Shariatmadari, Hamidreza; Iraji, Sassan; Jantti, Riku; Popovski, Petar; Li, Zexian; Uusitalo, Mikko A.

Fifth-Generation Control Channel Design

Published in:
IEEE Vehicular Technology Magazine

DOI:
10.1109/MVT.2018.2814378

Published: 01/06/2018

Document Version
Peer-reviewed accepted author manuscript, also known as Final accepted manuscript or Post-print

Please cite the original version:
Shariatmadari, H., Iraji, S., Jantti, R., Popovski, P., Li, Z., & Uusitalo, M. A. (2018). Fifth-Generation Control Channel Design: Achieving Ultrareliable Low-Latency Communications. *IEEE Vehicular Technology Magazine*, *13*(2), 84-93. https://doi.org/10.1109/MVT.2018.2814378

FIFTH-GENERATION CONTROL CHANNEL DESIGN

Achieving Ultrareliable Low-Latency Communications

Hamidreza Shariatmadari, Sassan Iraji, Riku Jäntti, Petar Popovski, Zexian Li, and Mikko A. Uusitalo

The fifth generation (5G) of wireless systems holds the promise of supporting a wide range of services with different communication requirements. Ultra-reliable low-latency communication (URLLC) is a generic service that enables mission-critical applications, such as industrial automation, augmented reality, and vehicular communications [1]. URLLC has stringent requirements for the reliability and latency of delivering both data and control information. To meet these requirements, the Third Generation Partnership Project (3GPP) has been introducing new features in the upcoming releases of the cellular system standards, i.e., Releases 15 and beyond. This article reviews some of these features and introduces new enhancements for designing the control channels to efficiently support URLLC. In particular, a flexible slot structure is presented as a solution to detect a failure in delivering the control information at an early stage, thereby allowing timely retransmissions of the control information. Finally, some remaining challenges and envisioned research directions are discussed for shaping the 5G new radio (NR) as a unified wireless access technology for supporting different services.

Digital Object Identifier 10.1109/MVT.2018.2814378
Date of publication: 25 April 2018

Implementing Services for a Range of Applications

According to the 3GPP, the main generic services for 5G include enhanced mobile broadband (eMBB), massive machine-type communications (mMTC), and URLLC [2], [3]. For eMBB, high data rates are targeted, which were considered a common objective for previous generations of cellular systems. The service mMTC provides connectivity for a large number of devices, which can further the development of the Internet of Things. The transmission links for the mission-critical applications enabled by URLLC can be either one-to-one, one-to-many, or many-to-many. For instance, augmented reality and remote surgery applications require one-to-one communication links. Vehicular communications, on the other hand, need one-to-one, one-to-many, and many-to-many links to deliver connectivity among vehicles and road infrastructures.

The 3GPP considers two paths for enabling the URLLC. The first path is based on long-term evolution (LTE) and entails backward compatibility with legacy

LTE systems. The other path is based on 5G NR and compels forward compatibility with 5G evolution. This paves the way for fundamental changes to the NR, which can better support the URLLC. While these two paths lead to different network designs, they might benefit from similar techniques for integrating URLLC [4].

However, URLLC can only be implemented if the high-reliability and low-latency features are addressed in the whole system [5]. Given the dynamics of wireless channels, the biggest challenge is to meet these requirements in radio access networks (RANs). The RAN consists of physical channels that carry various types of information, generally categorized as data and control channels. These channels exhibit different impacts on the overall communication performance. Thus, different reliability and latency constraints are imposed on the channels according to the given communication service [4]. Since these constraints are usually stringent for URLLC, new approaches and designs are needed for the data and control channels.

This article describes reliability tradeoffs between the data and control channels, which can help to identify the reliability requirements for these channels. To meet the reliability constraints, various solutions are presented that could be applicable in the design of the 5G NR. Specifically, these solutions ensure high reliability for delivering the scheduling request (SR), resource grant (RG), channel quality indicator (CQI) report, and hybrid automatic repeat request (HARQ) feedback. Furthermore, a flexible slot structure is proposed to identify a failure in delivering the control information at an early stage. This reduces the latency by taking the relevant actions in a timely manner.

URLLC Requirements and Enablers

The goal of the 3GPP is to support a communication reliability corresponding to a block error rate (BLER) of 10^{-5} and up to 1-ms radio latency for delivering short packets of up to 32 B. This target is specified by setting a user plane latency of 0.5 ms for both the uplink and downlink. The latency requirement is relaxed to 3–10 ms for supporting the enhanced vehicle-to-everything, which facilitates autonomous driving, with larger packet sizes of up to 300 B [2]. While these requirements are satisfactory for many mission-critical applications, more stringent requirements might be essential to support some other envisioned applications, particularly in the realm of industrial automation and vehicular communications.

The 3GPP has introduced new techniques for LTE Release 14 and 15 to support URLLC. These include fast uplink access, short transmission time interval (sTTI), and shortened processing time, thereby reducing the user plane latency. In the legacy LTE, a user equipment (UE) needs to send an SR to be granted the radio resources for transmitting its data. However, fast uplink access enables reserving radio resources for the UE, which can be used for uplink data transmissions whenever the UE has something to send. This reduces the latency because the UE does not need to send an SR and wait for the RG. Employing the sTTI is another approach for reducing the transmission latency. The legacy LTE defines a subframe spanning 14 symbols, resulting in a transmission time interval (TTI) of 1 ms. An sTTI can be formed by reducing the transmission duration, i.e., using a minislot that spans two to seven symbols. The shortened processing time can further reduce the latency by sending the HARQ feedback faster than the legacy LTE, by which the feedback is sent after at least four subframes from the time of receiving the data. A potential enhancement for improving the reliability is the dual connectivity. In such a case, the UE can simultaneously communicate with multiple access nodes.

The 5G NR offers promising features that better support URLLC. Some of the relevant features include access to the high bandwidths; support for massive multiple input, multiple output (MIMO) antennas; enabling device-to-device (D2D) communications; the introduction of new channel coding schemes; and configurable subcarrier spacing [3], [6]. The NR can access a wide range of spectrum, including the millimeter wave (mm-wave), which provides abundant radio resources for different services. In addition, employing the mm-wave allows for massive MIMO antenna systems comprising a large number of antennas accommodated at a base station, referred to as a *next-generation NodeB (gNB)* in 5G. This leads to better channel qualities and an increase in the system capacity. The communication latency can be reduced by using the D2D communications, in which UEs communicate directly without passing data through the gNB [5]. The NR supports both low-density parity checks (LDPCs) and polar coding schemes. Specifically, the LDPC is applied to both uplink and downlink data transmissions, which exhibit good BLER performance for URLLC. One of the best features of the NR is its subcarrier spacing configurability with the values of 15, 30, 60, 120, and 240 kHz [6]. This accommodates a different number of slots within a 1-ms subframe and obtains the TTI of 1, 0.5, 0.25, 0.125, and 0.0625 ms, respectively. However, the highest subcarrier spacing that supports data transmissions is 120 kHz, corresponding to a TTI of 0.125 ms.

In addition, a large variety of slot formats is introduced that bring flexibility to the scheduling. The slot configurations can be categorized according to the symbol types, as illustrated in Figure 1. There are three different symbol types: uplink, downlink, and flexible. A UE assumes a downlink transmission through the downlink or flexible symbols, while it transmits by the uplink or flexible symbols [6]. The support of both downlink and uplink symbols within a slot is a promising feature for

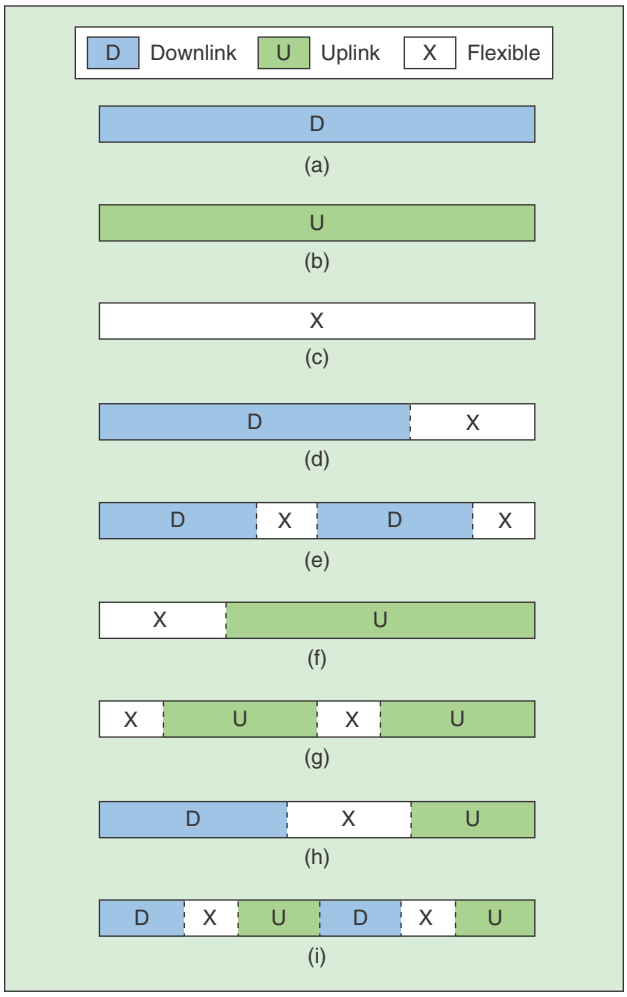

FIGURE 1 (a) A slot containing only downlink symbols; (b) a slot containing only uplink symbols; (c) a slot containing only flexible symbols; (d) a slot containing downlink and flexible symbols; (e) a slot containing downlink and flexible symbols; (f) a slot containing flexible and uplink symbols; (g) a slot containing flexible and uplink symbols; (h) a slot containing downlink, flexible, and uplink symbols; and (i) a slot containing downlink, flexible, and uplink symbols.

supporting URLLC, which reduces the latency. For instance, using the slot format shown in Figure 1(h) for a downlink transmission enables the UE to receive data at the beginning of the slot and report the corresponding HARQ feedback at the end of the same slot. The same format can be employed for an uplink transmission; the UE receives the uplink grant at the beginning of the slot and sends its data at the end of the slot.

URLLC provides reliable data and control channels. To better understand the effects of data and control channels on the overall communication reliability, we consider schedule-based communications for uplink and downlink data transmissions, as shown in Figure 2. For the uplink transmissions, a UE must send an SR to a gNB to access the radio resources. When the SR is detected, the gNB allocates the radio resources for the uplink data transmission. The gNB informs the UE about the allocated resources by sending the RG. The UE can transmit uplink data once the RG is decoded. If the gNB cannot retrieve the message correctly, it triggers the UE to retransmit the data. For adaptive data retransmissions, the gNB sends a new RG to the UE, indicating the allocated radio resources for the data retransmission. The procedure of data retransmissions continues until either the message is decoded successfully or the maximum number of retransmissions is reached. The maximum number of retransmissions depends on the different parameters, such as the latency requirement, the TTI duration, and the processing time. However, there is a common consensus that a maximum number of retransmissions should not be more than one due to the latency constraint [2], [4].

For downlink transmissions, the gNB needs an estimate of the downlink channel quality for handling the link adaptation, which is done by using the CQI report sent by the UE. The gNB allocates radio resources for the downlink data transmission, according to the CQI report, and instructs the UE by sending the RG to monitor the radio resources for retrieving the message. Upon decoding the RG, the UE decodes the message and sends either an acknowledgment (ACK) or a negative-acknowledgment (NACK) signal to indicate the success or failure in the data reception. If the gNB does not receive an ACK signal, it retransmits the data. The gNB again instructs the UE to monitor the allocated resources for the data retransmission by sending a new RG. The procedure of data retransmissions continues until either the gNB finally receives an ACK signal or the maximum number of retransmissions is reached. Similar to the uplink transmissions, a maximum of one retransmission is envisioned due to the latency constraint.

The uplink and downlink communications rely on transmitting the data and on the control information. Both data and control channels are prone to errors, which can affect the overall communication reliability. However, the effects of the errors in the data and control channels are different. For instance, one source of error is missing the RG that results in not sending the data in the uplink or listening to the incoming downlink data. This error might happen during the initial and/or transmission round. In the uplink, the gNB distinguishes this event when it does not receive any data from the UE; in the downlink, the gNB identifies this event when it does not receive an ACK or a NACK signal, known as *discontinuous transmission* (*DTX*). In case the gNB identifies the missing RG for the initial transmission round, it can allocate more radio resources for the retransmission round to compensate for the loss of the initial transmission. However, if the gNB detects the DTX erroneously as an ACK signal, no retransmission is triggered.

Another type of error is related to the CQI report, which carries an index derived according to the

measured signal-to-interference plus noise ratio (SINR) and BLER target for the data transmission. The gNB might decode the CQI report erroneously as a higher or lower value. Decoding the CQI report as a lower value results in employing an excessively robust modulation and coding scheme (MCS) for data transmission, thereby not degrading the communication reliability. However, incorrectly decoding the CQI report as a higher value leads to an MCS with a high transmission rate, which is less reliable. Another type of error is related to misinterpretation of ACK/NACK signals. The erroneous decoding of an ACK as a NACK triggers unnecessary data retransmission, which results in wasted resources, while the erroneous decoding of a NACK as an ACK leads to the absence of a necessary retransmission. The errors of ACK/NACK signals affect only the retransmission round.

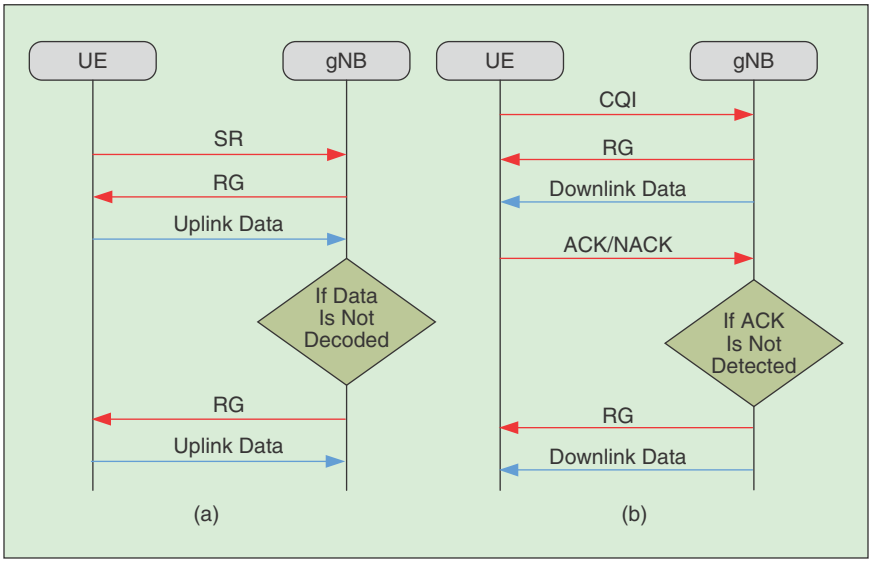

FIGURE 2 The schedule-based data transmissions in the (a) uplink and (b) downlink.

Consider uplink data transmissions. The failure rates of delivering the SR and the RG are ϵ_{SR} and ϵ_{RG}, respectively. The initial data transmission is performed with the BLER of P_1. The BLER of decoding the message using the received information from the both the initial data transmission and the retransmission is $P_{1,2}$. The BLER of P_2 is considered for decoding the message when the initial transmission is not triggered due to the missing RG. Considering the errors of data and control channels, the probability of successfully delivering a message can be expressed as [4]

$$\begin{aligned} P_{UL} = {} & (1-\epsilon_{SR})(1-\epsilon_{RG})\{(1-P_1)+P_1(1-\epsilon_{RG})(1-P_{1,2})\} \\ & +\epsilon_{SR}(1-\epsilon_{SR})(1-\epsilon_{RG})(1-P_1) \\ & +(1-\epsilon_{SR})\epsilon_{SR}(1-\epsilon_{RG})(1-P_2). \end{aligned}$$

Figure 3(a) illustrates the reliability requirements for the control information to meet the reliability of $1-10^{-5}$ in the uplink. The initial transmission is performed with three different reliabilities, while the retransmission ensures the BLER of 10^{-5}, i.e., $P_{1,2}=10^{-5}$ is achieved. It is assumed that $P_2=P_1$. The target of communication reliability can be met only if the error rates of the control information are within the reliability regions. There are tradeoffs between the reliabilities of data and control channels. For instance, ϵ_{SR} and ϵ_{RG} should be less than 10^{-4} if the initial data transmission confirms the BLER of 10%. These requirements can be relaxed by performing the initial transmission more reliably using a more robust MCS; however, this results in utilizing more radio resources for data transmissions [7], [8]. For example, the initial data transmission with the BLER of 1% entails that ϵ_{SR} and ϵ_{RG} be less than 10^{-3}.

Now consider downlink transmissions and assume that the gNB has the perfect knowledge of the downlink channel quality. The failure rate of delivering the RG is ϵ_{RG}. The initial transmission ensures the BLER of P_1. The probabilities of erroneously decoding a NACK as an ACK and a DTX are ϵ_{NA} and ϵ_{ND}, respectively, while the probabilities of incorrectly detecting a DTX as an ACK and a NACK are correspondingly ϵ_{DA} and ϵ_{DN}. The BLER of decoding a message using the received information from the initial transmission and retransmission rounds is $P_{1,2}$. If the gNB detects a DTX, it assumes that the UE could not receive any data information from the initial transmission round; therefore, it can perform the retransmission more robustly. The BLER of decoding the message for this case is P_{2D}. However, if the gNB decodes a DTX erroneously as a NACK, it retransmits data assuming that the UE has received the data from the initial transmission round, although it cannot decode the message successfully. In this case, the BLER of decoding the message is reduced to P_{2N}. The probability of successfully delivering a message can be expressed as [4]

$$\begin{aligned} P_{DL} = {} & (1-\epsilon_{RG})\{(1-P_1)+P_1(1-\epsilon_{NA}-\epsilon_{ND})(1-P_{1,2}) \\ & +\epsilon_{ND}(1-\epsilon_{RG})(1-P_{2D})\}+\epsilon_{RG}(1-\epsilon_{RG}) \\ & \times\{\epsilon_{DN}(1-P_{2N})+(1-\epsilon_{DN}-\epsilon_{DA})(1-P_{2D})\}. \end{aligned}$$

Figure 3(b) illustrates the reliability requirements for the control information to achieve the reliability of $1-10^{-5}$ in the downlink. The initial transmission round is performed with three different reliability targets. The data retransmission ensures the

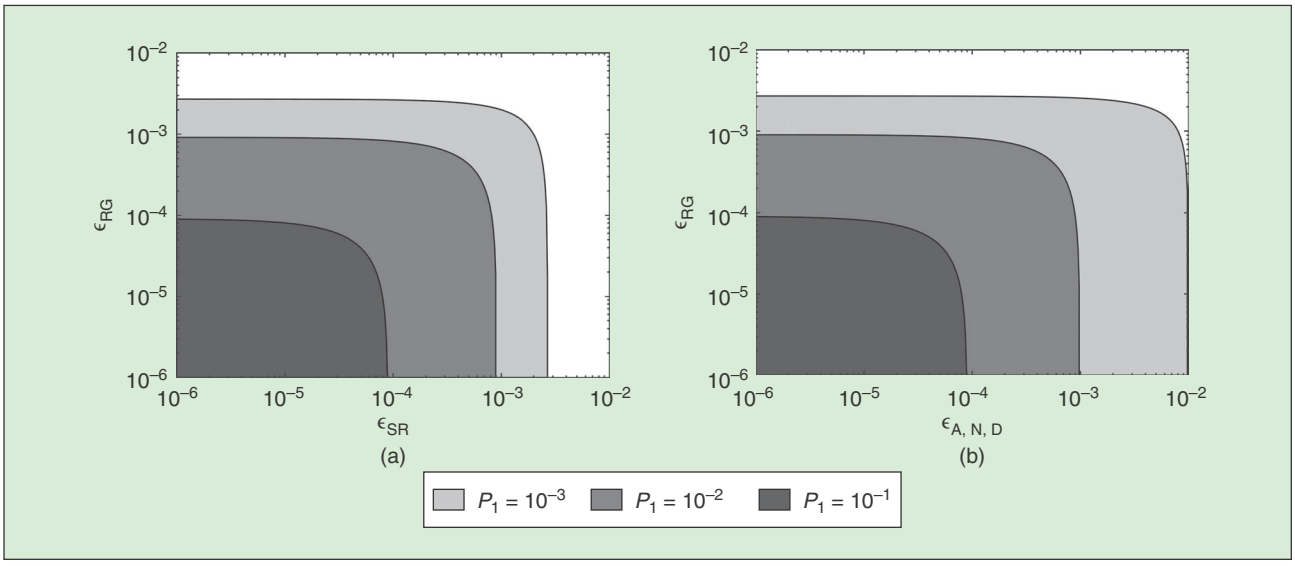

FIGURE 3 The reliability requirements for the control information in the (a) uplink and (b) downlink.

remaining BLER of 10^{-5}, i.e., $P_{1,2} = P_{2D} = 10^{-5}$. In addition, it is assumed that $P_{2N} = P_1$. For simplicity, we presume that $\epsilon_{A,N,D} = \epsilon_{NA} = \epsilon_{DA} = \epsilon_{DN}$. The results show the similar tradeoffs between the reliabilities of data and the control channels. However, the reliability constraint for the HARQ feedback, i.e., $\epsilon_{A,N,D}$, is quite different from that of the RG. This is because decoding the RG is a prerequisite for both the initial transmission and the retransmission rounds, while the ACK/NACK signals can affect only the retransmission round.

These observations indicate that URLLC entails higher reliability constraints for data and control channels than those offered by the legacy LTE (for instance, LTE complies 1% BLER for the RG, 1% for the probability of the ACK misdetection, and 1% BLER for CQI [9]). In the next section, we describe approaches that help improve the reliability of control channels and offer better communication performance for supporting URLLC.

Technical Challenges for Control Channels and Proposed Solutions

Future cellular systems need to provide higher levels of reliability for data and control channels to support URLLC. While using redundant resources is a trivial solution for improving the reliability, it also significantly reduces the communication efficiency. This prompts the usage of new approaches for designing the data and control channels to improve the reliability without degrading the communication efficiency. In addition, the new design should be able to support other services, such as eMBB and mMTC, at the same time. We present possible solutions for improving the reliability and the performance of delivering the control information. The promising solutions are provided separately for each type of control information.

The SR

A UE in a connected mode needs to send an SR to a gNB to be scheduled for the uplink data transmission. In LTE, the SR is carried over the physical uplink control channel (PUCCH), and the base station uses energy detection to identify it. Each UE is configured with periodic orthogonal resources on the PUCCH. The UE can send the SR using only predefined resources. When the UE wants to send data, it needs to wait until it has access to the PUCCH, introducing a random delay before the UE can access the channel. If the SR is not detected, the UE will not receive the RG for the uplink transmission. Consequently, the UE needs to retransmit the SR, resulting in further delay. This delay can be reduced by assigning PUCCH resources to the UE more frequently, e.g., every TTI; nevertheless, this results in wasting significant resources, particularly when the UE generates sporadic data traffic. To reduce the delay associated with the SR transmission while not wasting excessive radio resources, some of the following approaches can be considered:

- *Grant-free transmission*: Reserving radio resources for delivering the SR is not efficient for applications that generate sporadic data traffic. Instead, such applications can use grant-free transmission schemes to carry the data without sending the SR. For instance, the UE can send data along with the preamble used for establishing a link [10]. However, the main issue with such schemes is the transmission collisions from different UEs that reduce the communication reliability. This can be improved by sending a few replicas of the message, which increases the chance of receiving one of them successfully.
- *Quality of service (QoS)-based SR*: The SR in LTE does not carry any information about the constraints on the data delivery in terms of the latency and reliability. In

addition, the gNB does not know if the received SR is from the initial or the retransmission attempt. One enhancement is to include additional information regarding the communication requirements in the SR. For instance, the SR can carry information regarding the time budget and the required reliability for delivering the message. The gNB can use this information to allocate resources for transmission more efficiently. For example, the gNB would select more robust MCSs for the transmission if the time budget is low due to the buffer latency or missing the previous SR by the gNB. The inclusion of such information can also relax the reliability constraint on the SR [4].

- *Group-based SR*: The radio resources for the SR can be divided into different groups associated with different QoS. For instance, URLLC can access a set of resources to send the SR while allowing the eMBB access to another set of resources. Users accessing the former resources are scheduled using shorter TTI compared with other users. This allows multiplexing different services more efficiently.

The RG

The gNB delivers the downlink and uplink resource grants by sending the RG. In LTE, the RG is delivered over the physical downlink control channel (PDCCH). Decoding the RG is a prerequisite for sending and receiving data such that it requires high levels of reliability (see Figure 3). The following enhancements can be considered for delivering the RG:

- *Supporting higher aggregation levels*: LTE supports four different aggregation levels for PDCCH, which offer different reliability levels. For URLLC, the higher aggregation levels can be introduced to provide higher reliability. Another way is to send replicas of the RG using different resources in the PDCCH. This allows exploiting the frequency diversity gain.
- *In-resource control signaling*: To provide more flexibility for encoding the RG, it can be carried over the data channel [11]. This allows employing different code rates for the RG. However, the UE needs to monitor a wide spectrum to find the RG, resulting in high power consumption.
- *Joint data and control channel coding*: The efficiency of the coding scheme increases with the size of the input data [12]. However, the sizes of the RG and data for URLLC are quite small, which reduces the communication efficiency. For downlink transmissions, the coding scheme can be applied jointly on the RG and the data to improve the efficiency. Nevertheless, this approach might increase the complexity of the decoding procedure and the power consumption at the UE because it needs to decode both the RG and the data.
- *Semipersistent scheduling (SPS) and fast uplink access*: For periodic data transmission, a SPS can be applied. In this way, the UE is informed about a set of resources that are reserved for it, such that the UE can send or receive data without the need to receive the RG. If the initial transmission fails, the gNB allocates additional resources and informs the UE by sending the RG [4]. The fast uplink access, which is introduced in the new releases of LTE, can be used for nonperiodic data transmissions. This enables the UE to utilize the reserved resources only when it has data.
- *Advance (anticipative) RG transmission*: In LTE, the RG is sent for each data transmission or reception. In case a retransmission is required, a new RG is transmitted later. One of the solutions already agreed upon for 5G NR is that the RG carries the resource allocations for a set of transmission or reception instances. For instance, the RG can indicate the radio resources for both the initial transmission and retransmission. This approach improves the reliability of the RG detection while imposing more signaling overhead as the RG carries information regarding the multiple transmissions.

The CQI

The CQI carries the downlink channel quality information. The UE derives the CQI according to the estimated SINR. The UE estimates the SINR by measuring the reference signals transmitted by the gNBs in different cells. The UE reports the CQI to the gNB, which is ultimately used for the link adaptation. In LTE, the UE maps the SINR to the CQI by selecting the highest MCS that guarantees at least 10% BLER for a single transmission. In addition, there are 16 total CQI indexes that are collectively represented by 4 b. The CQI can be derived for the wideband, UE selected subbands, and the higher layer configured subbands. The wideband CQI is carried over the PUCCH, primarily using reserved radio resources periodically. In this case, the 4-b CQI value is encoded into 20 b for a protection against noise and interference. Generally, there are two different issues associated with the CQI report. One is related to the CQI decoding, i.e., decoding a CQI as a higher or a lower value. Another issue for the CQI report is the time gap between the channel measurement and the actual data transmission, during which the channel might change unfavorably [13]. Some of solutions for these issues are as follows:

- *Configurable CQI report*: A wideband CQI is carried over the PUCCH using the same number of resources. The lower coding rate can be utilized for the CQI report to provide higher protection. This can be done by allocating more radio resources to the UE for reporting the CQI. Another way is to reduce the content of the CQI report, e.g., using fewer than 4 b to represent the CQI values. The cost is the lower performance of the link adaptation, as only a subset of available MCSs can be used.
- *Delay-based link adaptation*: The delay between the channel report and the data transmission degrades

the accuracy of the CQI report. To obtain more accurate estimates of the channel quality, the UE can be configured to report the PUCCH more frequently [5]. This would increase the signaling overhead and the power consumption. To compensate for the effects of the outdated CQI report, the gNB can consider the CQI report delay while selecting the MCS for data transmissions. In this regard, a more robust MCS is selected when there is a long delay between the CQI report and downlink transmission [13]. This requires providing additional information for the scheduler, such as delay and channel variations.

- *HARQ feedback with an updated CQI*: To reduce the signaling overhead from the periodic CQI report, the UE can report an updated CQI after the initial downlink data transmission. For instance, the UE reports the CQI along with the NACK if the initial transmission fails.

ACK/NACK Signals

The UE needs to send either an ACK or a NACK signal after receiving the downlink data to indicate the success or failure of decoding the message. In LTE, these signals are carried over the PUCCH, using the same resource size for all UEs. An erroneous detection of a NACK as an ACK signal results in suppressing the data retransmission, degrading the overall communication reliability. However, the error in which an ACK is misinterpreted as a NACK results in unnecessary retransmissions of the data, wasting radio resources. LTE has a 1% target for the ACK misdetection probability at a low SINR level with a single antenna. This reliability level is not sufficient for URLLC, as shown in Figure 3. The following approaches can improve the reliability of ACK/NACK signal detection:

- *ACK/NACK repetition*: In LTE, the ACK/NACK repetition is supported to improve the detection reliability for the UEs with bad channel conditions. The UE sends the same ACK/NACK signal multiple times over the consecutive TTIs. The gNB can configure the repetition factor. This scheme is similar to the TTI bundling used for the physical uplink shared channel to improve the reliability of data transmissions, particularly for the edge users. Although the ACK/NACK repetition improves the reliability of the detection, it also introduces additional latency before the retransmission because the retransmission starts only after all of the ACK/NACK repetitions occur. To solve this issue, they can be performed during a single TTI while using different frequency resources.
- *Asymmetric ACK/NACK signal detection*: Protecting the NACK signal is more important than protecting the ACK signal because erroneous NACK detection degrades the communication reliability [4], [7]. This brings forward the idea of using enhanced NACK protection by applying an asymmetric signal detection. For this purpose, the threshold for the binary hypothesis testing can be set to favor the correct detection of the NACK. The cost of this approach is the higher rate of wrong detection of an ACK as a NACK compared to the case of employing a symmetric signal detection, in which the same probability is achieved for the missed detection of the ACK and NACK. This results in performing more unnecessary retransmissions.
- *Early ACK/NACK transmission*: One of the issues in LTE is the high processing time for decoding the data. This postpones the ACK/NACK transmission, i.e., at least four TTIs after receiving the data. This is because the ACK/NACK signal is transmitted after decoding the message. However, an early ACK/NACK transmission can be used by sending the ACK/NACK signal earlier based on the prediction of success or failure in decoding the message even before the message is decoded completely [14].
- *Multibit NACK*: LTE uses a single bit to carry the ACK/NACK signals. Hence, the transmitter does not know how close the receiver's decoder was when attempting to retrieve the message upon receiving the NACK. For URLLC, this can result in a significant decrease in communication efficiency due to the limited number of transmission attempts. One effective solution is to employ multibit NACK to adapt the redundancy of the data retransmission [15].

Flexible Slot Structure

A key challenge of URLLC is providing the high reliability for data transmissions with a limited number of transmission attempts; typically, only one retransmission attempt is envisioned. This situation is aggravated when the errors occur in delivering the control information. For instance, a UE misses the transmission/reception chance if it cannot decode the RG successfully. This motivates us to exploit the flexibility of the 5G NR slot structures to detect a failure in delivering the control information and take immediate compensating actions. We propose a flexible structure scheme that is applicable to both time-division duplex (TDD) and frequency-division duplex (FDD). However, in this section we only focus on the TDD implementation since it is preferred due to its lower complexity and cost for UEs.

Figure 4(a) illustrates schedule-based uplink data transmissions in a TDD system. It is assumed that data should be delivered within two consecutive slots. Employed slots contain downlink, flexible, and uplink symbols. With the conventional approach of using a symbol either for the uplink or the downlink, the flexible symbols can be configured to carry uplink data. Accordingly, the gNB can deliver the downlink control information (DCI) that contains an RG at the beginning of each slot to instruct the UE to deliver uplink data. However, the UE misses the DCI in slot 1 and does not transmit uplink

data. Hence, the gNB must send a new DCI in the next slot, which causes delay before the UE performs its first transmission. In addition, the gNB needs to allocate excessive radio resources for the data transmission in slot 2 because this is the last chance to deliver data within the time budget.

To reduce this time gap, we propose using the flexible symbols for both downlink and uplink transmissions. As shown in Figure 4(b), the gNB identifies that the UE has missed the RG since it does not transmit data in the uplink, i.e., DTX is detected. In this situation, the gNB retransmits the DCI using the flexible symbols. The UE decodes the retransmitted DCI and then starts transmitting data in the uplink. The retransmitted DCI can be the same as or different from the initial DCI to allocate extended resources in the frequency domain for compensating the shortened transmission time. This approach provides the opportunity to have two transmission attempts for delivering the data even if the DCI is missed.

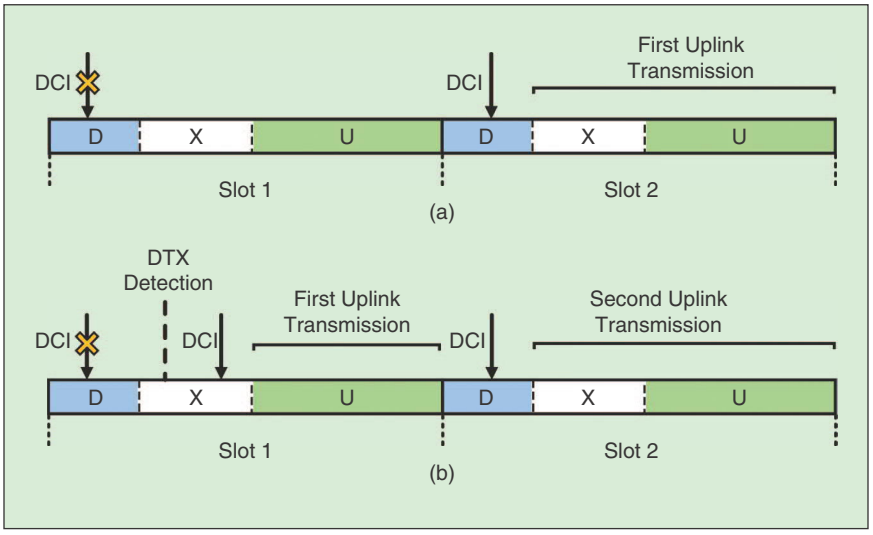

FIGURE 4 Uplink data transmissions with an error in detecting the DCI utilizing (a) the conventional slot structure and (b) the flexible slot structure.

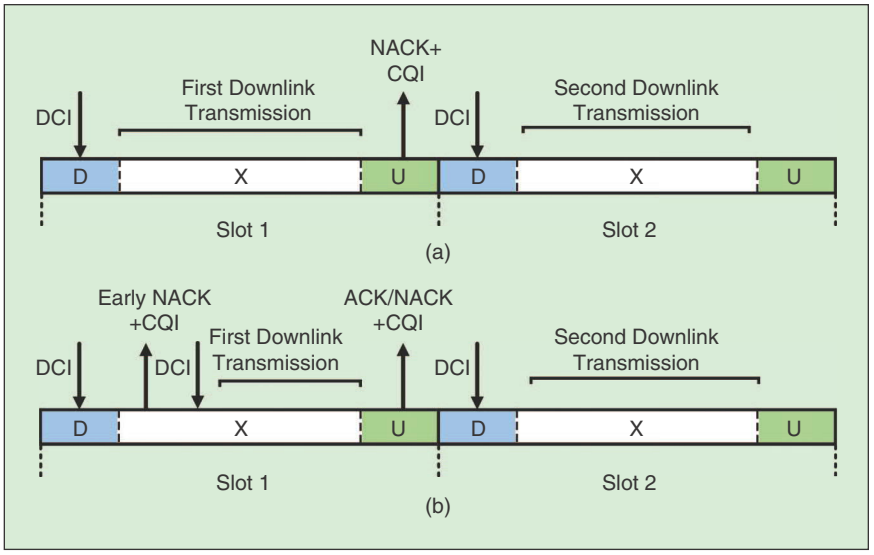

FIGURE 5 Downlink data transmissions with an inappropriate MCS with (a) the conventional slot structure and (b) the flexible slot structure with an early NACK transmission.

The proposed flexible slot structure also can be used for downlink data transmissions. One source of errors is the use of an inappropriate MCS for delivering the data. The gNB might select an inappropriate MCS if it has incorrectly decoded the CQI as a higher value or if the channel condition becomes worse. In such conditions, there is a high chance that the UE cannot decode the message successfully. Figure 5(a) illustrates schedule-based downlink data transmissions with the conventional approach, using a symbol either for the uplink or the downlink transmissions. In this scenario, the flexible symbols are configured for downlink data transmissions. The gNB performs the initial downlink transmission over slot 1 using an inappropriate MCS. The UE tries to decode the message after receiving all of the data and then sends the NACK signal along with the updated CQI for requesting the data retransmission. The gNB will retransmit the data using a more robust MCS.

To address the issue of data transmission with an inappropriate MCS, we propose using flexible symbols for both uplink and downlink transmissions, as shown in Figure 5(b). The UE decodes the DCI and determines the employed MCS and the resource allocations for the downlink transmissions. When the UE identifies that the employed MCS is not appropriate according to the current channel condition, it switches to the transmission mode immediately and sends an early NACK along with the updated CQI, using the resources allocated for its downlink transmission. When the gNB detects the early NACK signal, it terminates the concurrent data transmission and allocates additional resources for the UE according to the updated CQI. The gNB sends a new DCI along with the data information in the same slot using a more robust MCS. As the refined downlink transmission uses a more robust MCS in less time, the resource allocations should be expanded in the frequency domain.

The proposed flexible slot structure, which can be implemented by using flexible symbols for both the uplink and downlink transmissions, can reduce the latency and improve the communication efficiency. To employ the proposed scheme, the gNB should be able to operate in full-duplex mode to send and receive simultaneously. However, the UE can still operate in half-duplex mode, which does not impose higher complexity in designing the UE radio.

Conclusions and Future Directions

URLLC applications have different reliability and latency requirements. While 5G NR has the potential to meet these requirements, it can benefit from nontrivial enhancements to better support URLLC. This article presented solutions to improve the performance of delivering different control information utilized for the uplink and downlink transmissions. In addition, the proposed flexible slot structure allows for detecting a failure in delivering the control information at an early stage and taking immediate compensating actions.

It was shown that data and control channels have different effects on the overall communication reliability. In addition, there are tradeoffs between the reliability requirements for these channels. Hence, novel link adaptation and resource allocation schemes are required for the data and control channels. For instance, the resource allocations for the data channel should consider the reliabilities of control information as well as the link quality of the data channel. Another approach is to provide more flexibility for the control channels, so they can be configured to meet the communication requirements for different services. URLLC might be supported by both grant-based and grant-free transmission modes. The radio resources should be assigned for them optimally, and each user is configured to operate in one of these transmission modes according to its traffic type. For grant-based transmissions, the number of redundant transmissions, in the time and frequency domains, is a key parameter affecting the communication reliability and efficiency. The redundant transmissions can be combined with specific patterns to provide better performance.

Another concern for 5G NR is the multiplexing of different services while satisfying their communication requirements. This can bring new challenges, particularly when the system is faced with a sudden traffic surge from the URLLC users. One solution would be to puncture the radio resources allocated to other services to maintain the URLLC users. However, recovery mechanisms are essential for allowing other services to resume their communications. These challenges should be taken into consideration to ensure the efficient support of URLLC in 5G systems.

Author Information

Hamidreza Shariatmadari (hamidreza.shariatmadari@aalto.fi) received his B.Sc. degree in electrical engineering from the University of Tabriz, Iran, in 2009 and his M.Sc. degree (with distinction) in communications engineering from Aalto University, Finland, in 2013. He is currently pursuing a Ph.D. degree at Aalto University in the Department of Electrical Engineering. He also collaborates with Nokia Bell Labs, Finland, developing solutions for supporting ultrareliable low-latency communications in fifth-generation networks. He was a recipient of the Nokia Scholarship. His research interests focus on the development of wireless communication technologies for the efficient support of machine-type communications.

Sassan Iraji (sassan.iraji@aalto.fi) received his B.Sc. degree from Tehran University, Iran, his M.Sc. degree (with distinction) from Helsinki University of Technology, Finland, and his Ph.D. degree from Tampere University of Technology, Finland, in 1992, 1999, and 2005, respectively, all in electrical engineering. He was with Nokia from 1997 to 2012, holding various positions as a senior technology manager, a principal researcher, and a research leader. From 2012 until February 2018, he was with Aalto University, Finland, as a research fellow. Currently, he is with Intel Corporation. He holds numerous patents and publications in the field of wireless communications. His current research interests include wireless communication systems toward fifth generation.

Riku Jäntti (riku.jantti@aalto.fi) received his M.Sc. degree (with distinction) in electrical engineering in 1997 and his D.Sc. degree (with distinction) in automation and systems technology in 2001, both from Helsinki University of Technology (TKK), Finland. He is an associate professor (tenured) in communications engineering and the head of the Department of Communications and Networking at Aalto University's School of Electrical Engineering, Finland (formerly TKK). He is an associate editor of *IEEE Transactions on Vehicular Technology.* His research interests include radio-resource control and optimization for machine-type communications, cloud-based radio access networks, spectrum and coexistence management, and radio-frequency inference.

Petar Popovski (petarp@es.aau.dk) received his Dipl.-Ing./Magister Ing. degree in communication engineering from Sts. Cyril and Methodius University in Skopje, the Republic of Macedonia, and his Ph.D. in wireless communications from Aalborg University, Denmark, where he is currently a professor. He received a Consolidator Grant from the European Research Council in 2015 and the Danish Elite Researcher Award in 2016. He is an area editor of *IEEE Transactions on Wireless Communications.* His research interests are wireless communications/networks and communication theory. He is a Fellow of the IEEE.

Zexian Li (zexian.li@nokia-bell-labs.com) received his B.S. and M.S. degrees in electrical engineering from the Harbin Institute of Technology, China, and his Ph.D. degree in electrical engineering from Beijing University of Posts and Telecommunications, China. He has been a

senior specialist at Nokia Bell Labs, Finland, since 2005, focusing mainly on research and standardization activities on broadband wireless communications and, most recently, on fifth generation (5G) and long-term evolution-advanced Pro. He led the horizontal topic on direct device-to-device within the European Union METIS Project. His research interests include 5G, machine-type communications, the Internet of Things, and future wireless connectivity for improving human life.

Mikko A. Uusitalo (mikko.uusitalo@nokia-bell-labs.com) received his M.Sc. degree in English and his Dr. Tech. and his B.Sc. in economics in 1993, 1997, and 2003, respectively, all from predecessors of Aalto University, Finland. He is the head of the Research Department on Wireless Advanced Technologies at Nokia Bell Labs, Finland. He is a founding member of the *Celtic-Plus—Eureka* ICT Cluster for a Smart Connected World and the Wireless World Research Forum, chairing the latter from 2004 to 2006.

References

[1] P. Popovski, "Ultra-reliable communication in 5G wireless systems," in *Proc. 5G Ubiquitous Connectivity*, 2014, pp. 146–151.
[2] 3GPP, "Study on scenarios and requirements for next generation access technologies," 3GPP, Sophia Antipolis, France, Tech. Rep. 38.913, v14.2, 2017.
[3] I. Akyildiz, S. Nie, S. Lin, and M. Chandrasekaran, "5G roadmap: 10 key enabling technologies," *Comput. Netw.*, pp. 17–48, Sept. 2016.
[4] H. Shariatmadari, Z. Li, S. Iraji, M. A. Uusitalo, and R. Jäntti, "Control channel enhancements for ultra-reliable low-latency communications," in *Proc. IEEE International Conference on Communications Workshops (ICC Workshops)*, 2017, pp. 504–509.
[5] C. She, C. Yang, and T. Q. S. Quek, "Radio resource management for ultra-reliable and low-latency communications," *IEEE Commun. Mag.*, vol. 55, no. 6, pp. 72–78, 2017. doi: 10.1109/MCOM.2017.1601092.
[6] 3GPP, "NR; physical channels and modulation (Release 15)," 3GPP, Sophia Antipolis, France, Tech. Rep. 38.211, V15.0.0, 2017.
[7] H. Shariatmadari, R. Duan, S. Iraji, Z. Li, M. Uusitalo, and R. Jäntti, "Resource allocations for ultrareliable low-latency communications," *Int. J. Wireless Inform. Netw.*, vol. 24, no. 3, pp. 317–327, Sept. 2017.
[8] H. Shariatmadari, Z. Li, M. A. Uusitalo, S. Iraji, and R. Jäntti, "Link adaptation design for ultra-reliable communications," in *Proc. IEEE ICC*, 2016, pp. 1–5.
[9] 3GPP, "LTE; evolved universal terrestrial radio access (E-UTRA); base station (BS) radio transmission and reception," 3GPP, Sophia Antipolis, France, Tech. Rep., 36.104.v10.2.0, 2011.
[10] Y. Beyene, R. Jäntti, and K. Ruttik, "Random access scheme for sporadic users in 5G," *IEEE Trans. Wireless Commun.*, vol. 16, no. 3, pp. 1823–1833, Mar. 2017.
[11] K. I. Pedersen, G. Berardinelli, F. Frederiksen, P. Mogensen, and A. Szufarska, "A flexible 5G frame structure design for frequency-division duplex cases," *IEEE Commun. Mag.*, vol. 54, no. 3, pp. 53–59, Mar. 2016.
[12] G. Durisi, T. Koch, and P. Popovski, "Towards massive, ultrareliable, and low-latency wireless communication with short packets," *Proc. IEEE*, vol. 104, no. 9, pp. 1711–1726, Sept. 2016.
[13] U. Oruthota, F. Ahmed, and O. Tirkkonen, "Ultrareliable link adaptation for downlink MISO transmission in 5G cellular networks," *Inform.*, vol. 7, no. 1, pp. 1–18, Mar. 2016.
[14] G. Berardinelli, S. R. Khosravirad, K. I. Pedersen, F. Frederiksen, and P. Mogensen, "Enabling early HARQ feedback in 5G networks," in *Proc. IEEE Vehicular Technology Conf.*, May 2016, pp. 1–5.
[15] V. Braun, U. Doetsch, A. Zimaliev, M. Bonomo, and L. Vangelista, "Performance of asymmetric QPSK modulation for multi-level ACK/NACK in LTE uplink," *Eur. Wireless*, pp. 1–6, 2014.

VT