Physical layer design for Ultra-Reliable Low-Latency Communications in 5G

Dissertation
submitted to Sorbonne Université in partial fulfillment of the requirements
for the degree of Doctor of Philosophy

Author:

LE Trung Kien

Reviewers

Prof.	**Petar POPOVSKI**	**Aalborg University, Denmark**
Dr.	**Raphael VISOZ**	**Orange Labs, France**

Examiners

Prof.	**David GESBERT**	**EURECOM, France**
Prof.	**Ingrid MOERMAN**	**Ghent University, Belgium**
Prof.	**Merouanne DEBBAH**	**CentraleSupélec, France**

Director of Thesis

Asst. Prof.	**Florian KALTENBERGER**	**EURECOM, France**

Co-Director of Thesis

Dr.	**Umer SALIM**	**InterDigital, France**

La conception de la couche physique pour “Ultra-Reliable Low-Latency Communications” en 5G

Thèse
soumise à Sorbonne Université pour l'obtention du Grade de Docteur

Auteur:

LE Trung Kien

Rapporteurs		
Prof.	**Petar POPOVSKI**	**Aalborg University, Denmark**
Dr.	**Raphael VISOZ**	**Orange Labs, France**
Examinateurs		
Prof.	**David GESBERT**	**EURECOM, France**
Prof.	**Ingrid MOERMAN**	**Ghent University, Belgium**
Prof.	**Merouanne DEBBAH**	**CentraleSupélec, France**
Directeur de Thèse		
Asst. Prof.	**Florian KALTENBERGER**	**EURECOM, France**
Co-Directeur de Thèse		
Dr.	**Umer SALIM**	**InterDigital, France**

Abstract

The advent of new use cases and new applications such as augmented/virtual reality, industrial automation, autonomous vehicles, etc. in 5G has made the Third Generation Partnership Project (3GPP) specify Ultra-reliable low-latency communications (URLLC) as one of the service categories. To support URLLC with the strict requirements of reliability and latency, 3GPP Release 15 and Release 16 have specified the URLLC features in licensed spectrum. The ongoing 3GPP Release 17 extends the URLLC features to unlicensed spectrum to target the new use cases in the industrial scenario.

In the first part of the thesis from Chapter 2 to Chapter 4, we focus on the problems in URLLC physical layer design and the URLLC features in licensed spectrum.

The first study deals with the problem of ensuring the configured number of uplink (UL) configured-grant (CG) repetitions of a transport block. To that end, the thesis proposes three schemes: the use of reserved resources for the repetitions, the use of an explicit Hybrid automatic repeat request (HARQ) feedback structure and the use of an additional scheduling request.

Secondly, one key feature of 5G systems is the flexible structure being able to accommodate the three service paradigms. This may lead to the collision of an eMBB UL transmission of a user equipment (UE) and an URLLC UL transmission of another UE on the CG resources that causes a degradation of URLLC transmission's reliability. This study presents two schemes to enhance URLLC transmission's reliability in case of a collision. The first scheme consists of an overlap indication to trigger an explicit HARQ feedback structure at the URLLC UE in case of the collision. The second scheme uses an overlap indication to ask the URLLC UE to transmit an additional scheduling request if its transmission collides with an eMBB transmission of another UE.

Thirdly, for an efficient operation of periodic resource allocation, 5G systems have many flavors of semi-persistant resource allocation standardized. The focus of this study is the downlink (DL) transmission where the feedback of the DL semi-persistent scheduling transmission is dropped due to the conflict of the DL/UL symbols. This investigation resulted in proposing a dynamic indication of resources for feedback embedded in DL data so that resources for feedback are pointed to the UL symbols.

Following the work of URLLC design in licensed spectrum, we focus on URLLC operation in unlicensed spectrum in the second part from Chapter 5 to Chapter 8. In unlicensed spectrum, a 5G device is required to access to a channel by using load based equipment (LBE) where it does channel sensing whenever it has data to transmit or frame based equipment (FBE) where it only does channel sensing per fixed period. The uncertainty of obtaining channel access through LBE or FBE can impede the achievement of the URLLC latency requirements. Therefore, the study of impact of LBE and FBE on URLLC transmission and the enhancements of LBE and FBE are needed.

Firstly, we analyze the impact of LBE on URLLC performance by using Markov chain model to calculate transmission probability and channel access latency then introduce the new channel access tables to reduce channel access latency in LBE for URLLC transmission.

After that, we study the impact of FBE on URLLC performance. To improve the probability of channel

access in the URLLC latency constraint, we propose two schemes: the first scheme allows the transmitters to use multiple fixed frame period (FFP) configurations while the second scheme configures the FFP's starting point of each transmitter based on its priority.

Consequently, the models for LBE and FBE are combined to model a system where the LBE devices and the FBE devices coexist. The devices are allowed to switch dynamically from FBE to LBE to serve data with high priority such as URLLC or data with high arrival rate and from LBE to FBE to serve data with low priority such as eMBB or data with low arrival rate.

Finally, after a UE acquires a channel, its transmission might be interrupted due to the DL symbols in time division duplex configuration or an orphan symbol and the UE might have to do an additional channel access process to acquire the channel causing latency. The proposed schemes allow removing the gap by dynamically switch DL symbols to UL symbols or filling the orphan symbol with the signal such as Demodulation reference signal.

Acknowledgements

First of all, I would like to thank my supervisor Florian KALTENBERGER who has supported me for a long time since I did my master and my internship then now when I do my PHD. It would be impossible for me to have the achievements as today without his unconditional support. Subsequently, I would like to thank my co-supervisor Umer SALIM for his valuable ideas and support during the time of my PHD.

Besides, I also would like express my gratitude towards my former colleagues in TCL: Pierre BONNARD, Bruno JECHOUX, Sebastian WAGNER, Virgile GARCIA, Achraf KHSIBA, Hongzhi WANG who provided a perfect working environment and gave me the first working experience in industry.

Finally, I would like to thank my parents and my sister who have been with me in every moment of my life and given me the conditions so that I am able to succeed in my life.

Contents

List of Figures

List of Tables

Abbreviations

3GPP	Third Generation Partnership Project
5G	5th generation
ACK	Acknowledgment
AL	Aggregation level
AWGN	Additive white Gaussian noise
BLER	Block error rate
CCE	Control channel element
CG	Configured grant
CI	Cancellation indication
COT	Channel occupancy time
CQI	Channel quality indicator
CRC	Cyclic redundancy check
CSI	Channel state information
CS-RNTI	Configured scheduling-radio network temporary identifier
DCI	Downlink control information
DG	Dynamic grant
DL	Downlink
DMRS	Demodulation reference signal
eMBB	Enhanced mobile broadband
FAR	False alarm rate
FBE	Frame based equipment
FDD	Frequency Division Duplex
FFP	Fixed frame period
gNB	5G base station
HARQ	Hybrid automatic repeat request
HP	High priority
IoT	Internet of things
LBE	Load based equipment
LBT	Listen before talk
LDPC code	Low-density parity-check code
LP	Low priority
LTE	Long-Term Evolution
mMTC	Massive Machine-Type Communication
MCS	Modulation and coding scheme
NACK	Negative acknowledgment
NDI	New data indicator
OFDM	Orthogonal frequency division multiplexing
PDCCH	Physical downlink control channel
PDSCH	Physical downlink shared channel
PHICH	Physical HARQ Indicator Channel

PRB	physical resource block
PUCCH	Physical uplink control channel
PUSCH	Physical uplink shared channel
QPSK	Quadrature phase shift keying
QoS	Quality of service
RAN	Radio access network
RRC	Radio resource control
RTT	Round trip time
SC-RNTI	Single cell-radio network temporary identifier
SCL	Successive cancellation list
SCS	Sub-carrier spacing
SFI	Slot format indication
SIC	Successive interference cancellation
SINR	Signal-to-interference-plus-noise ratio
SPS	Semi-persistent scheduling
SR	Scheduling request
TA	Timing advance
TB	Transport block
TDD	Time division duplex
TDRA	Time domain resource assignment
TSC	Time sensitive communications
UCI	Uplink control information
UE	User equipment
UL	Uplink
UL-SCH	Uplink shared channel
URLLC	Ultra-reliable low-latency communications

Chapter 1

Introduction

1.1 5G New Radio overview

The emergence of new paradigms like connected self-driving cars, automated industrial control, augmented/virtual reality, etc. led the wireless standardization bodies to take these into account. To that respect, 3GPP has defined three service paradigms for 5G: Enhanced Mobile Broadband (eMBB), Massive Machine-Type Communication (mMTC) and Ultra-Reliable Low-Latency Communications (URLLC).

eMBB targets high data rate transmission with high requirements for bandwidth such as virtual reality, real time security, 3-dimensional image and 4K-resolution video streaming. The goal in new radio design is to increase system throughput.

mMTC aims to support a large number of low-power devices for a long life-time requiring highly energy efficient communication. They demand the improvements of latency, reliability, massive connection density and energy efficiency.

URLLC is for devices requiring low latency and high link reliability. There are a huge number of URLLC applications including:

- Smart grid: an electrical grid with smart meters that transmit electricity consumption from the clients to the providers in real time.
- Self-driving car: a car operates without the control from human being. It has the ability to detect the obstacles and make a decision to avoid them.
- Industrial automation: the factory is operated by the autonomous machines and robots to manufacture the products. The telling examples are motion control, factory automation and process automation. Motion control is about real-time control of machines with moving parts and isochronous transmission of sensory and actuation information in the uplink and downlink that requires latency below 1 ms. Factory automation requires ultra high reliability of 10^{-6} or even 10^{-9}. Process automation about production of goods in bulk quantities requires a reliability of 10^{-6}, E2E latency of 50 ms and jitter of 20 ms.
- Health care: remote surgery can be carried out by a surgeon in a different place with the help of robot that receives the instructions in real time.
- Augmented reality: the environment in real world is recreated by the computers with the support of audio, video and geographic information.
- V2X: a network supports the communication among vehicles and other objects.

- Tactile Internet: a network ensures the tactile sensing with a common feature being the existence of haptic feedback that requires strict latency and reliability requirements.

They are real time applications that require the immediate actions so time requirement is very strict being in the order of ten-hundreds of microseconds.

Among these three service categories, URLLC raises the most challenge because it has to deal with two requirements that have trade-off: reliability and latency. Basically, 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.

1.2 URLLC requirements

In [1], the term reliability is defined as follows: "Reliability can be evaluated by the success probability of transmitting X bytes within a certain delay, which is the time it takes to deliver a small data packet from the radio protocol layer 2/3 service data unit ingress point to the radio protocol layer 2/3 service data unit egress point of the radio interface, at a certain channel quality (e.g., coverage-edge)". In [2], The 3rd Generation Partnership Project (3GPP) defines URLLC requirements to support use cases such as smart grid, augmented and virtual reality in entertainment industry: "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". 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: factory automation, transport industry including the remote driving use case and electrical power distribution [8].

All URLLC standardized features in Release 15 and Release 16 are for the transmission in licensed spectrum. However, an increase of traffic demand in licensed spectrum causes a shortage of bandwidth for the transmission. For this reason, unlicensed spectrum attracts the attention due to its low cost, high flexibility and availability of bandwidth. Because of these advantages and new use cases in the industrial scenario, URLLC in unlicensed spectrum has become one of work items in the ongoing Release 17. URLLC in unlicensed spectrum poses a challenge of latency because the transmitter must do Listen Before Talk (LBT) procedure to acquire the channel before a transmission which might increase latency. Therefore, the impact and improvements of LBT procedures must be studied to guarantee URLLC latency requirement.

1.3 Physical layer design for URLLC in 3GPP Releases 15, 16, and 17

Table 1.1: 3GPP Release timeline

3GPP Release	**End of physical layer work**	**Release date**
Release 15	End 2017	End 2018
Release 16	End 2019	2020
Release 17	~ End 2021	~ End 2021

1.3.1 3GPP Release 15 - the foundation for URLLC in 5G

3GPP Release 15 is the first release with a full set of the 5G standards where the physical layer work of Release 15 was completed in December 2017. 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 in 5G to achieve these stringent requirements.

1.3.1.1 Flexible numerology and sub-slot-based transmission

A key new feature in 5G is the introduction of flexible sub-carrier spacing (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 a 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.

1.3.1.2 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.

1.3.1.3 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.

1.3.1.4 Uplink (UL) configured-grant (CG) transmission

In LTE, UL dynamic-grant (DG) transmission requires scheduling request (SR) from the UE and UL grant from the gNB that occupies 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. 1.1. 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,

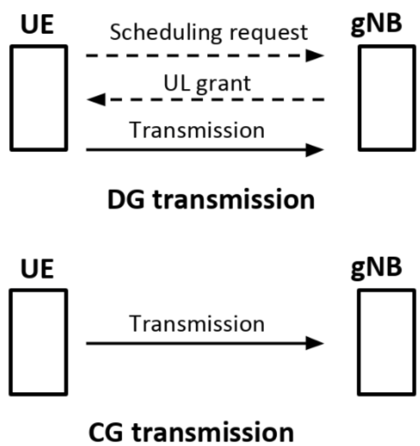


Figure 1.1: UL DG and CG transmission.

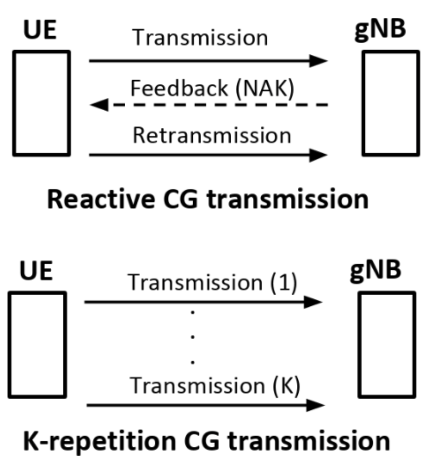


Figure 1.2: UL repetition CG transmission.

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. 1.2.

1.3.2 3GPP Release 16 features for URLLC in 5G

Release 16 where the physical layer work was completed in December 2019 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: factory automation, transport industry including the remote driving use case and electrical power distribution.

1.3.2.1 Physical downlink control channel (PDCCH) enhancements

1.3.2.1.1 PDCCH monitoring capability enhancements

As presented in Section 1.3.1.1, 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 1.2. Moreover, the resource for PDCCH in a slot is also limited as shown by the number of control channel elements (CCEs) in Table 1.2. A CCE consists of 6 resource element groups. A resource element group equals to 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 1.2. 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 1.2: UE monitoring capability in a slot in Release 15 [5]

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 with the maximum number of PDCCH candidates and non-overlapping CCEs defined per slot as in Release 15 or per span as in Release 16.

1.3.2.1.2 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. For example, 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.

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 1.3.2.5.2. Invalid symbol pattern indicator field with 0 or 1 bit is added to indicate the invalid symbols for PUSCH repetition Type B mentioned in Section 1.3.2.4.

1.3.2.2 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.

1.3.2.3 Uplink control information (UCI) enhancements

1.3.2.3.1 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.

1.3.2.3.2 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 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.

1.3.2.4 PUSCH enhancements

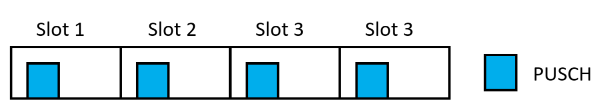


Figure 1.3: PUSCH repetition Type A.

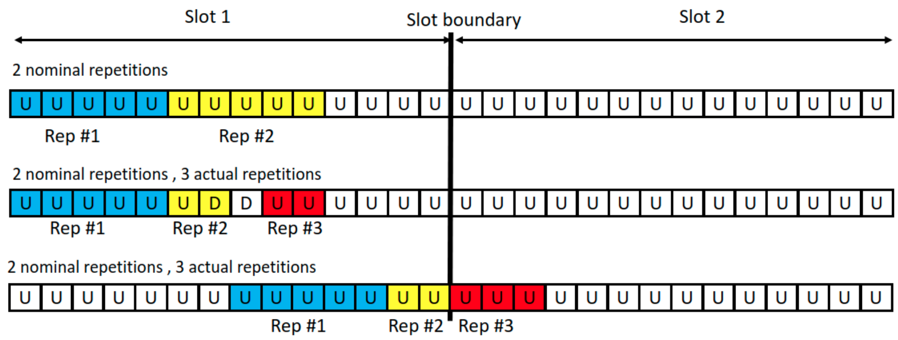


Figure 1.4: PUSCH repetition Type B.

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 (TB) is the same in those slots as shown in Fig. 1.3.

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. 1.4 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. 1.4, 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.

1.3.2.5 Enhanced inter-UE multiplexing in UL transmission

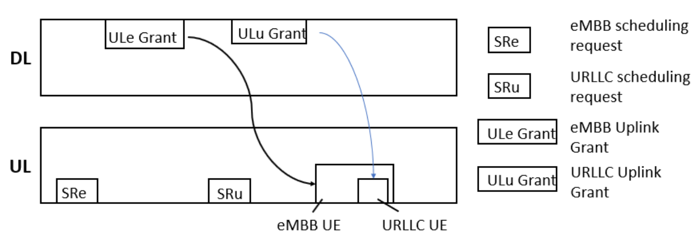


Figure 1.5: A collision of UL DG URLLC transmission with DG eMBB transmission.

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. 1.5. 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 the 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 resources 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.

1.3.2.5.1 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.

1.3.2.5.2 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.

1.3.2.6 Enhanced UL CG transmission

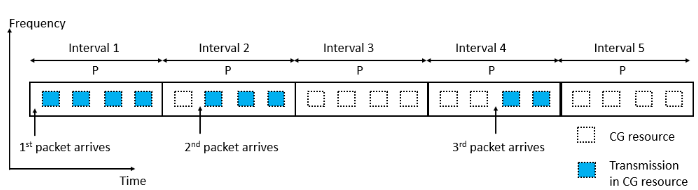


Figure 1.6: Less than K repetitions in CG UL transmission.

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. 1.6 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. 1.7 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.

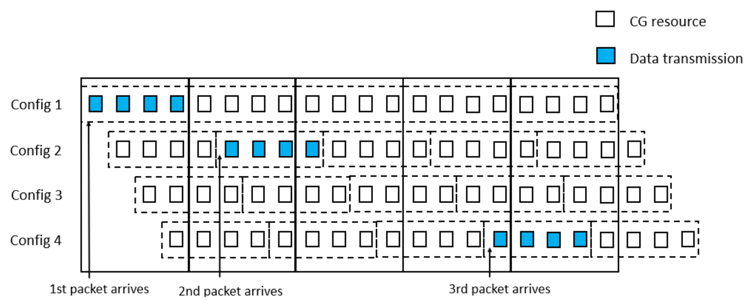


Figure 1.7: Multiple configurations to ensure K repetitions.

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.

1.3.3 3GPP Release 17 features for URLLC in 5G

1.3.3.1 Physical layer feedback enhancements

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 16. 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.

In Release 17, if the feedback for a DL SPS transmission is pointed to a DL symbol instead of an UL symbol in time division duplex (TDD) configuration, the SPS feedback is deferred to the next available PUCCH based on semi-static configuration of slot format. There is a limit on the maximum deferral of the SPS feedback that is configured by the gNB based on data requirements.

1.3.3.2 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 is 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.

1.3.3.3 URLLC enhancements in unlicensed spectrum

In Release 15 and 16, URLLC is specified to operate only in licensed spectrum. However, due to new use cases in the industrial scenario, unlicensed spectrum becomes a complement to URLLC operation in licensed spectrum. 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.

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 frame based equipment (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 UE is able to determine whether a scheduled UL transmission is transmitted according to the shared gNB-initiated COT or UE-initiated COT based on a predetermined rule. If the transmission is confined within a gNB FFP before the idle period of that gNB FFP and the UE has already determined that gNB initiated that gNB FFP, the UE assumes that the configured UL transmission corresponds to gNB-initiated COT. Otherwise, the UE assumes that the configured UL transmission corresponds to a UE-initiated COT. The FFP parameters of the UE-initiated COT such as the period of FFP, offset of FFP's starting point (having a symbol granularity) can be provided to the UE by dedicated RRC.

In unlicensed spectrum, when a UE transmits data in CG resources, it can be configured to transmit UCI containing redundancy version, HARQ identity and new data indicator in parallel with data in PUSCH. The gNB also can transmit ACK feedback to the UE when it decodes correctly the packet. Both the uses of CG UCI and CG ACK are enabled or disabled for unlicensed using one RRC parameter i.e. $cg - RetransmissionTimer - r16$.

1.4 Thesis outline and contributions

Chapter 1: Introduction. This chapter provides an overview of 5G New Radio where URLLC is one of three service categories. We introduce the applications of URLLC then the URLLC requirements from 3GPP to support these applications. Subsequently, we present the URLLC features in 3GPP Release 15, 16 and 17 to make URLLC achieve the specified requirements.

The content of this chapter is published in the journal:

- Trung-Kien Le, Umer Salim and Florian Kaltenberger, "An Overview of Physical Layer Design for Ultra-Reliable Low-Latency Communications in 3GPP Releases 15, 16, and 17," in IEEE Access, vol. 9, pp. 433-444, 2021.

Chapter 2: Ensuring Latency and Reliability of the UL configured-grant transmissions. This chapter presents the schemes to enhance the UL CG transmissions where the number of repetitions that the UE transmits is less than the configured number due to the boundary of the HARQ process. The first proposed scheme uses the reserved resources so that the UE can transmit the repetitions being out of the initial HARQ process in these resources. The size of reserved resources is optimized based on the location of each reserved resource in a HARQ process, the number of the UE in the system, data rate. The second proposed scheme allows the use of an explicit HARQ feedback structure when the UE cannot transmit all repetitions as configured due to the boundary of the HARQ process. This scheme permits the UE to transmit an additional repetition if the packet is not detected at the gNB so the errors due to a drop of packet decrease. The third proposed scheme requests the UE to transmit an additional SR in parallel with the transmission of TB when the UE transmits a smaller number of repetitions than configured. The additional SR increases the probability that the gNB detects the transmission from the UE and schedules a retransmission if necessary to reduce the errors due to packet loss.

The content of this chapter is published in the following papers and patents:

- Trung-Kien Le, Umer Salim and Florian 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, pp. 554-558.
- Trung-Kien Le, Umer Salim and Florian 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, pp. 597-601.
- Trung-Kien Le, Umer Salim and Florian Kaltenberger, "Enhancing URLLC Uplink Configured-grant Transmissions," The 2021 IEEE 93rd Vehicular Technology Conference (VTC2021-Spring), virtual conference, April 2021, pp. 1-5.
- Umer Salim, Trung-Kien Le, WO2020125473A1, "Uplink harq in cellular wireless communication networks", 2018.
- Umer Salim, Trung-Kien Le and Sebastian Wagner, WO2020164495A1, "RS For PUSCH Repetitions", 2019.
- Umer Salim, Trung-Kien Le, US patent application No. 63/061982, "Configurable uplink transmissions in a wireless communication system", 2020.

Chapter 3: UL eMBB and URLLC multiplexing. This chapter is about the problem of an overlap between UL eMBB and URLLC transmissions from different UE. Two schemes are proposed to improve the reliability of URLLC transmission when it collides with an eMBB transmission of another UE. The first scheme triggers the URLLC UE to use an explicit HARQ feedback structure by an overlap indication when the URLLC transmission overlaps with another eMBB transmission. The second scheme demands the URLLC UE to transmit an additional SR in parallel with the data transmission by an overlap indication when the URLLC transmission is in the same resources with the eMBB transmission. These two schemes reduce the errors of URLLC packet loss because the gNB cannot detect the URLLC UL transmission.

The content of this chapter is published in the following paper and patents:

- Trung-Kien Le, Umer Salim and Florian 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, pp. 1-6.

- Trung-Kien Le, Umer Salim and Virgile Garcia, WO2020143700A1, “Radio node device and method for inter-user equipment multiplexing”, 2019.
- Umer Salim, Trung-Kien Le, WO2020125472A1, “Uplink multiplexing in cellular wireless communication networks”, 2018.

Chapter 4: Feedback Enhancements for Downlink Semi-Persistent Scheduling Transmissions in Ultra-Reliable Low-Latency Communication. This chapter deals with the problem of feedback’s drop in URLLC DL SPS transmission due to the presence of DL slot/sub-slot in the designated slot/sub-slot of UL feedback in TDD configuration. A scheme is proposed to guarantee the transmission of HARQ feedback and potential data retransmission when there is UL-DL slot/sub-slot conflict at the indicated feedback resource. The scheme comprises a dynamic indication of feedback resource without using an associated DCI. An acknowledgement (ACK)-only feedback protocol is proposed to best suit the scenario in question. The combination of dynamic indication of feedback resource and ACK-only feedback structure guarantees higher reliability and brings flexibility to the transmission of HARQ feedback for DL SPS transmission.

The content of this chapter is published in the following papers and patent:

- Trung-Kien Le, Umer Salim and Florian Kaltenberger, “Feedback Enhancements for Semi-Persistent Downlink Transmissions in Ultra-Reliable Low-Latency Communication,” 2020 European Conference on Networks and Communications (EuCNC), virtual conference, June 2020, pp. 286-290.
- Trung-Kien Le, Umer Salim and Florian 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, pp. 1982-1986.
- Trung-Kien Le, Umer Salim and Florian Kaltenberger, WO2021098719A1, “Feedback for periodic resources”, 2019.

Chapter 5: Load based channel access enhancements in unlicensed spectrum for NR URLLC transmissions. This chapter analyzes latency of channel access when a transmitter operates in unlicensed spectrum and uses load based equipment (LBE) to access to the channel. LBE is analyzed by a Markov chain model and the closed-form equation of channel access latency is derived. Based on this equation, it is shown that URLLC latency requirement is not satisfied in some scenarios when the transmitter uses LBE. Therefore, the new tables of parameters used in LBE are proposed to ensure URLLC latency requirement.

The content of this chapter is published in the following paper:

- Trung-Kien Le, Umer Salim and Florian Kaltenberger, “Channel Access Enhancements in Unlicensed Spectrum for NR URLLC Transmissions,” GLOBECOM 2020 - 2020 IEEE Global Communications Conference, Taipei, Taiwan, December 2020, pp. 1-6.

Chapter 6: Frame based channel access enhancements in unlicensed spectrum for NR URLLC transmissions. This chapter analyzes channel access latency when the transmitter uses FBE to access to a channel in unlicensed spectrum. The URLLC latency constraint limits the number of channel sensing carried out by the transmitter and causes a decrease of transmission reliability because the channel cannot access to the channel to transmit data in the allowed time budget. Two schemes are proposed to improve channel access for URLLC transmission in FBE. The first scheme allows the transmitters to use multiple FFP configurations while the second scheme configures the FFP’s starting point of each transmitter based on its priority.

The content of this chapter is in the following paper:

- Trung-Kien Le, Umer Salim and Florian Kaltenberger, “Frame based equipment channel access enhancements in unlicensed spectrum for NR URLLC transmissions,” submitted to IEEE Transactions on Wireless Communications.

Chapter 7: Dynamic switching between load based and frame based channel access mechanisms in unlicensed spectrum. This chapter provides a model to analyze a system in unlicensed spectrum where the transmitters using LBE coexist with the transmitter using FBE. Based on channel access time and transmission probability from a Markov chain model, a scheme is proposed so that the devices are able to switch dynamically from FBE to LBE to serve data with high priority such as URLLC and from LBE to FBE to serve data with low priority such as eMBB.

The content of this chapter is in the following paper:

- Trung-Kien Le, Florian Kaltenberger and Umer Salim, "Dynamic switching between load based and frame based channel access mechanisms in unlicensed spectrum," submitted to GLOBECOM 2021 - 2021 IEEE Global Communications Conference.

Chapter 8: Enhancements of PUSCH repetitions for URLLC in unlicensed spectrum. This chapter focuses on the design of PUSCH repetitions for URLLC. In PUSCH repetition, the nominal PUSCH repetitions might be segmented into the smaller actual repetitions due to UL/DL directions in TDD configuration or slot boundary. This segmentation causes a degradation of the URLLC performance due to a smaller number of valid symbols for PUSCH repetitions, a drop of the repetition in an orphan symbol and an increase of LBT overhead. To enhance the URLLC performance, two schemes are proposed. The first scheme deals with segmentation due to UL/DL directions by dynamically switching the chosen semi-static DL symbols to UL symbols. The second scheme deals with orphan symbols by transmitting signal in these symbols in order to maintain a continuous PUSCH transmission and avoid an additional LBT.

The content of this chapter is published in the following paper and patents:

- Trung-Kien Le, Umer Salim and Florian Kaltenberger, "Enhancements of PUSCH repetitions for URLLC in licensed and unlicensed spectrum," 2021 17th International Symposium on Wireless Communication Systems (ISWCS), virtual conference, September 2021.
- Trung-Kien Le, Umer Salim and Florian Kaltenberger, US patent application No. 63/061979, "Performance enhancement of PUSCH repetition method in wireless communication systems", 2020.
- Trung-Kien Le, Umer Salim and Florian Kaltenberger, US patent application No. 63/061970, "Efficient scheduling in a wireless communications system", 2020.

1.5 Thesis perspective from practical (3GPP works) and more fundamental aspects

The work in the thesis follows the evolution of URLLC design in the 3GPP works and one of the goals has been to make the contributions to 3GPP. Therefore, the thesis starts with the problems that have been discussed in Release 16 by 3GPP. Some schemes were standardized by 3GPP to solve these problems in Release 16 and the ongoing Release 17 while some problems still have been unsolved by the standardized schemes.

Chapter 2 targets a problem of URLLC design in Release 16 that is less-than-the-configured-number-of-repetitions in the UL CG transmission. Release 16 adopted the use of multiple CG configurations to solve this problem as well as to serve the different UL traffic types at the UE. The use and details of multiple CG configurations are described in Section 1.3.2.6 of this thesis. The comparison between multiple CG configurations and the proposed reserved resource scheme in terms of resource consumption is presented in Section 2.6.

Chapter 3 studies the problem of the collisions in the UL CG resources of the URLLC transmission with the eMBB transmission of another UE in Release 16. Due to a shortage of time, Release 16 only standardized the schemes to solve the collision between different UE in the UL DG resources as described in Section 1.3.2.5. While the schemes to solve the collision in the UL CG resources have not been standardized. In the ongoing

Release 17, the problem of the collision in the UL CG resources between different UE also has not been included in the discussed subjects.

Chapter 4 deals with the problem of feedback for the DL SPS transmission in Release 16. Release 16 did not standardize a scheme to prevent a drop of feedback for the DL SPS transmission in the conflicting DL/UL symbols in TDD configuration so the work has been continued in Release 17. Release 17 adopted a scheme that defers the SPS feedback to the next available UL occasion based on semi-static configuration of slot format as described in 1.3.3.1. The advantage of this scheme is that it does not required additional signal to indicate resources for the dropped feedback so it saves resources. The UE decides by itself to defer the feedback to the next UL occasion as long as this occasion is within the allowed window. The proposed scheme allows the gNB to indicate dynamically the resources of feedback so the gNB can schedule feedback to resources in order to avoid the DL sub-slot/slot and multiplexing many feedback in one resource. However, this scheme requires a dynamic signal embedded in each SPS PDSCH that consumes resources and might reduce PDSCH reliability. Therefore, due to the simplicity, the deferring feedback scheme is standardized in the ongoing Release 17.

After Release 16 has been finalized, with the advent of new use cases in the industrial scenario, Release 17 starts to study URLLC operation in unlicensed spectrum in order to specify URLLC features in unlicensed spectrum to serve the applications in the industrial scenario. In one direction, we work on the theoretical aspect to study the impact of LBT on URLLC performance in unlicensed spectrum. From Chapter 5 to Chapter 7, we study the impact of two LBT mechanisms: LBE and FBE on URLLC transmission then propose new enhancements of LBE and FBE to support URLLC transmission. In another direction, we work with a practical problem of 3GPP in Chapter 8. The problem of PUSCH repetitions is being discussed in Release 17 to deal with the segmentation of PUSCH repetitions in unlicensed spectrum. The work of the ongoing Release 17 focusing on FBE is presented in Section 1.3.3.3 with the updated standards. Therefore, the theoretical aspect of FBE's impact on URLLC is the fundamental to the design of URLLC in Release 17 and the future releases. Besides that, the theoretical aspect of LBE and the model of the coexistence of LBE and FBE devices will also be the reference for the design of URLLC and other communication types with low latency requirement in the future releases.

Chapter 2

Ensuring Latency and Reliability of the UL Configured Grant transmissions

This chapter studies a problem in URLLC transmission in 3GPP Release 16 where the UE cannot transmit the number of repetitions as configured in the CG resources due to the boundary of a HARQ interval.

2.1 Problem formulation

In UL transmission, the gNB configures the UEs with high priority and strict requirements to transmit transport blocks in the CG regions. In addition, it also configures the number of repetitions K that these UEs need to carry out by a parameter $repK$ from higher layer in order to guarantee transmission's reliability and latency. K has values 1, 2, 4 and 8 as standardized in [7]. The UEs transmit the repetitions automatically in the CG regions without waiting for HARQ feedback or UL grant from the gNB. However, the UEs are only allowed to do repetitions in one interval with a periodicity P ranging from several symbols to several slots (a set of allowed periodicities P is defined in [7]) and prohibited to retransmit packet as configured by $repK$ crossing boundary of that interval. This constraint is to help the gNB avoid a confusion in HARQ identities (IDs) of different HARQ processes. Therefore, depending on the arrival time of data in relation to the periodicity P, the number of repetitions might be smaller than the configured number because the UEs need to stop their transmission at the last transmission occasion in the period P.

Fig. 2.1 illustrates a situation when the number of configured repetitions is not ensured due to the constraint of the boundary of a period P. In Fig. 2.1, an interval P contains 4 CG occasions, the UE is configured to do 4 repetitions for a packet. In the first period, data comes before all the 4 CG occasions so the UE is able to do 4 repetitions for the first packet as configured. However, when data comes in the second period, there are only 3 CG occasions left in that period. This means that the UE only can carry out 3 repetitions that are less than the configured number. Similarly, in the fourth period, the UE only can transmit the packet 2 times.

It is evident that when the packet comes after the first CG occasions in a period, the UE transmits the packet with a smaller number than the number configured by $repK$. It degrades the reliability of the UL transmission. The situation becomes more severe for the URLLC UEs with high reliability requirement. Moreover, latency of a transmission also increases because with a smaller number of repetitions, the gNB has a higher probability of failing to decode the packet and needs to schedule a retransmission. In that case, the

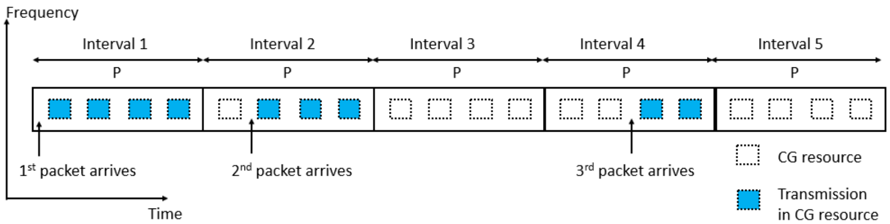


Figure 2.1: Less than K repetitions in CG UL transmission.

UE needs to wait the gNB to decode the repetitions of the packet and transmit an UL grant to reschedule a retransmission if necessary and it has a huge impact on latency.

2.2 Related works

In 3GPP Release 15, the UE only can wait until the next interval to transmit all K repetitions if data arrives lately. The waiting time might be large if SCS is small or the packet arrives only after some CG occasions as the second packet in Fig. 2.1. Latency requirement may not be satisfied in those cases.

In [10], 3GPP agreed that multiple configurations are used to enhance reliability and reduce latency. Ensuring K repetitions is included in the goal of multiple configurations. The UE can choose the configuration with the closest starting point to transmit all K repetitions as shown in Fig. 2.2. Two drawbacks of this scheme are overhead of signal to schedule multiple configurations and resource consumption of multiple configurations. For this reason, the main motivation for multiple configuration is to serve different applications with different requirements.

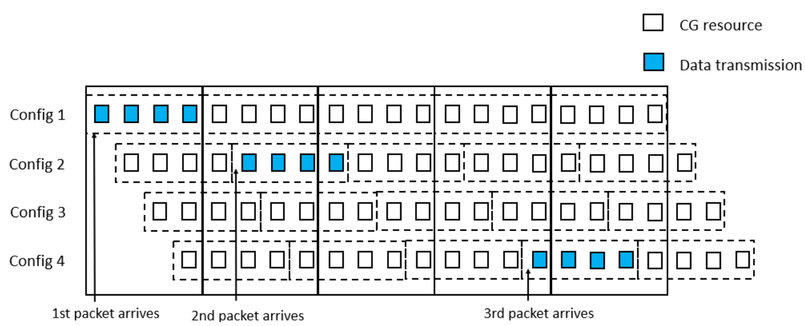


Figure 2.2: Multiple configurations to ensure K repetitions.

In [11] and [12], the UE is able to transmit the repetitions across the consecutive HARQ intervals. It requires lots of effort in standardization to avoid the confusion between HARQ IDs at the gNB such as a mechanism to communicate HARQ IDs to the gNB or different DMRS sequences in the repetitions.

In [13] and [14], the UEs transmits the repetitions in the shared resource. However, the constraint of HARQ process boundary is not considered. Moreover, the size of all resources for repetitions are the same. Both factors cause a degradation of reliability, latency and resource consumption.

2.3 Optimal reserved resources to ensure K repetitions

2.3.1 Reserved resources

To make the URLLC UEs achieve the strict requirements, a strategy to ensure that the UEs can transmit the number of repetitions as configured by $repK$ from higher layer is indispensable. Reserved periodic resources are proposed to be created and assigned to multiple UEs by the gNB so that they are likely to retransmit data in case the transmissions in the reserved resources are necessary to ensure the configured number of repetitions. These reserved resources have the same periodicity as the CG resources.

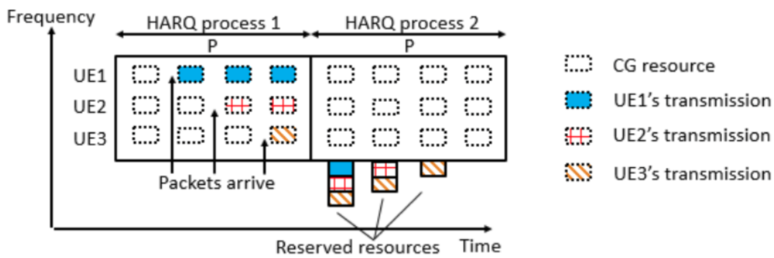


Figure 2.3: Reserved resources for repetitions.

The use of reserved resources is shown in Fig. 2.3. There are 3 UEs considered with CG resources in different bandwidth and each UE is configured to transmit 4 repetitions per a TB. The UE1's data comes after the first CG occasion in the first period so it only can do 3 repetitions in that period. In order to attain 4 repetitions, the UE1 retransmits data in the first reserved resource of the next period. Similarly, UE3's data arrives at the last CG occasion and only one repetition can be made. Thus, the UE3 uses the 3 reserved resources in the next period to achieve the configured number of repetitions.

In the example, 4 repetitions are configured so 3 reserved resources are needed in the next period. To increase the efficiency of resource consumption, the reserved resources are shared among the UEs. The first reserved resource in Fig. 2.3 with 3 blocks is likely to be shared with more than 3 UEs while still attain the target collision probability being approximate to the collision probability in the CG resources. Similarly, the second and third reserved resources are also shared by a group with more than 3 UEs. The equation showing the relation among the number of UEs, the size of the reserved resources and the collision probability is derived in the next sections.

2.3.2 System model

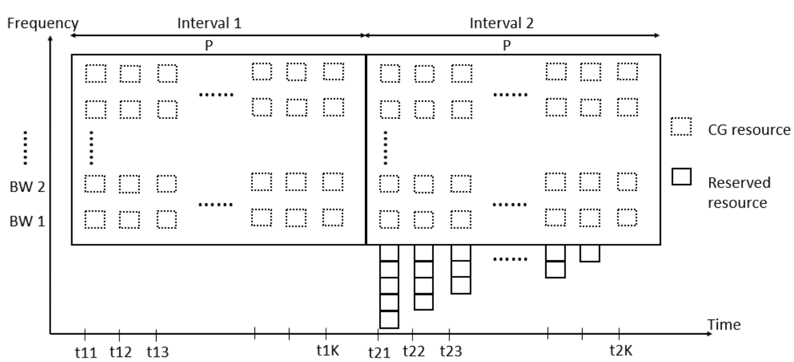


Figure 2.4: UL transmission resources' distribution.

In Fig. 2.4, N UEs are configured to transmit K repetitions in the shared consecutive CG resources. A HARQ process with time interval P contains K CG transmission occasions. The results derived below are also valid if the number of CG transmission occasions in an interval P are bigger than K. $K-1$ reserved resources are configured with the same period as the CG resources in each period P. All N UEs can use the reserved resources to attain the configured number of repetitions. The reserved resource at the ith transmission occasion in a period has M_i blocks where each block has the size of CG resource in one transmission occasion. Packet arrival follows a Poisson process. In an interval of T between two consecutive CG resources, there are λ packets arriving at a UE.

2.3.3 Collision probability in reserved resources

With random access, the N UEs in system are allowed to use any block in the reserved resources of a specific transmission occasion if they need to do the transmissions in order to fulfill the configured number of repetitions.

The collision probability in the reserved resource at the first transmission occasion of a period (at t21 in Fig. 2.4) is calculated as follows by considering 1 UE of interest having a transmission in the reserved resource at t21 and the rest of $N-1$ UEs.

It is assumed that the packet of the UE of interest cannot be decoded by the gNB if there is a collision with packets of other UEs in the reserved resources. The calculations below only focus on the error due to collision and do not count the radio errors.

In time T between two CG resources, the probability that one UE has one or more random transmissions is

$$P_{data} = 1 - e^{-\lambda}, \tag{2.1}$$

The CG resources in one bandwidth can be shared between a group of the UEs so collision probability in the CG resources between the UE of interest and other UEs of group is

$$P_{c\_CG} = 1 - e^{-\lambda(N_{\text{UE\_group}}-1)}, \tag{2.2}$$

where $N_{\text{UE\_group}}$ is the number of the UEs that use the same frequency band of CG resources.

The reserved resource at t21 is used by a UE if its data comes after the first CG transmission occasion at t11. The probability that one UE has transmission after the first CG transmission occasion in a period P is

$$P_d = (1 - P_{data})(1 - (1 - P_{data})^{K-1}). \tag{2.3}$$

There is no collision in the first reserved resource of a period P at t21 if no UE rather than the UE of interest has a transmission after the first CG transmission occasion at t11. The probability that no other UE from the set of $N-1$ UEs has a transmission after the first CG occasion is calculated by

$$P_0 = (1 - P_d)^{N-1}. \tag{2.4}$$

In case other UEs in a set of $N-1$ UEs has a transmission after the first CG occasion, the probability that n UEs have such transmission is

$$P_n = \binom{N-1}{n} P_d^n (1 - P_d)^{N-1-n}. \tag{2.5}$$

The probability that the UE of interest and n UEs do not access the same resource block in the first reserved resource at t21 is

$$P_{a0\_n} = \left(\frac{M_1 - 1}{M_1}\right)^n . \tag{2.6}$$

The probability that the UE of interest does not collide with any other UE in the first reserved resource at t21 is calculated by

$$P_{sum} = \sum_{n=1}^{N-1} P_n P_{a0\_n}. \tag{2.7}$$

From (2.4) and (2.7), the collision probability in the first reserved resource for the UE of interest is derived as

$$\begin{aligned} P_{c1} &= 1 - P_0 - P_{sum} \\ &= 1 - \left(\frac{M_1 - e^{-\lambda} + e^{-K\lambda}}{M_1}\right)^{N-1} . \end{aligned} \tag{2.8}$$

Based on the same calculating process, a general equation of collision probability for the reserved resource at any transmission occasion in a period can be derived as

$$P_{ci} = 1 - \left(\frac{M_i - e^{-i\lambda} + e^{-K\lambda}}{M_i}\right)^{N-1} , \tag{2.9}$$

where $i \in [1, K-1]$ is index indicating the position of the reserved resource based on the position of transmission occasion in a period.

According to (2.9), if P_{ci} is set to have the same value in all the reserved resources, the sizes of the reserved resources in an interval decrease from the first one to the last one: $M_1 > M_2 > M_3 > ... > M_{K-1}$. This means that the sizes of the reserved resources are optimized based on their positions. This optimization reduces resource consumption compared to using the same size for all reserved resources.

2.3.4 Group access to the reserved resources

In group access approach, one UE is only allowed to access and uses a specific part of the reserved resources pre-configured by the gNB. One part of the reserved resource is assigned and shared among a group of the UEs. For example, the UE1-3 is only permitted to access to the resource blocks as shown in Fig. 2.5. On the other hand, other UE groups are also prohibited to access to the resource blocks pre-configured to the UE1-3.

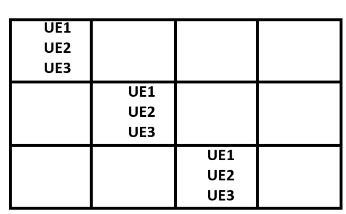


Figure 2.5: Group access to the reserved resource.

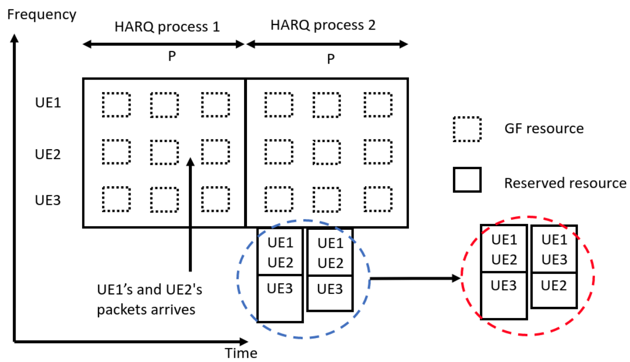


Figure 2.6: Method to group the UEs in different transmission occasions.

For a group of the UEs accessing to a part of the reserved resources, the collision probability is calculated by (2.9) as random access approach. The size of the whole reserved resource in each transmission occasion is sum of all parts assigned to the UE groups to guarantee a target probability.

Simulation in Section 2.6.1.1 shows that the sizes of the reserved resources for both random access and group access approaches are the same to achieve an equal target probability. However, group access approach reduces the decoding burden in the gNB. To decode a retransmitted packet, the gNB only needs to search a part of the reserved resource instead of the entire reserved resource as random access approach. Thereby, power consumption and processing time drop dramatically.

When two or more UEs of the same group having access to a part of the reserved resources have data coming at the same time and need to use the reserved resource for repetitions, they will compete for the same part of the reserved resources at the consecutive transmission occasions as show in Fig. 2.6. The UE 1 and 2 have 3 repetitions and data coming after the first two CG occasions so they compete two times at both the first and second reserved resources with the arrangement in the blue circle.

To avoid that problem, the UEs are assigned to different groups of the reserved resources for different transmission occasions as illustrated in the red circle. The UE 1 only has the competition in both two reserved resources when both the UE 2 and 3 need to use the part of the reserved resources assigned to the UE 1 so the probability of having two consecutive collisions decreases.

In each reserved resource, the collision probability is still the same as calculated from (2.9) but the overall reliability of a UE transmitting in the reserved resources at different transmission occasions will be improved with the hopping of the UEs to various groups in the different reserved resources.

2.3.5 Optimal reserved resources with a successive interference cancellation (SIC) receiver at the gNB

In the calculation in Section 2.3.3, a packet is assumed not to be decoded if it collides with the packets of other UEs in the reserved resources. However, even if there is a collision between the repetitions of the UEs in the system, the gNB equipped with a SIC receiver still can decode correctly the repetitions. When the gNB decodes correctly a packet in the CG resources or the previous reserved resources, it stores that packet. After that, if the gNB encounters a collision between the successful packet and another packet in the reserved resources, it can cancel the successful packet from the received signal to remove the interference. Thereby, the gNB decodes the other packet without interference and has higher successful probability. With a big number of repetitions (for example, 4 or 8 repetitions), there is low probability that the gNB has a collision among all non-decoded packets so the SIC receiver is useful in improving performance of repetitions in the reserved resources.

The successful probability of a packet in the transmission with a SIC receiver in the gNB is complex to calculate. It depends on the channel condition of other UEs, the successful probability of other packets competing for the resources, the time arrival of data. Therefore, a model of SIC receiver in physical layer called K-multipacket reception is used in [15]. In this model, the gNB is assumed to be able to decode correctly all the UE transmissions in the same block of reserved resources if the number of the UEs in that block are smaller than a threshold L. On the contrary, if the number of the collided UEs are bigger than L, all the packets in the collided resources cannot be decoded.

Using the system model in Section 2.3.2, the equations calculated in Section 2.3.3 and the model of SIC receiver in [15], the error probability of a packet due to the collision is calculated for the first reserved resource at t21 in Fig. 2.4.

We consider a UE of interest and other $N-1$ UEs. The probability that n UEs in $N-1$ UEs have a transmission is calculated in (2.5). The probability that l UEs in these n UEs access to the same block in the first reserved resource as the UE of interest is

$$P_{al\_n} = \binom{n}{l} \left(\frac{1}{M_1}\right)^l \left(\frac{M_1-1}{M_1}\right)^{n-l}. \tag{2.10}$$

If the value of l is smaller than $L-1$, the gNB still can decode all the packets in that block. The probability that the UE of interest collides with any other UEs ($l < L-1$) but the packet is still decodable in one block is

$$\begin{aligned} P_s &= \sum_{n=1}^{N-1} P_n \sum_{l=1}^{L-1} P_{al\_n} \\ &= \sum_{n=1}^{N-1} P_n \sum_{l=1}^{L-1} \binom{n}{l} \left(\frac{1}{M_1}\right)^l \left(\frac{M_1-1}{M_1}\right)^{n-l}. \end{aligned} \tag{2.11}$$

From (2.4), (2.7) and (2.11), the error probability of a packet from the UE of interest due to collision in the first reserved resource is

$$\begin{aligned} P_{col\_SIC\_1} &= 1 - P_0 - P_{sum} - P_s \\ &= 1 - (1-P_d)^{N-1} - \\ &- \sum_{n=1}^{N-1} P_n \sum_{l=0}^{L-1} \binom{n}{l} \left(\frac{1}{M_1}\right)^l \left(\frac{M_1-1}{M_1}\right)^{n-l}. \end{aligned} \tag{2.12}$$

(2.8) is a case of (2.12) where L is 1 and the packet cannot be decoded if there is a collision.

Similarly, collision probability for the reserved resource at any transmission occasion in a period can be derived as

$$P_{di} = (1-P_{data})^i (1-(1-P_{data})^{K-i}). \tag{2.13}$$

$$P_{ni} = \binom{N-1}{n} P_{di}^n (1-P_{di})^{N-1-n}. \tag{2.14}$$

$$\begin{aligned} P_{col\_SIC\_i} &= 1 - (1-P_{di})^{N-1} - \\ &- \sum_{n=1}^{N-1} P_{ni} \sum_{l=0}^{L-1} \binom{n}{l} \left(\frac{1}{M_i}\right)^l \left(\frac{M_i-1}{M_i}\right)^{n-l}. \end{aligned} \tag{2.15}$$

where $i \in [1, K-1]$ is index indicating the position of the reserved resource based on the position of transmission occasion in a period.

The presence of SIC receiver reduces resource consumption of the reserved resources, make the system support more UEs with higher data arrival rate.

2.4 Explicit HARQ feedback structure to reduce packet loss in the less-than-K-repetition situation

In case the PUSCH resources for UL transmission are scarce and the network is overloaded with the UEs, the gNB cannot configure reserved resources to guarantee both the number of repetitions and their reliability. The network faces with the fact that the number of repetitions are impossible to be achieved as configured. The priority in this situation is to reduce the impact of the repetitions that are not transmitted on system's reliability. The schemes described Section in 2.4 and Section 2.5 will attain this goal.

2.4.1 Operation of the explicit HARQ feedback structure

The HARQ structure in UL CG transmission is timer-based. This means that there is no explicit acknowledgement (ACK) feedback sent from the gNB to the UE if data is decoded correctly. Instead, the UE uses a timer. If it does not receive an UL grant to schedule a retransmission at the end of time configured in the timer, the transmission is considered as successful. This structure has a drawback because the UE cannot differentiate between a successful transmission and a miss-detection when the gNB cannot decode DMRS to identify the UE and transmits UL grant. Therefore, in case of miss-detection, the UE does not receive any signal from the gNB, it assumes a successful transmission and drops data in buffer. This behavior impacts the reliability of a transmission and becomes more severe in less-than-K-repetitions situation because the gNB has a smaller number of configured repetitions to detect successfully DMRS. Once the UE is not able to carry the configured number of repetitions, the quality of service (QoS) of transmission is badly affected and there is a higher probability that the gNB cannot decode DMRS sequence. It leads to a degradation of URLLC transmission's reliability due to the packet dropped by the UE. To handle the issue of the CG transmissions with less than K repetitions, we propose to use an explicit HARQ feedback from the gNB.

An UL transmission results in three scenarios. In the explicit HARQ feedback structure, the UE's behavior in each scenario will be analyzed as follows.

The first scenario is about correct data decoding. The gNB tries to combine all the repetitions of a TB to facilitate data decoding and the number of these repetitions can be less than K as per the previous discussion. For a normal operation, the gNB is capable of identifying the repetitions concerning a specific TB. Thus, whenever the gNB is able to correctly decode a TB, and it sees that it was sent with less than K repetitions it will send an explicit ACK for this TB to the transmitting UE.

The second scenario is about a failure of data decoding with a successful UE identification. When the UE transmits less than K repetitions, it is possible that the data decoding is not successful but the gNB is able to identify the UE transmitting the TB with less than K repetitions through identification of UE specific DMRS sequence which it was configured with as a part of CG configuration. In this case, the gNB will reschedule a retransmission of the previously transmitted TB.

The third scenario is related to UE identification failure. It is where the proposed explicit feedback becomes pivotal. The bad quality of received data may lead to a situation where the gNB is unable to identify the transmission from the UE through DMRS detection. This situation is the most damaging for the URLLC UEs/applications due to their tight constraints on latency and reliability. With a timer-based HARQ structure, which is currently used for URLLC transmissions in 3GPP Release 15, this situation leads to different understanding at the gNB and at the UE. The gNB, being unable to identify the transmission from the UE, cannot schedule the re-transmission. The UE, upon receiving no UL grant for re-transmission, considers the packet successfully decoded at the gNB and discards the buffer upon the expiry of HARQ timer.

Table 2.1: Comparison of different feedback structures

Case	**Timer-based feedback**	**Explicit feedback**
DMRS: detected TB: decoded	No ACK/UL grant	ACK
DMRS: detected TB: failed	UL grant: reschedule	UL grant: reschedule
DMRS: failed	No ACK/UL grant: packet lost	No ACK/UL grant: packet retransmitted automatically

Although the situation when the gNB cannot identify the UE may be caused by a number of reasons such as the very bad channel conditions, large amount of interference or insufficient number of actual repetitions, the configuration parameters of CG transmission, in particular MCS and the number of repetitions K, are designed to combat most of these adverse effects. On the other hand, if the configured number of repetitions cannot be made, this brings the CG operation point to a lower QoS target than the desired operating point.

In the proposed technique, whenever the UE transmits less than K repetitions, the TB in question is supposed to operate with the explicit HARQ based feedback. In general, the gNB can identify transmissions with less than K repetitions thanks to DMRS detection and CG window boundary knowledge. When the gNB fails to identify the transmitting UE and sends no ACK or UL grant to this UE, the UE, upon expiry of configured HARQ timer, re-transmits automatically the TB. The retransmission timing and resources can be configured as part of the explicit HARQ feedback configuration. One suitable option is to retransmit in the closest CG periodic window after the expiry of HARQ feedback timer. The HARQ feedback timer should include the time for the gNB to decode the data and find the suitable occasion for potential DL transmission of HARQ ACK or UL grant. The explicit HARQ feedback structure is only activated in the extreme cases as less-than-K-repetition transmission when the probability of DMRS miss-detection becomes high. The selective activation of the explicit feedback structure helps to increase the system's performance in the extreme case while does not increase the overhead of ACK feedback in the normal cases. In those cases, reliability of URLLC transmission is ultra high so an ACK feedback is not necessary and only wastes resource. Table 2.1 summarizes the operation of the conventional timer-based feedback structure and the explicit feedback structure.

2.4.2 Design of the explicit HARQ feedback

In the proposed strategy, the UEs are allowed to request an explicit HARQ feedback for certain TBs. Thus, a design for the explicit HARQ feedback in general may be needed. One strategy can be to define a channel where HARQ ACK can be transmitted. This can be similar to Physical HARQ Indicator Channel (PHICH) as specified in 4G LTE but this requires a lot of specification effort and high resource overhead. The rationale is that in typical operation mode, the UEs will not request explicit HARQ feedback to reduce overhead and only in exceptional cases it will be required.

With this in view, the proposal is to use the UL grant (which is DCI) as an explicit HARQ feedback. This DCI can be sent with UE specific configured scheduling-radio network temporary identifier (CS-RNTI) which is used with the dynamic grant transmissions. If the gNB is able to successfully decode the data, it sends DCI to this UE with the same HARQ process number (HARQ ID) as of the successfully received TB. To avoid any confusion between DCI used as feedback and DCI used as UL grant, new data indicator (NDI) field can be set to 0. Further, some of the fields in the DCI such as the time and frequency resource assignment fields are set to 0 to help the UE differentiate DCI used as feedback and DCI used for other purposes.

2.5 Additional SR to reduce packet loss in the less-than-K-repetition situation

In this section, another scheme is proposed to deal with the problem of less-than-K repetitions. In this scheme, to improve the reliability of UL CG transmissions, whenever the UE transmits less than the configured number of repetitions for a TB, it sends SR to the gNB in parallel to transmission of TB with less than K repetitions. This SR provides another mean for the gNB to detect UE ID and compensates the drop of reliability.

3GPP Release 15 does not allow transmission of Physical uplink control channel (PUCCH) and Physical uplink shared channel (PUSCH) simultaneously. The UE transmits UCI encoding SR, HARQ feedback, etc on PUCCH. Therefore, the UE multiplexes UCI and PUSCH if it wants to transmit uplink control information (UCI) while sending PUSCH. This strategy allows the UE to transmit SR in case of less than K repetitions. However, in UCI and PUSCH multiplexing, if the gNB cannot detect DMRS of PUSCH due to bad channel, there is high probability that the gNB also cannot decode UCI (SR) to find the UE ID. Thus, multiplexing strategy might not enhance the performance of UE ID detection. For this reason, SR should be configured to be transmitted on the configured PUCCH resources. The gNB upon receiving PUCCH and PUSCH from the same UE will understand that the SR in PUCCH is for the same TB sent in PUSCH for the UE that is only able to make less than K repetitions.

Table 2.2 shows in tabular format the UE and gNB actions for strategy of SR transmission in parallel to TB transmission.

The parallel transmission of PUCCH and PUSCH may be slightly onerous for certain UEs because the UEs need to create the gap in the resource grid between PUCCH and PUSCH to protect them from interference. Nevertheless, considering that the main focus is here on URLLC type of UEs with strict latency and reliability targets, this overhead may be acceptable.

If the UE is transmitting different types of traffic at the same time, the gNB can differentiate the proposed SR transmitted in parallel with PUSCH in CG resources from a standalone classic SR which is sent to the gNB to have the UL resources scheduled through the HARQ ID of the transmitted TB included in the SR. As upon receiving the TB, the gNB will know its HARQ ID from its timing window, it will be able to conclude that the SR concerns the same TB or not.

Under certain situations, it may be beneficial to allow a hybrid scheme where the UEs can flexibly choose between an explicit HARQ feedback structure or sending a SR in parallel to the transmission of a TB when they are unable to transmit K repetitions. The simplest scheme would be the one that the gNB configures the UEs to follow one of these two schemes. Alternatively, the UEs can be configured to choose one these two schemes. In that case, it would make sense to have an explicit indication in the TB for the explicit HARQ feedback. If the UEs choose to transmit SR in parallel to the transmission of the TB, they do not trigger an explicit feedback with the TB transmission. This can be advantageous in the situations when there is at least a suitable SR transmission occasion available where the UEs can transmit SR for the TB in question. On the contrary, the UEs do not transmit SR in parallel to the transmission of the TB but send an indication to tell that it triggers an explicit HARQ feedback. This can be more advantageous if there is no suitable SR transmission occasion and a SR transmission may harm the latency budget.

For the traffic with the extremely stringent latency-reliability constraints, it can be foreseen that both mechanisms, explicit HARQ feedback and transmission of SR, are triggered in parallel to maximize the reliability within a short time interval when the UE transmits the TB with less than K repetitions.

Table 2.2: SR Transmission with TB and Actions for the gNB and the UE

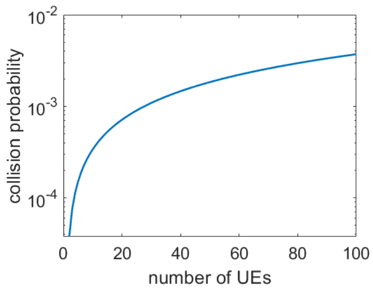
Case	**CG PUSCH**	**SR in PUCCH**	**gNB understanding**	**gNB action**	**UE action**
1	Correctly decoded at the gNB	Correctly decoded at the gNB	The gNB knows that SR is for the decoded TB	Indicate a correct detection (ex: using UL grant with the same HARQ ID)	Discard data upon receiving the gNB indication
2	Correctly decoded at the gNB	Incorrectly decoded at the gNB	The gNB upon correctly decoding the data and seeing less than K rep knows about missing SR. This case should be rare as SR is sent with strong coding	Indicate correct detection (ex: using UL grant with same HARQ ID)	Discard the data upon the gNB indication
3	Incorrectly decoded at the gNB but UE Identified through DMRS	Correctly decoded at the gNB	the gNB understands that UE sent SR along with the TB that it failed to decode	The gNB sends the UL grant for re-transmission	The UE follows the UL grant for re-transmission
4	Incorrectly decoded and UE Identification Failure at the gNB	Correctly decoded at the gNB	The gNB completely misses the CG transmission due to failure in UE identification but it receives SR. From the timing of SR and CG configurations, the gNB knows its decoding failure	The gNB sends the UL grant for transmission	The UE follows the UL grant for transmission and re-transmits the data
5	Incorrectly decoded at the gNB but UE Identified through DMRS	Incorrectly decoded at the gNB	The gNB identifies the UE from PUSCH. If it can identify the case of less than K repetitions, it knows also about SR detection failure	The gNB sends the UL grant for re-transmission	The UE follows the UL grant for re-transmission
6	Incorrectly decoded at the gNB and UE Identification Failure at the gNB	Incorrectly decoded at the gNB	The gNB has no indication about UE transmission	No action	The UE can be configured to retransmit in the subsequent CG resources or SR

2.6 Numerical results and performance evaluation

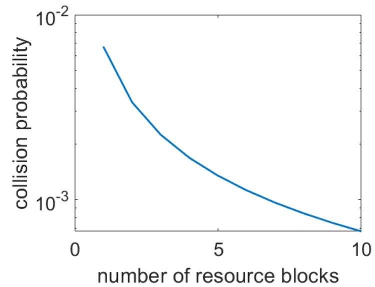
2.6.1 Optimal reserved resources

2.6.1.1 No SIC receiver at the gNB

2.6.1.1.1 Random access to the reserved resources



(a) Collision probability with respect to the number of the UEs.



(b) Collision probability of the second reserved resource.

Figure 2.7: Collision probability.

From (2.9), the number of the UEs that the system can support can be found. Besides, the number of resource blocks in each reserved resource are also calculated to sustain that system.

The simulation starts with the reserved resource in the first transmission occasion of a period. The set of parameters is $M_1 = 10, K = 4, \lambda = 1.25 \times 10^{-4}$.

Fig. 2.7(a) shows collision probability in the reserved resource in the first transmission occasion in terms of the number of the UEs sharing that resource. From the graph, we see that if the target collision probability (P_{c1}) is 10^{-3}, the system can support 28 UEs in the first reserved resources.

As mentioned in Section 2.3.2 and Section 2.3.3, the CG resources in a frequency band can be shared by a group of the UEs. 28 UEs calculated above can be divided in to 4 groups with 7 UEs in each group. The collision probability in the CG resources calculated from (2.2) is 7.5×10^{-4} that is approximate to the collision probability of 10^{-3} in the reserved resource.

Fig. 2.7(b) illustrates collision probability with respect to the reserved resource's size in the second transmission occasion. When all the parameters (λ, N, K and P_{ci}) are kept the same, the sizes of the reserved resources in the second and the third transmission occasions are calculated as shown in Table 2.3.

Table 2.3: Sizes of the reserved resources with $K = 4$ and random access

Position of reserved resources	1	2	3
Number of blocks	10	7	3

The percentage of resources saved in comparison to using the same size of 10 resource blocks for all the resources in [13] and [14] is: $(1 - (10 + 7 + 3)/(10 \times 3)) \times 100\% = 33\%$.

The proposed scheme also consumes much less resources than the scheme of multiple configurations in [10]. As shown in Fig. 2.2, if 4 repetitions are configured by the gNB, 4 configurations must be configured to ensure that the UEs always can transmit at the beginning of a period and reach 4 repetitions as configured. For a group of the UEs sharing the CG resources, 4 configurations are needed. Each configuration consists of 4 CG resources in one period. Thus, one group of the UEs requires $4 \times 4 = 16$ resource blocks in a period. As mentioned in the first scenario, there are 4 groups of the UE so in total, $16 \times 4 = 64$ resource blocks are demanded in a period. While the scheme with reserved resources only requires 16 CG resource blocks and 20 reserved resource blocks in a period that are 36 resource blocks in total. Resource consumption decreases by $36/64 \times 100\% = 56.25\%$.

One more factor taken into account when multiple configurations is applied is an increase of DMRS port. The distinction of configurations at the gNB is based on DMRS detection. Each UE transmits a specific DMRS sequence when using a configuration. Therefore, if 4 configurations are used, the number of orthogonal DMRS ports required are 4 instead of one port in the single configuration with reserved resources.

Another scenario is considered with a bigger number of repetitions: $M_1 = 10, K = 8, \lambda = 1.25 \times 10^{-4}$.

The system can support 12 UEs in the first reserved resources to achieve Pc target of 10^{-3}. These UEs can be divided into 2 groups with 6 UEs in one group that each group uses the CG resources in one bandwidth part. The collision probability in CG regions of a group of 6 UEs as scheduled above is: 6.25×10^{-4}. With 12 UEs, the sizes of the reserved resources in the transmission occasions are shown in Table 2.4.

Table 2.4: Sizes of the reserved resources with $K = 8$ and random access

Position of reserved resources	1	2	3	4	5	6	7
Number of blocks	10	8	7	6	4	3	2

The percentage of resources saved in comparison to using the same size of 10 resource blocks for all the resources in citeref10 and [14] is: $(1 - (10 + 8 + 7 + 6 + 4 + 3 + 2)/(10 \times 7)) \times 100\% = 42.86\%$.

In comparison to multiple configurations in [10], resource consumption decreases by $(8 \times 8 \times 2)/(8 \times 2 + 10 + 8 + 7 + 6 + 4 + 3 + 2) \times 100\% = 43.75\%$.

Furthermore, 8 orthogonal DMRS ports are required to distinguish the configurations. Generally, the required number of DMRS resources are equal to the number of configurations. In case the number of configurations are big such as 8, DMRS resources might not be enough. This means that multiple configurations cannot be used and it affects reliability and latency of URLLC transmission.

2.6.1.1.2 Group access to the reserved resources

Group access is simulated with the parameters as the first scenario in random access: $N = 28, K = 4, \lambda = 1.25 \times 10^{-4}$.

28 UEs are divided into 4 groups of 7 UEs to access to the reserved resources. Each group is only allowed to access to the part of the reserved resources pre-configured to them.

To satisfy Pc=10^{-3}, the sizes of the reserved resources are shown in Table 2.5.

Table 2.5: Sizes of the reserved resources with K=4 and group access

Position of reserved resources	1	2	3
Number of blocks per group of reserved resources	2	2	1
Total number of blocks of reserved resources	8	8	4

The resource consumption is equal for random access and group access to the reserved resources as shown in Table 2.3 and Table 2.5. However, with group access, the complexity and processing time of decoding in the gNB reduce substantially because in the reserved resource at the first transmission occasion, the gNB only needs to find data of a UE in 2 reserved blocks instead of finding blindly data of a UE in 10 reserved blocks as in random access approach.

2.6.1.2 SIC receiver is equipped at the gNB

The proposed scheme with optimal reserved resources guarantees reliability because all configured repetitions are transmitted in the target latency. Moreover, it consumes less resources than other schemes in [10], [13] and [14] as illustrated in Fig. 2.8. In Fig. 2.8, the simulation is done to compare the resource consumption of the proposed scheme and prior art that are used to guarantee the configured number of repetitions in two scenarios of K=4 and K=8 explained above.

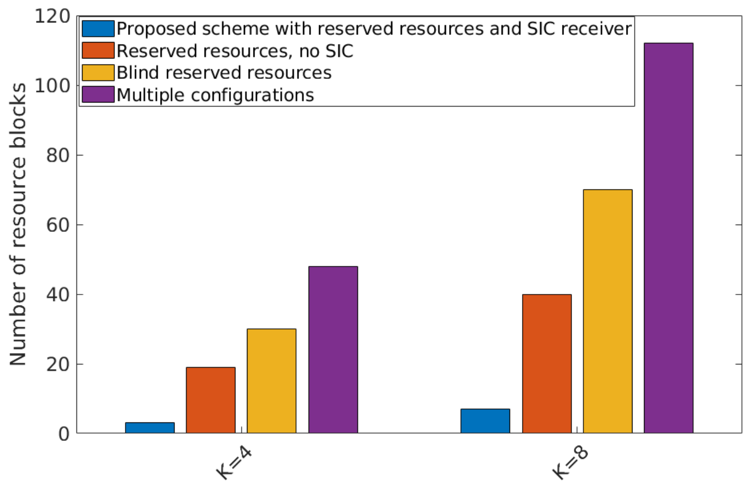


Figure 2.8: Comparison of resource consumption in different schemes.

From Fig. 2.8, when K is 4, the scheme with a SIC receiver at the gNB requires 16 CG resource blocks and 3 reserved resource blocks in a period that are 19 resource blocks in total. Reserved (additional) resource consumption decreases by $(1 - 3/48) \times 100\% = 93.75\%$ and total resource consumption decreases by $(1 - 19/64) \times 100\% = 70.31\%$ compared to multiple configurations in [10]. The proposed scheme also consumes 84.21% and 90% less reserved resources than the scheme without a SIC receiver at the gNB and the scheme in [13] and [14], respectively.

The second scenario considered in Fig. 2.8 has a higher configured number of repetitions where the UEs are configured to transmit 8 repetitions. The set of the parameters is: $M_1 = 10, K = 8, \lambda = 1.25 \times 10^{-4}$ and $P_{c1} = 10^{-3}$. The result also shows a significant decrease of resource consumption of the proposed scheme compared to prior art.

Fig. 2.9 shows the error probability due to collision in the first reserved resource ($P_{col\_SIC\_1}$ in (2.12)) in terms of the average number of random access events λ in an interval of T between two consecutive CG resources in licensed spectrum. When the gNB can decode more packets in the same block (L increases), the system can support much higher data rates while still achieving the same target reliability of 10^{-3} due to packet collision. In the scheme without a SIC receiver at the gNB, with $L = 1$, the system only can support $\lambda = 1.25 \times 10^{-4}$. But with $L = 2$ or $L = 3$ in the proposed scheme, the system can support $\lambda = 5.8 \times 10^{-3}$ or $\lambda = 2.53 \times 10^{-2}$, respectively.

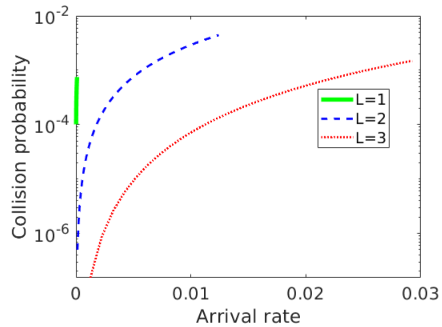


Figure 2.9: The arrival rate vs collision probability.

2.6.2 Explicit feedback structure and additional SR in less-than-K-repetition transmission

Table 2.6: Simulation parameters

Parameters	**Values**
Waveform	CP-OFDM
Subcarrier spacing	60kHz
Channel model	Rician
K factor	1
Number of allocated PRB	8
DMRS detection mechanism	Time-domain correlation

The performance of DMRS detection is simulated with the parameters in Table 2.6. Fig. 2.10 shows simulation results with different false alarm rate (FAR). For each DMRS detection, the correlation result is compared with a threshold to determine whether DMRS exists or not. This threshold is chosen according to a target FAR indicating the cases that the gNB determines the existence of DMRS while in reality there is no DMRS transmitted. A higher threshold is required for a lower FAR but also results in more missed detection. Because channel estimation to decode data as well as recognizing UE ID to reschedule a retransmission if necessary in conventional scheme depend on DMRS detection, a degradation of DMRS detection due to a smaller number of repetitions than configured makes the system not be able to support reliability URLLC requirement.

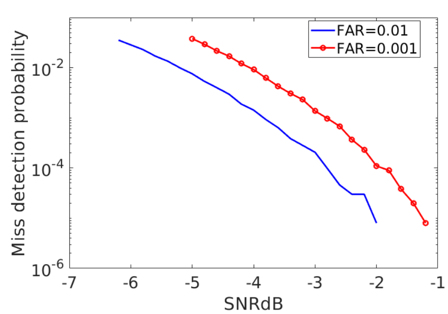


Figure 2.10: DMRS detection performance.

In the simulation showed in Table 2.8, HARQ process has periodicity P equal to 4 slots. With SCS of 60 kHz, 4 slots spread in 1 ms. The configured number of repetition are 4. 3GPP Release 15 standardized that each slot has only one repetition [9]. Thereby, 4 repetitions are carried out in 4 slots equal to one HARQ process. The sum of gNB processing time, feedback transmission time, UE processing time is one slot. The UE waits after one slot to decide whether dropping or retransmitting data.

Table 2.8 shows the performance of DMRS detection at SNR of -5dB and FAR of 0.001 in different schemes and arrival time of data.

As can be seen, due to the constraint of HARQ process boundary, the UE might not transmit all 4 repetitions as configured if data comes late. Less repetitions transmitted means that DMRS miss-detection at the gNB increases. It leads to an increase of packet loss because the packet is assumed to be successful in timer-based feedback structure.

If the UE waits until the next HARQ process to transmit data, it might also not be able to do all the configured repetitions because of URLLC latency requirement.

An explicit feedback structure makes the UE retransmit data even if the gNB misses DMRS. It increases DMRS detection's performance and reduces packet loss. However, in one scenario as shown in Table 2.8, after discovering that there is no ACK or UL grant, the UE cannot retransmit data because latency budget of 1 ms is reached.

An additional SR helps to reduce the probability of UE ID miss-detection. Nevertheless, if the actual number of repetitions are small as in in Table 2.8, the miss-detection probability is still high compared to the case of full repetitions transmitted.

The utilization of reserved resources always guarantees the number of repetitions so the target reliability is ensured. However, resource consumption is higher than other schemes.

The advantages and disadvantages of three proposed schemes: using reserved resources, using explicit HARQ feedback, using an additional SR are summarized in Table 2.7.

Table 2.7: Comparison of three proposed schemes

	Advantage	**Disadvantage**
Reserved resources	+ Ensure the configured number of repetitions + Optimal sizes of reserved resources	+ Resource consumption of reserved resource + Overhead of signal to configure reserved resources
Explicit feedback structure	+ Avoid packet loss due to DMRS miss-detection	+ Resource consumption of ACK feedback + Latency constraint limits the number of retransmissions
Additional SR	+ Enhance UE ID detection	+ Resource consumption of SR + High UE ID miss-detection if the number of packets transmitted are small + Latency constraint limits the number of retransmissions

Table 2.8: Performance comparison of different schemes at $SNR = -5dB$ and $FAR = 0.001$

Case	**Scheme**	**Starting time offset (ms)**	**Number of repetitions**	**UE ID miss-detection probability**	**Number of retransmissions in UE ID miss-detection**	**Total UE ID miss-detection probability**
TB comes after the 1 st CG occasion	Conventional transmission	0	3	5.5×10^{-5}	0	5.5×10^{-5}
	Conventional transmission with the UE waiting the next period	0.75	1	0.038	0	0.038
	Transmission with explicit feedback	0	3	5.5×10^{-5}	0	5.5×10^{-5}
	Transmission with SR	0	3	2.1×10^{-6}	0	2.1×10^{-6}
	Transmission with reserved resources	0	4	2.1×10^{-6}	0	2.1×10^{-6}
TB comes after the 2 nd CG occasion	Conventional transmission	0	2	10^{-3}	0	10^{-3}
	Conventional transmission with the UE waiting the next period	0.5	2	10^{-3}	0	10^{-3}
	Transmission with explicit feedback	0	2	10^{-3}	1	5.5×10^{-5}
	Transmission with SR	0	2	5.5×10^{-5}	0	5.5×10^{-5}
	Transmission with reserved resources	0	4	2.1×10^{-6}	0	2.1×10^{-6}
TB comes after the 3 rd CG occasion	Conventional transmission	0	1	0.038	0	0.038
	Conventional transmission with the UE waiting the next period	0.25	3	5.5×10^{-5}	0	5.5×10^{-5}
	Transmission with explicit feedback	0	1	0.038	2	5.5×10^{-5}
	Transmission with SR	0	1	10^{-3}	0	10^{-3}
	Transmission with reserved resources	0	4	2.1×10^{-6}	0	2.1×10^{-6}

2.7 Conclusion

This chapter works with the Release 16's problem where the UE is unable to transmit the configured number of repetitions in the CG resources due to the boundary of a HARQ interval. To deal with this problem, three schemes have been proposed. In the first scheme, reserved resources are configured to the UE and a SIC receiver is equipped at the gNB to resolve the collision of repetitions of the different UEs in the same reserved resources so that the configured number of repetitions in UL CG transmission is guaranteed. The scheme optimizes an amount of resources to be reserved while assuring the reliability of URLLC transmission. Besides the use of the reserved resources, this chapter presents two other strategies to help the UEs achieve the target QoS in case of less than K repetitions as shown in the results. These approaches relating to an explicit HARQ structure and an additional SR can be used individually or combined together based on different scenarios.

Chapter 3

UL eMBB and URLLC multiplexing

This chapter focuses on a Release 16 problem where the URLLC transmission of a UE in the CG resources collides with the eMBB transmission of another UE.

3.1 Problem of multiplexing URLLC and eMBB in the CG resources

In 5G, as mentioned in Section 1.3.1.4, the gNB can schedule the CG resources for the URLLC UEs but it does not have any prior information which of these CG resources will actually be used by the URLLC UEs or which of the UEs in the group configured to the resources will use a specific resource. If the cell is loaded and the gNB schedules some eMBB UEs on the resource overlapping with CG occasion, as shown in Fig. 3.1, there is going to be transmission collision between dynamically scheduled eMBB and URLLC CG transmissions.

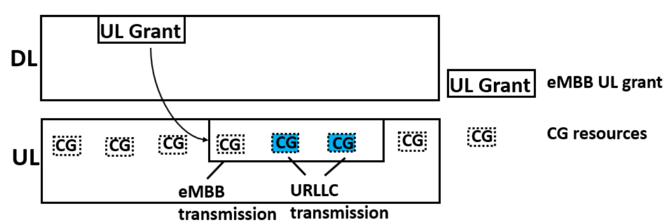


Figure 3.1: A collision between UL URLLC CG transmission and dynamic eMBB transmission in case of Frequency Division Duplex (FDD).

When the transmissions from the eMBB and URLLC UEs in the CG resources overlap, it results in lower decoding probability due to a lower resulting Signal-to-interference-plus-noise ratio (SINR) for both UEs. This can be a serious problem for the URLLC UEs in particular due to their tight latency and reliability targets.

In case the gNB is able to identify the URLLC UE from its DMRS sequence, it may try to quickly reschedule the UE over non-overlapping resources if an error happens.

The increased interference due to the overlapping transmissions of eMBB and URLLC UEs may lead to a catastrophic situation when the gNB may not even identify the URLLC UE (DMRS miss-detection). The current HARQ structure in UL transmission for NR is timer based, which means that upon transmission of packet, the UE will start the HARQ timer. If it receives an UL grant for the re-transmission of the same TB from the gNB, it does the retransmission over the resources scheduled by the UL grant. If it receives no UL grant from the gNB and the HARQ timer expires, it considers that the TB was successfully decoded at the gNB and discards the data in the buffer.

The timer-based HARQ feedback and UL dynamic grant retransmission are standardized because this minimizes the control overhead for sending HARQ feedback. This is reasonable in general but in the cases of dynamic UL multiplexing, giving rise to overlapping transmissions with the CG UEs, when the gNB will not be able to identify the UEs transmitting on CG resources, the UEs will discard their packets and consider the successful detection that leads to serious performance's degradation for the URLLC UEs.

3.2 Related works

In [16] and [17], when there is an overlap, the gNB asks the URLLC UE to apply a different pre-configured transmission power that is higher than power level used in case of no overlap. However, an increase of power causes an interference among the neighboring cells. Secondly the cell-edge UEs may be power limited and cannot raise their transmission power. This is also the problems of [18] when the gNB assigns URLLC PUSCH with updated transmission parameters such as resource, MCS, transmission power, etc. on CG resources that are occupied by eMBB PUSCH.

In [19], the group common control channel reveals resource range allocated to dynamic grant eMBB UE so the CG UE can exclude the occupied resource. Nevertheless, the CG UE has less resources left to transmit than the original configured resources so this partial transmission may not be decodable at the gNB.

In [20], the gNB informs the URLLC UE which of the CG resource set has the overlapping eMBB transmission such that the URLLC UE can initiate the CG transmission over resources not occupied by the ongoing eMBB transmission. It might result in high latency if all resources in the current occasion are full and the UE must wait until the next transmission occasion.

A preemptive scheme in [21] cannot be applied in URLLC CG transmission because the gNB does not know the URLLC transmission in advance to preempt the eMBB transmission.

A SIC receiver in [22] only benefits the eMBB UE rather than the URLLC UE because URLLC data is decoded first due to the latency requirement.

3.3 Strategy to multiplex the eMBB and URLLC UEs in the CG resources

3.3.1 The overlap indication and the explicit HARQ ACK feedback

This chapter proposes a two-step strategy to overcome the problem of the multiplexed UL eMBB and URLLC transmissions in the CG resources. In the first step, upon scheduling a dynamic grant transmission of the eMBB UE over the CG resources, the gNB sends an indication of the overlapped resource to the URLLC UEs. As the gNB does not know which of the URLLC UEs configured for the CG resources may become active in the current interval, this indication needs to be sent to all the UEs who have been configured with the CG resources in the overlapping interval as illustrated in Fig. 3.2. Upon receiving this overlap indication, the URLLC UEs are aware of the resources which have been dynamically scheduled for other UEs and in case of transmission, their transmissions will be received with an increased interference.

The second step of the proposed strategy comprises making the overlapping transmissions use an explicit HARQ feedback structure rather than the legacy timer-based feedback. The resource overlap indication can serve this purpose by containing a 1-bit flag to tell if the feedback becomes explicit or not. Upon receiving this indication, the URLLC UEs, who transmit on the overlapping resources, expect to receive the explicit HARQ feedback from the gNB for their transmissions.

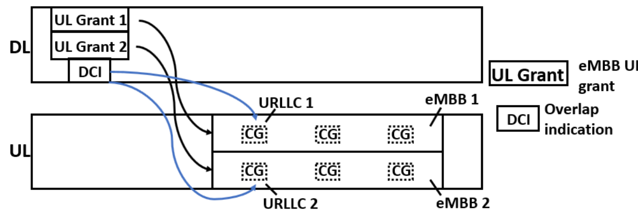


Figure 3.2: Signalling the URLLC UEs about an overlap with the eMBB UEs in CG regions.

Thereby, within a configured time period, the CG UEs receive either explicit HARQ ACK indicating the successful detection of their TB or UL grant for re-transmission in case the gNB failed to decode the TB but was able to identify the transmitting UE. For the third case, if the gNB even fails to identify the UE through DMRS due to high interference (DMRS miss-detection), it cannot schedule the UE for a re-transmission in the conventional timer-based feedback so the UE assumes that a packet is decoded correctly and drops it from buffer to transmit the next packet. This is the scenario where our proposed explicit HARQ feedback structure becomes the most promising. If the CG UE receives neither an ACK nor an UL grant within a configured time, the UE does not consider that its data is successful, rather it considers that the gNB failed to identify its identity (ID) due to high interference in the overlapping transmissions and retransmits the TB on the subsequent CG resources. Table 3.1 summarizes the operation of the conventional timer-based feedback structure and the explicit feedback structure.

Table 3.1: Comparison of different feedback structures

Case	**Timer-based feedback**	**Explicit feedback**
DMRS: detected TB: decoded	No ACK/UL grant	ACK
DMRS: detected TB: failed	UL grant: reschedule	UL grant: reschedule
DMRS: failed	No ACK/UL grant: packet lost	No ACK/UL grant: packet retransmitted automatically

In addition, the URLLC transmission can be made to stop immediately after receiving an explicit HARQ feedback. Thus, the URLLC UE does not need to carry out all repetitions as configured by parameter *repK* from higher layer. It helps the URLLC UE save power and resources. Besides, the eMBB UE also avoids suffering from an interference from the URLLC UE's repetitions and there is more chance that the gNB is still able to decode correctly eMBB data despite the collision with URLLC data at the beginning.

Block error rate (BLER) of a packet in timer-based feedback structure and explicit feedback structure is shown in (3.1) and (3.2) in Section 3.4.

The timer value, for which the UE should wait to receive ACK or UL grant, can be configured by radio resource control (RRC) parameters. This value can be selected as a function of reliability and latency targets of the UEs. For some UEs with extremely high latency and reliability targets, this timer value can be put to zero which means to do the automatic transmission in case of overlapping transmissions to maximize the chance of correct data detection at the gNB.

3.3.2 The overlap indication and the additional SR

In a variation of the proposed scheme in Section 3.3.1, the second step of making the transmissions use an explicit HARQ feedback structure can be replaced by using an additional SR. The gNB can indicate the URLLC CG UEs by overlap indication in the first step described in Section 3.3.1 to send a SR in parallel for the TB transmitted over the overlapping CG resources. The SR sent to the gNB will provide another way besides DMRS detection to detect the ID of the UE transmitting in the interfered CG transmission. When the gNB is unable to identify the UE making the CG transmission because of DMRS miss-detection, the gNB might still be able to identify that UE by decoding the additional SR and react fast to the received SR by sending an UL grant to this UE. Thanks to the UL grant, the UE is likely to retransmit the packet and has a successful transmission in the latency budget instead of assuming a successful transmission and dropping the packet as in the conventional scheme. When the gNB is able to decode the data successfully, it still reacts to SR to send some indications to the UE about the successful detection.

PUCCH and PUSCH are not allowed to be transmitted simultaneously. In case the UE needs to transmit UCI while transmitting UL data on PUSCH, it sends the UCI on PUSCH. When the UE is not transmitting PUSCH, it sends the UCI carrying SR, feedback for DL data, etc. on the PUCCH resources using appropriate PUCCH format as a function of UCI content and the PUCCH configurations. The simplest strategy to send SR (which is UCI) when the URLLC UE is transmitting over overlapping CG resources would be to transmit it over PUSCH CG resources along with the transmission of the TB. This can be simple from implementation perspective but from the performance point of view, it does not help to improve significantly the performance of the UE ID detection. If DMRS of PUSCH is not detected by the gNB because of bad channel and interference from eMBB transmission, there is high chance that the gNB also cannot decode the SR to obtain the UE ID multiplexed with PUSCH. For this reason, to achieve a better UE ID detection, the SR is proposed to be transmitted using PUCCH configuration on the specified PUCCH resources separated from PUSCH. As PUCCH resources are dedicated resources on different frequency physical resource blocks (PRBs) and OFDM symbols, this provides additional diversity advantage to the SR transmitted in these resources compared to multiplexing and transmitting it over UL CG resources along with the TB. Therefore, the UE is configured to transmit this SR on PUCCH resources in parallel to the transmission of TB on the overlapping resources.

The overlap indication may have an explicit indication, in the form of a single bit flag, which may require the UEs transmitting over overlapping CG occasions to send SR. In fact, more flexibility and better performance can be achieved by having the flexible control of the explicit HARQ feedback structure and the transmission of SR. The gNB can then choose which strategy to use in different situations. As an example, if the periodicity of current CG occasions is not very fast, it may make sense to indicate the overlapping CG UEs to send a SR. And in the cases, if the resources configured for SR are not sufficient or not in close proximity compared to the subsequent CG occasions, it may be suitable to prioritize the explicit HARQ feedback and automatic retransmission in case the UEs receive no ACK/UL grant indication for the transmitted TB.

3.3.3 Configuration and Signalling for the Overlap Indication

One very important feature of the proposed scheme is that the overlap indication is sent to the UEs who have been pre-configured for the CG resources. As the CG resources may be shared by multiple UEs, and the gNB has no idea in prior which of these UEs may transmit their data over the CG resources, this overlap indication is sent in a group-common manner. Thus, the proposal is to send the overlap indication in a group-common downlink control information (DCI).

For the UL overlap indication sent in a group-common DCI, a DCI format similar to the DCI format 2_1 in [4] can be used. DCI 2_1 is also used for DL pre-emption indication. The size of DCI format 2_1 is configurable by higher layers up to 126 bits and each indication has 14 bits.

The UL overlap indication sent to the URLLC UEs comprises the indication of UL CG resources typically scheduled for the eMBB UEs. The eMBB UEs would be normally scheduled for the slot duration or most part of the valid UL symbols so it would be judicious to have more bits of UL reference resource field defining the frequency granularity. To keep a format close to the DL pre-emption indication and adapted to indicate the UL overlap resource, there are two possibilities of UL overlap indication. In the first design, all 14 bits are used to indicate the frequency PRB region for the whole slot. Thus, each bit in the 14-bit long bitmap indicates 1/14 of the frequency PRBs of the carrier as shown in left part of Fig. 3.3. The second option is to split the time-frequency grid of the slot in 7 frequency zones each spanning one half slot. Thus, each bit may indicate an overlap over 1/7 of the frequency PRBs for a half slot time duration as illustrated in right part of Fig. 3.3.

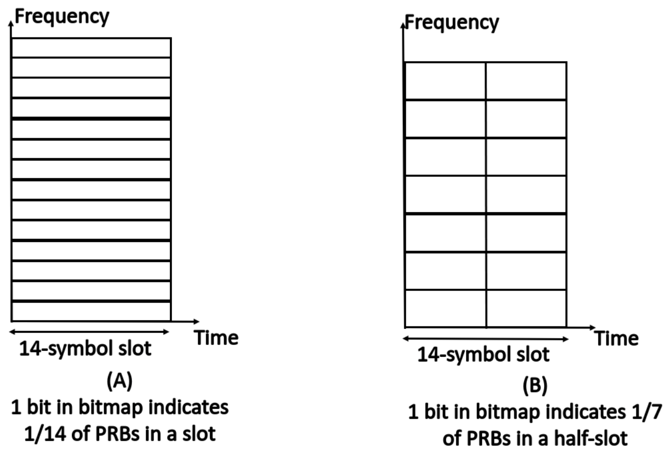


Figure 3.3: Resource Indication in Uplink Overlap Indication.

The UEs configured with the CG resources can be configured to listen to and decode the overlap indication as part of CG configuration. The activation or de-activation of this operation can be done in different ways for Type 1 and Type 2 CG as specified in [6]. For RRC configured Type 1, the activation and de-activation can be done by RRC signalling. For configured grant Type 2, where some parameters of CG can be updated by DCI, the activation or de-activation of overlap indication can be made through DCI signalling which is used to update other CG parameters.

The overlap indication can be sent at the same time when the gNB sends the dynamic UL grant scheduling a UE over the CG resources as shown in the left part of Fig. 3.4.

When the gNB sends an UL grant scheduling an eMBB UE, the scheduled resources are not necessarily in the same slot where the UL grant is sent. Rather typically the UL grant will be for the resources located in one of the subsequent slots. If the gNB transmits the overlap indication along with the UL grant, it needs to indicate the slot where this overlap will occur. To avoid this additional signalling and to keep the treatment of the UL grant simple, the overlap indication can be transmitted in the slot where overlap occurs as shown in the right part of Fig. 3.4. However, it may not be preferable to send and receive at the same time, so transmitting the overlap indication in the DL direction in the same slot as of the overlapping CG occasions may not be very interesting. In some other cases like TDD operation mode, it may be completely impossible. Thus, based on the system design, one of these two transmission schemes is determined.

One big advantage of sending the overlap indication to the URLLC UEs is that the overlap indication is primarily indicating the overlap caused by dynamic scheduling of the eMBB UEs over the CG resources. As the eMBB UEs may be scheduled only once during a slot, the indication periodicity can be kept to be once per slot and no mini-slot monitoring is needed to receive the overlap indication. This is advantageous in the sense that it does not overload the UEs with additional DCI monitoring and decoding burden.

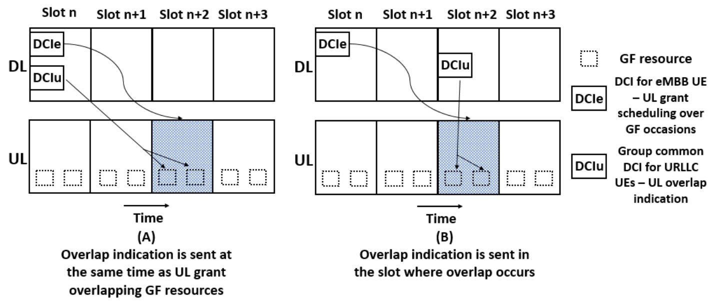


Figure 3.4: Timing to send the UL Overlap Indication.

3.3.4 Design of the Explicit HARQ Feedback

In the proposed strategy, the gNB shifts the CG transmissions in the overlap region from timer-based approach to explicit-HARQ-feedback approach. Thus, a design for the explicit HARQ feedback in general may be needed. The proposal is to use DCI as an explicit HARQ feedback sent with UE specific CS-RNTI which is used with the dynamic grant transmissions. If the gNB is able to successfully decode the data despite the overlap, it can send an UL grant to this UE with the same HARQ process number (HARQ ID) as of the successfully received TB, and the UE upon receiving this UL grant would know that this is in fact not a retransmission request but an explicit ACK for the previously transmitted TB. To avoid any confusion, NDI field can be set to zero. Further, some of the fields in the DCI which are actually not needed, such as the time and frequency resource assignment fields, may be sent with fixed known values which can be pre-decided to be used in the ACK indication.

3.4 Numerical results and performance evaluation

Table 3.2: Simulation parameters

Parameters	**Values**
Waveform	CP-OFDM
Subcarrier spacing	60kHz
Channel model	Rician
K factor	1
Number of allocated PRB	8
DMRS detection mechanism	Time-domain correlation

Simulation parameters are shown in Table 3.2. Fig. 3.5 illustrates the performance of DMRS detection in three cases: the URLLC UE transmits in the CG resources without any collision with another eMBB UE, the URLLC UE has a collision with an eMBB UE at the same power level and the URLLC UE has a collision when increasing power 1dB higher than an eMBB UE. For each DMRS detection, the correlation result is compared with a threshold to determine whether DMRS exists or not. This threshold is chosen according to a target FAR indicating the cases that the gNB determines the existence of DMRS while in reality there is no DMRS transmitted. A higher threshold is required for a lower FAR but also results in more missed detection.

Due to a collision between DMRS of the URLLC UE of interest and data/DMRS of another eMBB UE, the performance of DMRS detection of the URLLC UE degrades significantly as can be seen in Fig. 3.5 and cannot achieve the miss detection probability 10^{-5} at the same SNR level of the case without collision.

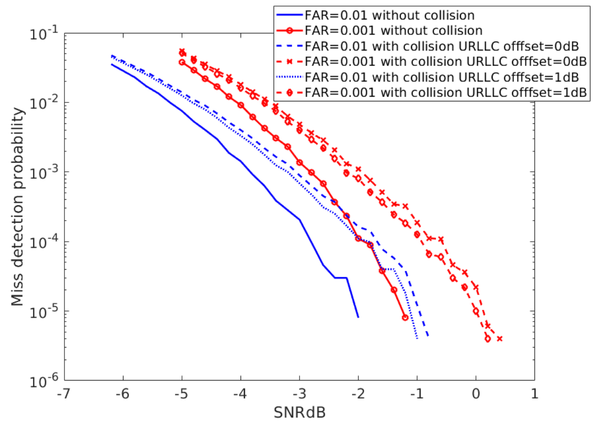


Figure 3.5: DMRS detection performance.

At FAR of 0.001 and SNR of -1.4dB, the miss detection probability increases from 10^{-5} of the case without collision to 3.4×10^{-4} of the case with a collision with an eMBB UE at the same power. Fig. 3.5 also shows that even with power control scheme in [16] and [17] when the URLLC UE's power increases 1dB higher than the eMBB UE's power, miss detection probability is still 2.44×10^{-4} that is higher than 10^{-5} of the case without collision. As DMRS detection is mandatory for channel estimation to decode data as well as for recognizing UE ID to reschedule a retransmission if necessary in conventional scheme, a degradation of DMRS detection makes the system unable to support reliability URLLC requirement. BLER of an UL transmission with one potential retransmission in the conventional scheme or power control scheme (P^e_1) is calculated as

$$\begin{aligned} P^e_1 &= P^e_{DMRS1}+ \\ &+ (1 - P^e_{DMRS1})P^e_{d1}(P^e_{DMRS2} + (1 - P^e_{DMRS2})P^e_{d2}), \end{aligned} \tag{3.1}$$

where P^e_{DMRS1}, P^e_{DMRS2} are the miss detection probabilities of the initial (with collision) and retransmitted (without collision) DMRS (see Figure 3.5), and P^e_{d1}, P^e_{d2} are the error probabilities of the initial (with collision) and retransmitted (without collision) PUSCH.

In (3.1), the first term is the error probability when the gNB cannot detect DMRS to decode or reschedule data so the UE does not retransmit data and data is lost. The second term is the error probability when the gNB detects DMRS and identifies UE ID but fails to decode data so it reschedules data. However, it cannot decode the retransmission and an error still occurs.

The usage of an explicit HARQ feedback as explained in Section 3.3.1 solves the problem of DMRS miss-detection in the overlapping region because it allows the UE to carry out the retransmissions in the interference-free regions even if DMRS is not detected by the gNB. BLER of an UL transmission with one potential retransmission in the proposed scheme with explicit feedback (P^e_2) is calculated as

$$\begin{aligned} P^e_2 &= P^e_{DMRS1}(P^e_{DMRS2} + (1 - P^e_{DMRS2})P^e_{d2})+ \\ &+ (1 - P^e_{DMRS1})P^e_{d1}(P^e_{DMRS2} + (1 - P^e_{DMRS2})P^e_{d2}). \end{aligned} \tag{3.2}$$

Compared to (2.1), the first term in (2.2) is enhanced because of a retransmssion while the second terms are the same.

The final column in Table 3.3 shows a remarkable enhancement of the first term of error probability in UL transmission with explicit HARQ feedback in case of URLLC and eMBB multiplexing in comparison to the conventional scheme with timer-based feedback and power control scheme in [16] and [17]. Moreover, the proposed scheme can be applied to all UEs in a cell while the power control scheme to increase URLLC UEs' power cannot be applied to the cell-edge UEs because of power limitation.

Table 3.3: Performance comparison between different scenarios and schemes at SNR=$-1.4dB$, FAR=0.001, P^e_{d2}=P^e_{SR}=0.01

Case	URLLC UE ID miss detection probability	Retransmission in UE ID miss detection	URLLC UL transmission's BLER due to the first UE ID miss detection
No collision	10^{-5}	No	10^{-5}
Collision in the conventional scheme	3.4×10^{-4}	No	3.4×10^{-4}
Collision with power control ([16] and [17])	2.44×10^{-4}	No	2.44×10^{-4}
Collision with explicit feedback (proposed)	3.4×10^{-4}	Yes	3.4×10^{-6}
Collision with additional SR (proposed)	3.4×10^{-6}	No	3.4×10^{-6}

Table 3.3 also shows an improvement of UE ID detection (the first term in (3.3)) when an additional SR is transmitted in the separate PUCCH in parallel with data (PUSCH) in CG resources. As can be seen in (3.3), the SR provides another chance for the gNB to detect UE ID. Therefore, the error probability because of DMRS miss-detection decreases. BLER of an UL transmission with one potential retransmission in the proposed scheme with an additional SR (P^e_3) is calculated as

$$\begin{aligned} P^e_3 &= P^e_{DMRS1} \times P^e_{SR} + \\ &+ (1 - P^e_{DMRS1} \times P^e_{SR}) P^e_{d1} \times \\ &\times (P^e_{DMRS2} + (1 - P^e_{DMRS2}) P^e_{d2}), \end{aligned} \tag{3.3}$$

where P^e_{SR}: the error probability of SR

The selection between two proposed schemes is explained in Section 3.3.2.

The presence of retransmission in explicit feedback leads to latency and resource consumption but guarantees target reliability in case of DMRS miss-detection, while the conventional scheme stops the transmission straightaway and causes packet loss so latency and resource consumption have no meaning when a packet is already failed to be decoded correctly. Moreover, with SCS 60kHz and the decoding time of one transmission being 0.1ms for a packet spreading in 4 OFDM symbols, even with one retransmission, the system consumes 0.5ms in total and still satisfies the latency requirement of 1ms.

Compared to the conventional scheme, overhead of the explicit feedback structure is higher due to ACK signal but is limited because the explicit feedback structure is only used when there is an overlap in CG resources (a cause of high interference) and an overlap indication triggers this structure. In addition, as can be seen in Table 3.3, the error probability in the overlapping transmissions increases so the number of ACK feedback decrease while the number of the cases without any feedback because of DMRS miss-detection increase. For this reason, a mechanism to trigger a retransmission as the proposed explicit feedback structure becomes imperative. It is the same reason that the overhead of SR transmitted in parallel with PUSCH is acceptable in eMBB and URLLC multiplexing.

3.5 Conclusion

This chapter presents a strategy to multiplex the dynamic grant eMBB and CG URLLC transmissions while guaranteeing the strict requirements of URLLC. An overlap indication is used and combined with an explicit HARQ feedback structure or an additional SR to help reduce the error probability of the URLLC UE's transmission in case of DMRS miss-detection. The proposed scheme provides a mechanism to meet very stringent URLLC constraints with minimal additional control overhead.

Chapter 4

Feedback Enhancements for Downlink Semi-Persistent Scheduling Transmissions in Ultra-Reliable Low-Latency Communication

This chapter copes with a problem in Release 16 and Release 17 where feedback for DL SPS transmissions is dropped due to the conflict of the UL and DL symbols in TDD configuration.

4.1 Feedback cancellation in DL SPS transmission in TDD configuration

As mentioned in Section 1.3.2.2, to reduce overhead, alignment time, transmission time and processing time of PDCCH, the gNB configures DL SPS resources which occur periodically to a set of UEs. In DL SPS transmission, the gNB configures partially the parameters of DL SPS configurations using higher layer signaling such as RRC signaling. After having made the configurations, the gNB activates the desired DL SPS configuration by sending a DCI using a single cell-radio network temporary identifier (SC-RNTI). Thereby, the UEs are aware of the locations and the parameters of the potential PDSCH transmission such that the gNB can transmit PDSCH in the DL SPS resources to the UE without an associated PDCCH scheduling the particular transmission.

Similar to the DL dynamic transmissions, when the gNB configures and activates the DL SPS resources by a DCI, the activation DCI contains PDSCH-to-HARQ-feedback indicator K1 indicating the number of slots/sub-slots from the end of PDSCH to the beginning of feedback on PUCCH. This parameter indicating the PDSCH-to-HARQ-feedback timing is then continually used throughout the active time of that SPS configuration. If the UE receives PDSCH in slot/sub-slot n, the UE will transmit HARQ feedback in slot/sub-slot n+K1. However, in TDD configuration, if the slot/sub-slot n+K1 is a DL slot/sub-slot, the UE cannot transmit PUCCH carrying HARQ feedback in that slot/sub-slot and cancels that PUCCH transmission.

In 3GPP Release 15, the periodicity of DL SPS is bigger than 10 ms so DL SPS transmission does not happen frequently. However, in 3GPP Release 16, the periodicity may be set as low as a slot, thus making them suitable for URLLC transmission. Nevertheless, the value K1 is only indicated once at the activation of DL SPS configuration. This is different from dynamic DL scheduling where the scheduling DCI will indicate the feedback timing for the scheduled transmission. Given a lower SPS resources' periodicity and a flexible format design in 5G where dynamic slot/sub-slot format update using slot-based DCI is supported to transmit low latency transmission [5], the collision probability between the UL resources of HARQ feedback of DL SPS PDSCH and DL/flexible slot/sub-slot in semi-static and dynamic TDD configuration increases that results in an increase of feedback cancellation. Therefore, when the DL SPS configurations with short periodicity are used for URLLC services/applications, the HARQ relevant conflicts can become a serious harm to URLLC QoS.

In Fig. 4.1, the DL SPS resources are configured to a UE with a periodicity of P in TDD configuration. The K1 value indicated by the activation DCI is 3. A PDSCH is sent to the UE in the first SPS resources. Based on K1 equal to 3, a HARQ feedback is expected to be transmitted in the fourth slot in Fig. 4.1 that is 3 slots after the first slot containing SPS PDSCH. However, this slot is a DL slot and cannot be used for an UL transmission of HARQ feedback. Following Release 15, this HARQ feedback is dropped without being resumed to be transmitted in the next PUCCH.

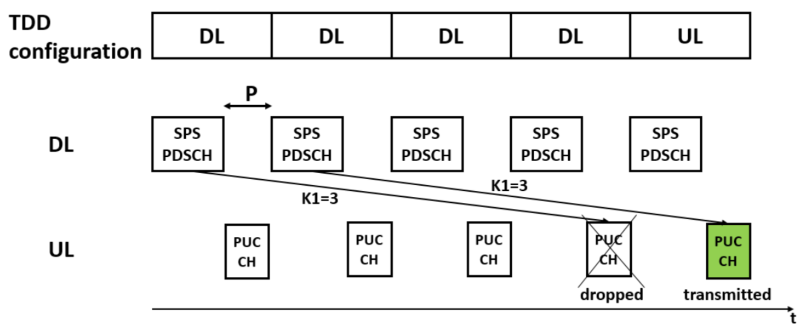


Figure 4.1: HARQ feedback cancellation in DL SPS transmission in TDD.

If the UEs drop HARQ feedback due to UL-DL slot/sub-slot conflict as shown in Fig. 4.1, the reliability of PDSCH degrades due to missing HARQ feedback and might not achieve the requirement because the gNB does not have the information to carry out the retransmissions in case of the first PDSCH's failure. On the other hand, if the gNB retransmits TB in case of feedback's absence, it causes a poor spectrum utilization due to the unnecessary retransmissions when the UE decodes correctly the first TB but an ACK feedback is dropped due to TDD conflict.

4.2 Related works

In [23], two options are supported. The first one is to defer HARQ-ACK until the first available valid UL slot (or PUCCH resource) but only semi-static TDD configuration is considered. The second one is to indicate K1 value for each SPS transmission but no scheme is mentioned.

In [24] and [25], two options are proposed to ensure the transmission of HARQ feedback. In the first option, multiple K1 values are indicated for one SPS configuration to guarantee SPS HARQ feedback to be pointed to an UL slot/sub-slot. However, the activation DCI has to carry multiple values of K1 so it causes an increase of overhead and a decrease of PDCCH reliability. In the second option, the UE postpones the collided SPS HARQ-ACK to the UL slot/sub-slot available for HARQ transmission. The UL slots/sub-slots are the semi-static UL slots/sub-slots that the gNB and the UE have an aligned understanding. This option is also proposed in [26]. However, this option is only for semi-static TDD configuration and does not discuss the UE behaviour in dynamic TDD configuration where the flexible slots/sub-slots can be updated to the UL slots/sub-slot by slot format indication (SFI).

In [27], it is argued that deferring HARQ feedback to the first available UL slot/sub-slot is a simple option but causes the imbalanced HARQ feedback because many HARQ feedback postponed in the previous slots might be transmitted in the same UL slot/sub-slot. Therefore, they propose that K1 is indicated to the UE in each DL SPS transmission. However, no scheme of K1 indication is proposed.

In [28], two options are proposed. The first one is to defer HARQ feedback to the next available UL slot/sub-slot within a higher-layer configured time window. There is no discussion about using the flexible slots/sub-slots switched to the UL slots/sub-slots by SFI. In the second option, HARQ slots/sub-slots for DL SPS are reserved periodically. When the HARQ occasion is not available in UL occasion, HARQ feedback is postponed to the latest periodically configured HARQ slot/sub-slot. Nevertheless, it consumes much resources for the reserved slots/sub-slots.

4.3 Enhancements for HARQ feedback in DL SPS transmission in TDD

4.3.1 Dynamically indication of K1 value for each DL SPS transmission

As discussed above, PDSCH-to-HARQ-feedback timing indicator is indicated to the UE by the activation DCI for DL SPS configuration transmitted from the gNB. The bits in the PDSCH-to-HARQ feedback timing indicator field specify a row index in the table configured by RRC parameter: *dl-DatatoUL-ACK in PUCCH-Config*. Based on the index from the activation DCI, the UE knows the value of K1 in the row corresponding to that index. In DCI format 1_0, 3 bits in the PDSCH-to-HARQ feedback timing indicator are mapped to K1 values from 1 to 8. In DCI format 1_1, there are from 0 bit to 3 bits used to indicate the row index of K1 from the mapping table. The number of bits is determined by $log_2 I$ where I is the number of rows in the table.

The value of K1 stays fixed throughout the SPS active time once the gNB activates DL SPS resources. DL SPS PDSCH is transmitted without an associated PDCCH so the gNB cannot change K1 at the time of PDSCH transmission to avoid the UL-DL slot/sub-slot conflict.

To overcome such shortcoming in DL SPS feedback mechanism, the SPS-PDSCH-to-HARQ-feedback timing K1 is indicated dynamically with each SPS transmission. In this case, the activation DCI may include a default value which may then be overridden by a later specific value. To avoid confusion, this initial value of K1 can be even removed from the activation DCI. Thus, a non-numerical value of K1 is used in the activation DCI. It is a signal to tell the UE that K1 value is indicated dynamically so it should expect and decode a specific numerical K1 value in the next data transmission. A row in the table specified in the *PUCCH-Config.dl-DatatoUL-ACK* has a non-numerical value. When the gNB chooses to transmit K1 dynamically in each SPS transmission, it sends the activation DCI with the sequence in the PDSCH-to-HARQ feedback timing indicator mapped to the row containing a non-numerical value. For example, as shown in Table 4.1, the gNB transmits a sequence 111 as PDSCH-to-HARQ-feedback timing indicator in the activation DCI to indicate a non-numerical value of K1 which is interpreted by the UE to instruct it to wait for a specific K1 value for each transmission. With the other sequences, the UE interprets the K1 value appropriately and applies that value for the relevant DL SPS resources.

Because there is no PDCCH associated with SPS PDSCH, dynamic K1 value utilized for a specific transmission is specified by PDSCH itself. Some bits (1-3 bits) in PDSCH are used to indicate K1 value. These bits are mapped to K1 value in a table as when K1 is indicated by DCI. Thus, the SPS-PDSCH-to-feedback-timing indication is transmitted with data on the PDSCH where the data packet is added the bits of K1 indication then they are encoded as one TB to be transmitted on the PDSCH. The embedded K1-indication bits exploit cyclic redundancy check (CRC) and redundant bits of the encoded data packet so the reliability of K1 detection is guaranteed without an increase of the separate CRC and redundant bits for K1 indication.

The operation of DL SPS transmission with dynamic K1 indication is shown in Fig. 4.2. At the beginning, an activation DCI specifying a non-numerical K1 is transmitted from the gNB to the UE to activate DL SPS

Table 4.1: Mapping table of PDSCH-to-HARQ-fedback timing indicator with a non-numerical K1 value

Row index	**PDSCH-to-HARQ-fedback timing indicator**	**K1 value**
0	000	1
1	001	2
2	010	3
3	011	4
4	100	5
5	101	6
6	110	7
7	111	FFFF

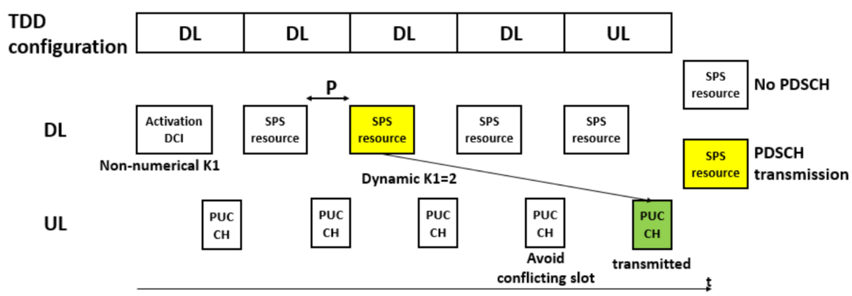


Figure 4.2: Dynamic K1 indication in DL SPS transmission.

configuration. When the gNB transmits PDSCH in DL SPS resource, it also adds dynamic K1 indication in PDSCH to indicate K1 value. In this case, with the latest information of slot configuration, the gNB chooses K1 to be 2 in order to avoid a DL slot and points the feedback to the closest available UL slot.

In TDD configuration, if the UE receives several SPS PDSCHs in the consecutive DL slots/sub-slots, the UE might have to transmit all the feedback of the previous SPS PDSCHs in the same PUCCH resource. However, multiplexing many feedback in the same PUCCH resource degrades reliability of feedback. Therefore, based on the size of PUCCH resources and the payload of feedback, the gNB should choose the dynamic values of K1 to avoid the multiplexing of feedback surpassing the capacity of PUCCH resources. As in Fig. 4.3, the feedback of SPS PDSCH 3 and 4 is dynamically indicated to be transmitted in PUCCH2 instead of PUCCH1 - the closest UL slot after the DL slots to avoid an overload in PUCCH1.

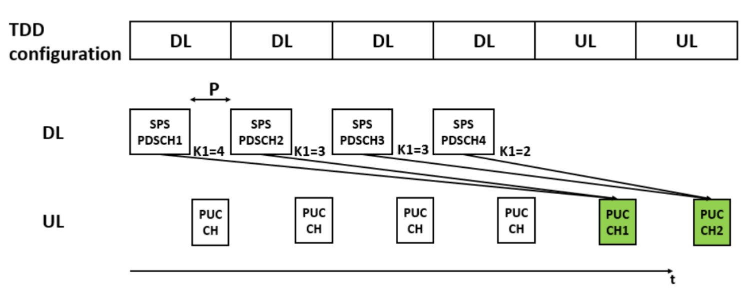


Figure 4.3: Dynamic K1 indication in the multiplexing of UL feedback for DL SPS transmission.

4.3.2 ACK-only feedback structure

A UE is required to transmit HARQ feedback for every DL SPS occasion. Even if there is no TB transmitted in a DL SPS occasion, the UE still transmits negative acknowledgement (NACK) feedback to the gNB. However, this scheme results in very high overhead with a shorter periodicity of DL SPS occasions. A further drawback arises when K1 is specified dynamically in PDSCH since the UE does not have a K1 value for an occasion with no transmission. To reduce the feedback burden and support dynamic K1 indication, an ACK-only feedback structure is used. The UE is configured to transmit only the ACK feedback if it decodes correctly PDSCH in the DL SPS occasions and obtains K1 value in PDSCH. In all other cases, no feedback is sent. The gNB is aware of which occasions it has made a transmission in and the K1 values set for those transmissions it makes. The gNB is thus aware when feedback should be received, and if no feedback is received at that time, can assume the transmission failed and make a retransmission.

The operation of the ACK-only feedback structure in the dynamic-K1-indication-by-PDSCH scheme is shown in Table 4.2.

Table 4.2: ACK-only feedback structure in DL SPS transmission

Case	**UE behavior**	**gNB behavior**	**Note**
No TB transmitted	No feedback	No action	No K1 is dynamically indicated to the UE
TB transmitted but not decoded	No feedback	Retransmit TB in SPS or dynamic resources (PDSCH associated with PDCCH)	K1 value is dynamically indicated to the UE in PDSCH but the gNB does not receive any feedback
TB transmitted and decoded correctly but ACK not decodable at the base station	ACK transmission	Retransmit TB in SPS or dynamic resources (PDSCH associated with PDCCH)	The UE transmits an ACK but the gNB fails to decode (due to bad channel conditions, etc). The gNB will retransmit but the UE will detect retransmission due to new data indicator (NDI) setting
TB transmitted and decoded correctly	ACK transmission	No action	K1 value is dynamically indicated to the UE in PDSCH, decoded correctly and used to determine the UL slot for ACK transmission

The ACK-only feedback structure is used when K1 is indicated dynamically. Dynamic-K1-indication scheme is activated if a non-numerical K1 value is indicated in the activation DCI. Therefore, a non-numerical K1 value in the activation DCI is interpreted by the UE as an indication to activate the ACK-only feedback structure in DL SPS transmission.

4.4 Numerical results

The first simulation is done to compare the reliability of the transmissions with an initial PDSCH and a potential retransmission in two schemes. In the conventional scheme, the feedback is cancelled if it is scheduled in a DL slot/sub-slot of TDD configuration. Therefore, if the first PDSCH is not decoded correctly by the UE, a retransmission is not done and the packet is lost. In the proposed scheme, the feedback's resources is indicated to avoid DL-UL slot/sub-slot conflict and an ACK-only feedback structure is used so a retransmission is carried out if necessary. A packet of 160 bits encoded by low-density parity-check (LDPC) code and modulated in quadrature phase shift keying (QPSK) is transmitted with MCS2 in the first transmission and with MCS1 in the retransmission following the failure of the first transmission in additive white Gaussian noise (AWGN) channel.

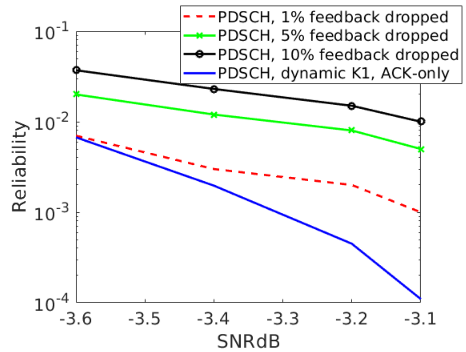


Figure 4.4: Performance of PDSCH with retransmission ensured in the proposed scheme and PDSCH with potential feedback cancellation.

In the conventional scheme, the percentages of the indicated resources for feedback in DL slot/sub-slot of TDD configuration are 1%, 5% and 10% corresponding to the rate of feedback cancellation. As be shown in Fig. 4.4, because of feedback cancellation, no retransmission is carried out in case of the first PDSCH's failure so PDSCH transmission's error (packet loss) increases remarkably. In contrast, PDSCH transmission with the proposed scheme achieves a much lower error rate. The benefits of the proposed scheme become more important with a high probability of slot/sub-slot conflict.

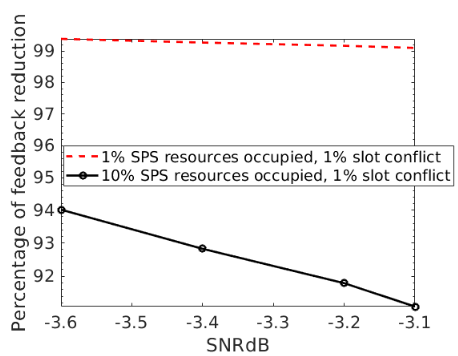


Figure 4.5: The reduction of feedback in ACK-only feedback structure.

The use of ACK-only feedback structure not only makes dynamic K1 indication feasible but also reduces significantly feedback overhead. Fig. 4.5 shows the percentage of feedback reduction of the ACK-only feedback structure compared to the conventional feedback structure. Two cases where 1% and 10% SPS resources are used for PDSCH transmission are considered. The number of feedback is counted after the first transmissions of the packet with MCS2 are done.

The second simulation is carried out to compare the performance (measured by BLER) of the conventional PDSCH without dynamic K1 indication embedded and the proposed PDSCH with dynamic K1 indication embedded in Section 4.3.1. In the first scenario of PDSCH without K1 indication, a packet of 160 bits is encoded by LDPC code and transmitted on PDSCH. In the second scenario of PDSCH with K1 indication, a packet of 160 bits is added 3 indication bits to create a TB with 163 bits. The TB is encoded by LDPC code and transmitted on PDSCH. In both scenarios, modulation technique is QPSK and channel is AWGN.

Fig. 4.6 shows the performance of PDSCH with MCS1 and MCS2. In the both two MCSs, a decrease of K1-embedded PDSCH's reliability is insignificant compared to a full PDSCH's reliability. The difference is negligible. Therefore, an use of dynamic K1 indication multiplexing with PDSCH does not affect the performance of PDSCH.

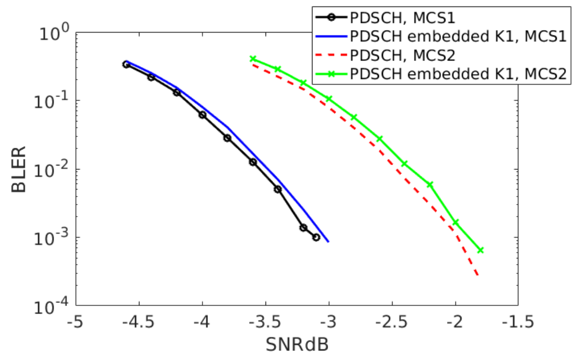


Figure 4.6: Performance of normal PDSCH and PDSCH embedded dynamic K1 indication.

The reliability of K1 detection is also the reliability of K1-embedded PDSCH as shown in Fig. 4.6. The scheme with PDSCH embedded with K1-indication bits makes K1 indication's detection achieve high reliability as PDSCH without an increase of separate CRC and redundant bits to encode the indication bits. The successful detection of K1 avoids an unnecessary PDSCH transmission and helps reduce resource consumption as well as increase spectrum utilization.

Fig. 4.7 compares BLER of a standard activation DCI containing one K1 (3 bits) with 24 data bits and 24 CRC bits used in the proposed method and BLER of an activation DCI containing multiple K1 with three bits for each K1 as proposed in [24] and [25]. 384 resource elements are used to transmit the DCI encoded by Polar code and modulated in QPSK in AWGN channel. The UE decoder is min-sum Successive cancellation list (SCL) decoder with list size 8.

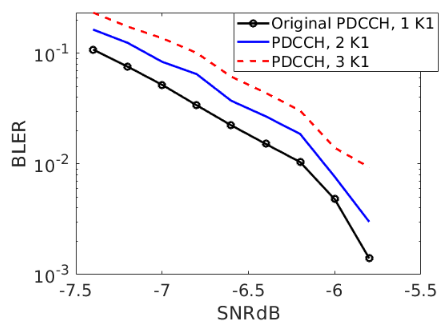


Figure 4.7: Performance of PDCCH with multiple K1 value.

The more K1 values DCI contains, the higher BLER PDCCH suffers from that makes the URLLC transmission unable to achieve the strict requirements and low spectrum efficiency.

Moreover, the method with multiple K1 indications in DCI of [24] and [25] only reduces the probability of the feedback conflict but cannot guarantee the transmission of feedback and the data retransmission. A set of K1 values cannot cover all the possibilities of UL slots/sub-slots' positions, especially if the slot format is updated periodically by DCI.

4.5 Conclusion

The work enhances HARQ feedback transmission for the DL SPS PDSCH transmission. A scheme consisting of a dynamic K1 indication embedded in PDSCH and an ACK-only feedback structure is presented.

The dynamic K1 indication embedded in PDSCH guarantees high reliability of K1 detection without an increase of CRC and redundant bits as well as ensures the feedback transmission. The ACK-only feedback structure assures the operation of the scheme with dynamic K1 indication, reduces feedback overhead and ensures the necessary retransmission.

Chapter 5

Load based channel access enhancements in unlicensed spectrum for NR URLLC transmissions

5.1 Load based channel access mechanism

The URLLC features in Release 15 and 16 are specified in licensed spectrum. Due to the new use cases in the industrial scenario, the operation of URLLC in unlicensed spectrum has become one of the main objectives in the ongoing Release 17.

In unlicensed spectrum, a transmitter is required to do Listen Before Talk (LBT) in order to acquire a channel prior to a transmission. There are two channel access mechanisms in LTE and 5G: load based equipment (LBE) and frame based equipment (FBE) [29]. In LBE, a transmitter attempts to access to a channel whenever it has data to transmit. While in FBE, a transmitter attempts to access to a channel and starts a transmission at the fixed occasions. This chapter focuses on LBE and presents a Markov chain model to analyze a 5G system with the transmitters using LBE. Based on the channel access latency calculated from the model, the enhancements of LBE are proposed to ensure the URLLC requirements of the transmission in unlicensed spectrum. FBE will be analyzed in Chapter 6 by another Markov chain model where all the transmitters use FBE. New schemes will be proposed in Chapter 6 to make the transmitters using FBE to access to the channel and transmit URLLC data achieve the URLLC requirements. Chapter 7 will present a model based on the models in Chapter 5 and Chapter 6 to analyze a system where the transmitters using LBE and FBE coexist.

In LBE, there are two types of LBT procedures: Type 1 channel access procedures used to acquire a channel in a new channel occupancy time (COT) (duration of maximum channel occupancy time (MCOT) is $T_{mcot,p}$ in Table 5.1 and Table 5.2) and Type 2 channel access procedures used to allow the UE/gNB to share the COT previously acquired by the gNB/UE.

Type 1 channel access procedures have two steps [9]. In the first step called initial Clear Channel Assessment (iCCA), the transmitter senses the channel in a defer duration T_d to be

$$T_d = t_f + m_p \times t_{sl}. \tag{5.1}$$

where t_f is 16 μs, t_{sl} is 9 μs that is duration of a sensing slot. m_p is the number of consecutive sensing slots depending on channel access priority class p defined in Table 5.1 and Table 5.2.

Table 5.1: Channel access priority class for DL

Channel access priority class (p)	m_p	$T_{mcot,p}$	**Allowed CW_p sizes**
1	1	2ms	{3, 7}
2	1	3ms	{7, 15}
3	3	8 or 10ms	{15, 31, 63}
4	7	8 or 10ms	{15, 31, 63, 127, 255, 511, 1023}

Table 5.2: Channel access priority class for UL

Channel access priority class (p)	m_p	$T_{mcot,p}$	**Allowed CW_p sizes**
1	2	2ms	{3, 7}
2	2	4ms	{7, 15}
3	3	6 or 10ms	{15, 31, 63, 127, 255, 511, 1023}
4	7	6 or 10ms	{15, 31, 63, 127, 255, 511, 1023}

Upon the success of the iCCA, the transmitter performs an extended Clear Channel Assessment (eCCA) in the second step. The transmitter senses the channel in N additional sensing slots. The counter N is a random number uniformly distributed between 0 and CW_p where CW_p is the value of contention window size in Table 5.1 and Table 5.2. The size of CW_p is set to be the smallest allowed value in a priority channel access class and might be increased to a next higher allowed value depending on the percentage of ACK feedback corresponding to the latest transmission burst that ACK feedback is available.

In each sensing slot, if the transmitter senses an idle channel, it decrements the counter by 1. If the transmitter senses a busy channel in a sensing slot, it goes to a defer stage. In this stage, it senses the channel in an additional defer duration. The transmitter exits this stage and decrements the counter by 1 if the channel is sensed to be idle during all sensing slots of the additional defer duration. Otherwise, the transmitter is still in the defer stage and has to sense the channel in another defer duration. When the counter reaches 0, the transmitter transmits the packet and occupies the channel in a duration of MCOT.

In Type 2 channel access procedures, when the receiver receives a packet from the transmitter in the COT initiated by the transmitter, it can use this COT to send data in the opposite direction. If the gap between received data and transmitted data is bigger than 16 us, the receiver must sense the channel as idle in 25 us before transmitting data in the opposite direction.

Type 1 channel access procedures are more complex and cause higher latency than Type 2 channel access procedures so this chapter focuses on the impact and the enhancements of Type 1 channel access procedures in unlicensed spectrum to ensure URLLC latency requirement.

5.2 Related works

[30] discusses URLLC in unlicensed spectrum with the potential enhancements in physical and Media Access Control layers to satisfy the URLLC requirements without going in detail the impact and enhancement of LBT procedures.

Several different types of LBT procedure are analyzed in [31], [32]. These types of LBT procedure do not contain defer duration after a busy sensing slot as Type 1 channel access procedures analyzed in this paper.

The expressions of LBT latency in [33], [34] neglect the contribution of the additional defer duration after a busy sensing slot. LBT latency based on these expressions has less impact on transmission latency so the channel and LBT conditions required to satisfy URLLC requirements cannot be calculated to be close to reality.

[35] does not provide in detail the transmission probability of a transmitter and a closed-form expression of latency of Type 1 channel access procedure.

5.3 Analysis of Type 1 channel access procedures in unlicensed spectrum

5.3.1 System model

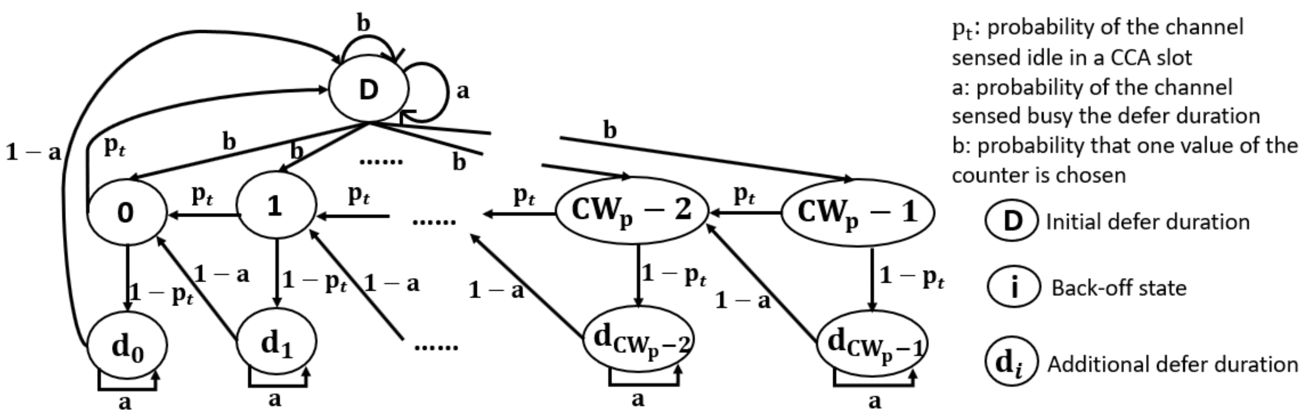


Figure 5.1: Markov chain for Type 1 channel access procedures.

Markov chain for Type 1 channel access procedures is shown in Fig. 5.1. At each transmission by a transmitter, regardless of the number of retransmissions, the probability that the channel is sensed to be busy in a sensing slot is assumed to be p_c. The probability that the channel is sensed to be idle in a sensing slot is $p_t = 1 - p_c$.

In a t_f gap of 16 μs, energy measurement is done for a total of at least 5 μs with at least 4 μs of sensing falling within the sensing slot of 9 μs immediately before the transmission [9]. Therefore, the idle and busy probabilities in a t_f gap are approximated to be p_t and p_c as in a sensing slot of t_{sl}, respectively.

In the Markov chain in Fig. 5.1, $a = 1 - p_t^{m_p+1}$ is the probability that the channel is busy in one of the sensing slots of the defer duration. Thereby, a is the probability of sensing a busy channel in a defer duration. $b = \frac{1-a}{CW_p+1}$ is the probability that one value of the counter N is chosen after the initial successful defer duration.

5.3.2 Probabilities of the states in Markov chain for Type 1 channel access procedures

We denote π_D, π_i, π_{d_i} where $i \in [0, CW_p - 1]$ to be stationary distribution of Markov chain in Fig. 5.1.

For the probability of the defer states from π_{d_0} to $\pi_{d_{CW_p-1}}$, we have

$$\pi_{d_i} = \frac{1-p_t}{1-a}\pi_i \qquad i \in [0, CW_p - 1]. \tag{5.2}$$

For the probability of the initial defer state π_D, we have

$$\pi_D = \frac{1}{1-a-b}\pi_0. \tag{5.3}$$

For the probability of the back-off states from π_0 to π_{CW_p-1}, we have

$$\pi_i = (1 - \frac{bi}{1-a-b})\pi_0 \qquad i \in [0, CW_p - 1]. \tag{5.4}$$

By applying the normalization condition, we have

$$\begin{aligned} 1 &= \pi_D + \sum_{i=0}^{CW_p-1} \pi_i + \sum_{i=0}^{CW_p-1} \pi_{d_i} \\ &= \frac{1}{1-a-b}\pi_0 + (1 + \frac{1-p_t}{d})\times \\ &\times (CW_p - \frac{b}{1-a-b}\frac{CW_p(CW_p-1)}{2})\pi_0. \\ \iff \pi_0 &= (\frac{1}{1-a-b} + (1+\frac{1-p_t}{d})\times \\ &\times (CW_p - \frac{b}{1-a-b}\frac{CW_p(CW_p-1)}{2}))^{-1}. \end{aligned} \tag{5.5}$$

From (5.2) to (5.5), we can calculate the probabilities of all states in Markov chain.

The transmitter transmits in a transmission occasion if the counter N is chosen randomly to be 0 immediately after the initial defer duration or a value N different to 0 is chosen randomly then the counter reaches zero after the back-off states. The probability that a transmitter transmits in a transmission occasion is

$$\begin{aligned} P_{Tx} &= \pi_0 + b\pi_D \\ &= (1 + \frac{b}{1-a-b})\pi_0. \end{aligned} \tag{5.6}$$

We consider a system with H transmitters. Every transmitter can detect each other. The hidden nodes do not impact the capability of channel access of a transmitter so they are excluded from the model. The transmitters share the same frequency resource in sub-6GHz bands, use omnidirectional sensing (omni-LBT) to sense and acquire a channel by following LBE Type 1 channel access procedures then transmit data by using omnidirectional transmission. A transmitter of interest encounters an idle channel in a sensing slot when no transmitter out of $H-1$ transmitters transmits in that sensing slot. We have the relation between p_t and P_{Tx}

$$p_t = (1 - P_{Tx})^{H-1}. \tag{5.7}$$

From (5.5) to (5.7), we can calculate p_t and P_{Tx}.

5.3.3 Transmitter's average channel access time in Type 1 channel access procedures

Based on the model of Type 1 channel access procedures in Section 5.3.1, the closed-form expression of the average time that a transmitter needs to access a channel in unlicensed spectrum at different probabilities p_t is derived.

In a defer duration, if the transmitter senses a channel and detects a busy sensing slot, it considers a busy defer duration and gets out of that defer duration then moves to sense another defer duration. We have T to be the average time that the transmitter spends in a busy defer duration:

$$\begin{aligned} T = &(1-p_t)t_f + p_t(1-p_t)(t_f+t_{sl}) + \\ &+ p_t^2(1-p_t)(t_f+2t_{sl}) + ... + p_t^{m_p}(1-p_t)(t_f+m_p t_{sl}). \end{aligned} \tag{5.8}$$

When a transmitter senses the channel in the defer duration, taking into account time spent in busy defer duration, the average time spent by the transmitter until it senses an idle defer duration (the channel is idle in the t_f gap and and all t_{sl} slots) and get out of the deferring state is

$$\begin{aligned} T_{D-out} &= p_t^{m_p+1}T_d + p_t^{m_p+1}(1-p_t^{m_p+1})(T_d+T) + \\ &+ p_t^{m_p+1}(1-p_t^{m_p+1})^2(T_d+2T) + ... \\ &= p_t^{m_p+1}T_d(1+(1-p_t^{m_p+1}) + (1-p_t^{m_p+1})^2 + ...) + \\ &+ p_t^{m_p+1}T((1-p_t^{m_p+1}) + 2(1-p_t^{m_p+1})^2 + ...) \\ &= \frac{p_t^{m_p+1}T_d}{1-(1-p_t^{m_p+1})} + \frac{(1-p_t^{m_p+1})T}{p_t^{m_p+1}} \\ &= T_d + \frac{T}{p_t^{m_p+1}} - T. \end{aligned} \tag{5.9}$$

The counter N decreases one unit in two cases. The first case is when the transmitter senses an idle channel in a sensing slot. The second case is when the transmitter senses a busy channel in a sensing slot then senses the channel to be idle in an additional defer duration. The average time for the counter to decrease by 1 unit is

$$T_{1-backoff} = p_t t_{sl} + (1-p_t)(t_{sl} + T_{D-out}). \tag{5.10}$$

The counter is counted down by sensing the channel from a value distributed uniformly between 0 and $CW_p - 1$ so the average value is

$$\tilde{N} = \frac{CW_p}{2} \tag{5.11}$$

The average time of the eCCA that the transmitter decrements the counter to 0 and acquires the channel to transmit data is:

$$T_{all-backoff} = \frac{CW_p}{2}(p_t t_{sl} + (1-p_t)(t_{sl} + T_{D-out})). \tag{5.12}$$

Type 1 channel access procedures consists of the iCCA and eCCA so the average time that the transmitter spends for Type 1 channel access procedures is

$$T_{LBT} = T_{D-out} + T_{all-backoff}. \tag{5.13}$$

5.3.4 DL and UL transmissions' latency in unlicensed spectrum

Type 1 channel access procedures' average time in (5.13) is applied to calculate DL and UL transmission time in unlicensed spectrum.

If the transmitter transmits a transport block and the receiver decodes correctly this transport block after one transmission, latency of the transmission in unlicensed spectrum is

$$T_{Tx} = T_{LBT} + T_{align} + T_{trans} + T_{gNB} + T_{UE}. \tag{5.14}$$

where T_{LBT} is time for channel access procedures at the transmitter, T_{align} is the alignment delay, T_{trans} is time length of a transmission occasion, T_{gNB} are T_{UE} are the processing time at the gNB and the UE.

When a packet arrives, it must wait until the beginning of a transmission occasion to be transmitted. The alignment delay is uniformly distributed among the symbols between two consecutive transmission occasion. Thus, T_{align} is $\frac{TTI}{2}$ where TTI is a transmission time interval. T_{trans} is one TTI for one transmission.

(5.14) provides latency of a single shot transmission for both DL and UL with any set of parameters specified in the system.

In DL transmission, transmission time for an initial transmission and a retransmission is

$$T_{Tx\_DL} = 2T_{Tx} + T_{HARQ}. \tag{5.15}$$

We have T_{HARQ} to be

$$T_{HARQ} = T_{LBT\_Type2} + T_{K1} + T_{UE} + T_{trans} + T_{gNB}. \tag{5.16}$$

where T_{LBT_Type2} is 25 μs that is the sensing duration of Type 2 channel access procedures defined in [9]. T_{K1} is the duration from the end of DL data to the beginning of feedback occasion.

The UE might be configured to transmit K repetitions of a transport block in the consecutive transmission occasions without feedback. This configuration increases the transmission time and processing time so margin to latency requirement decreases. Transmission time of an UL transmission with K repetitions is

$$T_{Tx\_UL} = T_{LBT} + T_{align} + K \times T_{trans} + K \times T_{gNB} + T_{UE}. \tag{5.17}$$

(5.17) reflects time that the UE spends with CG transmission when it does not take into account SR and UL grant.

To satisfy URLLC latency requirement, T_{Tx}, T_{Tx_DL} and T_{Tx_UL} must be smaller than 1 ms so it leads to constraint on T_{LBT} that the UE can spend for channel access procedures.

5.4 Conditions and enhancements in using Type 1 channel access procedures

5.4.1 Numerical results of the impact of channel access on URLLC transmission

Based on (5.13), the average time delay in Type 1 channel access procedures T_{LBT} is calculated. Fig. 5.2 shows the average time delay that the gNB needs to access to the channel in DL transmission with 4 priority classes in Table 5.1. Fig. 5.3 shows the average time delay that the UE needs to access to the channel in UL transmission with 4 priority classes in Table 5.2. When priority classes 3 and 4 are used, even only time delay in Type 1 channel access procedures exceeds the latency budget of 1 ms with low probability of sensing idle channel p_t.

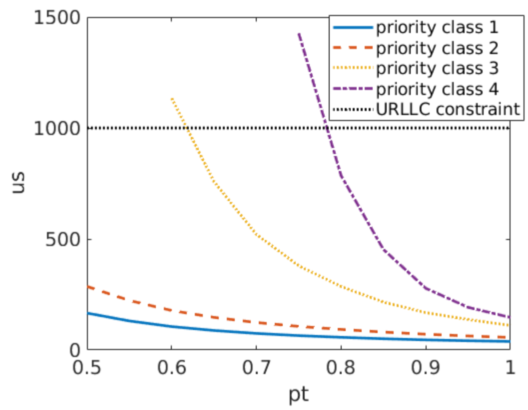


Figure 5.2: Type 1 channel access procedures average time in DL transmission.

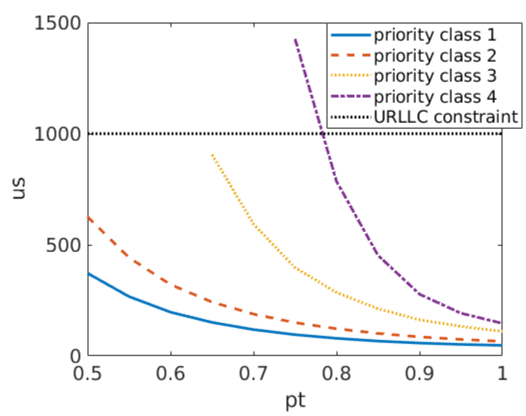


Figure 5.3: Type 1 channel access procedures average time in UL transmission.

Fig. 5.4 shows DL transmission time with an initial transmission and a retransmission as expressed in (5.15). TTI is 2 OFDM symbols. SCS is 30 kHz. The processing time at the gNB and the UE is 1 TTI. URLLC requirement of 1 ms is only met with the probability of sensing idle channel bigger than 0.7 for channel access priority class 1, 0.85 for channel access priority class 2.

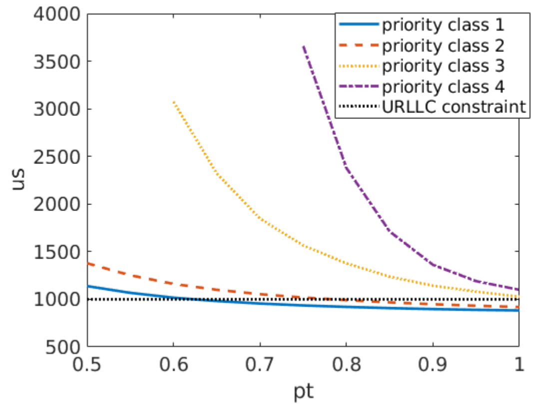


Figure 5.4: Total transmission time in DL transmission with an initial transmission and a retransmission in SCS of 30kHz.

Fig. 5.5 shows DL transmission time with an initial transmission and a retransmission similar to Fig. 5.4 but SCS is 60 kHz. URLLC requirement is only met with the probability of sensing idle channel bigger than 0.5 for channel access priority class 1, 0.55 for channel access priority class 2, 0.85 for channel access priority class 3, 0.9 for channel access priority class 4.

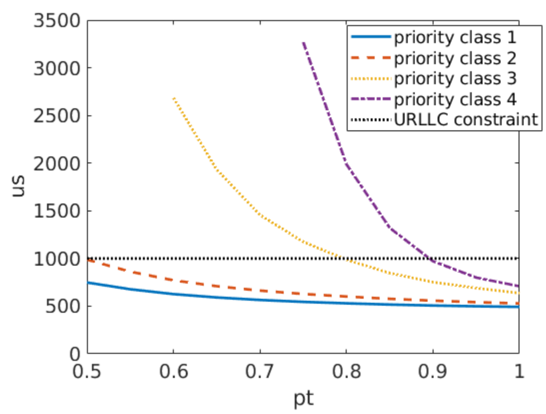


Figure 5.5: Total transmission time in DL transmission with an initial transmission and a retransmission in SCS of 60kHz.

Fig. 5.6 shows the total UL transmission time as expressed in (5.17). TTI is 2 OFDM symbols. SCS is 30 kHz. The processing time at the gNB and the UE is 1 TTI. The UE is configured to transmit 4 repetitions of a TB in 4 consecutive TTIs.

From Fig. 5.6, to meet latency budget of 1 ms, the probability of sensing idle channel must be bigger than 0.55 for channel access priority class 1, 0.65 for channel access priority class 2, 0.85 for channel access priority class 3, 0.9 for channel access priority class 4.

Fig. 5.7 shows the total UL transmission time of a transmission with 4 repetitions in SCS of 60 kHz. To

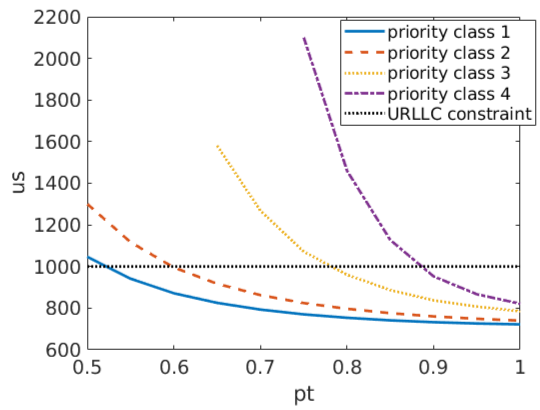


Figure 5.6: Total transmission time in UL transmission with 4 repetitions in SCS of 30kHz.

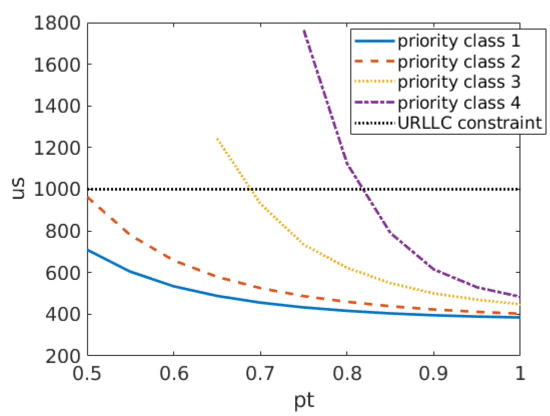


Figure 5.7: Total transmission time in UL transmission with 4 repetitions in SCS of 60kHz.

meet 1 ms, the probability of sensing idle channel must be bigger than 0.65 for channel access priority class 2, 0.7 for channel access priority class 3, 0.85 for channel access priority class 4.

As can be seen in Fig. 5.4, Fig. 5.5, Fig. 5.6 and Fig. 5.7, channel access priority classes 3 and 4 cause high latency in channel access and are not suitable for URLLC, especially with low probability of sensing idle channel. However, at this moment, channel access priority class is not dedicated to the specific services. Therefore, different channel access priority classes are proposed to be used for different services. Priority classes 3 and 4 in Table 5.1 and Table 5.2 that provide longer COT but higher latency of channel access are assigned to eMBB service because eMBB deals with the transmission with high payload without latency constraint. URLLC focusing on short time transmission is assigned to use priority classes 1 and 2 in Table 5.1 and Table 5.2 due to its shorter channel access time. Furthermore, URLLC with latency requirement of 1 ms does not need long COT. In DL transmission, the gNB chooses priority class dedicated to the arrival data of URLLC or eMBB service so that it can access to the channel and transmit data within the allowed requirements. In UL transmission, the UE distinguishes eMBB and URLLC to choose the appropriate priority class by decoding priority class field in UL grant from the gNB. When the gNB sends UL grant to schedule an UL transmission, it includes bits in UL grant to make the UE know the priority class of the transmission. In Type 1 CG transmission, all transmission parameters are configured by RCC signaling and there is no UL grant or activation DCI. In this case, the dedicated priority class for a specific service is configured by a RRC parameter.

5.4.2 New proposed tables of channel access priority class for URLLC DL and UL transmission

Based on the simulation results and prior discussion, we propose to define new set of access classes which are suitable for URLLC services, in terms of their physical characteristics and quality of service requirements. COT for priority classes dedicated to URLLC is assigned new values being smaller than the existing values. URLLC latency budget is 1 ms so COT in priority classes dedicated to URLLC should not be larger than 1 ms. A shorter COT reduces delay for other UEs in the system so the use of these new priority classes become fairer to the co-channel UEs. A smaller value of m_p is used in new priority classes for UL. New tables of channel access priority class with new entries dedicated to URLLC are shown in Table 5.3 and Table 5.4. New added entries are highlighted in bold text. When Table 5.3 and Table 5.4 are used, priority class 1, 2, 4, and 5 are assigned to URLLC service. The priority class chosen by the gNB is based in the channel condition of the probability of sensing idle channel to satisfy the URLLC latency requirement and guarantee a fair use of channel with other UEs by using a suitable COT.

Fig. 5.8 compares total UL transmission time with 4 repetitions in SCS of 30 kHz between the use of original URLLC dedicated priority class 1, 2 in Table 5.2 and the use of new URLLC dedicated priority class 1, 4 in Table 5.4. As can be seen, new priority classes bring a shorter transmission time and guarantee latency of 1 ms even with low probability of idle channel p_t.

5.5 Conclusion

This chapter presented Markov chain model for Type 1 channel access procedures in unlicensed spectrum and calculated the probability of each state and transmission probability. A closed-form expression of average time delay of Type 1 channel access procedures is derived and used to calculate DL and UL transmission time in unlicensed spectrum where channel access delay has an impact. To satisfy URLLC requirements, the specific priority classes are dedicated to URLLC service. The entries in Type 1 channel access procedures' tables are extended to distinguish and serve better different services.

Table 5.3: New channel access priority class for DL

Channel access priority class	m_p	$T_{mcot,p}$	**allowed** CW_p **sizes**
1	**1**	**0.5ms**	**{3, 7}**
2	**1**	**1ms**	**{3, 7}**
3 (original priority class 1)	1	2ms	{3, 7}
4	**1**	**0.5ms**	**{7, 15}**
5	**1**	**1ms**	**{7, 15}**
6 (original priority class 2)	1	3ms	{7, 15}
7 (original priority class 3)	3	8 or 10ms	{15, 31, 63}
8 (original priority class 4)	7	8 or 10ms	{15, 31, 63, 127, 255, 511, 1023}

Table 5.4: New channel access priority class for UL

Channel access priority class	m_p	$T_{mcot,p}$	**allowed** CW_p **sizes**
1	**1**	**0.5ms**	**{3, 7}**
2	**1**	**1ms**	**{3, 7}**
3 (original priority class 1)	2	2ms	{3, 7}
4	**1**	**0.5ms**	**{7, 15}**
5	**1**	**1ms**	**{7, 15}**
6 (original priority class 2)	2	4ms	{7, 15}
7 (original priority class 3)	3	6 or 10ms	{15, 31, 63, 127, 255, 511, 1023}
8 (original priority class 4)	7	6 or 10ms	{15, 31, 63, 127, 255, 511, 1023}

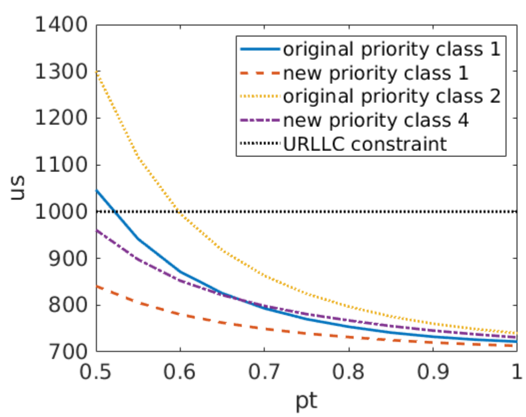


Figure 5.8: Comparison of total transmission time between original priority class and new priority class in UL transmission with 4 repetitions in SCS of 60kHz.

Chapter 6

Frame based channel access enhancements in unlicensed spectrum for NR URLLC transmissions

6.1 Frame based channel access mechanism

As mentioned in Chapter 5, LBE and FBE are two channel access mechanisms in LTE and 5G. Following the analysis of LBE in Chapter 5, this chapter focuses on FBE.

In FBE channel access mechanism, when a transmitter has data to transmit, it senses a channel to check its availability for a transmission only per fixed period called fixed frame period (FFP) with a duration of 1, 2, 2.5, 4, 5 or 10 ms. As shown in Fig. 6.1, a FFP consists of a channel occupation time (COT) and an idle period with the durations of T_{COT} and T_{idle}, respectively. The maximum duration of T_{COT} is 95% of T_{FFP}. The duration of T_{idle} is at least 5% of T_{FFP} but not smaller than 100 μs. In the idle period, there is a single observation slot where a CCA within T_{CCA} of 25 μs is performed. If a transmitter has data and senses an idle channel at a CCA occasion, the transmitter starts immediately a transmission at the beginning of a FFP after the end of that CCA occasion. The transmitter occupies the channel in T_{COT} and stops the transmission in T_{idle}. The transmitter can share the channel within T_{COT} with the receiver so that the receiver is also able to use that COT to transmit data. In contrast, if the transmitter senses a busy channel, it does not transmit data and has to wait until the CCA occasion in the next FFP to perform another channel sensing. In a network, both the gNB and the UE are capable of initiating its own COT.

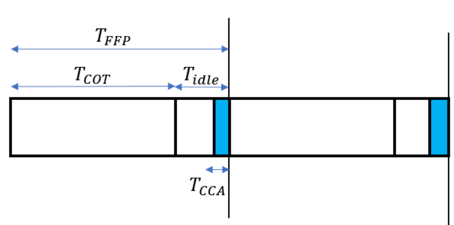


Figure 6.1: Fixed frame period in FBE.

A transmitter only senses the channel in the fixed occasions so the number of channel sensing is limited in an interval. This reduces channel access probability of the transmitter in a latency budget that is harmful to the transmission with a strict latency constraint as URLLC. Therefore, the performance of URLLC transmission in FBE will be analyzed in Section 6.3. The targeted scenario in this chapter is the industrial scenario in the factories where the absence of any other technologies sharing the channel such as WIFI and devices operating in LBE is guaranteed on a long-term basis. It can be done by the facility owner when he installs the devices to prevent an unexpected interference from other systems and radio access technologies. This environment is called an unlicensed controlled environment [46], [47]. The industrial scenario with an unlicensed controlled environment is chosen to follow the work of the ongoing 3GPP Release 17 where one of the objectives is the UL enhancements for URLLC in an unlicensed controlled environment [48]. However, this does not mean that LBT is not required prior to a transmission in an unlicensed controlled environment. Although the URLLC nodes work in a controlled environment without WIFI and devices operating in LBE, the uncertainty of FBE LBT due to a competition among the nodes still makes the URLLC nodes unable to attain the requirements. Subsequently, two methods to enhance URLLC performance in FBE are presented in Section 6.4 and Section 6.5.

6.2 Related works

[36] shows the throughput of a system using FBE on 5 GHz unlicensed band. [37] and [38] show the performance of FBE when LTE and WIFI coexist. These models in [36], [37] and [38] do not include the constraints of the URLLC transmission so no enhancements of FBE for URLLC operation are considered. In [39], the model to analyze the coexistence of LTE using FBE and WIFI is only for DL transmission from one base station to several UEs in unlicensed spectrum while UL data is transmitted in licensed spectrum.

[40] proposes “Enhanced FBE” and “Backoff and Idle Time Reduction FBE”. In Enhanced FBE, a backoff procedure used after a clear channel access (CCA) increases delay in the URLLC transmission. In Backoff and Idle Time Reduction FBE, the idle period is eliminated after an unsuccessful CCA so the transmitters sense the channel after the channel occupancy time in the next frame instead of waiting the entire frame. If the channel occupancy time is long, it still limits the number of channel sensing at the transmitter. Moreover, a backoff procedure is also used that increases latency.

[41] proposes that a transmitter senses periodically or continuously the channel in a subframe. The idle period is removed in this subframe compared to the original FBE frame. This method changes the design of the conventional frame that might make it incompatible in the system that the transmitters use both the conventional frame and the proposed frame. Moreover, to increase the sensing opportunities in the URLLC latency budget, the duration of subframe must be reduced that affects the duration of a transmission.

In [42] and [43], a transmitter might not transmit after a successful CCA if channel quality is bad to avoid a higher required transmission power. However, when a transmitter does not transmit after a successful CCA, it misses its transmission opportunity and without transmitting data, it has no chance to have a successful transmission of a packet in the URLLC latency budget. Therefore, the latency constraint of URLLC limits the use of this scheme.

In [44], a transmitter does channel sensing on several parallel channels and can switch among the channels to avoid the interference and the busy channels. The transmitter also can change the idle period in a frame when there is no available channel. This scheme consumes more resources for one transmission because the transmitter is scheduled with multiple resources in the parallel channels for a transmission instead of only one resource in a channel in the conventional scheme.

In [45], a central entity is used in a fully coordinated FBE approach to configure a common Time Division Duplex (TDD) configuration among the nodes in the system so that the UE’s UL transmission to a gNB is not blocked by the neighbor gNB DL transmission due to the misalignment of UL and DL slots among the gNBs. A common TDD configuration among the gNB nodes might not satisfy the specific requirements of each gNB network about the ratio of DL and UL transmissions.

6.3 Analysis of FBE in unlicensed spectrum

6.3.1 System model

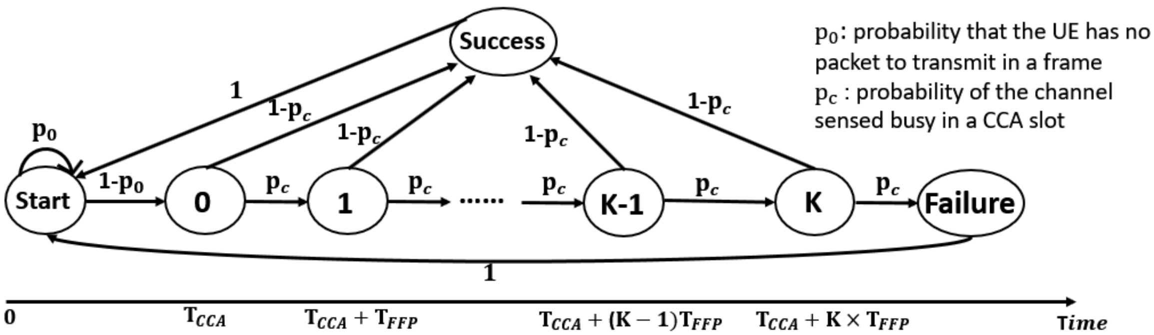


Figure 6.2: The Markov chain for FBE channel access.

FBE channel access mechanism is modeled by a Markov chain in Fig. 6.2. p_0 is the probability that a transmitter has no packet to transmit in a FFP. If the transmitter has a packet to transmit in a FFP, it jumps from State Start to State 0 with a probability of $1 - p_0$ and senses the channel in a CCA occasion at the end of that frame. Ignoring the alignment time, time consumption for the first channel sensing is T_{CCA}. If the channel is sensed idle, the transmitter obtains the channel to transmit at T_{CCA} and jumps to State Success with a probability of $1 - p_c$ then goes back to State Start where p_c is the probability of sensing a busy channel. State Success means that the transmitter succeeds in acquiring the channel. It does not means that after acquiring the channel, the transmitter transmits data and data is decoded correctly at the receiver. If the channel is busy, the transmitter must wait until the CCA occasion in the next FFP to continue to sense the channel and jumps from State 0 to State 1 with a probability of p_c. If the sensing is successful, the transmitter obtains the channel at $T_{CCA} + T_{FFP}$. The process continues until the transmitter accesses to the channel or gives up the sensing process for that packet due to a time constraint. In the model, State K presents the last channel sensing at the (K+1)th frame since the first sensing frame. K can be infinite.

The Markov chain in Section 6.3 is used to model FBE channel access mechanism of the URLLC transmitters or other types of the 5G NR transmitters coexisting with the URLLC transmitters in the 5G NR system. The transmitters share the same frequency channel in sub-6GHz bands, use omnidirectional sensing (omni-LBT) to sense and acquire a channel in FBE channel access mechanism then transmit data by using omnidirectional transmission. Every of the transmitters can detect each other. The hidden nodes are not included because they do not affect the ability of a transmitter to access to a channel that is the main focus of this Markov chain model. The transmitter cannot sense the hidden nodes so it is not blocked to access to the channel by these hidden nodes. The receivers use omnidirectional reception to receive data. The URLLC transmitters have the transmissions with the strict requirements of latency and reliability so they work in an unlicensed controlled environment to satisfy these requirements because the unlicensed controlled environment guarantees the absence of other access technologies such as WIFI and the devices in 5G network operating in LBE on a long-term basis [46], [47]. The unlicensed controlled environment is also applied in the scenarios in Section 6.4 and Section 6.5.

6.3.2 Probabilities of the states and channel access in Markov chain for FBE channel access

We denote $\pi_{Start}, \pi_{Success}, \pi_{Failure}, \pi_i$ where $i \in [0, K]$ to be stationary distributions of Markov chain in Fig. 6.2. We have

$$\pi_{Start} = \pi_{Success} + \pi_{Failure}. \tag{6.1}$$

The probability of State 0 is

$$\pi_0 = (1 - p_0)\pi_{Start}. \tag{6.2}$$

The probabilities from States 1 to States K are:

$$\pi_i = p_c^i \times \pi_0 \qquad i \in [1, K]. \tag{6.3}$$

The probability that a transmitter succeeds in obtaining a channel is also the probability that a transmitter sends data. The probability of a channel sensing success $\pi_{Success}$ is

$$\begin{aligned}\pi_{Success} &= \pi_0(1 - p_c)\sum_{i=0}^{K} p_c^i \\ &= \pi_0(1 - p_c^{K+1}) \\ &= \pi_{Start}(1 - p_0)(1 - p_c^{K+1}).\end{aligned} \tag{6.4}$$

The probability is calculated when a packet already entered the process so π_{Start} equals to 1. In other words, the transmission probability is calculated given that π_{Start} is 1. By substituting $\pi_{Start} = 1$ to (6.4), we have the probability that a transmitter transmits data

$$P_{trans} = (1 - p_0)(1 - p_c^{K+1}). \tag{6.5}$$

When a transmitter has a packet and senses the channel to transmit that packet, the probability that the transmitter fails to obtain the channel after all allowed channel sensing and must drop the packet is

$$P_{failure} = p_c^{K+1}. \tag{6.6}$$

6.3.3 Relation between the probability of no data and the probability of sensing a busy channel

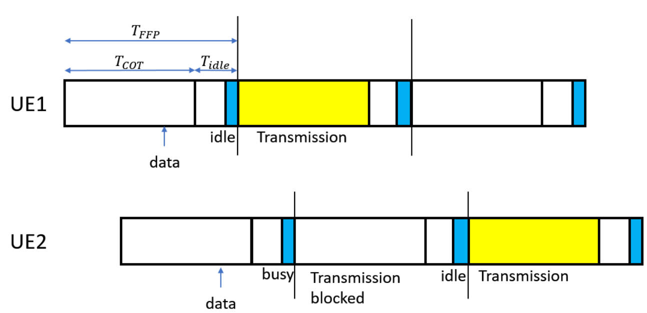


Figure 6.3: Channel sensing and data transmission in FBE.

The FFP of each transmitter can be configured with an offset to start at different time in the same frequency resources as Fig. 6.3 so a CCA occasion of a transmitter might overlap with a COT of another transmitter. If a transmitter is sending data in a COT then another transmitter senses the channel at that time, the channel is sensed busy as the case of UE2 in Fig. 6.3. The UE2 must wait the next FFP to sense the idle channel and transmit data. This offset ensures two UE not to sense an idle channel then transmit at the same time so as to avoid the interference between two simultaneous transmissions causing a degradation of transmission's reliability.

We consider a system that has Q transmitters with the same $K+1$ sensing states but different starting points of FFPs. These transmitters belong to a 5G NR network and send data in the same frequency resource in an unlicensed controlled environment. A CCA occasion of a transmitter overlaps with the COTs of the other transmitters. A transmitter of interest senses an idle channel in a CCA occasion if there is no UE out of $Q-1$ UE transmitting at that time. We have the relation between p_0 and p_c by using the transmission probability of a transmitter in (6.5)

$$p_c = 1 - (1 - (1 - p_0)(1 - p_c^{K+1}))^{Q-1}. \tag{6.7}$$

From (6.7), p_c is calculated when p_0, K, Q are known.

6.3.4 URLLC operation with FBE in unlicensed spectrum

As shown in Fig. 6.3, a transmitter does channel sensing to transmit a packet in the fixed moments with the gap of T_{FFP} between two consecutive moments. The URLLC transmission has a latency budget of 1 ms while the smallest duration of a fixed frame period T_{FFP} is 1 ms. Therefore, in the URLLC latency budget, the transmitter only can do one channel sensing because the second channel sensing is at $T_{CCA} + T_{FFP}$ that is bigger than 1 ms. If it fails in this only chance, it cannot access to the channel to transmit data in the URLLC latency requirement and the URLLC packet is dropped.

When a transmitter has a URLLC packet with only one chance of channel sensing, the value of K in the Markov chain is 0. Substituting $K = 0$ to (6.6), we have the probability that the transmitter fails to access to a channel in order to transmit an URLLC packet is:

$$P_{failure\_URLLC} = p_c. \tag{6.8}$$

When Q transmitters are configured to transmit the URLLC packets in a controlled environment, from (6.7), we have the relation between p_0 and p_c of the URLLC transmitters.

$$p_c = 1 - (1 - (1 - p_0)(1 - p_c))^{Q-1}. \tag{6.9}$$

The limit of channel sensing opportunity due to the time constraint increases the probability of channel access failure. It causes an increase of the dropped packets and reduces the URLLC transmission's reliability.

This section provided a Markov chain to model channel access in unlicensed spectrum when the transmitters in the system use FBE to acquire a channel. Subsequently, the URLLC latency constraint is applied to the model to show the impact of channel access in FBE on URLLC performance.

6.4 Multiple configurations of FFP in FBE for URLLC in unlicensed spectrum

6.4.1 Multiple configurations of FFP

The conventional FBE scheme provides only one opportunity of channel sensing in the URLLC transmission due to the time constraint. To provide more opportunities of channel access in the URLLC latency budget of 1 ms and reduce the errors due to the packets dropped out of the latency budget in UL transmission, we propose a new FBE scheme where a transmitter is configured with multiple configurations of FFP. These FFP configurations are in the same frequency resources but overlap in time and have different

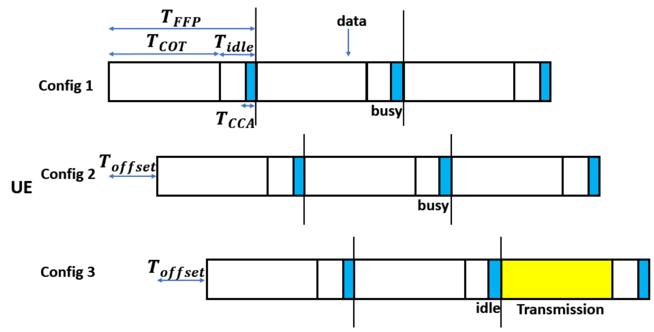


Figure 6.4: Multiple configurations of FFP.

starting points so they do not require the additional resources compared to the conventional scheme with one FFP configuration as shown in Fig. 6.4. The number of FFP configurations that a transmitter can use is configured by the gNB through DCI or RRC based on arrival rate of data, the probability of channel access, the priority and requirements of data transmission. The beginning of a FFP in a configuration is shifted an amount of time defined as T_{offset} compared to the beginning of a FFP in the previous configuration. T_{offset} is configured by gNB through DCI or RRC based on the number of configurations that a transmitter has.

The transmitter with high priority data such as URLLC uses multiple FFP configurations in the attempts to access to the channel and transmit a high priority packet while the transmitter with low priority data uses only one FFP configuration in the attempts to access to the channel and transmit a low priority packet. This means that a transmitter might have multiple FFP configurations but only uses one configuration in the attempts to access to the channel and transmit low priority data as the conventional FBE scheme. FFP periodicity in all configurations of a transmitter is the same and configured by the gNB through DCI or RRC.

When a transmitter has high priority data such as URLLC, it senses a channel in the CCA occasions of different FFP configurations. It starts to sense the channel from the configuration with the closest CCA occasion from the arrival time of data in order to reduce the waiting time. If it fails to access to the channel in a configuration, it does another attempt in the closest CCA occasion of the next configuration instead of waiting one frame period in the same configuration as the conventional scheme. Subsequently, if it succeeds in channel access in a configuration, it uses that configuration to start the transmission in T_{COT} of a FFP immediately after that successful CCA occasion. This COT is also shared with the receiver to transmit data in the opposite direction.

As shown in Fig. 6.4, a UE is configured with three FFP configurations having different starting points in the same frequency resources. When data arrives, the UE starts to do channel sensing in a CCA occasion of the first configuration because this configuration has the closest CCA occasion from the moment that data arrives. This reduces the waiting time before the first sensing. The UE senses a busy channel in the CCA occasion of the first configuration then it moves to the second configuration to do channel sensing in the next CCA occasion. Consequently, the channel is still busy so the UE moves to the third configuration to do channel sensing. This time the channel is idle so the UE chooses this configuration to transmit data. Data is transmitted immediately in a FFP after the successful CCA. With the multiple FFP configuration scheme in this example, the UE has three opportunities of channel sensing in an interval of T_{FFP} instead of only one opportunity in the conventional scheme.

6.4.2 The Markov chain of FBE channel access with multiple configurations of FFP

Fig. 6.5 shows the Markov chain of FBE channel access with multiple FFP configurations. n is the number of the FFP configurations that a transmitter uses to do channel access for a packet's transmission. T_{offset}

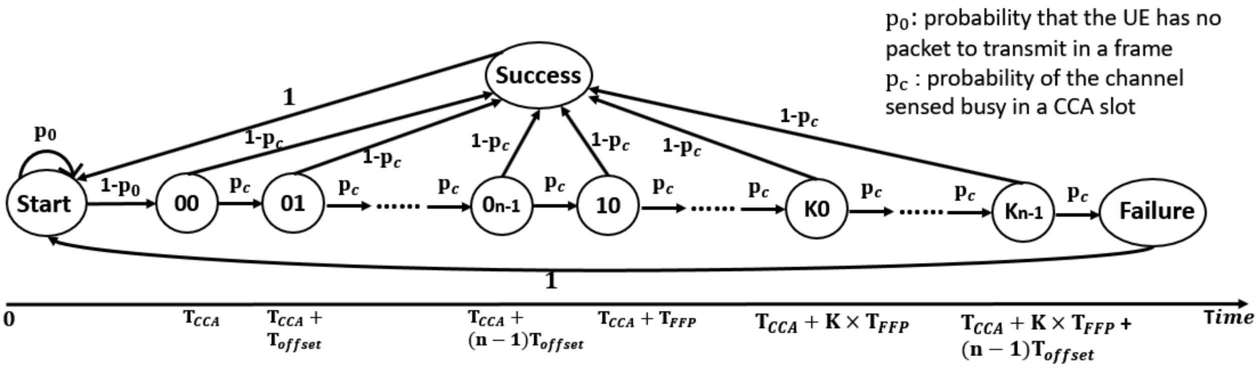


Figure 6.5: The Markov chain for multiple configurations of FFP in FBE channel access.

equals to $\frac{T_{FFP}}{n}$. Other parameters are defined in Section 6.3.1. A transmitter has a packet and jumps from State Start to State 00 with a probability of $1 - p_0$. State 00 presents the first sensing frame in the first chosen configuration. If the channel is idle, the transmitter jumps from State 00 to State Success with a probability of $1 - p_c$ then moves back to State Start. State Success means that the transmitter succeeds in acquiring the channel. It does not means that after acquiring the channel, the transmitter transmits data and data is decoded correctly at the receiver. If the channel is busy, the transmitter jumps from State 00 to State 01 with a probability of p_c. State 01 presents the first sensing frame in the next configuration. If the channel continues to be busy, the transmitter goes to the next states. After going through the first sensing frame of all configurations, the transmitter comes back to the first chosen configuration and senses the channel in the second sensing frame as presented by State 10. The process continues until the channel is obtained successfully and the transmitter jumps to State Success or all CCAs in the allowed time fail and the transmitter jumps to State Failure. Finally, the transmitter comes back to State Start.

Following the same calculations in Section 6.3.2, when π_{Start} equals to 1 for a packet that already entered the process, the probability of transmission for a transmitter using FBE with multiple FFP configurations is

$$P_{trans\_nconfig} = (1 - p_0)(1 - p_c^{n(K+1)}). \tag{6.10}$$

The probability of channel access failure when the transmitter has a packet to transmit is

$$P_{failure\_nconfig} = p_c^{n(K+1)}. \tag{6.11}$$

If Q transmitters have n configurations in a controlled environment, the relation between p_c and p_0 is

$$p_c = 1 - (1 - (1 - p_0)(1 - p_c^{n(K+1)}))^{Q-1}. \tag{6.12}$$

From (6.12), p_c is calculated when p_0, K, n, Q are known.

In the multiple FFP configuration scheme, a transmitter has maximum n attempts to access to a channel from T_{CCA} to $T_{CCA} + T_{FFP}$ (channel sensing at $T_{CCA} + T_{FFP}$ is not counted). Therefore, a transmitter with an URLLC packet has maximum n opportunities of channel sensing in the URLLC latency budget of 1 ms with T_{FFP} of 1 ms instead of only one opportunity in the conventional FBE scheme. We have m ($m \leq n$) to be the number of FFP configurations that the transmitter can use in the URLLC latency budget of 1 ms. The probabilities of transmission and channel sensing failure of a URLLC transmitter are

$$P_{trans\_nconfig\_URLLC} = (1 - p_0)(1 - p_c^m). \tag{6.13}$$

$$P_{failure\_nconfig\_URLLC} = p_c^m. \tag{6.14}$$

The relation between p_c and p_0 for Q URLLC transmitters that can use m configurations in the URLLC latency budget of 1 ms in a controlled environment is

$$p_c = 1 - (1 - (1 - p_0)(1 - p_c^m))^{Q-1}. \tag{6.15}$$

Multiple FFP configurations in FBE mitigate channel access failure in the URLLC transmission due to the time constraint. Moreover, the multiple FFP configuration scheme in FBE also reduces the alignment time compared to the conventional FBE with one configuration. When a packet arrives, a transmitter must wait until the closest CCA occasion to do channel sensing. The alignment time is uniformly distributed among two consecutive CCA occasions. For the conventional FBE scheme, the alignment time is $\frac{T_{FFP}}{2}$. While the alignment time is $\frac{T_{offset}}{2}$ for the multiple FFP configuration scheme. Because T_{offset} between two configurations is smaller than T_{FFP}, the alignment time of the multiple FFP configuration scheme is smaller than that of the conventional scheme. It benefits the URLLC transmission with a strict latency requirement. On the other hand, in the multiple FFP configuration scheme, the receiver must detect blindly the presence of a transmission more than one time in a duration of T_{FFP}. The receiver has to check each configuration to see whether there is a transmission or not because of the uncertainty of channel sensing and arrival of data at the transmitter. The number of blind detections in a duration of T_{FFP} is equal to the number of the FFP configurations. Therefore, to reduce the burden and energy consumption at the receiver, only the high priority transmitters such as URLLC type are allowed to use multiple FFP configurations in channel sensing and transmission.

6.5 FFP arrangement based on the transmitter's priority

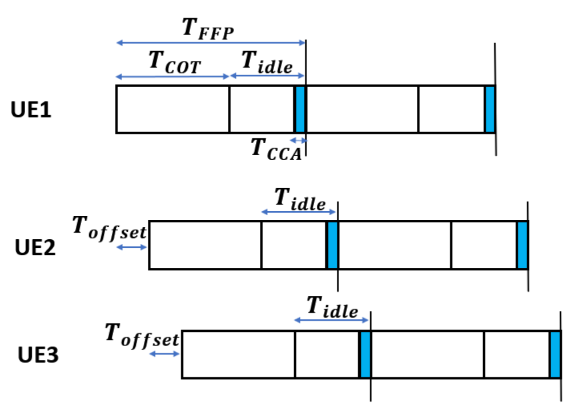


Figure 6.6: FFP arrangement in FBE based on the UE's priority.

In Section 6.3 and Section 6.4, the transmission in one COT initiated by a transmitter might block any transmitter in the network to initiate its own COT because a CCA occasion of a transmitter overlaps with a COT of another transmitter. It results in the same probability of sensing a busy channel for all transmitters in the network. However, in case there are the transmitters with different priorities such as latency and reliability in a network, it would be better if the high priority transmitters have a higher chance to access to the channel than the low priority transmitters. Therefore, we propose another FBE scheme in this section to support the UE with different priorities. FFP in each transmitter is configured with an offset so that a CCA occasion of a high priority transmitter overlaps with an idle period of a low priority transmitter. In other words, a low priority transmitter stops the transmission before a CCA occasion of a high priority transmitter so the high priority transmitter senses an idle channel to transmit data whenever it has data to transmit. While a CCA occasion of a low priority transmitter overlaps with a COT of a high priority transmitter so the low priority transmitter might be blocked to initiate its own COT. In this scheme, a transmitter is only blocked to initiate its COT by other COT initiators with higher priority and not blocked by other COT initiators with lower priority. The transmitters belong to 5G NR network and send data in the same frequency resources in an unlicensed controlled environment without interference from WIFI and other devices operating in LBE.

The priorities of the UE transmitters are defined based on the latency and reliability requirements of data that they transmit. Following the 3GPP standards, the priority of data is defined by two methods: a priority indicator field in DCI or RRC. A priority indicator field is added in the new DCI format to indicate the priority of the scheduled data on physical uplink shared channel (PUSCH). However, in CG PUSCH, priority of data on PUSCH is configured by RRC and is not written by the activation DCI.

Fig. 6.6 shows an example of a FFP arrangement based on the UE's priorities. The UE1 has the highest priority so its CCA occasion is configured to overlap with the idle periods of the UE2 and UE3. The UE1 always senses an idle channel to transmit data and fulfil its requirements. The UE2 has lower priority than the UE1 but higher priority than the UE3. Its CCA occasion is set by T_{offset} to overlap with the UE1's COT and the UE3's idle period. Due to this arrangement, the UE2 might be blocked to initiate its COT by the UE1 but is not affected by the UE3. Finally, the lowest priority UE3 might be blocked to initiate its COT by the other UE because its CCA occasion overlaps with the UE1 and UE2's COT.

The system model with the parameter p_0 defined in Section 6.3 is used to calculate the channel blocking probability of each UE in the FFP arrangement scheme based on the UE's priority. We extend the system in Fig. 6.6 to a system with Q transmitters. In this system, the UE1 has an absolute priority and is not blocked by the transmissions of the other UE in the network so the probability that the UE1 senses a busy channel is $p_{c1} = 0$.

The UE2 has lower priority than the UE1 so it might be blocked by the UE1's transmission. However, it has higher priority than the other UE except the UE1 and is not blocked by those transmissions. The probability that the UE2 senses a busy channel is

$$\begin{aligned} p_{c2} &= 1 - (1 - (1 - p_0)(1 - p_{c1})) \\ &= 1 - p_0. \end{aligned} \tag{6.16}$$

The UE3 might be blocked by the UE1 and UE2's transmissions with higher priorities so the probability that the UE3 senses a busy channel is

$$\begin{aligned} p_{c3} &= 1 - (1 - (1 - p_0)(1 - p_{c1}))(1 - (1 - p_0)(1 - p_{c2})) \\ &= 1 - p_0(1 - (1 - p_0)p_0). \end{aligned} \tag{6.17}$$

From (6.16) and (6.17), we can derive a general equation of the channel blocking probability for the i^{th} UE in the FFP arrangement scheme based on the UE's priority

$$p_{ci} = 1 - \prod_{j=1}^{i-1} (1 - (1 - p_0)(1 - p_{cj})). \tag{6.18}$$

The URLLC UE has only one chance to do channel sensing so the arrangement to obtain $p_c = 0$ for the URLLC UE guarantees the URLLC transmission in the latency budget. Therefore, this scheme benefits the URLLC UE where the URLLC UE coexists with other lower priority UE.

To arrange the CCA occasions of the high priority UE to overlap with the idle periods of the low priority UE, the requirement of the idle period T_{idle} must be stricter than the current requirement of T_{idle} where T_{idle} is at least 5% of T_{FFP} but not smaller than 100 μs. T_{idle} in a network with Q UE in the FFP arrangement scheme based on the UE's priority must satisfy

$$T_{idle} > max\{(Q-1)T_{offset} + T_{CCA}, 100\mu s\}. \tag{6.19}$$

T_{idle} increases when the number of the UE in the network increases. When T_{FFP} does not change, it results in a decrease of the transmission time T_{COT}. Therefore, the UE cannot transmit a long packet but has to fragment it into small segments and transmits them in the consecutive FFPs. Latency increases because there are more idle periods in the transmission of this long packet. Moreover, the UE needs to do channel access between FFPs that increases latency due to the uncertainty of channel access. Therefore, the arrangement of FFP to guarantee the transmission of the high priority UE is suitable to a network with a small number of the UE having different priorities and short packets to transmit.

6.6 Numerical and simulation results

Table 6.1: Simulation parameters for Fig. 6.7, Fig. 6.8, Fig. 6.9, Fig. 6.10, Fig. 6.11 and Fig. 6.12

Parameters	**Values**
Fixed frame period	1 ms
Channel occupancy time	900 μs
Number of configurations per UE	1-4
p_0	0.99, 0.95
Number of simulated frames	10^{10}
Bandwidth	20, 80 MHz

In this section, the analytic results are verified by the MATLAB simulations. The performance of channel access in three different schemes of FBE analyzed in Sections 6.3, 6.4 and 6.5 are compared: the conventional scheme where each transmitter uses one FFP configuration to sense a channel and transmit data, the proposed scheme where each transmitter uses multiple FFP configurations to sense a channel and transmit data, the proposed scheme where FFPs for the transmitters in the system are arranged based on their priorities.

The parameters in Table 6.1 are used for the simulations in Fig. 6.7, Fig. 6.8, Fig. 6.9, Fig. 6.10, Fig. 6.11 and Fig. 6.12. In these simulations, UL transmission is carried out where the transmitters being the UE transmits data to the receiver being the gNB in a single cell. The transmitters use omnidirectional sensing to sense and acquire a channel in FBE channel access mechanism then transmit data by using omnidirectional transmission. Every of the transmitters can detect each other. The hidden nodes are not included because they do not affect the ability of a transmitter to access to a channel that is the main focus of the proposed scheme. The transmitter cannot sense the hidden nodes so it is not blocked to access to the channel by these hidden nodes. The receivers use omnidirectional reception to receive data.

The FFP is set to 1 ms with COT of 900 μs. LBT channel access mechanism is done per a bandwidth of 20 MHz so the UE in the simulations uses the same channel with bandwidth of 20 MHz to do channel sensing and transmit data. This means that the number of the UE in the following graphs represents the number of the UE in one frequency resource unit of 20 MHz. This result can be extended to other systems with different bandwidth. In one system, if a bandwidth of 20 MHz is divided into several interlaces and each UE transmits in an interlace, the total number of the UE in the system is the product of the number of the UE shown in Fig. 6.7 and Fig. 6.8 and the number of interlaces in a bandwidth of 20 MHz. On the other hand, if NR-U carrier is wide band that is a multiple of 20 MHz and each UE in the system works in a channel band of 20 MHz, the total number of the UE in the system is a multiple of the number of the UE shown in Fig. 6.7 and Fig. 6.8. The results of a wide band system with a bandwidth of 80 MHz are shown in Fig. 6.9 and Fig. 6.10. The UE is assumed to transmit the URLLC packets that have a low arrival rate where p_0 is set to 0.99 or 0.95.

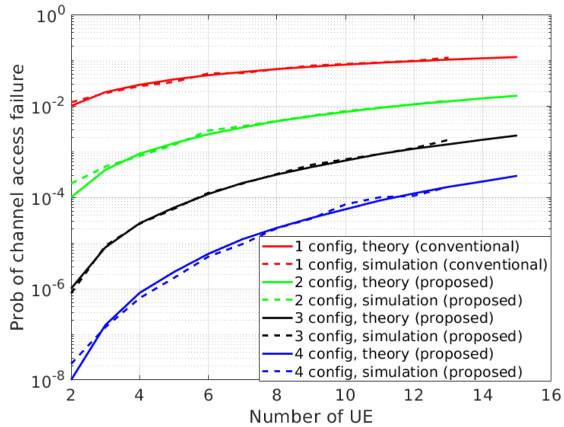


Figure 6.7: Channel access failure's probability in FBE in the conventional one-configuration scheme and the proposed multiple-configuration scheme with $p_0 = 0.99$ and bandwidth of 20 MHz.

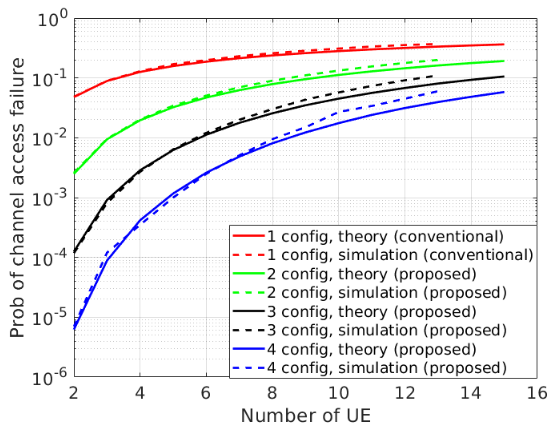


Figure 6.8: Channel access failure's probability in FBE in the conventional one-configuration scheme and the proposed multiple-configuration scheme with $p_0 = 0.95$ and bandwidth of 20 MHz.

Fig. 6.7 and Fig. 6.8 show the performance of channel access with different number of FFP configurations per UE in a bandwidth of 20 MHz where p_0 is set to 0.99 and 0.95, respectively. In Fig. 6.7, the conventional scheme where each UE uses only one FPP configuration in channel sensing has a high probability of channel sensing failure. This probability increases rapidly when the number of the UE in the system increases. Even if there are only two UE, the blocking probability of each UE is 10^{-2} that is much higher than URLLC reliability requirement. Fig. 6.8 shows a similar result for one configuration at p_0 of 0.95. Therefore, the conventional scheme is not suitable for the URLLC transmission.

Fig. 6.7 also shows the blocking probability in the multiple FFP configuration scheme at p_0 of 0.99 to compare with the blocking probability in the conventional scheme with one FFP configuration. If each UE has two configurations, the blocking probability of 10^{-4} for each UE in the system of two UE is much smaller than that of the conventional scheme. When more FFP configurations per UE are used, the UE achieves a smaller blocking probability even if there is a bigger number of the UE in the system. The benefit of multiple FFP configurations is also shown in Fig. 6.8, although the probability of channel access failure in Fig. 6.8 is higher than that in Fig. 6.7 because of a higher data rate. Therefore, this scheme is suitable for a system where several URLLC UE coexist.

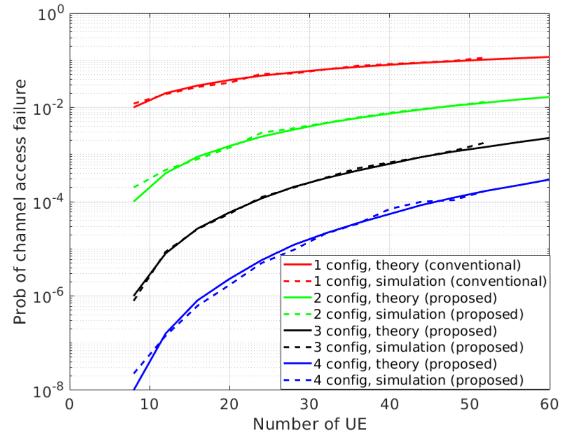


Figure 6.9: Channel access failure's probability in FBE in the conventional one-configuration scheme and the proposed multiple-configuration scheme with $p_0 = 0.99$ and bandwidth of 80 MHz.

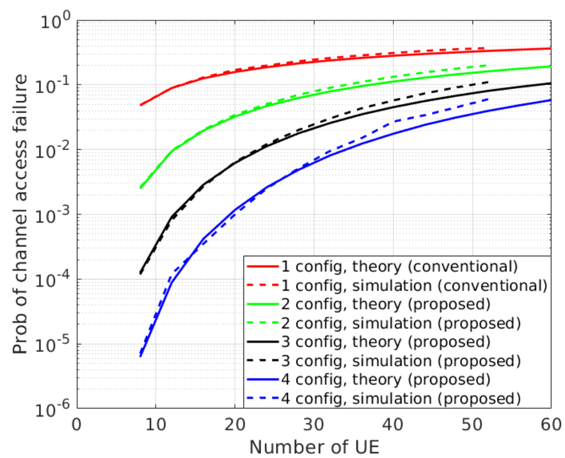


Figure 6.10: Channel access failure's probability in FBE in the conventional one-configuration scheme and the proposed multiple-configuration scheme with $p_0 = 0.95$ and bandwidth of 80 MHz.

In [40], when a UE senses a busy channel in a CCA occasion of 25 μs and cannot acquire the channel, the idle period in the following FFP is removed so the UE can sense the channel after the channel occupancy time instead of waiting the entire FFP. Channel occupancy time in Table 6.1 is 900 μs. This means that after an unsuccessful CCA, the UE does the second channel sensing after 900 μs. Therefore, within the URLLC latency budget of 1 ms, an URLLC UE has maximum two sensing opportunities. Similarly, in [41], the idle period does not exist in a frame so the UE also has maximum two channel sensing opportunities in the latency budget of 1 ms. Fig. 6.11 and Fig. 6.12 compare the performance of the schemes in [40], [41] with the proposed multiple FFP configuration scheme. As can be seen in the figures, the performance of the schemes in [40], [41] is equivalent to the proposed scheme with two FFP configurations. The probability of channel access failure in these schemes is still higher than the URLLC requirements. When the number of the FFP configurations increases to three and four, the multiple FFP configuration scheme achieves a better performance with the lower probabilities of channel access failure. The multiple FFP configuration scheme allows the number of the FFP configurations to be modified flexibly based on channel condition and data requirements without affecting the transmission duration in COT while the schemes in [40], [41] must reduce COT to increase channel sensing opportunities.

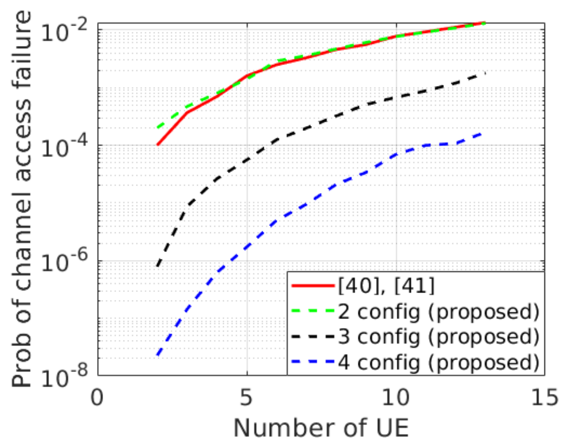


Figure 6.11: Channel access failure's probability in the schemes of [40], [41] and the proposed multiple-configuration scheme $p_0 = 0.99$ and bandwidth of 20 MHz.

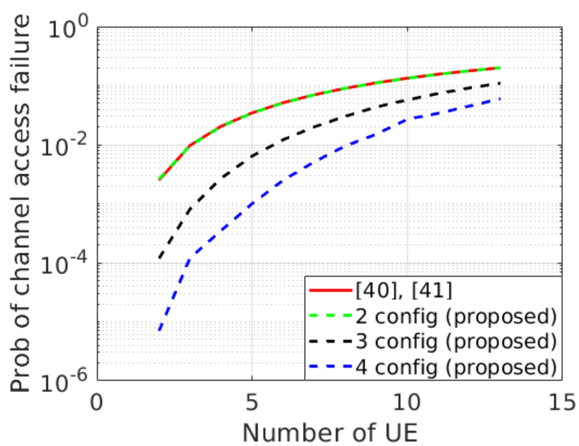


Figure 6.12: Channel access failure's probability in the schemes of [40], [41] and the proposed multiple-configuration scheme $p_0 = 0.95$ and bandwidth of 20 MHz.

The parameters in Table 6.2 are used for the simulations in Fig. 6.13. Fig. 6.13 shows the performance of channel access of the UE in FBE where FFPs are arranged based on the UE's priorities. The first UE is assumed with the highest priority then the priority of the UE decreases in terms of the ordinal number of the UE. The first UE always has the probability of channel access failure to be 0. In Fig. 6.13, each point represents the blocking probability of the i^{th} UE in the system. With p_0 of 0.99, the second UE has the blocking probability of 0.01 in the system with at least two UE. The third UE has the blocking probability of 0.02 in the system with at least three UE. Similarly, the blocking probability of the i^{th} UE is presented. The UE is in one bandwidth of 20 MHz and the number of the UE can be extended by using the interlaces

Table 6.2: Simulation parameters for Fig. 6.13

Parameters	**Values**
Fixed frame period	1 ms
Channel occupancy time	650 μs
Offset	40 μs
p_0	0.99, 0.95
Number of simulated frames	10^{10}
Bandwidth	20 MHz

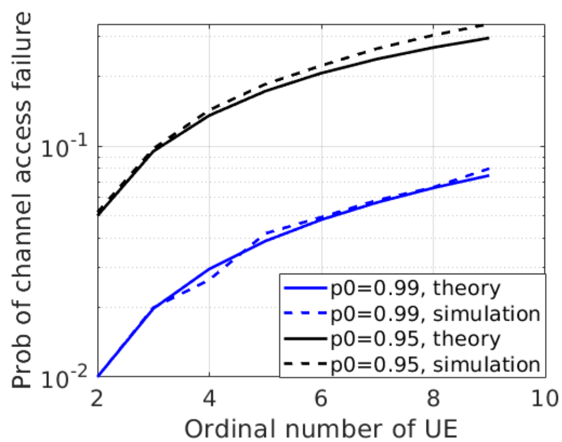


Figure 6.13: Performance of channel access in the FFP priority arrangement scheme.

or a wide band as explained above.

Fig. 6.14 compares the performance of channel access in three analyzed schemes. The scheme of FFP priority arrangement gives an approximate probability of channel access failure as the conventional scheme with one FFP configuration. The difference is that all UE in the conventional scheme have the same blocking probability. While the UE in the FFP priority arrangement scheme has the blocking probability depending on its priority. As can be seen in Fig. 6.14, for the line of the conventional FBE scheme, each point represents the whole number of the UE in the system. While for the line of the proposed FFP priority arrangement scheme, each point represents the blocking probability of the i^{th} UE (the ordinal number) in the system. For example, if there are three UE, in the conventional scheme, all three UE have the probability of 0.02. While in the FFP priority arrangement scheme, the first UE has the failure probability of 0 that is smaller than the probability of the conventional scheme. The second UE has the probability of 0.01 corresponding to the point (2, 0.01) in the graph that is smaller than the probability of the conventional scheme. The third UE has the probability of 0.02 corresponding to the point (3, 0.02) in the graph. Therefore, the FFP priority scheme is suitable for a system where the UE with different priorities including the URLLC UE coexist. This scheme does not increase the complexity of each UE and network design as the multiple FFP configuration scheme while the URLLC UE is provided an absolute priority at cost of the channel access probability of the other UE. On the other hand, the multiple FFP configuration scheme with four configurations in the simulation provides the best performance of channel access for all UE in the system, although it requires a more complex design of the receivers to detect a transmission in one of four FFP configurations.

The parameters in Table 6.3 are used to simulate a scenario where several URLLC UE coexist with a low priority UE such as an eMBB UE as shown in Fig. 6.15. The low priority UE has a higher arrival rate of data and no latency constraint. It can do channel sensing until it obtains the channel so K in the Markov chain goes to infinity for this low priority UE. In the multiple configuration scheme, each URLLC UE has four

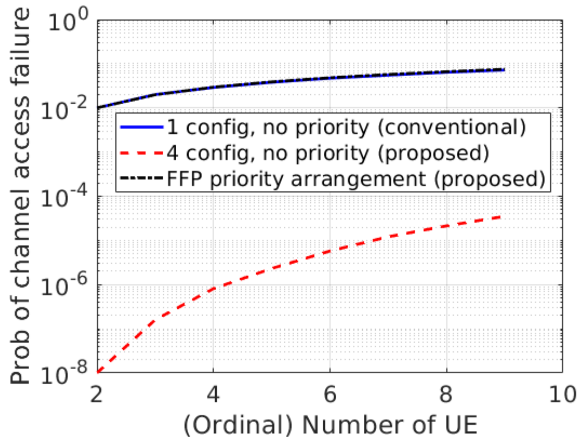


Figure 6.14: A comparison of the conventional one-configuration scheme, the multiple-configuration scheme and the FFP priority arrangement scheme at $p_0 = 0.99$.

Table 6.3: Simulation parameters for Fig. 6.15

Parameters	**Values**
Fixed frame period	1 ms
Channel occupancy time	900 μs
p_0 of URLLC UE	0.99
p_0 of low priority UE	0.5
Number of URLLC UE	1-9
Number of low priority UE	1
Bandwidth	20 MHz

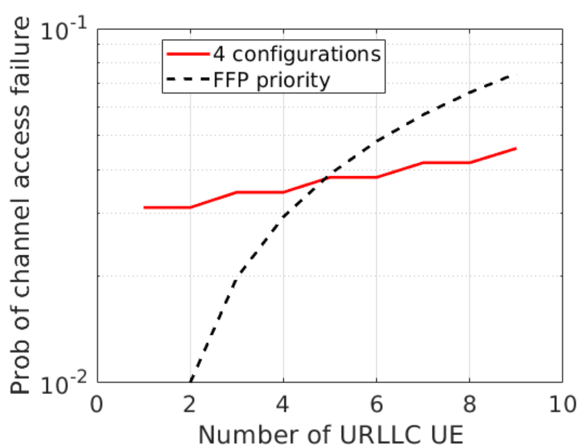


Figure 6.15: A comparison of FBE performance in the multiple-configuration scheme and the FFP priority arrangement scheme in a scenario of the URLLC UE coexisting with the low priority UE.

configurations to sense a channel while the eMBB UE only uses one configuration for channel sensing. In the FFP priority scheme, the eMBB UE is set to the lowest priority FFP. The FFP priority scheme brings the high priority UE a better channel sensing performance. From the first URLLC UE to the fifth URLLC UE in the FFP priority scheme achieve the lower channel access failure's probabilities compared to the URLLC UE in the multiple configuration scheme with the same number of the UE. When the number of the URLLC UE is bigger than 5, the multiple configuration scheme has a better channel access performance. The use of each scheme depends on the number of the UE and the UE's priorities in the system.

6.7 Conclusion

This chapter analyzes channel access process in an unlicensed controlled environment when the FBE channel access mechanism is used. The analysis through a Markov chain shows the limit of FBE to support URLLC due to a latency constraint. To improve the performance of channel access in FBE for URLLC, we propose two schemes. The first scheme allows the transmitter to use multiple FFP configurations to sense a channel and transmit data after a successful CCA. By using multiple FFP configurations, the transmitter has more chance to access to the channel in the URLLC latency budget. The second scheme configures FFPs of the transmitters in a system based on their priorities so that a high priority transmitter's transmission is not blocked by a lower priority transmitter's transmission. Therefore, by using one of two proposed schemes, the URLLC transmitter has a smaller channel blocking probability. Simulations have verified the analysis and shown the benefits of two proposed schemes compared to the current FBE schemes.

Chapter 7

Dynamic switching between load based and frame based channel access mechanisms in unlicensed spectrum

Chapter 5 analyzed a system with all LBE devices and Chapter 6 analyzed a system with all FBE devices. This chapter continues the work in the two previous chapters and focuses on the operation in unlicensed spectrum where the devices using LBE and FBE to access to a channel coexist in the 5G network. In Section 7.1, the behavior of the devices using LBE and FBE that coexist in unlicensed spectrum is modeled by a Markov chain model. The model allows calculating channel access time and transmission probabilities of the devices based on the parameters in the system such as the probability of data arrival, the number of the LBE and FBE devices. Based on these calculations, Section 7.2 proposes a scheme to dynamically switch between LBE and FBE at the devices to serve data with different priorities, requirements and data rates.

7.1 Markov chain model for the coexistence of the devices using LBE and FBE in unlicensed spectrum

7.1.1 System model

The devices in the model share the same frequency resource in sub-6GHz bands, use omnidirectional sensing (omni-LBT) to sense and acquire a channel by following LBE or FBE then transmit data by using omnidirectional transmission. Every of the transmitters (the UE) can detect each other through channel sensing. The receiver (the gNB) uses omnidirectional reception to receive data. q is the probability that a transmitter (using LBE or FBE) has data to transmit. At each transmission by a transmitter using LBE, regardless of the number of retransmissions, the probability that a transmitter senses a busy channel in a sensing slot of 9 μs is p_{c_LBE}. In a t_f gap of 16 μs, energy measurement is done for a total of at least 5 μs with at least 4 μs of sensing falling within the sensing slot of 9 μs immediately before the transmission. Therefore, the busy probability in a 16 μs gap is approximated to be p_{c_LBE} as in a sensing slot of 9 μs. At each transmission by a transmitter using FBE, regardless of the number of retransmissions, the probability that a transmitter senses a busy channel within 9 μs in an observation slot is p_{c_FBE}.

7.1.2 LBE's model

When a transmitter uses LBE to access to a channel, the counter N is chosen randomly between 0 and CW_p in eCCA where CW_p is the contention window size in Table 5.1 and Table 5.2. Based on the Markov chain model for LBE in [50] and Chapter 5, we can calculate transmission probability and channel access time of a transmitter in LBE. The probability that a transmitter using LBE acquires a channel to transmit data is

$$P_{t\_LBE} = \frac{2q(1-p_{c\_LBE})}{2(1-p_{c\_LBE})^2(1-q)+(CW_p-2p_{c\_LBE}+1)q} \tag{7.1}$$

The average time that the transmitter spends in a busy defer duration is

$$\begin{aligned} T = {} & p_{c\_LBE}t_f + p_{c\_LBE}(1-p_{c\_LBE})(t_f+t_{sl})+ \\ & + (1-p_{c\_LBE})^2 p_{c\_LBE}(t_f+2t_{sl}) + ... \\ & ... + (1-p_{c\_LBE})^{m_p} p_{c\_LBE}(t_f+m_p t_{sl}). \end{aligned} \tag{7.2}$$

When a transmitter senses the channel in the defer duration, taking into account time spent in the busy defer duration, the average time spent by the transmitter until it senses an idle defer duration (the channel is idle in the t_f gap and and all t_{sl} slots) and gets out of the defer state is

$$T_{D-out} = T_d + \frac{T}{(1-p_{c\_LBE})^{m_p+1}} - T. \tag{7.3}$$

The average time of the eCCA that the transmitter decrements the counter to 0 and acquires channel to transmit data is:

$$T_{all-backoff} = \frac{CW_p}{2}((1-p_{c\_LBE})t_{sl} + p_{c\_LBE}(t_{sl}+T_{D-out})). \tag{7.4}$$

LBE consists of the iCCA and eCCA so the average time that the transmitter spends to initiate a COT for a transmission is

$$T_{access\_LBE} = T_{D-out} + T_{all-backoff}. \tag{7.5}$$

7.1.3 FBE's model

Based on the Markov chain model in Chapter 6, the probability that a transmitter accesses to a channel to transmit data in FBE is

$$P_{t\_FBE} = q(1-p_{c\_FBE}^K) \tag{7.6}$$

where K is the number of channel sensing that is allowed for a transport block at the transmitter in FBE.

The average time that a transmitter needs to access to a channel in FBE is

$$\begin{aligned} T_{access\_FBE} &= T_{CCA} + p_{c\_FBE}T_{FFP} + 2p_{c\_FBE}^2 T_{FFP} + ... \\ &= T_{CCA} + T_{FFP}\sum_{i=1}^{\infty} i p_{c\_FBE}^i \\ &= T_{CCA} + T_{FFP}\frac{p_{c\_FBE}}{(1-p_{c\_FBE})^2}. \end{aligned} \tag{7.7}$$

7.1.4 Coexistence of LBE and FBE's model

When the transmitters using LBE and the transmitters using FBE coexist in the system, a model based on the model for LBE in Section 7.1.3 and the model for FBE in Section 7.1.4 is used to calculate the transmission and collision probabilities of the transmitters. In the model, there are $N1$ transmitters using LBE and $N2$ transmitters using FBE. For any transmitter using LBE, it senses a busy channel in a sensing slot when at least one transmitter (using LBE or FBE) in the system transmits at that time. The probability of sensing a busy channel in a sensing slot and transmission probability of a LBE transmitter in the coexisting model are

$$\begin{cases} p_{c\_LBE} = 1 - (1 - P_{t\_LBE})^{N1-1}(1 - P_{t\_FBE})^{N2} \\ P_{t\_LBE} = \frac{2q(1-p_{c\_LBE})}{2(1-p_{c\_LBE})^2(1-q)+(W-2p_{c\_LBE}+1)q} \end{cases} . \tag{7.8}$$

Similarly, for any transmitter using FBE, it senses a busy channel in an observation slot when at least one transmitter (using LBE or FBE) in the system transmits at that time. The probability of sensing a busy channel in an observation slot and transmission probability of a FBE transmitter in the coexisting model are

$$\begin{cases} p_{c\_FBE} = 1 - (1 - P_{t\_LBE})^{N1}(1 - P_{t\_FBE})^{N2-1} \\ P_{t\_FBE} = q(1 - p_{c\_FBE}^{K}) \end{cases} . \tag{7.9}$$

7.2 Dynamic switch between LBE and FBE at the UE in unlicensed spectrum

7.2.1 Switch from FBE to LBE

In FBE, a transmitter is only allowed to do channel sensing at the fixed occasions with a period of T_{FFP}. If the channel is busy in a Channel Clear Access (CCA) occasion, the transmitter must wait until the CCA occasion in the next FFP to sense the channel. This limits the opportunity of the transmitter to attempt to acquire the channel. In contrast, in LBE, a transmitter senses the channel continuously whenever it has data to transmit. Therefore, the average access time of the LBE transmitter with low channel access priority class is smaller than that of the FBE transmitter with the same probability of sensing a busy channel.

When a transmitter using FBE has a high priority packet such as an URLLC packet with a required latency of 1 ms, the performance of FBE with the current parameters $(q, CW_p, N1, N2)$ in the system might not satisfy the URLLC requirement. In this case, the FBE transmitter can dynamically switch to LBE mode to have a better performance of channel access. In another case, the latency requirement is satisfied by using FBE with the current set of parameters then the data rate increases leading to a higher value of q. This makes the probability of sensing a busy channel increase and the transmitter needs longer time to acquire a channel so the latency requirement is not ensured anymore. To overcome this problem, the transmitter is also allowed to dynamically switch to LBE so as to reduce the average channel access time.

In the system, at the beginning, there are $N1$ LBE transmitters and $N2$ FBE transmitters. We have the probabilities of sensing a busy channel for the LBE and FBE transmitters

$$\begin{cases} p_{c\_LBE} = 1 - (1 - P_{t\_LBE})^{N1-1}(1 - P_{t\_FBE})^{N2} \\ p_{c\_FBE} = 1 - (1 - P_{t\_LBE})^{N1}(1 - P_{t\_FBE})^{N2-1} \end{cases} . \tag{7.10}$$

When a FBE transmitter switches to LBE due to a high priority packet or a higher data rate, we have the probabilities of sensing a busy channel for $N1 + 1$ LBE transmitters and $N2 - 1$ FBE transmitters

$$\begin{cases} p'_{c\_LBE} = 1 - (1 - P'_{t\_LBE})^{N1}(1 - P'_{t\_FBE})^{N2-1} \\ p'_{c\_FBE} = 1 - (1 - P'_{t\_LBE})^{N1+1}(1 - P'_{t\_FBE})^{N2-2} \end{cases} . \tag{7.11}$$

P_{t_LBE} and P'_{t_LBE} are calculated from (7.1). P_{t_FBE} and P'_{t_FBE} are calculated from (7.6).

For the FBE transmitter of interest switching to LBE, it has more chances to do channel sensing and access to the channel in an interval. Therefore, the transmitter needs less time to access to a channel and attains the latency requirement in LBE with low channel access priority class as calculated from (7.5) and (7.7).

The decision to switch from FBE to LBE at a transmitter (a UE) is made by the gNB or the transmitter. If the transmitter transmits data on the configured resources, it calculates channel access time based on the parameters including $q, CW_p, N1, N2$. If channel access time is higher than the data requirement, it switches to LBE then informs this switch to the gNB through UCI or RRC. Subsequently, the gNB updates the number of the LBE and FBE transmitters in the system and informs them to all the transmitters through DCI or RRC so that the transmitters can use this information to calculate transmission probability, sensing busy channel's probability and channel access time for the upcoming packets. If the transmitter transmits data on the dynamic resources, it sends a scheduling request to the gNB. The gNB can calculate the variables related to the UL transmission requested and makes the decision to switch by itself. The gNB demands the transmitter to switch from FBE to LBE if necessary through DCI playing the role of UL grant.

7.2.2 Switch from LBE to FBE

When a transmitter using LBE has the low priority packets without a strict latency requirement, the transmitter can switch to FBE with a longer channel access time to reduce the number of channel sensing. This also mitigates the detecting burden at the gNB because the gNB only needs to detect data at the fixed moments at the beginning of the FFP. Even if the transmitter has the high priority packets such as URLLC packets then the URLLC data rate decreases leading to a smaller value of q, the transmitter also can switch to FBE to save energy while still achieving latency requirement. This switch is helpful for the power limited devices.

In the system, at the beginning, there are $N1$ LBE transmitters and $N2$ FBE transmitters. We have the probabilities of sensing a busy channel for the LBE and FBE transmitters

$$\begin{cases} p_{c\_LBE} = 1 - (1 - P_{t\_LBE})^{N1-1}(1 - P_{t\_FBE})^{N2} \\ p_{c\_FBE} = 1 - (1 - P_{t\_LBE})^{N1}(1 - P_{t\_FBE})^{N2-1} \end{cases} . \tag{7.12}$$

When a LBE transmitter switches to FBE to reduce energy consumption when priority of data or data rate is low, we have the probabilities of sensing a busy channel for $N1-1$ LBE transmitters and $N2+1$ FBE transmitters

$$\begin{cases} p'_{c\_LBE} = 1 - (1 - P'_{t\_LBE})^{N1-2}(1 - P'_{t\_FBE})^{N2+1} \\ p'_{c\_FBE} = 1 - (1 - P'_{t\_LBE})^{N1-1}(1 - P'_{t\_FBE})^{N2} \end{cases} . \tag{7.13}$$

P_{t_LBE} and P'_{t_LBE} are calculated from (7.1). P_{t_FBE} and P'_{t_FBE} are calculated from (7.6).

The transmitter of interest switching from LBE to FBE does channel sensing less frequently and needs less steps to acquire a channel. The gNB also only needs to detect the transmission at the beginning of the FFP. Therefore, energy consumption at the gNB and the transmitter is reduced.

Similarly to Section 7.2.1, the decision to switch from LBE to FBE can be made by the gNB or the transmitter. The information related to the switch is exchanged between the gNB and the transmitter through

UCI, DCI or RRC. The information about periodicity and the starting point of the FFP is preconfigured by the gNB through the activation DCI or RRC in the configured grant transmission or included in DCI playing the role of UL grant in the dynamic grant transmission.

7.3 Numerical results

The simulations in this section are done to show the performance of a UE switching between LBE and FBE and the benefits of this switch. The parameters of the simulations are shown in Table 7.1.

Table 7.1: Simulation parameters

Parameters	**Values**
Number of the LBE UEs	5
Number of the FBE UEs	5
Channel access priority class	2
Contention window size (CW_p)	7
Fixed frame period	1 ms
The allowed number of channel sensing (K)	2

The first simulation is done to show the performance of a UE switching from FBE to LBE. At the beginning, there are 5 LBE UEs and 5 FBE UEs in the system. The UE of interest uses FBE to access to the channel and transmits URLLC data. The UE of interest might switch from FBE to LBE to make the URLLC transmission achieve the latency requirement when probability of arriving data (data rate) increases. After the UE of interest switches from FBE to LBE, there are 6 LBE UEs and 4 FBE UEs in the system.

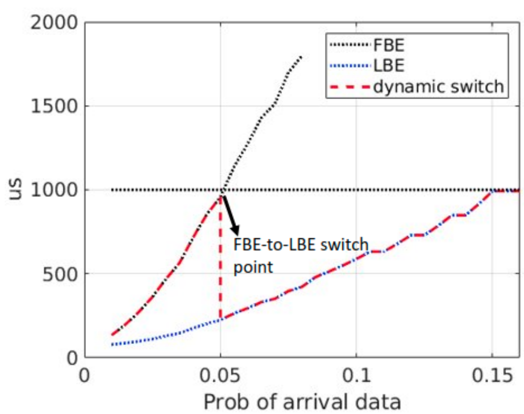


Figure 7.1: Channel access time of a UE using FBE, LBE and dynamic-FBE-to-LBE-switch scheme.

Fig. 7.1 shows channel access time when the UE of interest uses FBE, channel access time when the UE of interest uses LBE and channel access time when the UE of interest uses a dynamic switch scheme to switch from FBE to LBE. URLLC has a latency requirement of 1 ms so when the UE transmits URLLC packets, channel access time must not be bigger than 1 ms. If channel access time in FBE is bigger than 1 ms, the UE must switch to LBE to attain the URLLC requirement. As shown in Fig. 7.1, if the probability that the UE has data to transmit q is smaller than 0.05, channel access time in FBE is smaller than 1 ms so the UE can use FBE to access to the channel and transmits URLLC data because FBE process is less complex and requires a smaller number of channel sensing that leads to lower energy consumption at the transmitter and receiver. On the other hand, if q is bigger than 0.05, channel access time in FBE is bigger than 1 ms so the UE must switch to LBE. Fig. 7.2 demonstrates that the UE has a higher probability of transmission after the switch, although the probabilities of sensing a busy channel before and after the switch are nearly equal. By

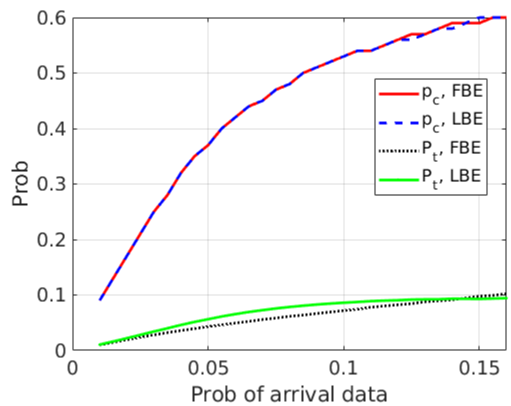


Figure 7.2: Probability of sensing a busy channel and transmission probability of a UE using FBE and LBE.

using LBE when data rate increases and makes q higher than 0.05, the UE has channel access time smaller than 1 ms to satisfy the URLLC requirement as shown in Fig. 7.1. When the UE uses LBE, it has to do more channel sensing and the gNB also has to do more blind detection to detect a transmission. Therefore, a switch from FBE to LBE requires higher energy consumption at the UE and the gNB.

The second simulation is done to show the performance of a UE switching from LBE to FBE. At the beginning there are 5 LBE UEs and 5 FBE UEs in the system. The UE of interest uses LBE to access to the channel and transmits URLLC data. The UE might switch from LBE to FBE to reduce the sensing burden at the UE and the detecting burden at the gNB while still achieving the URLLC latency requirement. After the UE of interest switches from LBE to FBE, there are 4 LBE UEs and 6 FBE UEs in the system.

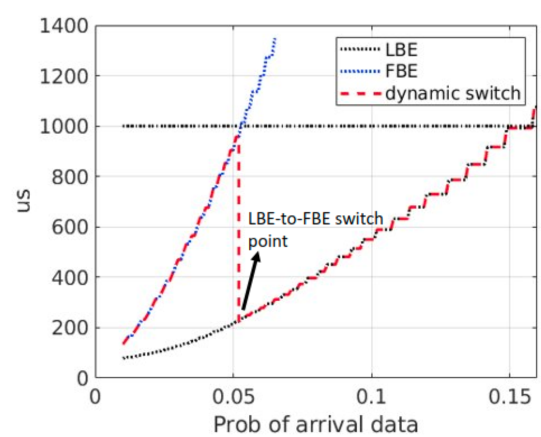


Figure 7.3: Channel access time of a UE using FBE, LBE and dynamic-LBE-to-FBE-switch scheme.

Fig. 7.3 shows channel access time when the UE of interest uses FBE, channel access time when the UE of interest uses LBE and channel access time when the UE of interest uses a dynamic switch scheme to switch from LBE to FBE. The UE is configured to use LBE so that it can access to the channel in the URLLC latency budget of 1 ms. As can be seen in Fig. 7.3, if the probability that the UE has data to transmit q is smaller than 0.052, the UE can use FBE and still achieves the URLLC requirement. Therefore, if data rate decreases and q is smaller than 0.052, the UE switches from LBE to FBE in order to access to the channel. By using FBE, the sensing burden at the UE is reduced that leads to lower energy consumption. As shown in Fig. 7.4, when q is smaller than 0.052 and the UE switches from LBE to FBE, it needs to do a

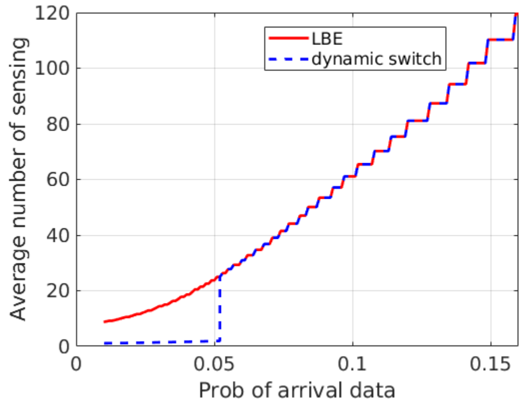


Figure 7.4: Number of channel sensing of a UE using LBE and dynamic-LBE-to-FBE-switch scheme.

smaller number of channel sensing to access to channel so energy consumption decreases while the URLLC requirement is still ensured.

7.4 Conclusion

This chapter presented a Markov chain for the coexistence of the LBE and FBE devices in the system. To serve data with different priorities and arrival rates, we proposed that the devices are capable of dynamically switching between LBE and FBE based on the parameters calculated from the Markov chain. When a device has high priority data such as URLLC with a strict latency requirement or data with high arrival rate, it switches from FBE to LBE to reduce channel access time and increase transmission probability. On the other hand, when a device has low priority data such as eMBB or data with low arrival rate, it switches from LBE to FBE to mitigate the sensing burden at the transmitter and the detecting burden at the receiver. The benefits of the dynamic switch between LBE and FBE have been shown in the numerical results.

Chapter 8

Enhancements of PUSCH repetitions for URLLC in unlicensed spectrum

8.1 Gap in the middle of PUSCH repetitions

8.1.1 Gap due to UL/DL directions

In 3GPP Release 15 and 16, UL transmission of a TB is supported with multiple repetitions to achieve the strict URLLC requirements. Time domain resource is indicated for the first "nominal" repetition. The time domain resources for the remaining repetitions are derived based at least on the resources for the first repetition and UL/DL direction of symbols. The number of valid symbols for PUSCH of a TB that the UE can use are in the time window $K \times L$ where K is the number of nominal repetitions, L is the length of a nominal repetition.

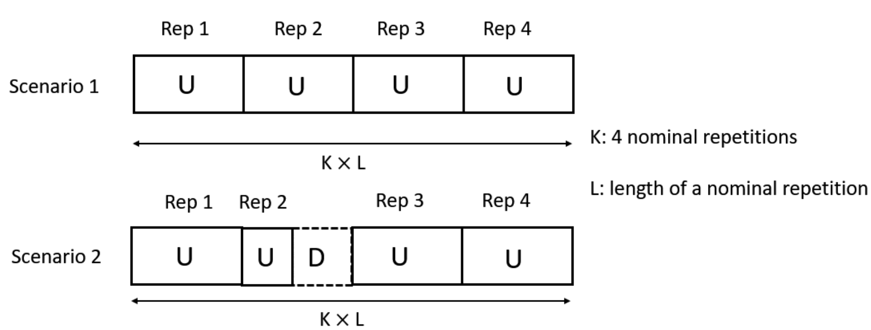


Figure 8.1: UL transmission in TDD configuration licensed spectrum.

In PUSCH repetition Type B, the actual number of repetitions might be bigger than the nominal K repetitions. If a nominal repetition encounters slot boundary or DL symbols in TDD configuration, this repetition is fragmented into multiple actual repetitions. However, the time window of valid symbols is still $K \times L$ so the presence of DL symbols in the middle of PUSCH repetitions reduces the number of valid symbols for PUSCH that the UE can use in TDD. In Fig. 8.1, a part of the second repetition in scenario 2 is dropped due to DL symbols in the middle of the repetitions. This reduces reliability of UL transmission of a TB when PUSCH repetition Type B is used in licensed spectrum.

In unlicensed spectrum, when PUSCH repetition Type B is used, the interruption in UL transmission of the repetitions even has a bigger impact due to LBT overhead. Fig. 8.2 shows two scenarios of UL

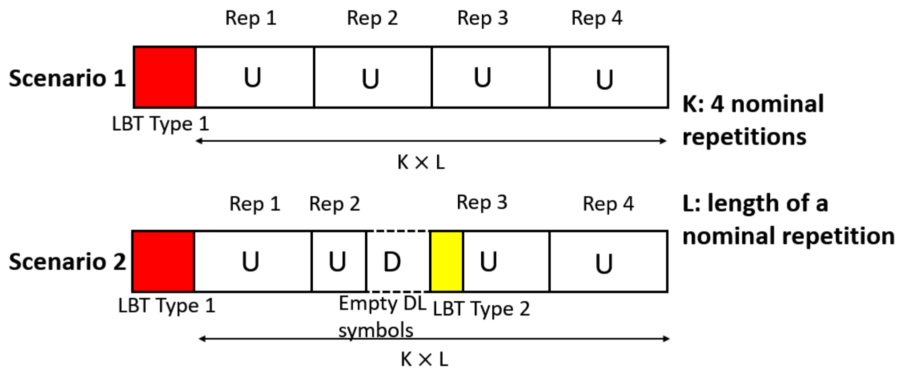


Figure 8.2: UL transmission in TDD configuration unlicensed spectrum.

transmission of a TB in TDD configuration in unlicensed spectrum. In both scenarios, the UE is scheduled by the gNB to transmit 4 nominal repetitions in Type B (DG or CG transmission). The time window within which valid symbols are used for transmission is $K \times L$ where K is 4 nominal repetitions, L is the length of a nominal repetition. Before the transmission, the UE must do LBT Type 1 to acquire the channel and it is allowed to use this channel in COT duration. This COT is shared with the gNB so the gNB can transmit DL data in DL symbols within this COT. Depending on the gap between DL and UL bursts, the gNB would transmit without or with LBT Type 2 channel access procedures.

In scenario 1, there are 4 consecutive UL sub-slots so the UE can transmit the whole 4 repetitions without an interruption after a successful LBT Type 1. However, in scenario 2, due to the URLLC latency budget of 1 ms, the UL repetitions must be scheduled in the UL resources segmented by the DL sub-slots when the transmission cannot wait the next continuous UL sub-slots for all repetitions. The UL transmission is interrupted because of DL symbols in the middle of the second repetition. The UE cannot transmit UL data in a DL transmission occasion so the second repetition is fragmented into 2 actual repetitions. With sub-carrier spacing (SCS) from 15 kHz to 120 kHz and a sub-slot of 2, 4, 7 symbols, the time length of DL sub-slot is larger than 16 μs. In some cases, DL symbols are configured in a semi-static way in advance. If the gNB does not have DL data to transmit at those semi-static DL symbols, there will be a gap between two UL bursts. When the gap between two UL transmission occasions is larger than 16 μs due to the empty semi-static DL symbols, the UE must do an additional LBT Type 2 to continue to transmit the remaining repetitions in the UE-initiated COT. This additional LBT Type 2 consumes at least 25 μs or more if the channel is busy. The time window $K \times L$ is fixed so an increase of LBT time reduces time of data transmission. The UE transmits less data in the repetition than configured that causes a decrease of reliability.

8.1.2 Gap due to orphan symbols

The gap in the transmission of PUSCH repetitions Type B is also caused by orphan symbols. In scenario 2 of Fig. 8.3, two repetitions Type B of a TB are scheduled. Each repetition has 4 symbols. After the first repetition, the second repetition is fragmented by the slot boundary between slot 1 and slot 2. There is one symbol (orphan symbol) in slot 1 and 3 symbols in slot 2. However, an actual repetition with a single symbol is not transmitted [6]. Thus, the UE does not transmit a repetition in the last symbol of slot 1 and this symbol is empty. The UE only starts to transmit the second repetition from the first symbol of slot 2. It creates a gap of one symbol between two repetitions. With SCS from 15 kHz to 60 kHz, the length of a symbol is larger than 16 μs. Therefore, in unlicensed spectrum, at least, a short LBT of 25 μs is needed before the transmission of the second repetition. It causes LBT overhead and reduces the time interval that the UE can transmit the second repetition.

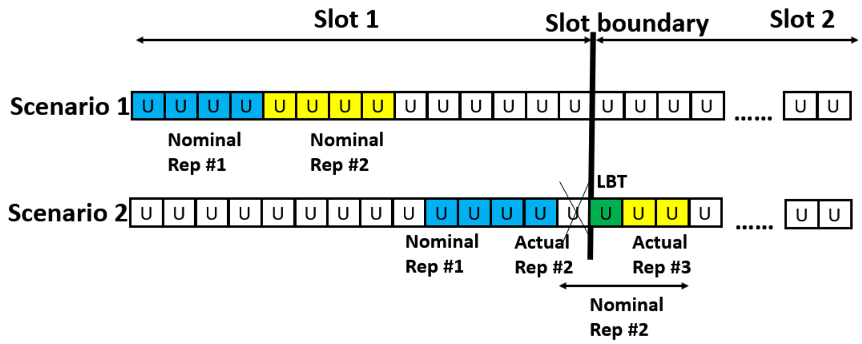


Figure 8.3: Interruption in the transmission of UL repetitions due to orphan symbol in unlicensed spectrum.

8.2 Related works

In [51], PUSCH repetition Type B is harmonized with CG PUSCH in unlicensed spectrum but there is no technique to avoid the gap in PUSCH repetitions.

In [52] and [53], the time window within which valid symbols are used for PUSCH transmission is extended to be longer than $K \times L$ at least in case of semi-static DL symbols. The extension of the time window of a transmission might cause a delay of the next transmission or collision with other UE. Moreover, this extension may cause ambiguity on the ending of the window if the UE misses the dynamic signal to update the flexible symbols in dynamic channel configuration.

In [54], the UE is proposed to transmit the last symbol of the repetition before the orphan symbol or the first symbol of the repetition after the orphan symbol instead of leaving an empty orphan symbol. It helps the UE to keep the channel but changes code rate of the repetitions.

In [55], the number of UL and DL symbols in TDD configuration is calculated based on the arrival rates of DL and UL data. It mitigates DL symbols in the middles of PUSCH repetitions but cannot avoid entirely the problem of gap between UL bursts.

8.3 Enhancements of PUSCH repetitions in licensed and unlicensed spectrum

8.3.1 Handling gap due to UL/DL directions

In TDD configuration, DL symbols might be configured in a semi-static way so that the gNB can transmit DL data (PDCCH, PDSCH, synchronization signal block). In case the gNB must schedule a PUSCH transmission with DL symbols between the repetitions in Scenario 1 of Fig. 8.4 due to time constraint of URLLC, the semi-static DL symbols cause a segmentation in the PUSCH repetitions even if the gNB has no DL data to transmit in that occasion. This DL occasion is scheduled in a semi-static way in advance so the gNB does not know whether it has DL data at that occasion or not. Therefore, when the gNB schedules the repetitions of PUSCH, if it recognizes that it does not have DL data to transmit in the DL symbols causing segmentation, it switches semi-static DL symbols to UL symbols as in Scenario 2 of Fig. 8.4. This scenario is for a transmission in unlicensed spectrum. There are two benefits of this switch. Firstly, the UE can transmit PUSCH repetitions without segmentation so that the valid symbols for UL transmission is ensured to be $K \times L$ symbols where K is the number of nominal repetitions, L is the length of a nominal repetition. Secondly, no additional LBT Type 2 is required due to segmentation. As can be seen in Fig. 8.4, the UE can transmit full data in the second and third repetitions in Scenario 2 instead of dropping a part of the second and third repetitions in Scenario 1.

The update of empty semi-static DL symbols to UL symbols is also applied to the PUSCH repetitions in licensed spectrum so the UE can transmit all full repetitions as configured in the time window without segmentation and dropping data in DL symbols to guarantee reliability.

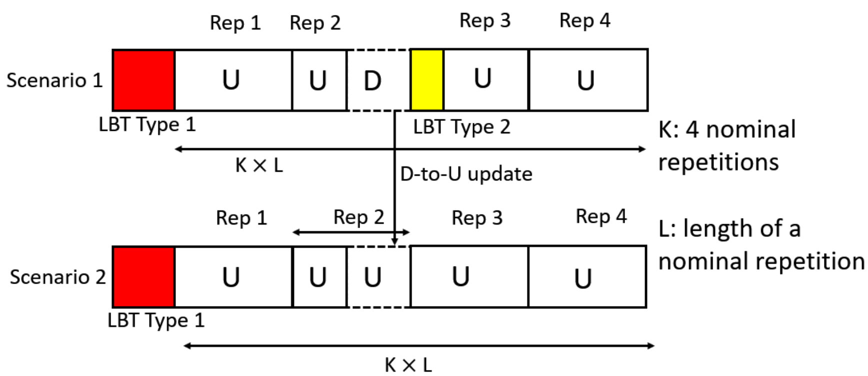


Figure 8.4: Semi-static DL symbols are switched to UL symbols.

In case the gNB has DL data to transmit in the semi-static DL symbols among the UL symbols that it intends to schedule to an UL transmission, if the UL data has higher priority than DL data, the gNB also switches semi-static DL symbols to UL symbols and delays the DL transmission to the next occasion so that the UE can transmit a continuous high priority UL transmission. As in Fig. 8.4, if DL symbols contain a low priority transmission compared to an UL transmission, these DL symbols are switched to UL symbols.

The update of the chosen semi-static DL symbols to UL symbols are applied to both DG and CG transmission. For DG transmission, after the gNB decides the semi-static DL symbols that are switched to UL symbols, when the gNB sends UL grant to schedule PUSCH repetitions, it also sends a dynamic SFI signal such as DCI format 2_0 to the UE to switch the chosen DL symbols in the middle of PUSCH repetitions to UL symbols. The gNB also can indicate the update of semi-static DL symbols to UL symbols to the UE in the UL grant DCI by using one additional bit. Another way is to indicate implicitly through priority index in UL grant DCI. The UE that is scheduled with high priority UL transmission uses the semi-static DL symbols to transmit PUSCH. The gNB knows that pre-configuration so it does not transmit DL data in those symbols.

For CG transmission, after activating CG configurations, the gNB can send periodically DCI format 2_0 or a RRC signal to switch the chosen DL sub-slots/slots to UL sub-slots/slots. The period of DCI format 2_0 or RRC signal depends on priority of the CG configurations, the arrival rate of UL data at the UE using the configurations and the arrival rate of DL data at the gNB.

The method to update semi-static DL symbols can be extended to update semi-static UL symbols. This means that the gNB can update semi-static UL symbols to DL symbols to use in DL transmission. It is useful when the gNB has a high arrival rate of URLLC DL data and needs to use the updated symbols to meet URLLC requirements for DL transmission. The gNB sends dynamic SFI to notify the UE about the update of semi-static UL symbols to DL symbols so that the UE decodes DL transmission in these symbols instead of transmitting UL data.

In dynamic configuration, a symbol can be configured as a semi-static flexible symbol then is updated dynamically to UL or DL symbol by DCI format 2_0 indicating SFI from the gNB to the UE. For CG transmission in licensed and unlicensed spectrum, if dynamic SFI is configured but the UE cannot decode DCI format 2_0 indicating SFI to update the semi-static flexible symbols due to channel condition, an actual PUSCH repetition in those non-updated flexible symbols is dropped. This causes a decrease of PUSCH transmission' reliability.

To increase reliability of URLLC UL transmission, if dynamic SFI is not received, the UE will use the semi-static flexible symbols in the scheduled resource as the UL symbols to transmit the high priority repetitions without any gap instead of dropping the repetitions in those symbols. In other words, a repetition is not fragmented around the semi-static flexible symbols or dropped because of a conflict with the semi-static flexible symbols but an actual repetition is transmitted continuously in time in that case. Therefore, in both licensed and unlicensed spectrum, the configured number of repetitions is ensured. Moreover, in unlicensed spectrum, no additional LBT is required due to the dropped repetition in the middle of the transmission of PUSCH repetitions.

8.3.2 Handling orphan symbols

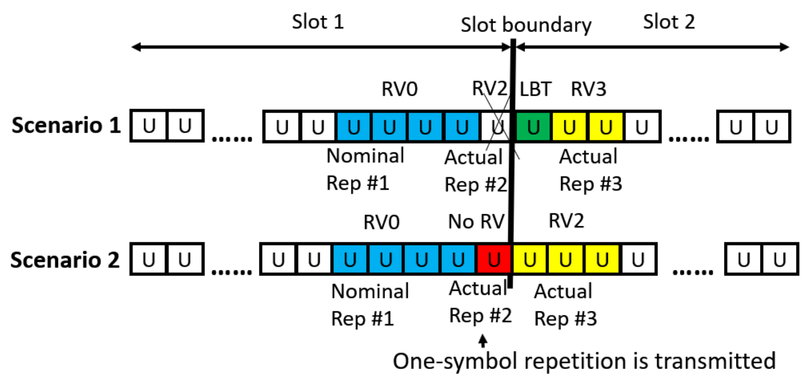


Figure 8.5: Orphan-symbol actual repetition containing DMRS is transmitted in unlicensed spectrum.

An actual PUSCH repetition in an orphan symbol due to segmentation is dropped that causes a gap between two burst of PUSCH repetitions. In unlicensed spectrum, the presence of this gap leads to LBT overhead in PUCSH transmission. To guarantee the performance of high priority transmission as URLLC, the discontinuous PUSCH repetition's transmission due to one-symbol fragment must be avoided.

In the proposed scheme, the UE still transmits the one-symbol repetition so the transmission of PUSCH repetitions is continuous and LBT Type 2 channel access procedures after one-symbol fragment is not required before the UE continues to transmit the PUSCH repetitions as shown in Fig. 8.5. Thanks to a continuous channel without LBT, the UE has more symbols in the scheduled time window to transmit PUSCH repetitions so that PUSCH transmission of a TB attains a higher reliability. The transmission of one-symbol repetition is initiated based on priority of the transmission.

The one-symbol repetition contains only DMRS and does not contain any TB data. The transmission of additional DMRS provides an opportunity to do channel estimation. It improves the performance of high priority UL transmission. Because this one-symbol repetition does not contain data, it is not taken into account in the calculation of redundancy version (RV) sequence of the repetitions as shown in Fig. 8.5. With a RV sequence of $\{0, 2, 3, 1\}$, in Scenario 2, only the first and third repetitions have RVs to be 0 and 2, respectively, while the second repetition is not assigned RV. In other words, the one-symbol repetition is transmitted but not considered as an actual repetition.

To indicate the activation of a transmission in an orphan symbol, in UL DG transmission, the gNB adds one bit to UL grant DCI. This bit tells the UE to transmit DMRS in the one-symbol fragment or drop the transmission in the orphan symbol. If UL transmission is high priority with the strict requirements as URLLC, DMRS transmission on one–symbol repetition is activated. Otherwise, the transmission of one-symbol repetition is not activated.

In UL CG transmission, the transmission of DMRS in the one-symbol fragment to avoid LBT is activated by a new RRC parameter for Type 1 CG or 1 bit in the activation DCI in Type 2 CG.

The transmission of one-symbol repetition also can be indicated implicitly to the UE by high priority index in DCI or RRC. When DCI or RRC contains high priority index for a PUSCH transmission, by pre-configuration, the UE transmits one-symbol repetition instead of dropping it. Otherwise, the UE does not transmit one-symbol repetition.

8.4 Simulation results

8.4.1 Performance of the scheme to handle UL/DL directions

Table 8.1: Simulation parameters

Parameters	**Values**
Waveform	CP-OFDM
Subcarrier spacing	30kHz
Channel model	Additive white Gaussian noise (AWGN)
Channel coding	Low-density parity-check (LDPC) code
TB length	160 bits
Number of repetitions/TB	4
Number of symbols/repetition	4
MCS Index	1

Using the parameters in Table 8.1, the first simulation is done in licensed spectrum to compare the performance of PUSCH repetitions in the current scheme where PUSCH repetitions might be dropped because of DL symbols and in the proposed scheme where DL symbols are dynamically updated to UL symbols for PUSCH repetitions. It is set that one out of four repetitions conflicts with one DL symbol so this repetition is not transmitted in full. As can be seen in Fig. 8.6, block error rate (BLER) of a repetition conflicted with DL symbol increases because only three UL symbols can be used to transmit data instead of four symbols. In the proposed scheme, the DL symbol is updated to UL symbol so the repetition is transmitted in full and has a lower BLER.

To increase reliability of PUSCH transmission, the UE combines the repetitions of a TB so that it can decode the packet with lower code rate. In Fig. 8.6, even with soft-combining, the performance of PUSCH transmission is still degraded due to the conflict of a repetition with a DL symbol. When DL symbol is updated to UL symbol, soft-combining of the repetitions achieves better performance.

The second simulation is done in unlicensed spectrum to compare the current and proposed schemes. The influence of DL-UL directions is bigger in unlicensed spectrum than in licensed spectrum because the UE must do a LBT of 25 μs (equivalent to one symbol in simulation) after the empty DL symbols. Due to LBT and DL symbols, the UE has less time to transmit data in the repetitions. As can be seen in Fig. 8.7, when the UE decodes data by doing soft-combining of the repetitions, the BLER of PUSCH repetitions is higher in unlicensed spectrum than in licensed spectrum when a repetition collides with an empty DL symbol. The scheme of updating symbols brings more benefits to the performance of PUSCH repetitions in unlicensed spectrum.

The third simulation using the parameters in Table 8.1 is done in licensed and unlicensed spectrum to compare the performance of PUSCH repetitions in two schemes: the conventional scheme where the PUSCH repetition is dropped in the non-updated semi-static flexible symbols and the proposed scheme where the non-updated semi-static flexible symbols are used as UL symbols to transmit PUSCH repetition. In the simulation, one out of four repetitions has the non-updated semi-static flexible symbols so this repetition is

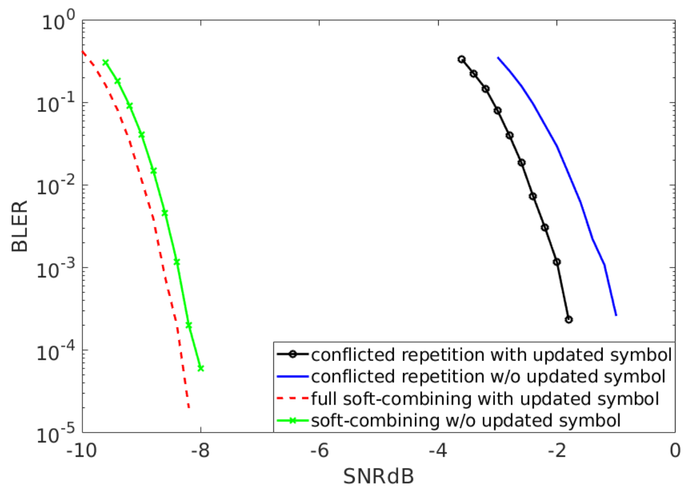


Figure 8.6: Performance PUSCH repetitions in licensed spectrum with updated DL-to-UL symbol.

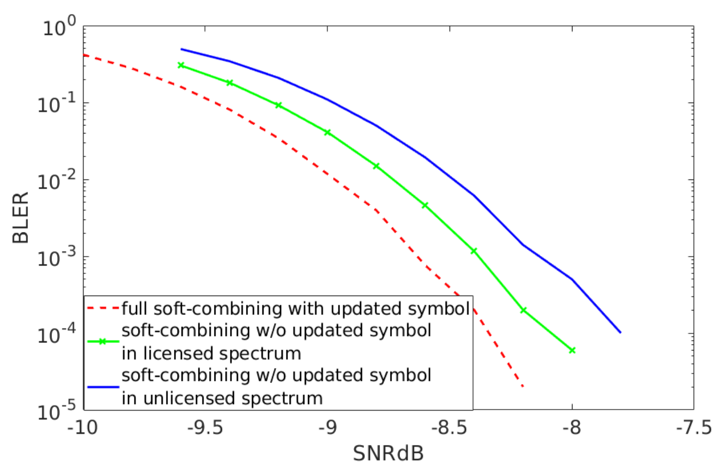


Figure 8.7: Performance PUSCH repetitions in unlicensed spectrum with updated DL-to-UL symbol

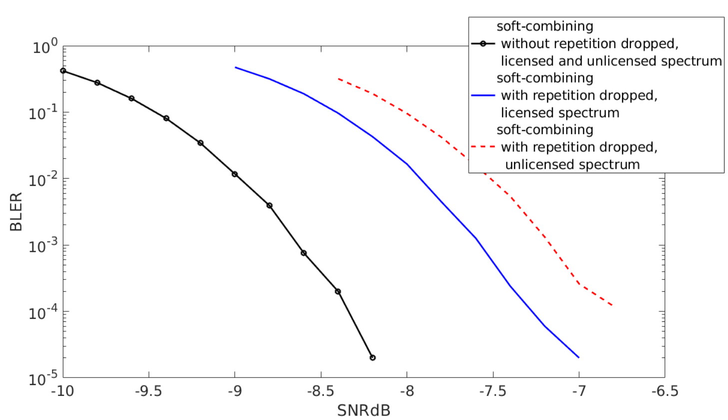


Figure 8.8: Performance PUSCH repetitions in licensed and unlicensed spectrum with flexible symbols used as UL symbols

dropped in the conventional scheme and the gNB only receives 3 repetitions to do soft-combining. As shown in Fig. 8.8, BLER of PUSCH transmission in licensed spectrum increases when one repetition is dropped and only three repetitions are transmitted. BLER even grows more significantly for PUSCH transmission in unlicensed spectrum because the UE must consume more time for an additional LBT after dropping a repetition and has less time to transmit PUSCH repetitions. When the non-updated flexible symbols are used for UL symbols and the repetition is not dropped in the proposed scheme and the gNB receives all four repetitions for soft-combining, BLER of PUSCH transmission in both licensed and unlicensed spectrum is lower than that of the conventional scheme.

8.4.2 Performance of the scheme to handle orphan symbols

Using the parameters in Table 8.1, the simulation is done to compare PUSCH performance in unlicensed spectrum between the conventional scheme of dropping the actual repetition in the orphan symbols and the scheme of transmitting DMRS in the orphan symbols. It is set that one repetition is segmented and creates an actual repetition in an orphan symbol. As can be seen in Fig. 8.9, the actual repetition is dropped and creates a gap in PUSCH repetition so the UE must consume more time for an additional LBT so it has less time to transmit data in the nominal repetition (2 symbols instead of 4 symbols). This repetition has a high BLER and becomes undecodable. A DMRS in orphan symbol helps the UE to keep channel and avoid an additional LBT so the nominal PUSCH repetition has a lower BLER. The performance of PUSCH transmission in soft-combining with DMRS in orphan symbol is also better that that of PUSCH in the conventional scheme.

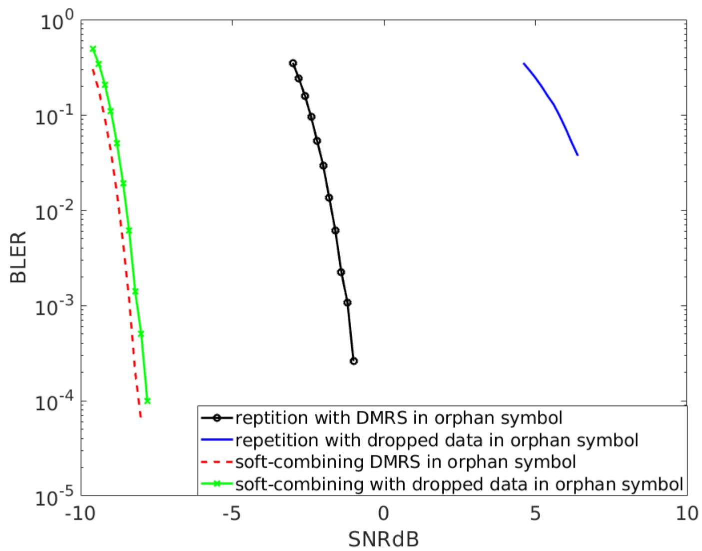


Figure 8.9: Performance PUSCH repetitions in unlicensed spectrum with DMRS in orphan symbol

8.5 Conclusion

The work enhances the performance of PUSCH Type B repetition for URLLC in licensed and unlicensed spectrum by proposing two schemes. The first scheme about dynamically updating DL symbols to UL symbols helps PUSCH repetitions avoid the unnecessary segmentation in the middle of transmission. The second scheme about transmitting DMRS in orphan symbols helps the UE keep channel so that it can resume PUSCH transmission without an additional LBT.

Chapter 9

Conclusions

9.1 Concluding remarks

In this thesis, we studied the performance of URLLC, the problems in physical layer design that impact URLLC and the schemes to solve these problems. We started the work on the URLLC operation in licensed spectrum. In Chapter 2, one problem identified in URLLC UL transmission is that when the UE transmits a transport block by a configured number of repetitions in the UL CG resources, the UE might not transmit all repetitions as configured due to the constraint of a HARQ interval. This causes a degradation of URLLC transmission's reliability. We proposed three schemes to guarantee the number of repetitions transmitted or mitigate the impact on URLLC performance when a smaller number of repetitions than configured is transmitted. The first scheme uses reserved resources so that the UE can transmit the repetitions out of the original HARQ interval to ensure the number of repetitions. The second scheme exploits an explicit HARQ feedback structure to trigger a retransmission if the gNB cannot detect the packet due to a smaller number of repetitions than configured. The third scheme allows the UE to transmit an additional scheduling request if it transmits a smaller number of repetitions than configured. The gNB can detect this scheduling request to trigger a retransmission if necessary.

In Chapter 3, another problem in URLLC UL transmission is identified. An UL eMBB transmission of a UE can be scheduled on the CG resources for the UL URLLC transmission of another UE to maximize resource utilization. This might cause a collision between an eMBB transmission and an URLLC transmission that reduces URLLC transmission' reliability because the gNB does not have prior information about the transmission of the URLLC UE on the CG resources. The first proposed scheme contains an overlap indication to initiate an explicit HARQ feedback structure used in case of an overlap so that the gNB can schedule a retransmission if necessary. The second proposed scheme consists of an overlap indication and an additional scheduling request transmitted in case of an overlap so that the gNB have another chance to detect the presence of a transmission that is not detected through DMRS to schedule a retransmission.

After the problems in UL transmission, Chapter 4 solves a problem in DL SPS transmission where the feedback of a DL SPS transmission might be dropped because it is pointed to the DL symbols instead of the UL symbols. To avoid the drop of the feedback of a DL SPS transmission, we proposed that the resource of feedback is indicated dynamically in PUSCH on SPS resources. Moreover, the UE only transmits feedback when it decodes correctly data in PUSCH. Otherwise, the gNB will retransmit the packet automatically.

Subsequently, we extended our research to the URLLC operation in unlicensed spectrum to target new use cases in the industrial scenario. The 5G transmitters in unlicensed spectrum are required to access to a channel through one of two channel access mechanisms: LBE and FBE. The uncertainty of acquiring a channel through these mechanism causes latency that is harmful to URLLC transmission. Chapter 5 analyzed the impact of LBE on URLLC performance by a Markov chain model. Based on the analysis, the new channel access tables in LBE are proposed to support URLLC. Chapter 6 analyzed the impact of FBE on ULLC performance by another Markov model. Two schemes are introduced to enhance URLLC performance in FBE: multiple configurations of FFP and the arrangement of FFP' starting point based on data priority. Chapter 7 continues the work of Chapter 5 and Chapter 6 by combining two models of LBE and FBE to use in a system where the LBE and FBE transmitters coexist. The transmitters are allowed to switch dynamically between LBE and FBE to reduce channel access latency and energy consumption.

Chapter 8 focuses on the gap in the middle of PUSCH repetition in unlicensed spectrum that increases latency due to additional LBT and reduces transmission reliability. In the proposed scheme, the gap is eliminated by updating dynamically DL symbols to UL symbols. The gap due to an orphan symbol is also filled by the transmission of a filling signal as DMRS in this symbol.

9.2 Future perspectives

The ongoing Release 17 will be finalized at the end of 2021. One of the subjects in Release 17 is URLLC operation in unlicensed spectrum. Therefore, the analysis of LBT in the theoretical perspective in the thesis is fundamental to evaluate the URLLC performance in unlicensed spectrum. Based on this analysis, Release 17 and the future releases are able to target the URLLC features to guarantee that the URLLC requirements are attained. Moreover, this analysis is not only useful to URLLC design but also other types of transmission, especially time-sensitive communications (TSC) in order to assess the capability of these types of transmission in unlicensed spectrum.

After Release 17 is finalized at the end of 2021, the work of Release 18 will start in 2022. In June 2021, Release 18 content, priorities and timeline are now moving into the discussion and decision phase. From high-level considerations, Release 18 is expected to consider eMBB evolution, non-eMBB evolution and cross-functionalities for both eMBB and non-eMBB evolution to meet short-term and long-term commercial needs. The time duration for the release in RAN is tentatively set at 18 months, with the final decision on that to be confirmed in September 2021. The work will culminate with Release 18 Package Approval in December 2021. TSC that is one of the subjects in Release 17 will be continued to be discussed in Release 18. TSC has some characteristics similar to URLLC. While URLLC traffic is characterized by a random packet generation process, often experiencing time intervals of several milliseconds between two adjacent packet arrivals, TSC is characterized by deterministic traffic with fixed inter-packet arrival times such as 0.5 ms, and packets must be delivered according to an agreed time-schedule with microsecond resolution [56]. This means that TSC has a higher data rate. To meet TSC requirements, the radio access network must support a one-way latency down to 0.5 ms with up to six-nines reliability in terms of packet error rate. Some URLLC features in the previous releases as well as the proposed schemes in this thesis can be applied to TSC to ensure the latency and reliability requirements of TSC.

Furthermore, in Release 18, tactile and multi-modality communication is also a study subject. Tactile and multi-modality communication services enable multi-modality interactions, combining ultra-low latency with extremely high availability, reliability and security that are applied in remote human interaction, tele-operation, social network of robots , industry and commercial Internet of things (IoT) services, etc. To support tactile and multi-modality communication services, the 5G system needs to address service requirement of different types of media steams with coordinated throughput, latency and reliability. The schemes in URLLC design are able to play a role in design of tactile and multi-modality communication to provide the required latency and reliability. Besides that, personal IoT networks with a huge number of consumer IoT devices such as door sensors, switch controls, cameras, thermostats, garage door openers, etc. in the home and cameras, headset, earphones, watch, etc. on person would also require some URLLC features to assure their operation.

The proposed schemes in this thesis allow increasing reliability and reducing latency in UL and DL communication would be able to be applied in the use cases of these topics. The methods to mitigate the impact of the collision of the transmission between different UE introduced in Chapter 3 could be used to deal with the collision in a personal IoT network with many devices. The schemes to enhance UL CG transmission in Chapter 2 or DL SPS transmission in Chapter 4 are useful to tactile and multi-modality communication services enable multi-modality interactions with the requirements of latency and reliability as URLLC.

Bibliography

[1] 3GPP TR 38.802 v14.2.0, “Study on new radio access technology physical layer aspects”, 2017.

[2] 3GPP TR 38.913 v16.0.0, “Study on scenarios and requirements for next generation access technologies”, 2020.

[3] 3GPP TS 38.211 v16.3.0, “Physical channels and modulation”, 2020.

[4] 3GPP TS 38.212 v16.3.0, “Multiplexing and channel coding”, 2020.

[5] 3GPP TS 38.213 v16.3.0, “Physical layer procedures for control”, 2020.

[6] 3GPP TS 38.214 v16.3.0, “Physical layer procedures for data”, 2020.

[7] 3GPP TS 38.331 v16.4.1, “Radio Resource Control (RRC) protocol specification”, 2021.

[8] 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.

[9] 3GPP TS 37.213 v16.0.0, “Physical layer procedures for shared spectrum channel access.”, 2020

[10] Chairman’s notes RAN1 #95, Spokane, USA, Nov 12–16, 2018.

[11] Ericsson, “Enhancement of Configured Grant for NR URLLC”, 3GPP R1-1812162, RAN1#95, Spokane, USA, November 12–16, 2018.

[12] Huawei, HiSilicon, “Enhanced UL configured grant transmissions”, 3GPP R1-1812226, RAN1#95, Spokane, USA, November 12–16, 2018.

[13] R. breu, G. Berardinelli, T. Jacobsen, K. Pedersen and P. Mogensen, “A Blind Retransmission Scheme for Ultra-Reliable and Low Latency Communications”, 2018 IEEE 87th Vehicular Technology Conference (VTC Spring), June 2018.

[14] Z. Zhou, R. Ratasuk, N. Mangalvedhe and A. Ghosh, “Resource Allocation for Uplink Grant-Free Ultra-Reliable and Low Latency Communications”, 2018 IEEE 87th Vehicular Technology Conference (VTC Spring), June 2018.

[15] C. Stefanovic, E. Paolini, and G. Liva, “Asymptotic performance of coded slotted aloha with multipacket reception,” IEEE Communications Letters, vol. 22, no. 1, pp. 105–108, 2018.

[16] Huawei, HiSilicon, “UL inter-UE transmission prioritization and multiplexing”, 3GPP R1-1810158, RAN1#94bis, Spokane, Chengdu, China, 8–12 October, 2018.

[17] ZTE, “UL inter-UE multiplexing between eMBB and URLLC”, 3GPP R1-1900074, RAN1 AH 1901, Taipei, 21–25 January, 2019.

[18] Sony, “Inter-UE uplink multiplexing of URLLC & eMBB traffics”, 3GPP R1-1812745, RAN1#95, Spokane, USA, November 12–16, 2018.

[19] Institute for Information Industry (III), “UL Inter UE Tx prioritization/multiplexing”, 3GPP R1-1813481, RAN1#95, Spokane, USA, November 12–16, 2018.

[20] NEC, “UL inter-UE multiplexing of eMBB and URLLC”, 3GPP R1-1812419, RAN1#95, Spokane, USA, November 12–16, 2018.

[21] C.-P. Li, J. Jiang, W. Chen, T. Ji, and J. Smee, “5G ultra-reliable and low-latency systems design,” in 2017 European Conference on Networks and Communications (EuCNC), Jun. 2017, pp. 1-5.

[22] D. Tse and P. Viswanath, Fundamentals of Wireless Communication. Cambridge University Press, 2005.

[23] Chairman’s notes RAN1 #98b, Chongquing, China, October 14–20,2019.

[24] Huawei, HiSilicon, SIA,“Other aspects for URLLC/IIOT enhancements”, 3GPP R1-1909317, Prague, Czech Republic, August 26–30, 2019.

[25] WILUS Inc.,“On SPS PDSCH for NR URLLC”, 3GPP R1-1911320, Chongqing, China, October 14– 20, 2019.

[26] Vivo,“Other issues for URLLC”, 3GPP R1-1908164, Prague, Czech Republic, August 26–30, 2019.

[27] ZTE,“Other enhancements for Rel-16 URLLC”, 3GPP R1-1910106, Chongqing, China, October 14– 20, 2019.

[28] CMCC,“Discussion on DL SPS enhancements”, 3GPP R1- 1910169, Chongqing, China, October 14– 20, 2019.

[29] ETSI EN 301 893, “5 GHz RLAN; Harmonised Standard covering the essential requirements of article 3.2 of Directive 2014/53/EU (v2.1.1)”, May, 2017.

[30] G. J. Sutton et al., “Enabling Ultra-Reliable and Low-Latency Communications through Unlicensed Spectrum,” in IEEE Network, vol. 32, no. 2, pp. 70-77, March-April 2018.

[31] G. Bianchi, “Performance analysis of the IEEE 802.11 distributed coordination function,” in IEEE Journal on Selected Areas in Communications, vol. 18, no. 3, pp. 535-547, March 2000.

[32] 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, 2019, pp. 1-7.

[33] T. Tao, F. Han and Y. Liu, ”Enhanced LBT algorithm for LTE-LAA in unlicensed band,” 2015 IEEE 26th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC), Hong Kong, 2015, pp. 1907-1911.

[34] R. M. Cuevas, C. Rosa, F. Frederiksen and K. I. Pedersen, ”On the Impact of Listen-Before-Talk on Ultra-Reliable Low-Latency Communications,” 2018 IEEE Global Communications Conference (GLOBECOM), Abu Dhabi, United Arab Emirates, 2018, pp. 1-6.

[35] Y. Zeng, Y. Wang, S. Sun and K. Yang, ”Feasibility of URLLC in Unlicensed Spectrum,” 2019 IEEE VTS Asia Pacific Wireless Communications Symposium (APWCS), Singapore, 2019, pp. 1-5.

[36] J. Um, S. Park and Y. Km, “Analysis of channel access mechanism on 5 GHz unlicensed band,” 2015 International Conference on Information and Communication Technology Convergence (ICTC), Jeju, 2015, pp. 898-902.

[37] A. Abdelfattah, N. Malouch and J. Ling, “Analytical Evaluation and Potentials of Frame Based Equipment for LTE-LAA/WIFI Coexistence,” 2019 IEEE Symposium on Computers and Communications (ISCC), Barcelona, Spain, 2019, pp. 1-7.

[38] J. Li, H. Shan, A. Huang, J. Yuan and L. X. Cai, "Modelling of synchronisation and energy performance of FBE- and LBE-based standalone LTE-U networks," in The Journal of Engineering, vol. 2017, no. 7, pp. 292-299.

[39] Gordon J. Sutton, Ren Ping Liu, Y. Jay Guo, "Coexistence Performance and Limits of Frame-Based Listen-Before-Talk," IEEE Transactions on Mobile Computing, pp. 1084-1095, vol. 19, May 2020.

[40] G. P. Wijesiri and F. Y. Li, "Frame Based Equipment Medium Access in LTE-U: Mechanism Enhancements and DTMC Modeling," GLOBECOM 2017 - 2017 IEEE Global Communications Conference, Singapore, 2017, pp. 1-6.

[41] B. Jia and M. Tao, "A channel sensing based design for LTE in unlicensed bands," 2015 IEEE International Conference on Communication Workshop (ICCW), London, 2015, pp. 2332-2337.

[42] N. Wei, X. Lin, Y. Xiong, Z. Chen and Z. Zhang, "An Optimal Stopping Approach to Listen-Before-Talk for Frame Based Equipment in Unlicensed Spectrum," GLOBECOM 2017 - 2017 IEEE Global Communications Conference, Singapore, 2017, pp. 1-6.

[43] N. Wei, X. Lin, Y. Xiong, Z. Chen and Z. Zhang, "Joint Listening, Probing, and Transmission Strategies for the Frame-Based Equipment in Unlicensed Spectrum," in IEEE Transactions on Vehicular Technology, vol. 67, no. 2, pp. 1750-1764, Feb. 2018.

[44] Z. Wang, H. M. Shawkat, S. Zhao and B. Shen, "An LTE-U coexistence scheme based on cognitive channel switching and adaptive muting strategy," 2017 IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC), Montreal, QC, 2017, pp. 1-6.

[45] R. Maldonado, C. Rosa and K. I. Pedersen, "A Fully Coordinated New Radio-Unlicensed System for Ultra-Reliable Low-Latency Applications," 2020 IEEE Wireless Communications and Networking Conference (WCNC), Seoul, Korea (South), 2020, pp. 1-6.

[46] Huawei, HiSilicon, "Uplink enhancements for URLLC in unlicensed controlled environments", 3GPP R1-2007568, 3GPP TSG RAN WG1 Meeting #103-e, Oct 26 – Nov 3, 2020.

[47] InterDigital, Inc., "Enhancements for unlicensed band URLLC/IIoT", 3GPP R1-2101291, 3GPP TSG RAN WG1 Meeting #104-e, Jan 25 – Feb 5, 2021.

[48] Nokia, Nokia Shanghai Bell, "Revised WID: Enhanced Industrial Internet of Things (IoT) and ultra-reliable and low latency communication (URLLC) support for NR", 3GPP RP-201310, 3GPP TSG RAN Meeting #88e, June 29 – July 3, 2020.

[49] S. Khoshabi Nobar, M. H. Ahmed, Y. Morgan and S. Mahmoud, "Joint Channel Assignment and Occupancy Time Optimization in Frame-Based Listen-Before-Talk," in IEEE Communications Letters, vol. 24, no. 3, pp. 695-699, March 2020.

[50] Y. Song, K. W. Sung, Y. Han, "Coexistence of Wi-Fi and Cellular With Listen-Before-Talk in Unlicensed Spectrum," IEEE Communications Letters, vol. 20, issue 1, pp.161-164, January, 2015.

[51] Qualcomm Incorporated, "Uplink enhancements for URLLC in unlicensed controlled environments", 3GPP R1-2006801, e-Meeting, August 17–28, 2020.

[52] Huawei, HiSilicon, "PUSCH enhancements for URLLC", 3GPP R1-1908053, Prague, Czech Republic, August 26–30, 2019.

[53] Nokia, Nokia Shanghai Bell, "On PUSCH enhancements for NR URLLC", 3GPP R1-1908438, Prague, Czech Republic, August 26–30, 2019.

[54] Apple Inc,"Discussion on Orphan symbol handling for unlicensed spectrum", 3GPP R1-2006518, e-Meeting, August 17–28, 2020.

[55] A. A. Esswie and K. I. Pedersen, "On the Ultra-Reliable and Low-Latency Communications in Flexible TDD/FDD 5G Networks", 2020 IEEE 17th Annual Consumer Communications & Networking Conference (CCNC), Las Vegas, NV, USA, 2020, pp. 1-6.

[56] R. B. Abreu, G. Pocovi, T. H. Jacobsen, M. Centenaro, K. I. Pedersen and T. E. Kolding, "Scheduling Enhancements and Performance Evaluation of Downlink 5G Time-Sensitive Communications," in IEEE Access, vol. 8, pp. 128106-128115, 2020.

[57] "Study on supporting tactile and multi-modality communication services", 3GPP SP-201039, e-Meeting, December 8–14, 2020.

[58] "Study on Personal IoT Networks", 3GPP SP-200592, e-Meeting, June 30 – July 3, 2020.