IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

An overview of physical layer design for Ultra-Reliable Low-Latency Communications in 3GPP Releases 15, 16, and 17

TRUNG-KIEN LE[1], UMER SALIM[2], AND FLORIAN KALTENBERGER.[1],
[1]EURECOM, Sophia Anitpolis, France
[2]TCL Communications, Sophia Antipolis, France

Corresponding author: Trung-Kien Le (e-mail: Trung-Kien Le@eurecom.fr).

This work was supported in part by TCL and H2020 project 5GENESIS (5genesis.eu).

ABSTRACT

Ultra-reliable low-latency communication (URLLC) has been introduced in 5G new radio for new applications that have strict reliability and latency requirements such as augmented/virtual reality, industrial automation and autonomous vehicles. The physical layer design of the first 5G release, Release 15, was finalized in December 2017. It provided a foundation for URLLC with new features such as flexible subcarrier spacing, a sub-slot-based transmission scheme, new channel quality indicator, new modulation and coding scheme tables, and configured-grant transmission with automatic repetitions.

The second 5G release, Release 16, was finalized in December 2019 and allows achieving improved metrics for latency and reliability to support new use cases of URLLC. A number of new features such as enhanced physical downlink (DL) control channel monitoring capability, new DL control information format, sub-slot physical uplink (UL) control channel transmission, sub-slot-based physical UL shared channel repetition, enhanced mobile broadband and URLLC inter-user-equipment multiplexing with cancellation indication and enhanced power control were standardized. This article provides a detailed overview of the URLLC features from 5G Release 15 to Release 16 by describing how these features allow meeting URLLC target requirements in 5G networks.

The ongoing Release 17 targets further enhanced URLLC operation by improving mechanisms such as feedback, intra-user-equipment multiplexing and prioritization of traffic with different priority, support of time synchronization and new quality of service related parameters. In addition, a fundamental feature targeted in URLLC Release 17 is to enable URLLC operation over shared unlicensed spectrum. The potential directions of URLLC research in unlicensed spectrum in Release 17 are presented to serve as a bridge from URLLC in licensed spectrum in Release 16 to URLLC in unlicensed spectrum in Release 17.

INDEX TERMS 5G, URLLC, physical layer design, 3GPP Release 15, 3GPP Release 16, 3GPP Release 17

I. INTRODUCTION

A. ULTRA-RELIABLE LOW-LATENCY COMMUNICATION (URLLC) OVERVIEW

To satisfy the requirements of emerging applications such as intelligent transportation, augmented/virtual reality, industrial automation, etc., Third Generation Partnership Project (3GPP) defined three main service categories in 5G New Radio: Enhanced mobile broadband (eMBB), Massive machine-type communication and URLLC. In these three service categories, the physical design of URLLC is the most challenging one because two conflicting factors of reliability and latency have to be coped with at the same time. In classic communication, one of two factors must be sacrificed to attain the other factor. To achieve a low latency, a shorter packet has to be used that causes a degradation in channel coding and results in a decrease of reliability. In contrast, to improve the reliability, while a bigger number of retransmissions can be used in eMBB transmission, latency requirement limits the number of retransmissions in URLLC transmission. Moreover, if more time domain resources are consumed due to an increase of parity check bits in the low code rates, it also increases latency and reduces the system efficiency.

In 3GPP Release 15, URLLC is targeted to support use cases such as smart grid, augmented and virtual reality in entertainment industry. Based on these use cases, 3GPP defines the requirements to be used in URLLC design: “A general

URLLC reliability requirement for one transmission of a packet is 10^{-5} for 32 bytes with a user plane latency of 1 ms" [1]. This reliability requirement poses a challenge in URLLC design because it is much higher than the typical block error rate of Long-Term Evolution (LTE) system that is 10^{-2}. Release 16 URLLC enhancements have further boosted the requirements setting 10^{-6} as reliability target and a latency further down in a range of 0.5 to 1 ms to support new use cases in industrial applications [2].

The focus of this works is to provide a comprehensive overview of the physical layer design for URLLC that has undergone fundamental changes compared to legacy LTE systems to satisfy strict requirements on latency and reliability.

B. PAPER ORGANIZATION

3GPP Release 15 is the first release of the 5G standard. It specified URLLC requirements that are much stricter than LTE requirements to support use cases such as smart grid, augmented and virtual reality in entertainment industry so this release built a foundation for URLLC design to achieve these stringent requirements by introducing higher subcarrier spacing (SCS), sub-slot transmission time intervals, configured grant resources, etc. The physical layer work of Release 15 was completed in December 2017. A summary of these features is presented in Section II.

Release 16 continued to develop further the physical layer design for URLLC to deal with the unsolved problems in Release 15 as well as support Industrial Internet of Things with more stringent requirements (higher reliability of 10^{-6}, lower latency of 0.5 to 1 ms) in some URLLC use cases of Release 16 as specified in [2], [4] and [5]: factory automation, transport industry including the remote driving use case and electrical power distribution. The physical layer work of Release 16 was completed in December 2019. An overview of the challenges and the techniques standardized in Release 16 are described in Section III.

The ongoing Release 17 is to enhance the Release 16 features and extend URLLC operation to unlicensed spectrum besides the operation in licensed spectrum in Release 15 and 16 so that URLLC transmission achieves a better performance [6]. Release 17 physical layer work has started and is expected to be completed at the end of 2021. The objectives of Release 17 and some techniques for these objectives expected to be specified in Release 17 are presented in Section IV. Some conclusions are drawn in Section V.

C. MAIN CONTRIBUTIONS

URLLC is a new service category and the work of physical layer design for URLLC starts from the first full-set-of-5G-standard release that is Release 15 and is still being continued in the ongoing Release 17. This work provides a full picture of URLLC physical layer design from Release 15 to the ongoing Release 17. The problems in each step of URLLC evolution are presented from the transition from LTE to 5G in Release 15 to Release 16 then to Release 17. The URLLC features standardized in Release 15 and 16 are described highlighting the difference from legacy and explaining how they help improve the URLLC performance. Some simulations are also done to show the benefits of the new techniques compared to the conventional techniques. The work also analyzes the principal research directions of the ongoing Release 17 and presents the techniques that are candidates to be standardized in Release 17 to enhance URLLC performance in both licensed and unlicensed spectrum.

II. 3GPP RELEASE 15 FOUNDATION FOR URLLC IN 5G

A. FLEXIBLE NUMEROLOGY AND SUB-SLOT-BASED TRANSMISSION

A key new feature in 5G is that the introduction of flexible SCS. Whereas in LTE the SCS was fixed to 15kHz, in 5G values of 15 kHz, 30 kHz, 60 kHz, 120 kHz and 240 kHz are allowed. This is one of the major differences between 5G and LTE that aims to reduce transmission latency by decreasing the time length of Orthogonal frequency division multiplexing (OFDM) symbols. By using flexible SCS, 5G changes OFDM symbol duration including cyclic prefix duration from a fixed value of 71.35μs to a set of $71.35, 35.68, 17.84, 8.92$ and 4.46μs.

In LTE, slot-based transmission (Physical downlink shared channel (PDSCH)/Physical uplink shared channel (PUSCH) mapping Type A) is used where one slot is a transmission time interval. The transmission only can start at the beginning of a slot so if a packet arrives after the starting point in a slot, it must wait until the next slot to be transmitted. This alignment time is harmful to URLLC with low latency requirement. Therefore, in 5G, to further reduce latency by shortening transmission time interval, sub-slot based transmission (PDSCH/PUSCH mapping Type B) is introduced where a packet is scheduled in a transmission time interval of 2, 4 or 7 OFDM symbols. A transmission can start at the beginning of the sub-slot transmission time interval so it has more occasions to start in one slot instead of only one occasion in a slot in LTE. It reduces the waiting time before an arriving packet is transmitted.

B. CHANNEL QUALITY INDICATOR (CQI) AND MODULATION AND CODING SCHEME (MCS) TABLES FOR URLLC

New CQI and MCS tables are specified to support the PDSCH and PUSCH transmission with URLLC requirement of 10^{-5} besides the CQI and MCS tables for eMBB with block error rate of 10^{-1}. These tables allow the transmission to have the appropriate code rate and modulation scheme for URLLC transmission.

C. PREEMPTION INDICATION IN DOWNLINK (DL) TRANSMISSIONS' MULTIPLEXING

In DL, when the base station (gNB) wants to schedule a URLLC transmission over the resources that are already allocated to an eMBB transmission, the gNB can puncture

the eMBB transmission's resources to schedule an URLLC transmission in those punctured resources. This means that the URLLC packet is transmitted as soon as possible after its arrival with eMBB and URLLC multiplexing instead of waiting until the end of the ongoing eMBB transmission to reduce latency. After puncturing a part of the eMBB transmission, the gNB transmits an preemption indication to the eMBB user equipment (UE) so as to inform that the resources indicated are punctured and contain data of URLLC transmission rather than its own eMBB transmission. Thus, the eMBB UE does not take into account the resources punctured when decoding data.

D. UPLINK (UL) CONFIGURED-GRANT (CG) TRANSMISSION

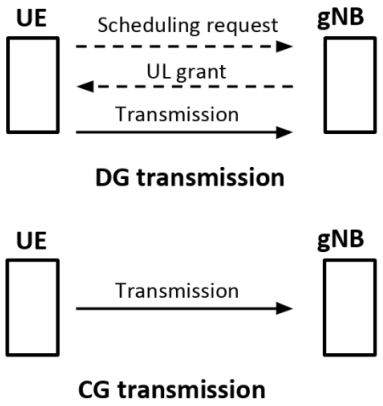

FIGURE 1. UL DG and CG transmission.

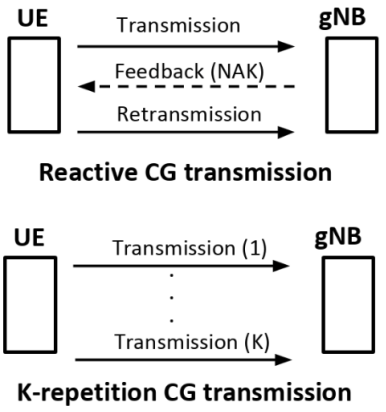

FIGURE 2. UL repetition CG transmission.

In LTE, UL dynamic-grant (DG) transmission requires scheduling request (SR) from the UE and UL grant from the gNB that occupy a large portion of time. To reduce transmission's latency in 5G, besides the conventional DG transmission, CG transmission is standardized to support time sensitive transmission. CG resources are configured to the UE by the gNB so that the UE uses these CG resources to transmit data on PUSCH directly to the gNB without SR and UL grant as shown in Fig. 1. There are two types of CG PUSCH transmission. In Type 1 CG PUSCH transmission, a radio resource control (RRC) signalling configures the time and frequency domain resource allocation including periodicity of CG resources, offset, start symbol and length of PUSCH, MCS, the number of repetitions, redundancy version, power level, etc. In Type 2 CG PUSCH transmission, only periodicity and the number of repetitions are configured by RRC signalling. The other parameters are configured through an activation downlink control information (DCI). Another technique to reduce latency as well as increase reliability in UL CG transmission is that the UE in 5G is configured to transmit automatically a number of repetitions in the consecutive available slots without waiting feedback from the gNB as in LTE as illustrated in Fig. 2.

Fig. 3 shows the higher reliability of K-repetition CG transmission compared to DG transmission and reactive CG transmission because a lower latency for one transmission in K-repetition CG transmission allows more repetitions of a transport block in the URLLC latency budget of 1ms. In the simulation, a packet of 160 bits is encoded by low-density parity-check code with MCS1 and quadrature phase shift keying modulation and transmitted in additive white Gaussian noise channel. SCS is 30 kHz. Due to latency of SR and UL grant, a packet is only transmitted one time in URLLC latency budget of 1ms in DG transmission. In reactive CG transmission, due to latency of feedback, there are maximum two repetitions (an initial transmission and a retransmission) of a packet transmitted in 1ms. In K-repetition CG transmission, there are four repetitions of a packet transmitted in 1ms.

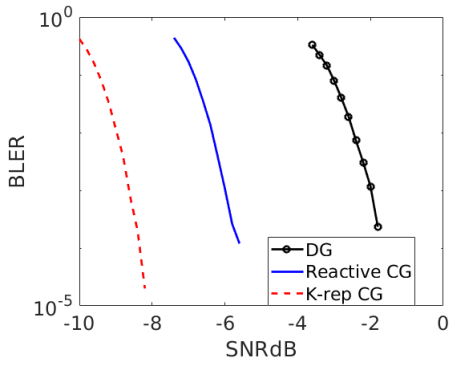

FIGURE 3. Comparison of UL transmission performance in DG transmission, reactive CG transmission and K-repetition CG transmission.

III. 3GPP RELEASE 16 FEATURES FOR URLLC IN 5G

The standardized techniques in Release 15 enhance the performance of URLLC but new use cases such as factory automation, transport industry including the remote driving use case and electrical power distribution with stricter requirements (higher reliability of 10^{-6}, lower latency of 0.5 to 1 ms) in Release 16 require more improvements in URLLC physical layer design. New studies were carried out

in Release 16 and led to new techniques standardized so that the URLLC performance in the targeted use cases is ensured. Section III-A and Section III-B are about the new features for DL transmission in Release 16. From Section III-C to Section III-F are about the new features for UL transmission in Release 16.

A. PHYSICAL DOWNLINK CONTROL CHANNEL (PDCCH) ENHANCEMENTS

1) PDCCH monitoring capability enhancements

As presented in Section II-A, sub-slot-based transmission is one of the features in Release 15 of URLLC. In DL transmission, this feature requires the UE to monitor DL data including PDCCH and PDSCH in sub-slot level. The location of PDSCH is indicated by PDCCH so the UE needs to decode PDCCH before decoding PDSCH. However, the UE does not know the exact location of PDCCH so it carries out blind decoding in a search space. Each possible location of PDCCH in the search space is called PDCCH candidate. However, in Release 15, the number of PDCCH candidates that the UE can monitor in a slot is limited as shown in Table 1. Moreover, the resource for PDCCH in a slot is also limited as shown by the number of control channel elements (CCEs) in Table 1. A CCE consist of 6 resource element groups. A resource element group equals one resource block during one OFDM symbol that contains 12 resource elements. The number of CCEs that a PDCCH has is defined as the aggregation level (AL) (for example, 1 CCE is AL 1, 2 CCEs are AL 2). The transmission might be in sub-slot level while PDCCH monitoring capability is only defined in slot level. This limit degrades the ability of the UE to operate in sub-slot-based transmission when not all PDCCHs can be transmitted from the gNB and monitored by the UE. For example, if the gNB transmits PDCCH in a sub-slot of 2 OFDM symbols with SCS of 60 kHz, the UE has 7 occasions to monitor PDCCH in a slot of 14 symbols. Therefore, the UE, on average, only can monitor 3 PDCCH candidates and 7 non-overlapping CCEs per sub-slot based on Table 1. When AL 8 (8 CCEs) is needed to guarantee PDCCH reliability, there is not enough CCEs for that PDCCH to be transmitted and monitored in a sub-slot. Moreover, with 3 PDCCH candidates per sub-slot, if the UE monitors 2 PDCCH candidates with AL 2 and 1 PDCCH candidate with AL 4, it is not capable of monitoring another PDCCH candidate with AL 8 so this PDCCH AL 8 is dropped or PDCCH with a lower AL is used that decreases reliability.

TABLE 1. UE monitoring capability in a slot in Release 15 [8]

SCS	15kHz	30kHz	60kHz	120kHz
Number of monitored PDCCH candidates	44	36	22	20
Number of non-overlapping CCEs	56	56	48	32

In Release 16, in order to solve this problem, 3GPP enhances PDCCH monitoring capability by defining the maximum number of monitored PDCCH candidates and non-overlapping CCEs per span of 2, 4 or 7 symbols instead of per slot.When monitoring capability is defined per span for sub-slot level transmission, the UE has more PDCCH candidates and non-overlapping CCEs that it can monitor in a sub-slot because the capability is not divided by the number of sub-slots in a slot as in the conventional scheme. Therefore, PDCCH with high AL can be used to guarantee reliability. As in the above example, there are enough CCEs in a sub-slot for PDCCH AL 8 and the UE is also able to monitor several PDCCH candidates with different ALs. Moreover, more PDCCHs are able to be transmitted in a slot that reduces the waiting time due to a bottleneck of PDCCH monitoring capability. The UE can be configured by the gNB to monitor PDCCH for the maximum number of PDCCH candidates and non-overlapping CCEs defined per slot as in Release 15 or per span as in Release 16.

2) New DCI format

In Release 15, DCI formats have a fixed number of bits in the information fields. In Release 16, with the introduction of new RRC parameters, new DCI formats where the number of bits in several fields are configurable based on time and frequency resources of data, frequency hopping, antenna ports etc. are introduced to schedule URLLC UL and DL transmission. Even in some fields, the number of bits can be set to 0 because new RRC parameters are introduced to convey that information or those fields are not required for a specific transmission. For example, in Release 16 DCI, redundancy version field is configurable from 0 bit to 2 bits compared to a fixed 2 bits in Release 15 DCI. Similarly, hybrid automatic repeat request (HARQ) process field is configurable from 0 bit to 4 bits compared to a fixed 4 bits in Release 15 DCI. Therefore, Release 16 DCI can be configured to use less bits than Release 15 DCI that helps improve DCI transmission's performance for URLLC. Using a Release 16 DCI with 24 bits increases reliability of DCI because this DCI with a smaller payload achieves higher reliability than a Release 15 DCI with 40 bits coded with the same codeword length as shown in Fig. 4. In the simulation, a DCI with payload of 24 or 40 bits is added 24 cyclic redundancy check bits and encoded in Polar code to generate a codeword AL 4 and 8 having 576 and 1152 bits, respectively. The codeword is modulated in quadrature phase shift keying and transmitted in additive white Gaussian noise channel. The decoder at the UE is min-sum Successive cancellation list decoder with list size 8.

In Release 16 DCI, some new fields are added to support new features. Priority indicator field with 0 or 1 bit is added to indicate the priority of a PDSCH or PUSCH scheduled. However, in SPS PDSCH and Type 2 CG PUSCH, priority of PDSCH and PUSCH is configured by RRC and is not overwritten by the activation DCI. Open loop power control set indication field with from 0 to 2 bits is added to control PUSCH transmission's power level in case of eMBB and URLLC multiplexing mentioned in Section III-E2. Invalid

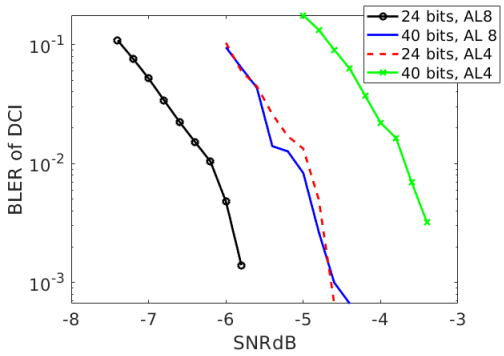

FIGURE 4. Block error rate of Release 15 DCI with 40 bits payload and Release 16 DCI with 24 bits payload.

symbol pattern indicator field with 0 or 1 bit is added to indicate the invalid symbols for PUSCH repetition Type B mentioned in Section III-D.

B. DL SEMI-PERSISTENT SCHEDULING (SPS) ENHANCEMENTS

In DL transmission, the gNB can configure SPS resources with a specific periodicity to the UE. When these SPS resources are activated by the gNB, the UE will expect to receive PDSCH in these resources. Therefore, the gNB can transmit PDSCH without an associated PDCCH to schedule PDSCH resources. A transmission of SPS PDSCH without PDCCH reduces control overhead so SPS PDSCH transmission becomes a promising technique to be used for URLLC. In Release 16, to support URLLC transmission with low latency, periodicity of SPS resources is supported down to one slot for all SCS. To serve different types of traffic, the gNB can configure multiple configurations of SPS resources with different periodicities, resource allocations, MCS, etc. and indicates the index of SPS configurations by RRC. For a given bandwidth of a serving cell, the maximum number of SPS configurations is 8. Each configuration is activated separately by a DCI from the gNB to the UE. On the other hand, SPS configurations can be released jointly or separately as indicated by a DCI.

SPS resources in different configurations might overlap in time domain. If the UE receives multiple SPS PDSCHs overlapped in time domain, the UE starts by decoding a SPS PDSCH with the lowest SPS configuration index in the first step. In the second step, any SPS PDSCHs in the received group that overlap with the chosen SPS PDSCH in the first step are excluded from the group and not decoded by the UE. The step one and two are repeated to resolve the overlap among the remaining SPS PDSCHs in the group until all overlapped SPS PDSCHs are resolved. The UE only sends HARQ feedback for the SPS PDSCHs chosen to be decoded.

If only HARQ feedback for SPS PDSCHs in multiple SPS configurations are reported, maximum 4 physical uplink control channel (PUCCH) resources are configured common for all SPS configurations per HARQ-ACK codebook. If HARQ feedback for SPS PDSCHs in multipled SPS configurations is multiplexed with HARQ feedback for dynamic scheduled PDSCH, HARQ bit location for SPS PDSCHs is based on the time domain resource assignment (TDRA) table row index and time from the end of PDSCH to the beginning PUCCH for HARQ feedback indicated in the activation DCI.

C. UPLINK CONTROL INFORMATION (UCI) ENHANCEMENTS

1) Multiple PUCCHs for hybrid automatic repeat request-acknowledgement (HARQ-ACK) within a slot

DL transmission in sub-slot level that is featured in Release 15 requires an improvement in feedback transmission. The UE is expected to transmit feedback on sub-slot level as DL data because a fast Negative acknowledgment (NACK) feedback on sub-slot level reduces the reception time of feedback at the gNB and guarantees a retransmission in latency budget of URLLC. However, in Release 15, a UE is able to transmit only one PUCCH with HARQ-ACK information in a slot. If the UE finishes decoding process of a packet after the PUCCH resource for HARQ feedback in a slot, it must wait until the next slot to transmit feedback that delays feedback transmission and a retransmission if necessary. Moreover, if HARQ-ACK for URLLC PDSCH occurs in the same slot as HARQ-ACK for other eMBB/URLLC PDSCHs, all the HARQ-ACK information will be multiplexed together and transmitted over the PUCCH resource indicated in the latest DL assignment. The multiplexing degrades the reliability of HARQ feedback.

In Release 16, therefore, sub-slot-based HARQ-ACK feedback procedure is supported where PUCCH resources are configured per sub-slot of 2 or 7 symbols so multiple PUCCHs for HARQ-ACK can be transmitted within a slot. Any sub-slot PUCCH resource is not across sub-slot boundaries and no more than one transmitted PUCCH carrying HARQ-ACK starts in a sub-slot. In this way, HARQ-ACK feedback is also transmitted in sub-slot level to match with DL transmission in sub-slot level.

2) UCI intra-UE multiplexing

In Release 15, the number of PUCCHs transmitted by a UE in a slot is limited to 2. Therefore, when the UE has multiple overlapping PUCCHs in a slot or overlapping PUCCHs and PUSCHs in a slot, the UE multiplexes different UCI types in one PUCCH/PUSCH. However, in URLLC transmission, low latency requires urgent schedules that cause an overlap of URLLC UCI with PUCCH/PUSCH of a different type services with lower priority where the multiplexing causes a degradation of the URLLC transmission. Moreover, if the ending symbol of the multiplexing PUCCH/PUSCH is later than the ending symbol of URLLC UCI, it causes an additional delay to URLLC transmission. For these reasons, the behavior of the UEs must be specified to guarantee URLLC service.

In Release 16, the behaviors of the UE are standardized following UCI prioritization based on two-level priority so

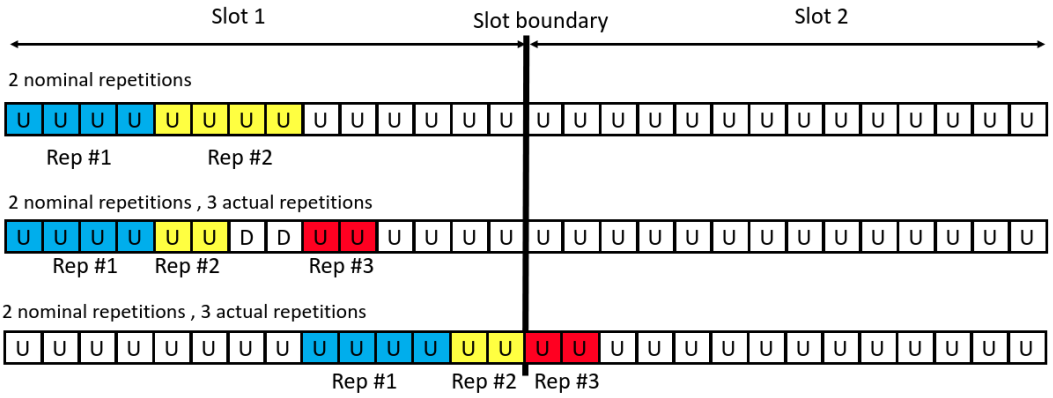

FIGURE 5. PUSCH repetition Type B.

that if there is an overlap between two low priority (LP) and high priority (HP) UL transmissions, the LP UL transmission such as eMBB PUSCH/PUCCH is cancelled instead of being multiplexed with the HP UL transmission such as URLLC PUSCH/PUCCH. In the non-overlapping cancelled symbols of the LP UL transmission, the UE is not scheduled to transmit. In case the UE encounters the intra-collision of more than two UL PUSCH/PUCCH transmissions, the UE resolves collision between UL transmissions with same priority by UCI multiplexing then resolves collision between UL transmission with different priorities by UCI prioritization.

D. PUSCH ENHANCEMENTS

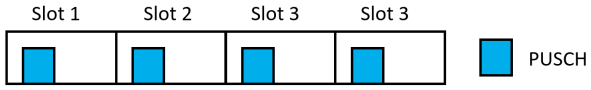

FIGURE 6. PUSCH repetition Type A.

In Release 15, one PUSCH transmission instance is not allowed to cross the slot boundary for both DG and CG PUSCH. Therefore, to avoid transmitting a long PUSCH across slot boundary, the UE can transmit small PUSCHs in several repetitions without feedback scheduled by an UL grant or RRC in the consecutive available slots. This method is called PUSCH repetition Type A. Each slot contains only one repetition and the time domain for the repetitions of a transport block is the same in those slots as shown in Fig. 6.

However, PUSCH repetition Type A causes big time gap among the repetitions and makes the system unable to achieve URLLC latency requirement. Therefore, in Release 16, PUSCH repetition Type B in Fig. 5 is developed to eliminate time gap among repetitions and ensures the configured number of repetitions in the time constraint because the repetitions are carried out in the consecutive sub-slots so one slot might contain more than one repetition of a transport block.

For PUSCH repetition Type B, the time domain resource is indicated by the gNB for the first "nominal" repetition while the resources for the remaining repetitions are derived based at least on the resources for the first repetition and UL/DL direction of symbols. The dynamic indication of the number of nominal repetitions for dynamic grant is jointly coded with start and length indicator of PUSCH in TDRA table by adding an additional column for the number of repetitions in the TDRA table. For CG PUSCH transmission, if the number of repetitions is not included in the TDRA table, it is provided by RRC parameter $repK$. If a "nominal" repetition goes across the slot boundary, invalid symbols or DL/UL switching point as in Fig. 5, this "nominal" repetition is split at the slot boundary or the switching point between UL symbols and DL/invalid symbols into multiple PUSCH repetitions. Therefore, the actual number of repetitions can be larger than the nominal number.

E. ENHANCED INTER-UE MULTIPLEXING IN UL TRANSMISSION

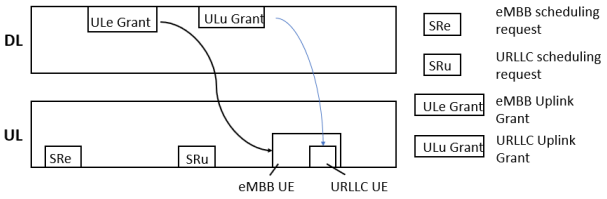

FIGURE 7. A collision of UL DG URLLC transmission with DG eMBB transmission.

To increase spectrum efficiency, latency critical communication service type and non-latency critical communication service type transmission of different UE are multiplexed in UL transmission so the gNB needs a mechanism to handle the collision and multiplexing of UL transmissions with different priorities such as the collision between LP DG eMBB and HP DG URLLC transmissions in Fig. 7. First, after receiving SR from an eMBB UE, the gNB schedules UL resources to the eMBB UE to transmit data. After that, another URLLC UE also sends a SR to ask for UL resources. Due to stringent latency requirement of URLLC transmission, if no resources are available in the latency budget, the gNB must schedule the URLLC transmission over the eMBB transmission's re-

sources that causes a collision between the transmission of two UEs.

However, no mechanism exists in Release 15 to solve this problem. Therefore, in Release 16, 3GPP supports UL cancellation indication (CI) and enhanced UL power control to handle the multiplexing between LP DG eMBB and HP DG URLLC transmissions.

1) UL cancellation indication

When the gNB allocates resources scheduled to the eMBB transmissions to another URLLC UE because of a strict latency requirement, it also transmits an UL CI as a group common DCI to the eMBB UEs in the group to ask them to stop their transmissions without resuming in the non-overlapping scheduled symbols. However, only sounding reference signal and PUSCH can be cancelled by UL CI. In case of PUSCH repetitions, UL CI is applied individually to each repetition overlapping the resource indicated by UL CI. The UE monitors UL CI in one occasion per slot or per span of 2, 4 and 7 symbols

The time and frequency resource for cancellation is jointly indicated in UL CI by a 2D-bitmap. In 2D-bitmap, time domain of the overlapping regions is divided into 1, 2, 4, 7, 14 or 28 partitions mapping to the corresponding number of bits. In time duplex division configuration, the DL symbols are excluded when the partitions of reference time region are chosen. The number of partitions in frequency domain of the overlapping regions is the division of the total number of indication bits and the number of bits indicating time domain. Each bit is used to to indicate whether a time-frequency partition is punctured or not.

2) Enhanced UL power control

Besides using UL CI in eMBB and URLLC multiplexing, the gNB has a second option by using power control scheme. The URLLC UE is indicated to increase power level of its PUSCH transmission which improves its decoding probability despite an overlap with an eMBB transmission of another UE. It helps the URLLC UE operate in a higher signal to noise ratio and compensates the effect from the interference of the eMBB transmission. For DG PUSCH, open-loop parameter set in Open loop power control set indication field of UL grant DCI is supported to control transmission power. One or two bits in UL grant are used to indicate whether a low or high power level in the open loop power control parameter set is used. However, power boosting is not applicable to the power limited UEs.

Fig. 8 compares the performance of URLLC packet detection by Demodulation reference signal (DMRS) detection in case of eMBB and URLLC transmission collision in three scenarios: no mechanism applied as in Release 15, using UL CI in Release 16 and using UL power control in Release 16. For each URLLC packet detection, the correlation result between the received DMRS and the known sequence is compared with a threshold based on target false alarm rate to determine whether the packet exists or not. In case of UL power control, transmission power of URLLC UE increases by 1dB compared to the other scenarios. As can be seen in Fig. 8, the performance of URLLC packet detection is improved by using CI or power control at the URLLC UE.

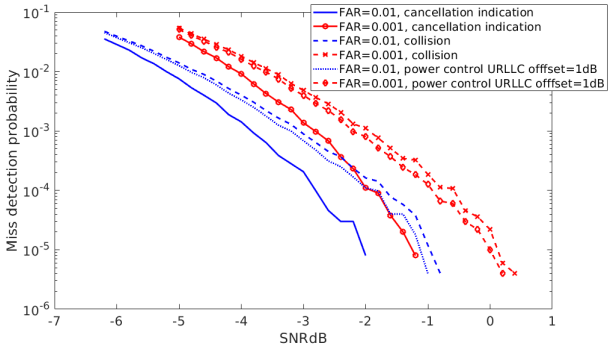

FIGURE 8. Performance of packet detection in Release 15 and 16 schemes.

F. ENHANCED UL CG TRANSMISSION

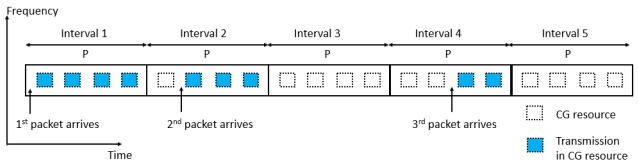

FIGURE 9. Less than K repetitions in CG UL transmission.

In Release 15, the UE is able to transmit blindly CG repetitions without feedback from the gNB. However, the UE is only allowed to transmit the repetitions in one HARQ process interval to avoid the confusion between the initial transmission and the retransmissions at the gNB. If the gNB misses the first transmission and only detects the retransmissions in a different HARQ process to that of the first transmission, the gNB will use the wrong UE HARQ identity in the UL grant to schedule a retransmission. Due to this constraint, the UE must stop to carry out the repetitions if it reaches the boundary of a HARQ process even if it still has not transmitted all repetitions configured as the second and the third packet in Fig. 9 where the UE is configured to transmit 4 repetitions.

In Release 16, to solve this problem, multiple active CG configurations for a given bandwidth part of a serving cell is supported. The number of CG configurations that a UE has is configured by RRC related to logical channel configuration with maximum 12 configurations per bandwidth part. The UE chooses the configuration with the earliest starting point to transmit data so that data is always transmitted at the beginning of a HARQ process interval and all configured repetitions are transmitted before reaching the HARQ process boundary as shown in in Fig. 10 for the case of four active configurations. One UE might have multiple configurations and one configuration might be shared among several UEs. Multiple CG configurations are also used to serve different traffic types at the UE.

TABLE 2. Performance comparison of different schemes at $SNR = -3.9dB$

Case	Scheme	Starting time offset (ms)	Number of repetitions	Error probability
Packet comes between the 1 st CG occasion and the 2 nd CG occasion	Release 15 scheme	0	3	$10^{-4.5}$
	Release 15 scheme with the UE waiting the next period	0.75	1	$10^{-1.5}$
	Release 16 scheme with multiple configurations	0	4	10^{-6}
Packet comes between the 2 nd CG occasion and the 3 rd CG occasion	Release 15 scheme	0	2	10^{-3}
	Release 15 scheme with the UE waiting the next period	0.5	2	10^{-3}
	Release 16 scheme with multiple configurations	0	4	10^{-6}
Packet comes between the 3 rd CG occasion and the 4 th CG occasion	Release 15 scheme	0	1	$10^{-1.5}$
	Release 15 scheme with the UE waiting the next period	0.25	3	$10^{-4.5}$
	Release 16 scheme with multiple configurations	0	4	10^{-6}

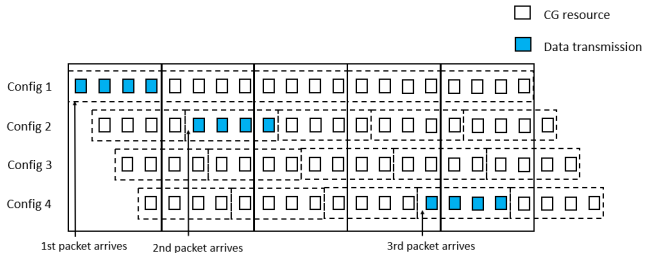

FIGURE 10. Multiple configurations to ensure K repetitions.

The gNB sends RRC or DCI to make the UE activate or release the configurations. In activation of the configurations, only separate activation is allowed. Each configuration is activated by a separate DCI. However, in release of the active configurations, both separate release and joint release are allowed. The gNB sends the Release DCI to indicate whether a single configuration or multiple configurations are released.

The benefit of Release 16 multiple configurations in the enhancement of PUSCH repetition's performance is shown in Table 2. Thanks to multiple configurations, it is ensured that the UE transmits all repetitions as configured so transmission error probability is smaller. In the simulation, with subcarrier spacing of 60 kHz, 4 slots spread in 1 ms. The configured number of repetition are 4. 4 repetitions are carried out in 4 slots equal to one HARQ process. Each repetition of 160 data bits is encoded by low-density parity-check code with MCS1 and quadrature phase shift keying modulation and transmitted in additive white Gaussian noise channel.

IV. URLLC ENHANCEMENTS IN RELEASE 17

In Release 15 and 16, the operation of URLLC is specified to operate only in licensed spectrum. However, due to an increase of demand for data transmission in 5G, unlicensed spectrum becomes a complement to URLLC operation in licensed spectrum because of availability and low cost of bandwidth. One important use case is the industrial automation in controlled environments with restricted access. The features of transmission in unlicensed spectrum have been specified since Release 13. However, the features of unlicensed spectrum do not take into account the features of URLLC specified in Release 15 and 16. This incompatibility requires the work in the ongoing Release 17 to harmonize the features of unlicensed spectrum and URLLC so that URLLC can operate in unlicensed spectrum and still attains the latency and reliability requirements. Section IV-A presents the potential research directions of Release 17 for URLLC in unlicensed spectrum.

The work of Release 15 and 16 for URLLC in licensed spectrum is also continued in the ongoing Release 17 to further improve URLLC performance. The objectives of Release 17 in licensed spectrum are presented in Section IV-B, Section IV-C and Section IV-D.

A. URLLC ENHANCEMENTS IN UNLICENSED SPECTRUM

1) Harmonize PUSCH repetitions in URLLC and unlicensed spectrum

PUSCH repetition scheme in unlicensed spectrum specified in Release 16 is similar to PUSCH repetition Type B in URLLC as presented in Section III-D. However, there is an importance difference where segmentation of PUSCH repetition due to slot boundary and DL/invalid symbols is not supported in unlicensed spectrum. PUSCH repetition is dropped if it collides with slot boundary and DL/ invalid symbols. This creates the gap between PUSCH repetitions in unlicensed spectrum while PUSCH repetition Type B in URLLC supports back-to-back repetitions. Therefore, this reduces scheduling flexibility, transmission's reliability and increases latency. On the other hand, PUSCH repetition scheme in unlicensed spectrum has a benefit when resource

for PUSCH repetitions can be repeated in several consecutive slots and indicated by RRC parameter. This gives the UE more opportunities to schedule multiple transport blocks by a single DCI.

One potential way to combine beneficial benefits of both schemes in unlicensed spectrum and URLLC is that the number of resources repeated across the slots is determined following the scheme of unlicensed spectrum while the start symbol, the segmentation, back-to-back repetitions and the number of repetitions follow PUSCH repetition Type B in URLLC.

2) Harmonize feedback for CG PUSCH transmission in URLLC and unlicensed spectrum

In Release 16, feedback for CG PUSCH transmission of URLLC follows timer-based feedback. If the gNB decodes correctly PUSCH, it does not send ACK feedback to reduce overhead when the gNB must not send ACK most of the time due to high reliability of URLLC transmission. The UE waits until the end of timer and if no feedback is received, the UE assumes a successful transmission. If the gNB fails to decode PUSCH, it sends an UL grant to schedule a retransmission in the scheduled resource.

In contrast, feedback for CG PUSCH transmission in unlicensed spectrum follows explicit-ACK feedback in Release 16 to counter the uncertainty of channel access in unlicensed spectrum. If the gNB decodes correctly PUSCH, it sends ACK. If the gNB fails to decode PUSCH, it sends an UL grant or NACK to schedule a retransmission in the scheduled resource. If the gNB cannot access to the channel to transmit ACK, NACK or UL grant, the UE does not receive any signal from the gNB and waits until the end of timer to retransmit data automatically in the CG resources.

Due to different feedback schemes in URLLC and unlicensed spectrum, a feedback scheme must be decided so that URLLC can operate in unlicensed spectrum. This scheme can be chosen from two existing schemes in Release 16. The gNB then tells the UE to use timer-based feedback in case of URLLC transmission in unlicensed spectrum. Otherwise, the UE uses explicit-ACK feedback in unlicensed spectrum for other types of transmission. The feedback scheme for URLLC in unlicensed spectrum might also be a new scheme that combines the benefits of two existing schemes.

3) Frame based equipment (FBE) enhancements

In unlicensed spectrum, a transmitter is required to do Listen before talk (LBT) through the channel access mechanisms to access to the channel and transmit data in the duration of channel occupancy time (COT). One of the channel access mechanisms is FBE where the transmitter is allowed to do LBT in the fixed moments. The periodicity between two consecutive LBT moments is a fixed frame period (FFP) from 1ms to 10ms. In Release 16, only the gNB is allowed to initiate a COT by doing LBT in the fixed moments. After obtaining the channel, the gNB might share the COT to the UE so that it can transmit the UL transmission. This may cause long latency in UL transmission due to two reasons. First, if LBT fails, the gNB must wait from 1ms to 10ms to do LBT in the next moment. In that interval, the UE also cannot start its UL transmission because no COT is initiated by the gNB. Second, if the gNB has no DL data to transmit, it does not initiate a COT. If the UE has UL data at that time, it also cannot transmit because of the absence of the gNB-initiated COT. Therefore, to reduce latency and support URLLC in unlicensed spectrum, in Release 17, the UE is allowed to initiate its own COT to transmit UL data.

The configuration of the UE's FFP should avoid blocking the gNB to initiate its own COT. It can be done by configuring offset and periodicity of the UE's FFP to be different from that of the gNB's FFP. Another problem is to make the UE choose to transmit PUSCH in the gNB-initiated COT or the UE-initiated COT. It can be indicated by the gNB through UL grant or RRC when it schedules PUSCH transmission. Another way is to preconfigure the UE by a rule to determine between the gNB-initiated COT and the UE-initiated COT based on some criteria such as transmission's priority, the PUSCH location in correlation to the beginning the gNB' FFP or the UE's FFP. Subsequently, the UE needs a mechanism to indicate to the gNB whether PUSCH transmission is in the gNB-initiated COT or the UE-initiated COT. This indication might be carried in an UCI multiplexed with PUSCH and some bits embedded in PUSCH. Furthermore, the UE also needs to indicate whether the UE's COT is shared with the gNB so that the gNB can transmit DL data in that COT to the initiating UE. These enhancements need to be included in Release 17 to allow the UE to initiate its own COT.

B. PHYSICAL LAYER FEEDBACK ENHANCEMENTS

In Release 17, HARQ feedback for DL SPS transmission needs an improvement to work in Time division duplex (TDD) configuration. In DL SPS transmission, the gNB transmits PDSCH to the UE in the pre-configured resources without an associated PDCCH so time from the end of PDSCH to the beginning of HARQ feedback is set since the SPS resources are activated. However, in TDD configuration, the configured time might point feedback to the DL slot so the feedback is dropped as illustrated in Fig. 11. The UE is configured to transmit HARQ feedback three slots after a SPS PDSCH transmission. This value is used from the activation to the release of a SPS configuration. The gNB cannot predict slot format of all slots in advance so HARQ feedback pointed to a DL slot is cancelled. It causes a degradation of URLLC transmission as shown in Fig. 12 because the gNB does not have information to trigger a retransmission if necessary. In this simulation, a packet of 160 bits is encoded with MCS2 in the MCS table by low-density parity-check code then transmitted in DL additive white Gaussian noise channel. The feedback is assumed to be dropped with the probabilities of 1% and 5% due to DL slot in TDD configuration. When NACK feedback is dropped, there is no retransmisson and the packet is not decoded correctly by the UE causing an error.

When NACK is transmitted, the gNB retransmits the packet with MCS1 to increase reliability.

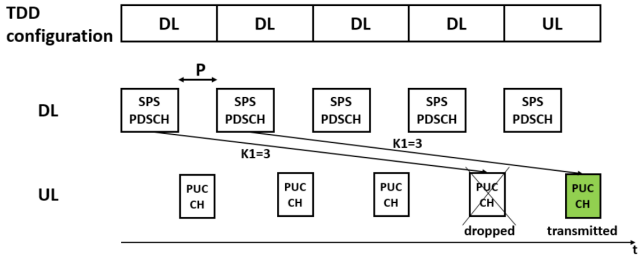

FIGURE 11. DL SPS transmission's HARQ feedback cancellation DL SPS transmission in TDD.

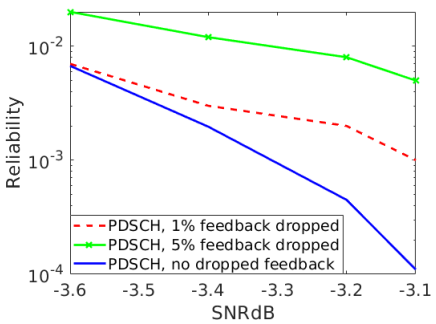

FIGURE 12. SPS PDSCH transmission's reliability in latency of 1ms with dropped feedback and no dropped feedback.

In Fig. 12, if SPS PDSCH's feedback is not dropped in the conflicting slots, reliability of SPS PDSCH is improved. Thereby, to avoid the drop of SPS PDSCH's feedback, several techniques are the candidates to be included in Release 17. The first technique is proposed to defer the dropped feedback in the conflict slot to the next available UL slot. For example, in Fig. 11, with the first technique, the feedback in the DL slot is not dropped but deferred to the next PUCCH resource in the UL slot so the feedback of two SPS transmissions is multiplexed in one PUCCH resource. The multiplexing of two feedback might decrease feedback's reliability. The second technique uses multiple values of time from the end of PDSCH to the beginning of HARQ feedback in the activation DCI so the UE can choose the most appropriate value based on slot format. For example, in Fig. 11, with the second technique, instead of K1 only being 3, the UE is configured with K1 to be 3 or 4. K1 being 3 points the feedback of the first SPS PDSCH to a DL slot while K1 being 4 points to an UL slot. Therefore, the UE chooses K1 being 4 to transmit the feedback of the first SPS PDSCH. One drawback of this technique is that multiple K1 values increase DCI length that reduces its reliability. In the third technique, the K1 value is indicated dynamically in each SPS occasion by RRC or the embedded bits in SPS PDSCH. For example, in Fig. 11, with the third technique, instead of K1 being fixed to be 3 in all SPS PDSCHs, each SPS PDSCH has its own value of K1 indicated by the bits in SPS PDSCH or RRC to avoid the DL slot. However, the dynamic signal in each SPS occasion causes an overhead in the system. Each technique has its own benefits and drawbacks that need to be analyzed in Release 17.

Besides the problem of dropped feedback, there exists another problem of HARQ feedback for SPS transmission. In Release 16, even if there is no PDSCH transmission in the SPS resource, the UE is still required to send NACK. The URLLC packets have a low arrival rate so the NACK transmission in the empty SPS occasions might cause a waste of resources and interference with other UE. On the other hand, URLLC transmission has high reliability so there is high probability that SPS PDSCH is decoded correctly. Most of the time the UE sends ACK for SPS PDSCH transmission that also leads to resource consumption and interference. Skipping ACK or NACK scheme in SPS transmission should be considered to reduce resource consumption and interference in Release 17.

In Release 16, PUCCH repetitions are done in slot level where there is only one PUCCH repetition per slot and PUCCH repetitions cannot cross slot boundary. The reliability of PUCCH can be enhanced by allowing PUCCH repetitions in sub-slot level as PUSCH repetition Type B in Release 17. There are more PUCCH repetitions allowed in URLLC latency constraint and a long PUCCH can also be segmented to small PUCCH repetitions to cross slot boundary.

Channel state information (CSI) feedback helps the gNB make the optimal scheduling decisions and is conducive to URLLC traffic types with sporadic traffic burst. Due to URLLC latency, latency of CSI feedback must also be taken into account in new URLLC features of Release 17. First, new schemes are required to trigger aperiodic CSI with lower latency. Second, CSI computation should be reduced to capture more accurate channel fading and interference.

C. INTRA-UE MULTIPLEXING

In Release 16, only UCI prioritization based on a two-level priority is standardized where the LP UCI is cancelled by the HP UCI when they overlap. In Release 17, multiplexing of UCI such as HARQ-ACK and SR on PUCCH with different priorities must be supported. In multiplexing, the target code rate and latency of the HP UCI could be guaranteed by using separate coding where two code rates for the HP UCI and the LP UCI are used based on their original PUCCH resources. The HP UCI is mapped to the multiplexing PUCCH before the LP UCI to guarantee the resource for the HP UCI. With separate coding, latency of the HP UCI decoding is also reduced because the gNB can start the decoding process after receiving the symbols in the HP UCI's resources instead of all symbols of the multiplexing PUCCH. Moreover, the multiplexing PUCCH should end no later than the PUCCH carrying the HP UCI.

Besides multiplexing of UCI on PUCCH, Release 17 also supports UCI multiplexing on PUSCH with different priorities. Similar to UCI multiplexing on PUCCH, separate coding also should be used in UCI multiplexing on PUSCH to guarantee the target code rate and latency of the

HP UCI/PUSCH. Furthermore, the ending symbol of the LP PUSCH should be no later than the ending symbol of PUCCH carrying the HP HARQ-ACK.

D. ENHANCEMENTS FOR SUPPORT OF TIME SYNCHRONIZATION

In Release 17, time synchronization requirements are defined in [10] where time synchronization budget (the time error contribution between ingress and egress of the 5G system on the path of clock synchronization messages) is set to 900 ns. The flow of clock synchronization messages traverses the air interface twice so the synchronization budget between Uu interface (the radio interface between the UE and the radio access network) should not exceed 450 ns. This time accuracy is affected by time alignment error at the gNB (requirements specified in [11]), timing error at the UE (requirements specified in [12]) and time delay caused by propagation delay. The UE estimates the downlink propagation delay to synchronize with the gNB as half of the timing advance (TA) value obtained from the gNB. However, TA measurement procedure, quantizations involved and additional errors leave residual error despite the TA based time compensation. The detailed analysis from [13] shows that despite Release 16 based TA compensation, time synchronization error between Uu interface may well exceed 450 ns. Therefore, improved propagation delay compensation is necessary to achieve time synchronization requirements in Release 17. There are two options listed for Release 17 based enhanced propagation delay compensation that need further study and analysis: TA-based propagation delay where a finer TA indication granularity is used for propagation delay estimation and round trip time (RTT)-based propagation delay where propagation delay estimation is based on a managed reception-transmission procedure.

V. CONCLUSION

URLLC has been specified as one of the three key services of 5G New Radio. In order to satisfy URLLC requirements, many new techniques in physical layer design have been specified. This article has described the features for URLLC in Release 15: new numerology with flexible SCS, new CQI and MCS tables, UL and DL transmissions at sub-slot level, preemption indication in DL eMBB and URLLC multiplexing, UL CG transmissions with automatic repetitions. Subsequently, the evolution of URLLC in Release 16 is analyzed with new features to improve URLLC performance in new use cases: increasing PDCCH monitoring capability, new DCI with the configurable number of bits in the fields, SPS enhancements, sub-slot PUCCH transmission, UCI intra-UE-multiplexing, PUSCH Repetition Type B with back-to-back repetitions, CI and power control in UL inter-UE multiplexing, multiple CG configurations. These features improve the performance of URLLC transmissions and serve as a bridge leading to the evolution of URLLC. In the current Release 17, the features of URLLC in both licensed and unlicensed spectrum are being standardized. Potential directions in URLLC research such as enhancements for unlicensed band URLLC, feedback enhancements, intra-UE multiplexing and prioritization of traffic with different priority and enhancements for support of time synchronization are discussed with the promising candidate techniques to be included in the next releases.

ACKNOWLEDGMENT

This work was supported in part by TCL and H2020 project 5GENESIS (5genesis.eu).

REFERENCES

[1] 3GPP TR 38.913 v15.0.0, “Study on scenarios and requirements for next generation access technologies.”
[2] Huawei, HiSilicon, Nokia, Nokia Shanghai Bell, “New SID on Physical Layer Enhancements for NR URLLC”, 3GPP RP-182089, TSG-RAN#81, Gold Coast, Australia, Sept 10–13, 2018.
[3] 3GPP TR 38.802 v14.2.0, “Study on New Radio Access Technology Physical Layer Aspects.”
[4] Huawei, HiSilicon, “Physical Layer Enhancements for NR Ultra-Reliable and Low Latency Communication (URLLC)”, 3GPP RP-191584, TSG-RAN#84, Newport Beach, CA, June 3–6, 2019.
[5] Nokia, Nokia Shanghai Bell, “Support of NR Industrial Internet of Things (IoT)”, 3GPP RP-192590, TSG-RAN#85, Newport Beach, USA, September 16–20, 2019.
[6] Nokia, Nokia Shanghai Bell, “Support of NR Industrial Internet of Things (IoT)”, 3GPP RP-201310, TSG-RAN#88e, June 29–July 3, 2020.
[7] 3GPP TS 38.211 v16.3.0, “Physical channels and modulation.”
[8] 3GPP TS 38.213 v16.3.0, “Physical layer procedures for control.”
[9] 3GPP TS 38.214 v16.3.0, “Physical layer procedures for data.”
[10] 3GPP TS 22.104 v17.4.0, “Service requirements for cyber-physical control applications in vertical domains.”
[11] 3GPP TS 38.104 v15.3.0, “Base Station (BS) radio transmission and reception.”
[12] 3GPP TS 38.133 v15.3.0, “ Requirements for support of radio resource management.”
[13] Nokia, Nokia Shanghai Bell, “Discussion on RAN1 involvement in propagation delay compensation”, 3GPP R1-2006341, TSG-RAN#102e, August 17–28, 2020.
[14] Chairman’s notes RAN1 AH1901, Taipei, January 21–25,2019.
[15] Chairman’s notes RAN1 #96, Athens, Greece, February 25–March 1,2019.
[16] Chairman’s notes RAN1 #96bis, Xi’an, China, April 8–12,2019.
[17] Chairman’s notes RAN1 #97, Reno, USA, May 13–17,2019.
[18] Chairman’s notes RAN1 #98, Pargue, CZ, August 26–30,2019.
[19] Chairman’s notes RAN1 #98b, Chongquing, China, October 14–18,2019.
[20] Chairman’s notes RAN1 #99, Reno, USA, November 18–22,2019.
[21] Chairman’s notes RAN1 #100, March 24–28,2020.
[22] Chairman’s notes RAN1 #100b, April 20–30,2020.
[23] Chairman’s notes RAN1 #101, May 15 – June 5,2020.
[24] Chairman’s notes RAN1 #102, August 17–28,2020.
[25] Chairman’s notes RAN1 #103, October 26 – November 13,2020.
[26] C. She, C. Yang, and T. Q. S. Quek, “Cross-layer optimization for ultra-reliable and low-latency radio access networks,” IEEE Trans. Wireless Commun., vol. 17, no. 1, pp. 127-141, Jan. 2018.
[27] G. Pocovi, H. Shariatmadari, G. Berardinelli, K. Pedersen, J. Steiner and Z. Li, “Achieving Ultra-Reliable Low-Latency Communications: Challenges and Envisioned System Enhancements,” in IEEE Network, vol. 32, no. 2, pp. 8-15, March-April 2018.
[28] G. Berardinelli et al., "Reliability Analysis of Uplink Grant-Free Transmission Over Shared Resources," in IEEE Access, vol. 6, pp. 23602-23611, April 2018.
[29] Ghosh. A, “5G New Radio (NR): Physical Layer Overview and Performance,” IEEE Communication Theory Workshop, May 2018.
[30] H. Ji, S. Park, J. Yeo, Y. Kim, J. Lee and B. Shim, “Ultra-Reliable and Low-Latency Communications in 5G Downlink: Physical Layer Aspects,” in IEEE Wireless Communications, vol. 25, no. 3, pp. 124-130, June 2018.
[31] H. Shariatmadari, S. Iraji, R. Jantti, P. Popovski, Z. Li and M. A. Uusitalo, “Fifth-Generation Control Channel Design: Achieving Ultrareliable Low-Latency Communications,” in IEEE Vehicular Technology Magazine, vol. 13, no. 2, pp. 84-93, June 2018.

[32] H. Shariatmadari, R. Duan, S. Iraji, R. Jantti, Z. Li and M. A. Uusitalo, "Asymmetric ACK/NACK Detection for Ultra - Reliable Low - Latency Communications," 2018 European Conference on Networks and Communications (EuCNC), Ljubljana, Slovenia, June 2018, pp. 1-166.

[33] M. Bennis, M. Debbah and H. V. Poor, "Ultra-reliable and Low-Latency Wireless Communication: Tail, Risk, and Scale," in Proceedings of the IEEE, vol. 106, no. 10, pp. 1834-1853, Oct. 2018.

[34] R. M. Cuevas, C. Rosa, F. Frederiksen and K. I. Pedersen, "Uplink Ultra-Reliable Low Latency Communications Assessment in Unlicensed Spectrum," 2018 IEEE Globecom Workshops (GC Wkshps), Abu Dhabi, United Arab Emirates, December 2018, pp. 1-6.

[35] H. Chen, R. Abbas, P. Cheng, M. Shirvanimoghaddam, W. Hardjawana, W. Bao, Y. Li, and B. Vucetic, "Ultra-reliable low latency cellular networks: Use cases, challenges and approaches," IEEE Commun. Mag., vol. 56, no. 12, pp. 119-125, December 2018.

[36] J. Thota and A. Aijaz, "On performance evaluation of random access enhancements for 5G uRLLC," in Proc. IEEE Wireless Commun. Netw. Conf. (WCNC), Apr. 2019, pp. 1-7.

[37] A. Z. Hindi, S. Elayoubi and T. Chahed, "Performance Evaluation of Ultra-Reliable Low-Latency Communication Over Unlicensed Spectrum," ICC 2019 - 2019 IEEE International Conference on Communications (ICC), Shanghai, China, May 2019, pp. 1-7.

[38] T. Le, U. Salim and F. Kaltenberger, "Optimal reserved resources to ensure the repetitions in Ultra-Reliable Low-Latency Communication Uplink Grant-free transmission," 2019 European Conference on Networks and Communications (EuCNC), Valencia, Spain, June 2019.

[39] D. Feng, C. She, K. Ying, L. Lai, Z. Hou, T. Q. S. Quek, Y. Li, and B. Vucetic, "Toward Ultra-reliable Low-Latency Communications: Typical Scenarios, Possible Solutions, and Open Issues," in IEEE Vehicular Technology Magazine, vol. 14, no. 2, pp. 94-102, June 2019.

[40] T. Le, U. Salim and F. Kaltenberger, "Strategies to meet the configured repetitions in URLLC Uplink Grant-Free transmission," 2019 16th International Symposium on Wireless Communication Systems (ISWCS), Oulu, Finland, August 2019.

[41] G. J. Sutton et al., "Enabling Technologies for Ultra-Reliable and Low Latency Communications: From PHY and MAC Layer Perspectives," in IEEE Communications Surveys & Tutorials, vol. 21, no. 3, pp. 2488-2524, third quarter 2019.

[42] T. Le, U. Salim and F. Kaltenberger, "Improving Ultra-Reliable Low-Latency Communication in multiplexing with Enhanced Mobile Broadband in grant-free resources," 2019 IEEE 30th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC), Istanbul, Turkey, September 2019.

[43] Z. Ma, M. Xiao, Y. Xiao, Z. Pang, H. V. Poor and B. Vucetic, "High-Reliability and Low-Latency Wireless Communication for Internet of Things: Challenges, Fundamentals, and Enabling Technologies," in IEEE Internet of Things Journal, vol. 6, no. 5, pp. 7946-7970, Oct. 2019.

[44] T. Le, U. Salim and F. Kaltenberger, "Control and data channel combining in Ultra-Reliable Low-Latency Communication," 2019 53rd Asilomar Conference on Signals, Systems, and Computers, Pacific Grove, CA, USA, November 2019.

[45] A. Anand, G. de Veciana and S. Shakkottai, "Joint Scheduling of URLLC and eMBB Traffic in 5G Wireless Networks," in IEEE/ACM Transactions on Networking, vol. 28, no. 2, pp. 477-490, April 2020.

[46] T. Le, U. Salim and F. Kaltenberger, "Feedback Enhancements for Semi-Persistent Downlink Transmissions in Ultra-Reliable Low-Latency Communication," 2020 European Conference on Networks and Communications (EuCNC), Dubrovnik, Croatia, June 2020.

[47] Y. Liu, Y. Deng, M. Elkashlan, A. Nallanathan, and G. K. Karagiannidis, "Analyzing grant-free access for URLLC service," 2020, https://arxiv.org/abs/2002.07842.

[48] J. Park, S. Samarakoon, H. Shiri, M. K. Abdel-Aziz, T. Nishio, A. Elgabli, M. Bennis,"Extreme URLLC: Vision, Challenges, and Key Enablers", 2020, https://arxiv.org/abs/2001.09683.

[49] C. She, C. Sun, Z. Gu, Y. Li, C. Yang, H. V. Poor, B. Vucetic, "A Tutorial of Ultra-Reliable and Low-Latency Communications in 6G: Integrating Theoretical Knowledge into Deep Learning", 2020, https://arxiv.org/abs/2009.06010.

[50] T. Le, U. Salim and F. Kaltenberger, "Channel Access Enhancements in unlicensed spectrum for NR URLLC transmissions", 2020 GLOBECOM, December 2020.

APPENDIX: LIST OF ABBREVIATIONS

3GPP	Third Generation Partnership Project
5G	5th generation
ACK	Acknowledgment
AL	Aggregation level
CCE	Control channel element
CG	Configured grant
CI	Cancellation indication
COT	Channel occupancy time
CQI	Channel quality indicator
CSI	Channel state information
DCI	Downlink control information
DG	Dynamic grant
DL	Downlink
DMRS	Demodulation reference signal
eMBB	Enhanced mobile broadband
FBE	Frame based equipment
FFP	Fixed frame period
gNB	5G base station
HARQ	Hybrid automatic repeat request
HP	High priority
LBT	Listen before talk
LP	Low priority
LTE	Long-Term Evolution
MCS	Modulation and coding scheme
NACK	Negative acknowledgment
OFDM	Orthogonal frequency division multiplexing
PDCCH	Physical downlink control channel
PDSCH	Physical downlink shared channel
PUCCH	Physical uplink control channel
PUSCH	Physical uplink shared channel
RAN	Radio access network
RRC	Radio resource control
RTT	Round trip time
SCS	Sub-carrier spacing
SPS	Semi-persistent scheduling
SR	Scheduling request
TA	Timing advance
TDD	Time division duplex
TDRA	Time domain resource assignment
UCI	Uplink control information
UE	User equipment
UL	Uplink
UL-SCH	Uplink shared channel
URLLC	Ultra-reliable low-latency communication

TRUNG-KIEN LE received MCS. degree specializing in Mobile Computing Systems from EURECOM, France. Currently, he is a PHD student in Sorbonne University and EURECOM. His research interest is Ultra Reliable Low Latency Communication in 5G New Radio.

UMER SALIM received his Ph.D. and M.S. degrees, specializing in communication theory and signal processing from Eurecom and Supelec, France, respectively. He is currently working at TCL Communications as 5G Systems Architect and 3GPP RAN1 delegate for 5G standardization activity. Before joining TCL, he worked at Intel Mobile Communications designing modems for high end smart phones and tablets. He has many years of research experience and holds several patents in physical layer of wireless communications.

FLORIAN KALTENBERGER (S05-M08) received his Dipl.-Ing. degree and his Ph.D. degree both in technical mathematics from the Vienna University of Technology in 2002 and 2007, respectively. He is an assistant professor at the Communication Systems Department of Eurecom, Sophia-Antipolis, France, and part of the management team of the real time open-source 5G platform OpenAirInterface.org. From 2003 to 2007 he was with the Wireless Communications Group of the Austrian Research Centers, where he was developing a real-time MIMO channel emulator. His research interests include 5G and MIMO systems at large, software defined radio, signal processing for wireless communications, as well as channel modeling and simulation.