatial diversity as a function of frequency. With multiple receive antennas, different versions of the same signal will be available at the receiver. It is less likely that all of these versions will be in deep fade; therefore, they can be combined to effectively increase the signal-to-interference-plus-noise ratio (SINR).

The design of the NR control channel provides flexibility to support different service requirements. To ensure high reliability and wide coverage for the physical DL control channel (PDCCH), NR supports sufficiently low code rate transmission for typical URLLC downlink control information (DCI) sizes. The NR physical UL control channel (PUCCH) supports both short formats of duration 1–2 OFDM symbols and long formats of duration 4–14 OFDM symbols, enabling low-latency features for mini-slot-based transmission and ultra-high reliability for UL control transmission. One relevant example is the two-symbol PUCCH, which also enables frequency hopping and thus increased reliability. In LTE, short PDCCH and short PUCCH have been introduced to support low-latency transmission based on sTTI, as illustrated in Fig. 2. Similar reliability enhancements such as support for repetition of data and control, compact DCI format, and high aggregation level can also be considered without significant impact on the existing LTE system.

To support very low BLER operation at reasonable SINR levels, some form of robust channel coding is required. In LTE, turbo codes are used for data channels and tail-biting convolutional or Reed-Müller codes for control channels. Reliability enhancement can be achieved by, for example, extending the existing modulation and coding schemes (MCSs) to support operations at lower code rates. In NR, 3GPP has chosen new channel coding techniques, namely low-density parity check (LDPC) codes and polar codes for data and control channels, respectively. The LDPC codes for NR support two base graphs for the parity check matrix. The use of two base graphs provides benefits in terms of performance and implementation complexity, supporting a wide range of code rates and information block sizes. In particular, the LDPC base graph #2 is extended to support a code rate of 1/5 without relying on repetition, unlike in LTE, where repetition is used for code rates below 1/3. This allows higher coding gains at low code rates, which are suitable for use cases requiring high reliability like URLLC.

Multi-Connectivity

In multi-connectivity, the UE is connected via multiple carriers to the radio network. Several flavors of multi-connectivity have been defined in 3GPP for LTE in recent years. While these features previously focused on improving user throughput by aggregating resources of the different used carriers, the focus in 3GPP has shifted recently, and new features are developed for LTE and NR to improve the transmission reliability.

3GPP introduced carrier aggregation (CA) in Release-10 as a method for the UE to connect via multiple carriers to a single base station. In CA, the aggregation point is the medium access control (MAC) entity, allowing a centralized scheduler to distribute packets and allocate resources, for example, according to the channel knowledge among all carriers, but also requiring tight integration of the radio protocols involved. In Release-12 dual connectivity (DC) was introduced for LTE, where a different protocol architecture for the resource aggregation is used: the aggregation point has been moved up to the Packet Data Convergence Protocol (PDCP). This way, two MAC protocols with their separate scheduling entities can be executed in two distinct nodes, without strict requirements on their interconnection. DC also provides the basis for LTE and NR integration, and as such will be the basis for the first NR release in 3GPP Release-15.

Currently, in 3GPP Release-15, both architecture concepts of CA and DC are reused in both LTE [12] and NR [15] to provide a method to improve the transmission reliability on higher layers, beyond reliability improvements intended on the physical layer (PHY). This is achieved by packet duplication, which has been decided to be employed at the PDCP layer. An incoming data packet, for example, of a URLLC service, is thereby duplicated on PDCP, and each duplicate undergoes procedures on the lower-layer protocols, radio link control (RLC) and MAC, and hence individually benefits from retransmission reliability schemes on RLC, MAC, and PHY. Eventually, the data packet will thus be transmitted via different frequency carriers to the UE, which ensures uncorrelated transmissions due to frequency diversity, and in the case of DC, from different sites, even macro diversity. The method is illustrated in Fig. 4 for both CA and DC.

It is noteworthy that due to the chosen protocol architectures for DC and CA, the duplicate packet transmissions do not need to be synchronized. The individual schedulers can independently transmit the packet as soon as possible, minimizing the transmission time difference among the packets and thus the observable latency.

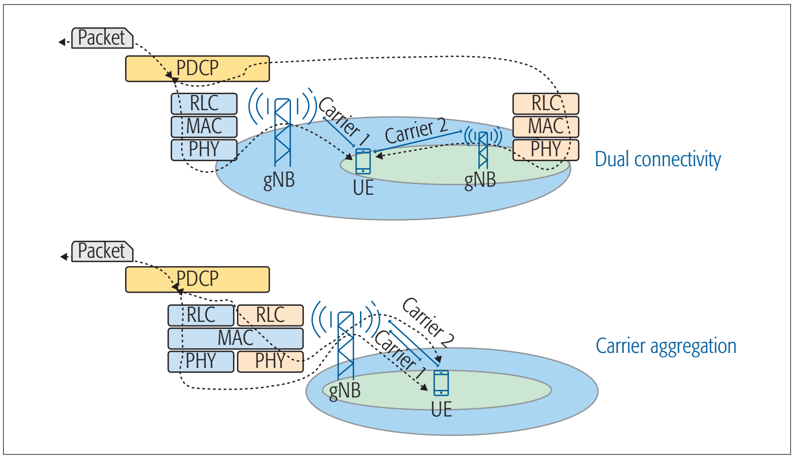


FIGURE 4. Packet duplication in DC and CA; conceptually applicable to LTE and NR.

Frequency diversity among carriers goes beyond diversity schemes offered by the PHY on the same carrier. Compared to time diversity (e.g., repetition schemes), it has the advantage of mitigating potential time correlations of the repetitions. Furthermore, carrier diversity allows, as shown in Fig. 4 for DC, placing transmission points at different locations, thus further reducing potential correlations of the transmission by the introduced spatial diversity.

Multi-connectivity with packet duplication on PDCP has the advantage of utilizing lower-layer retransmission schemes (HARQ and RLC retransmission) effectively, and by this lowering the latency to be guaranteed with a certain reliability. For example, let us assume the PHY achieves, for each HARQ transmission, a residual error probability of 0.1 percent. In 0.1 percent of the cases a retransmission is required, increasing the transmission latency by an extra HARQ round-trip time (RTT). With packet duplication, the probability that both uncorrelated HARQ transmissions fail is 0.1 percent * 0.1 percent. This means that in $1 - 10^{-6}$ of the cases, low latency without the extra HARQ RTT is achieved, since simply the first decodable packet duplicate is accepted and delivered, while the second is discarded.

Furthermore, multi-connectivity based on dual connectivity has the potential to enable reliable handovers without handover interruptions for user plane data. Thereby, the handover is done in two steps: one carrier is moved at a time from source to target node, and hence the UE always maintains at least one connection. During the procedure, packet duplication may be employed so that packets are available at both nodes for interruption-free transmission to the UE.

Spectral Efficiency of 5G URLLC

To achieve high reliability, the network can use different scheduling strategies. A packet can be encoded with a very low code rate and just be transmitted once, or alternatively be transmitted multiple times with a higher initial code rate. The most efficient option is to perform retransmissions based on feedback. However, this is only possible if the transmission duration and RTT are shorter than the latency constraint. The shorter a trans-

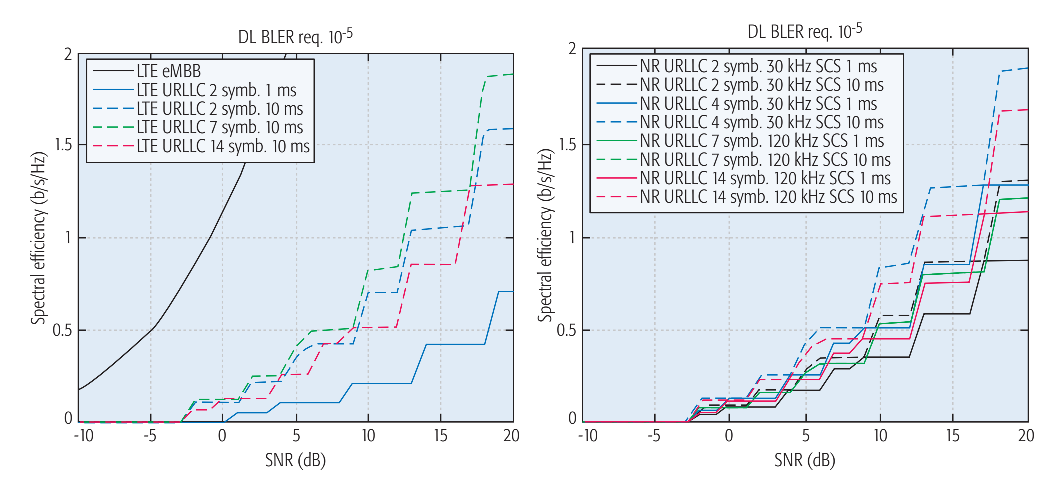


FIGURE 5. Spectral efficiency of NR FDD (right) and LTE FDD (left) URLLC with requirement of 10^{-5} maximum error rate and different latency constraints compared to LTE eMBB.

Comparing URLLC with eMBB, it is clear that the high reliability has a cost in terms of higher spectrum usage, as is expected from the more robust scheduling. But when considering URLLC services, it should also be expected that the corresponding URLLC traffic generated would be significantly lower than that of eMBB.

mission RTT is compared to the latency constraint, the higher the initial code rate can be set, giving higher spectral efficiency. An important limitation to this scheme is the reliability of the feedback channels; if the feedback fails, a retransmission is not triggered. The cost of ultra-reliable transmission in terms of required radio resources and transmission energy is reflected in the spectral efficiency of the transmission.

An evaluation of the spectral efficiency for DL packets, shown in Fig. 5, assumes a target error rate of 10^{-5} with 1 ms or 10 ms RAN latency constraints. The transmission latencies and HARQ timing was computed according to Table 1. The antenna configuration was 1 transmit antenna and 2 uncorrelated receive antennas. A 20 MHz wide carrier at 700 MHz was assumed. The NR subcarrier spacing was set to 30 kHz or 120 kHz and the TTI length to 2, 4, or 7 symbols. Realistic overhead was accounted for: 24 percent (LTE 14 symbols); 25 percent (LTE 7 symbols); 37 percent (LTE 2 symbols); 21 percent (NR 30 kHz 7 symbols); 25 percent (NR 30 kHz 4 symbols); 49 percent (NR 30 kHz 2 symbols); 33 percent (NR 120 kHz 14 symbols); and 52 percent (NR 120 kHz 7 symbols). A set of 13 MCS with a code rate between 1/20 and 2/3 and modulation between quadrature phase shift keying (QPSK) and 64-QAM was used, which encoded a 32 B packet with LDPC. The fading model TDL-C was used with a delay spread of 300 ns and an assumed speed of 3 km/h.

For LTE, only the shortest LTE TTI configuration can deliver the 1 ms latency constraint (Table 1). At a more relaxed constraint of 10 ms the longer TTIs also become usable, and provide higher spectral efficiency due to lower overhead. With NR the strictest latency requirements can be fulfilled with different SCS, and the gain from a short transmission (high SCS) is balanced by the overhead cost due to longer cyclic prefix required, as discussed earlier. Enabling more retransmissions within the latency budget through shorter transmissions improves the spectral efficiency, up to a limit set by the control channel performance for triggering a retransmission.

Comparing URLLC with eMBB, it is clear that the high reliability has a cost in terms of higher spectrum usage, as is expected from the more robust scheduling. But when considering URLLC services, it should also be expected that the corresponding URLLC traffic generated would be significantly lower than that of eMBB. For example, if URLLC constitutes 10 percent of the total traffic, the average spectral efficiency would be reduced by ~7 percent if the difference in spectral efficiency is a factor of 3. This means that the introduction of URLLC services would come at a moderate cost in terms of system capacity.

Summary

3GPP started the specification of the 5G radio technology in early 2017. Two tracks are being pursued: a backward-compatible evolution of LTE and the New Radio interface. One class of new services that 5G is targeted to address is ultra-reliable low-latency communication services. The requirements that have been defined for URLLC in ITU and 3GPP are the capability to transfer a small packet over the radio interface, where successful transmission can be guaranteed with a failure probability of 10^{-5} within a specified time bound. This article describes the 5G functionality and performance for URLLC services.

In order to enable low-latency communication, the following features are introduced in 5G. New short slot structures enable faster uplink and downlink transmission for URLLC, called *mini-slot* for NR and *short TTI* for the LTE radio interfaces. These short slots can be simultaneously transmitted with longer ordinary slots being used by other services, and can be prioritized over other traffic by pre-empting ongoing transmissions. NR also introduces higher OFDM sub-carrier spacings, which can scale the slots down in the time domain. For uplink transmission grant-free access is introduced to shorten the procedure for uplink resource assignments. An assessment of the achievable latency is made for different options of NR and LTE configurations. It is shown that all low-latency features can significantly reduce transmission latencies, and both LTE and NR can reach the ITU and 3GPP latency requirements.

In order to increase the reliability for URLLC services, robust coding and modulation and diversity schemes can be applied in accordance with the LTE and NR designs. Redundancy can further be provided over different carriers and transmission points by means of multi-connectivity extending -the dual connectivity and carrier aggregation frameworks specified for LTE. The spectral efficiency for different 5G URLLC configurations is evaluated. The support of URLLC services comes at the cost of reduced spectral efficiency compared to mobile broadband services without latency or reliability constraints. However, many URLLC services have lower traffic volumes than MBB services, and will in most cases only account for a fraction of the total traffic being served by the 5G network. In this case it can be expected that the URLLC impact on the overall network capacity is moderate.

> The support of URLLC services comes at the cost of reduced spectral efficiency compared to mobile broadband services without latency or reliability constraints. However, many URLLC services have lower traffic volumes than MBB services, and will in most cases only account for a fraction of the total traffic being served by the 5G network.

References

[1] J. Sachs *et al.*, "Machine-Type Communications," *5G Mobile and Wireless Communications Technology*, 2016; www.cambridge.org/9781107130098. ISBN 9781107130098.

[2] ITU-R Working Party 5D, "Minimum Requirements Related to Technical Performance for IMT-2020 Radio Interface(s)," draft new report ITU-R M.[IMT-2020.TECH PERF REQ], 22 Feb. 2017.

[3] E. Dahlman *et al.*, "5G Wireless Access: Requirements and Realization," *IEEE Commun. Mag.*, vol. 52, no. 12, Dec. 2014, pp. 42–47.

[4] 3GPP, "Feasibility Study on New Services and Markets Technology Enablers for Critical Communications," tech. rep. TR 22.862, Sept. 2016.

[5] M. A. Lema *et al.*, "Business Case and Technology Analysis for 5G Low Latency Applications," *IEEE Access*, vol. 5, 2017, pp. 5917–35.

[6] 3GPP, "Requirements for Further Advancements for Evolved Universal Terrestrial Radio Access," tech.l rep. TR 38.913, Mar. 2017.

[7] M. Simsek *et al.*, "5G-Enabled Tactile Internet," *IEEE JSAC*, vol. 34, no. 3, 2016, pp. 460–73.

[8] J. Vihril *et al.*, "Numerology and Frame Structure for 5G Radio Access," *27th IEEE Annual Int'l. Symp. Personal, Indoor, and Mobile Radio Communications*, Valencia, 2016.

[9] I. Aktas *et al.*, "LTE Evolution — Latency and Reliability Enhancements for Wireless Industrial Automation," *Proc. 28th IEEE Annual Int'l. Symp. Personal, Indoor, Mobile Radio Commun.*, Montreal, Canada, Oct. 8–13, 2017.

[10] P. Schulz *et al.*,"Latency Critical IoT Applications in 5G: Perspective on the Design of Radio Interface and Network Architecture," *IEEE Commun. Mag.*, vol. 55, no. 2, Feb. 2017.

[11] J. C. S. Arenas, T. Dudda, and L. Falconetti, "Ultra-Low Latency in Next Generation LTE Radio Access," *Proc. 11th Int'l. ITG Conf. Sys. Commun. Coding*, Hamburg, Germany, Feb. 2017 .

[12] 3GPP, Work Item on Ultra Reliable Low Latency Communication for LTE (URLLC), Dubrovnik, Croatia, March 6–9, 2017 [RP-170796].

[13] N. A. Johansson *et al.*, "Radio Access for Ultra-Reliable and Low-Latency 5G Communications," *Proc. IEEE ICC*, London, U.K., June 2015.

[14] Preliminary Draft New Report ITU-R M.[IMT-2020.Eval.], Revision 2 to Document 5D/TEMP/347-E, ITU WP5D 27, June 2017.

[15] 3GPP, Work Item on New Radio (NR) Access Technology, West Palm Beach, FL, June 5-8, 2017 [RP-171485].

Biographies

Joachim Sachs studied electrical and electronics engineering at RWTH Aachen University, Germany, ENSEEIHT, France, the Norwegian University of Science and Technology (NTNU), and the University of Strathclyde, Scotland. He received a diploma from RWTH Aachen University in 1997 and a doctorate degree from Technical University of Berlin in 2009. He joined Ericsson Research in 1997 and currently holds the position of principal researcher in network architectures and protocols. He has contributed to the development and evaluation of 3G and 4G, and today coordinates the research and standardization activities on machine-type communication and URLLC for 5G. He contributes to several cross-industry research collaborations. He has published several books and book chapters, over 80 papers in scientific journals and conferences, and has filed more than 120 inventions (patent families). He received the Ericsson inventor of the year award in 2006 and the research award of the Vodafone Foundation for Scientific Research in 2010. He was a visiting scholar at Stanford University in 2009. Since 1995, he has been active in the IEEE and the German VDE Information Technology Society (ITG), where he currently co-chairs the Technical Committee on Communication Networks and Systems. He has been a Program Committee member for numerous scientific conferences and is on the Editorial Boards of two scientific journals.

Gustav Wikstrüm is a master researcher at Ericsson Research. He has a background in experimental particle physics and received his Ph.D. from Stockholm University in 2009, after Master's studies in engineering physics in Lund, Uppsala, and Rennes. After postdoctoral studies in Geneva, he joined Ericsson in 2011. He has been driving the evolution of network performance studies, and worked with WLAN and LTE capacity enhancements. Since 2015 he has been the driver of latency and reliability improvements (URLLC) in LTE and NR. His research interests include network algorithms, signal processing, and communication theory for wireless communications systems.

Torsten Dudda is a senior researcher in Radio Networks Research located in Aachen, Germany. He joined Ericsson in 2012 and has focused on radio architecture and protocol design, working closely with standardization of LTE and 5G NR during 3GPP Releases 10–15. His current research is centered around enhancing LTE and 5G NR to enable 5G critical machine-type communication use cases. He graduated as a diploma engineer from RWTH-Aachen University in Germany.

Robert Baldemair received his Dipl.Ing. and Dr. degrees from Vienna University of Technology in 1996 and 2001, respectively. In 2000 he joined Ericsson where he initially was engaged in research and standardization of ADSL and VDSL. Since 2004 he has been working on research and development of radio access technologies for LTE and since 2011 on wireless access for 5G. Currently he holds a master researcher position at Ericsson. He received the Ericsson Inventor of the Year 2010 award, an award Ericsson awards to employees with substantial contributions to Ericsson's patent portfolio. In 2014 he and colleagues at Ericsson were nominated for the European Inventor Award, the most prestigious inventor award in Europe, for their contribution to LTE. His research interests include signal processing and communication theory for wireless communications systems.

Kittipong Kittichokechai received his B.Eng. degree in electrical engineering in 2007 from Chulalongkorn University, Thailand, and his M.Sc. and Ph.D. degrees, both in electrical engineering in 2009 and 2014 from KTH Royal Institute of Technology, Sweden. In 2012, he was a visiting scholar at the Information Systems Laboratory, Stanford University, California. From 2014 to 2016, he was a postdoctoral researcher at Technische Universität Berlin, Germany. Since 2016, he has been a researcher at Ericsson Research, Stockholm, Sweden, where he has been contributing to the development of new communication technologies of 5G. His research interests include network information theory, information theoretic security and privacy, distributed detection, and their applications in wireless communications. He was a recipient of the Ananda Mahidol Foundation Scholarship under the Royal Patronage of His Majesty the King of Thailand.