

Dynamic URLLC and eMBB Multiplexing Design in 5G New Radio

Wei Yang, Chih-Ping Li, Ali Fakoorian, Kianoush Hosseini, and Wanshi Chen
Qualcomm Technologies, Inc., San Diego, CA, 92121

Abstract—The fifth generation (5G) new radio (NR) of mobile communications is designed to support two major class of services with vastly heterogeneous requirements: ultra-reliable low-latency communication (URLLC) and enhanced mobile broadband (eMBB). To enable the coexistence of URLLC and eMBB in the same radio spectrum in a cost-effective manner while guaranteeing the latency and reliability performance of URLLC, 5G NR has developed an innovative preemption and superposition framework which allows a dynamic multiplexing of the two services in an interleaved manner. In this tutorial paper, we review the key technological innovations in this framework in NR Release 15 and 16, namely, preemption indication and enhanced power control, and explain the design principles underlying these technologies. System-level and link-level simulations are provided to illustrate the performance benefits of these designs.

Index Terms—Dynamic multiplexing, 5G NR, preemption indication, power control, ultra-reliable low-latency communications.

I. INTRODUCTION

The fifth generation (5G) new radio (NR) of mobile communications is designed to support two major class of services: ultra-reliable low-latency communication (URLLC) and enhanced mobile broadband (eMBB). On the one hand, eMBB will provide the next era of immersive, always-connected mobile experiences with fiber-like speeds (gigabit per second data rates) and moderate latency. On the other hand, URLLC will enable new services and applications in which data messages are time-sensitive and critical and must be delivered with high system reliability (e.g., 99.999%) and extremely low latency (e.g., 1 ms) [1], [2]. When the services of URLLC and eMBB in 5G are deployed in the same radio spectrum, how to enable their coexistence to best accommodate the mixed demands in a cost-effective manner becomes a critical problem that needs to be addressed. To solve this problem, the 5G NR standard has developed various innovative techniques that allow the eMBB and URLLC traffics to share the same spectrum by means of *dynamic multiplexing*. The goal of this paper is to provide an overview of dynamic URLLC and eMBB multiplexing design in 5G NR and study its performance.

Let us first review the main techniques in the physical and multiple-access control (MAC) layer to support URLLC in 5G NR Release 15 and Release 16 standards. In NR Release 15—the first release of 5G NR—the fundamental building blocks to reduce latency and ensure reliability are introduced [3]. To reduce latency, NR Release 15 introduces flexible and short transmission duration, scalable Orthogonal frequency-division multiplexing (OFDM) numerologies, flexible frame

structure for time division duplexing (TDD), short user processing timeline for hybrid automatic repeat request (HARQ) and channel state information feedback, and grant-free uplink transmissions. The key techniques to improve reliability at the 5G NR physical layer include better channel coding schemes with low error floor, channel state information reporting for lower block error rate (BLER) target (i.e., 10^{-5}), modulation coding scheme (MCS) table enhancement, and various techniques to exploit time, frequency, and spatial diversity.

Building on the framework laid in Release 15, the second release of 5G NR (i.e., Release 16) provides enhancements to further improve latency and reliability. In particular, the techniques that are currently being standardized are higher-capability downlink control channel monitoring, faster HARQ acknowledgement feedback, enhancements to uplink data channel repetitions, efficient inter- and intra-user uplink channel prioritization and multiplexing, transmission and reception via multiple transmission reception points (multi-TRP) for spatial diversity enhancement, and mechanisms catering for industrial internet of things (IIoT) applications [4]. These enhancements enable new use cases with more stringent reliability and latency requirements, such as electrical power distribution, factory automation, and transport industry including remote driving [5].

The techniques reviewed above enable the 5G NR system to support URLLC services with guaranteed reliability and latency. However, from a system design perspective, it is of vital importance to allow URLLC and other type of traffics such as eMBB to coexist in the same network. While a semi-static resource partition between URLLC and eMBB could guarantee the URLLC reliability and latency, it results in very inefficient resource utilization. Indeed, as is shown in [6], half of the URLLC resources will be wasted under the semi-static resource partition due to queuing effect and low trunking efficiency. Therefore, it is desirable to allow a dynamic resource sharing between the eMBB and URLLC traffics. Furthermore, the dynamic multiplexing of URLLC and eMBB services must be able to provide the same reliability and latency guarantee for URLLC as in the semi-static multiplexing case.

In this paper, we review the techniques developed in 5G NR to support dynamic URLLC and eMBB multiplexing, explain the main design principles underlying such techniques, and illustrate the performance tradeoffs between different dynamic multiplexing schemes using both system-level and link-level simulations. As will become clear in later sections, dynamic multiplexing design on the uplink is considerably more challenging than that in the downlink. Therefore, we

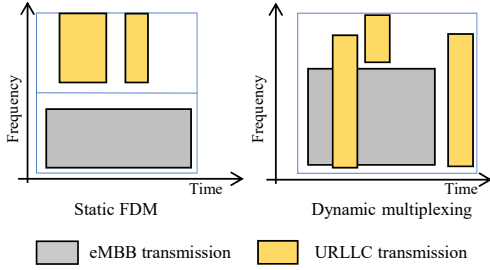


Figure 1. Static FDM versus dynamic multiplexing of eMBB and URLLC

focus on the multiplexing design in the uplink of 5G NR. Throughout the paper, we assume that the URLLC and eMBB transmissions are initiated from different users, i.e., we focus on the inter-user dynamic multiplexing design. The design of intra-user dynamic URLLC and eMBB multiplexing is out of the scope of this paper. Finally, we also remark that, dynamic multiplexing among heterogeneous services have also received extensively attentions in the academic community; we refer the readers to [7]–[9] and references therein.

The remainder of the paper is organized as follows. Section II provides the motivation and challenges for dynamic multiplexing. In Sections III and IV, we introduce the two dynamic multiplexing techniques adopted in 5G NR uplink in Release 16, namely, uplink preemption indication (PI) and enhanced power control, respectively. We also illustrate the performance of the two techniques via system-level and link-level simulations. Finally, Section V concludes the paper.

II. MOTIVATION AND PRELIMINARIES FOR DYNAMIC URLLC AND EMBB MULTIPLEXING

In Fig. 1, we illustrate two methods for the coexistence of URLLC and eMBB: static resource-partitioning (e.g., via frequency-division multiplexing (FDM)) and dynamic multiplexing. As shown in the figure, in the former approach, the resources are statically (or semi-statically) partitioned into orthogonal parts, and are pre-allocated to support the eMBB and URLLC services. Compared with the static FDM approach, dynamic multiplexing of URLLC and eMBB traffics provides large system benefits to both traffic types. On the one hand, it optimizes the capacity of mini-slot based URLLC transmissions by exploiting the statistical multiplexing gain over wider bandwidth. As shown in [6], the URLLC system capacity scales super-linearly and the spectral efficiency improves with the available bandwidth. This is because, wider bandwidth allows more URLLC uplink transmissions to be dynamically multiplexed in the frequency domain in the same mini-slot, thereby reducing the latency at the transmitter side and increasing the number of URLLC users that can be connected to the network. On the other hand, dynamic multiplexing improves the capacity of slot-based eMBB transmissions by fully utilizing the resources that would otherwise be wasted under a static resource reservation for URLLC. When the URLLC traffic load is at the capacity-achieving level, the resource utilization of URLLC transmissions remains relatively low [6],

[10], and the unused resources should be made available to eMBB transmissions (via dynamic multiplexing) to improve overall resource efficiency.

For the design of dynamic URLLC and eMBB multiplexing, it is important to observe that the two type of traffics operate at different transmission time intervals (TTI). URLLC requires short TTI duration, e.g., two OFDM symbols (or $71.43 \mu\text{s}$ under the subcarrier spacing of 30 KHz), to reduce both the HARQ round trip time (RTT) and scheduling delay. For the eMBB traffic that typically involves heavy data transfer, long TTI duration (e.g., 0.5 ms) is preferred to minimize the control overhead and exploit the coding gain of long codewords. Due to the different scheduling granularity, the eMBB and URLLC transmissions will be interleaved in the time domain (see Fig. 2 for an example). To achieve such an interleaved scheduling, new techniques at the physical layer of 5G NR are needed. This is further discussed next.

III. PREEMPTION INDICATION

Consider a scenario where the URLLC and eMBB services dynamically share the same resources. It may occur that at the time when URLLC traffics are generated at the transmitter, all resources have already been allocated to eMBB traffics, and no resources are immediately available to schedule the URLLC transmissions. To overcome this issue, 5G NR has developed an innovative *preemption* framework. In a nutshell, preemption allows the base station to signal the eMBB users to stop or pause their ongoing transmission/receptions, in order to free up resources for urgent URLLC transmissions. In this section, we explain the detailed mechanisms that enables the use of preemption in 5G NR.

A. Preemption in the downlink

In the downlink scenario, the preemption is performed by the base station. That is, the base station itself may pause an eMBB transmission (or part of it) and replace it with an urgent URLLC transmission. Such operation could meet the high and latency requirement of the URLLC service, and is transparent to the URLLC users. In order to reduce the impact of the preemption to an eMBB user, a new signaling known as *downlink preemption indication (PI)* is introduced in 5G NR. It works as follows. Each eMBB user can be configured by the base station to periodically monitor a downlink PI signaling, which indicates a subset of resources in the last monitoring period over which no transmission to the user is present.

A particular example of the downlink PI signaling is shown in Fig. 2. In this example, an urgent URLLC downlink transmission may be scheduled on resources that were previously scheduled for an eMBB downlink transmission. In this case, the base station may indicate to the eMBB user via the downlink PI that some of its resources is preempted by other (URLLC) transmission(s). This indication is signaled to the eMBB user after the actual eMBB downlink transmission took place. The downlink PI consists of a bitmap of 14 bits, each corresponds to a certain time-frequency resource unit.

Finally, the eMBB user may use the preemption information in the downlink decoding, as well as in subsequent HARQ

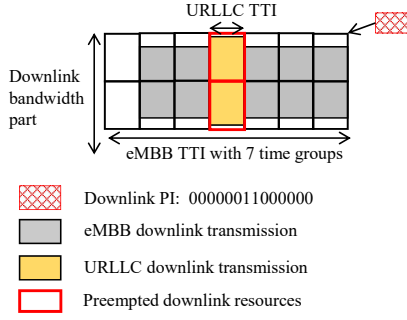


Figure 2. Dynamic eMBB and URLLC multiplexing in the downlink via downlink PI.

retransmissions. For example, the eMBB user may discard the signals (or more specifically the soft information, e.g., the log-likelihood ratios (LLRs)) received in the preempted resources, to avoid corrupting future retransmissions. This way, the damage caused by the preemption to the eMBB user is minimized.

B. Preemption in the uplink

In principle, the same preemption idea could be applied to the uplink scenario. Namely, the base station could signal an uplink PI¹ to the eMBB users to reclaim its uplink resources for an URLLC uplink transmission. However, there is a critical difference from the downlink scenario which makes the uplink preemption considerably more challenging to implement. In the uplink, the preemption is performed by the user instead of the base station. As such, the uplink PI must be signaled to the eMBB user prior to the URLLC uplink transmission, in order not to cause interference to the latter. This is in strong contrast to the downlink case, in which the downlink PI is transmitted *after* the actual downlink data transmission.²

In order for the eMBB user to stop its transmission in a timely manner, the eMBB user must satisfy two requirements.

- Firstly, it must monitor the uplink PI very frequently. For example, the uplink PI monitoring periodicity must be significantly less than 1 ms in order to meet the 1 ms latency requirement for URLLC. In contrast, the downlink PI could be monitored less frequently.
- Secondly, the eMBB user must be able to decode the uplink PI signaling and to stop the uplink transmission quickly. To understand the challenges of this, we note that 5G NR defines the minimum time required for an eMBB user to prepare an uplink transmission to be 12 OFDM symbols³ in the case of 30 KHz subcarrier spacing, or equivalently 0.43 ms. However, such a processing time may be too slow for the purpose of uplink preemption.

¹In the uplink scenario, the preemption indication is also known as uplink cancellation indication [5].

²In principle, the downlink PI may also be transmitted before or along with the URLLC downlink data transmission. However, the 5G NR standard chooses to send it after the URLLC (and eMBB) downlink data transmission.

³This number corresponds to the processing timing capability 1 defined in 5G NR [12, Sec. 6.4]. For users that support URLLC, a processing timing capability 2 is defined in 5G NR, under which the uplink data preparation time is reduced to 5.5 OFDM symbol for 30 KHz subcarrier spacing.

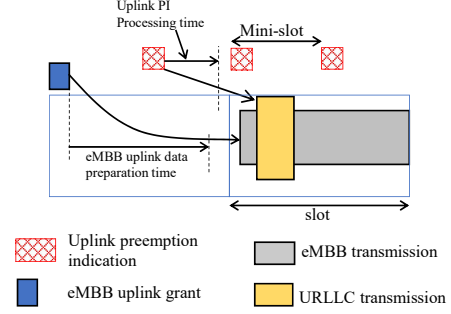


Figure 3. Dynamic eMBB and URLLC multiplexing in the uplink via uplink PI.

An example of the uplink PI which illustrates the two aspects listed above can be found in Fig. 3.

As discussed above, the uplink PI monitoring may bring considerable processing overhead to the eMBB user, which may in turn imply higher energy consumption. In order to reduce the cost of uplink PI, several power saving mechanisms can be considered. For example, to reduce the blind detection overhead, the base station may only configure one physical downlink control channel (PDCCH) candidate for each uplink PI monitoring occasion. This is in contrast to the monitoring of scheduling grant, in which a user may need to perform a large number of blind detections. Another efficient method for power saving is to allow a user to “skip” certain uplink PI monitoring occasions. For example, if the eMBB user has no uplink transmission on a given resource, the user can safely skip the corresponding uplink PI monitoring occasion.

C. Uplink PI for grant-free uplink transmissions

For URLLC uplink communications, the regular grant-based transmission requires a hand-shaking procedure by the user to request resources from the base station and to wait for the uplink grant from the base station. This procedure incurs additional delay overhead and reduces the latency budget for the actual transmission. As a result, it can be shown that under the 5.5 symbol minimum processing time defined in NR Release 15 and 16 (for 30 KHz subcarrier spacing), the URLLC user is not able to perform a HARQ retransmission within the 1 ms deadline. Grant-free uplink transmission (which is also known as configured grant in NR) allows the base station to configure periodic uplink resources for an URLLC user. When the uplink traffic arrives at the user, it can transmit on the configured resources without going through the handshaking procedure, thus reducing the transmission latency.

One may question whether the uplink PI is useful for grant-free uplink transmissions, as the base station is not aware of the traffic activity for the URLLC users *a priori*, and hence may not use the uplink PI to free-up resources prior to an URLLC grant-free uplink transmission. However, as we shall explain in this section, uplink PI can be very beneficial for grant-free uplink transmissions too.

To illustrate the above point, we note that there may be two scheduling methods for URLLC grant-free transmissions to achieve high reliability, as illustrated in Fig. 4. In the first

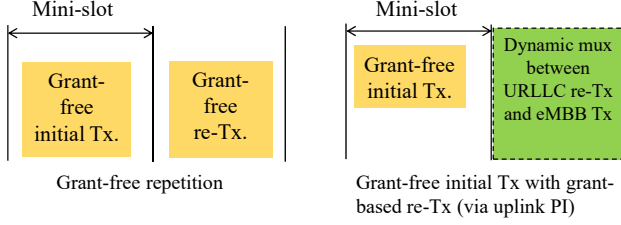


Figure 4. Two scheduling solutions for URLLC: the solution on the left is to schedule grant-free URLLC transmission with two repetitions, and the solution on the right is to schedule a grant-free URLLC initial transmission with a grant-based re-transmission.

approach, the base station configures autonomous repetitions for each grant-free transmission (or equivalently, it configures the grant-free transmission with very low coding rate). In this case, resources for both the initial transmission and the repetition(s) are reserved for the URLLC user. In the second approach, the base station configures a relatively higher coding rate for the grant-free initial transmission, which needs not be ultra reliable. Instead, it relies on grant-based HARQ retransmission(s) to achieve the desired reliability. It is important to note that, in the latter approach, the retransmission resources are *not* reserved for URLLC. Rather, the base station can schedule eMBB transmissions on these resources, and rely on the uplink PI to reschedule them for URLLC retransmissions whenever it is necessary.

The performance of the two scheduling methods is studied via link-level simulations in [10]. For the sake of example, let us assume that an URLLC packet has a payload of 32 bytes (256 bits), that the uplink SNR is -3 dB per receive antenna. It is observed in [10] that, in order to deliver such a packet over a TDL-A channel with reliability 10^{-5} using grant-free repetition, the base station must schedule 40 resource blocks (RB) in each grant-free transmission opportunity. This implies an overall resource consumption of 80 RBs. On the other hand, to achieve 10^{-5} using the second scheduling approach in Fig. 4, the base station could schedule 20 RBs in the initial grant-free transmission, which achieves a initial transmission BLER of 10%. For the retransmissions, 70 RBs are required to deliver the desired 10^{-5} BLER. Overall, it requires $20 + 70 \times 10\% = 27$ RBs on average, which is almost one third of the resources required in the first approach.

From the analysis above, we see that the re-transmission resources in the second scheduling approach is only used 10% of the time. Without dynamic resource sharing, i.e., without being able to reclaim eMBB resources for URLLC transmissions, these resources have to be reserved for the URLLC transmissions, which is highly inefficient. With uplink PI, however, the URLLC retransmission resources can be dynamically shared among eMBB and URLLC users. Thus, the base station may always schedule an eMBB transmission on the resource for potential URLLC retransmissions. When a retransmission for URLLC is indeed required, the base station may use the uplink PI to clean up the resource for the URLLC user.

IV. SUPERPOSITION VIA ENHANCED UPLINK POWER CONTROL

As explained in the previous section, uplink PI generally imposes higher requirements on the (eMBB) user's processing capability. Thus, it may not be feasible to let all users in the 5G NR system to support uplink PI. In addition, the uplink PI technology is not specified until NR Release 16 (partially due to its implementation challenges).⁴ This means that the NR Release 15 user will not have the capability to monitor uplink PI. Therefore, a dynamic multiplexing design that is performed only at the base station and the URLLC user and does not require any additional processing at the eMBB user could be helpful. This is achieved via an enhanced power control mechanism at the URLLC user, which we will introduce in this section.

The main idea of the enhanced power control is that, the base station could schedule an urgent URLLC transmission on top of an ongoing eMBB transmission, and indicate the URLLC user to boost its power to guarantee the reliability. In other words, the eMBB and URLLC transmissions are super-positioned on the overlapping resources (i.e., via non-orthogonal multiple access). As simple as it may seem, enhancements to the traditional uplink power control method is needed to make such idea useful for the desired scenario. To see why this is the case, we briefly review the power control method employed in NR Release 15. In principle, the power for a user to transmit an uplink packet is given by [11, Sec. 7.1]

$$P = \min\{P_{\max}, P_0 + \alpha \cdot PL + 10 \log_{10}(W) + \Delta_{TF} + f(\Delta_{\text{TPC}})\}. \quad (1)$$

Here, P_{\max} denotes the maximum transmit power of the user; P_0 denotes the target received power at the base station; PL denotes the pathloss that is estimated by the user; α denotes the fractional pathloss compensation exponent; W denotes the bandwidth of the uplink transmission; Δ_{TF} is a term that depends on the spectrum efficiency of the uplink transmission; finally, $f(\Delta_{\text{TPC}})$ denotes a power control parameter and is dynamically signaled from the base station to the user for each uplink transmission via transmit power control (TPC) indication. The parameters P_0 and α are known as the open-loop power control parameters, and $f(\Delta_{\text{TPC}})$ is known as the closed-loop power control parameter. In a typical mode of power control, the base station will set the open-loop power control to a properly chosen operating point, under the assumption of orthogonal intra-cell uplink transmission. The base station may then fine-tune the transmit power of the user via closed-loop power control. Since the close-loop power control is mainly used to adapt to the dynamic channel condition (including the fast fading), its control range is set to be from -1 to 3 dB with 1 dB granularity.

As can be inferred from the above discussion, such a power control method may not be sufficient to support dynamic superposition of eMBB and URLLC transmissions. Indeed, the URLLC user may need to boost its power by a significant amount in case its transmission is overlaid on top of an eMBB

⁴In contrast, the downlink PI is supported already in NR Release 15.

transmission. To achieve this goal, two methods have been proposed to 5G NR Release 16 standard: enhanced open-loop and enhanced closed-loop power control. The idea of the enhanced open-loop power control is to let the base station configure two sets of open-loop power control parameters to the (URLLC) user: one set for the interference-free scenario, and one set for the interference-limited scenario (e.g., with increased P_0 value). Furthermore, the base station will dynamically indicate to the URLLC user which set of parameters to use for each uplink transmission. With enhanced closed-loop power control, the idea is to increase the range of TPC parameters beyond the 3 dB limit, thereby allowing the URLLC user to power-boost its transmission if needed.⁵

One major limitation of the enhanced power control approach, compared with uplink PI, is that it is not useful if the URLLC user is already power-limited (e.g., if the users are located at the cell edge). This can be easily seen from (1), where the uplink power P is always limited to P_{\max} . From a system perspective, the URLLC capacity is typically dominated by cell-edge users. This renders the benefit of enhanced power control marginal.

Indeed, this observation is validated in [13] through system-level simulations. More specifically, [13] considers a layout of one URLLC serving cell surrounded by 20 eMBB cells with 500 meters inter-site distance (ISD). The URLLC serving cell comprises 10 URLLC users and 2 eMBB users, and each eMBB cell serves 2 eMBB users. The URLLC capacity⁶ is simulated for three scenarios:

- Baseline scenario: URLLC transmissions are scheduled on eMBB resources with very low code rate (to guarantee a high reliability).
- Uplink PI scenario: URLLC transmissions are scheduled on (intra-cell) interference-free resources with the help of uplink PI.
- Enhanced power control scenario: the URLLC transmissions are scheduled on eMBB resources with a boosted higher transmit power.

The detailed simulation assumptions can be found in [13].

The simulation results in [13] indicate that the uplink PI may improve the URLLC capacity by 173% over the baseline scenario. For comparison, the URLLC capacity improvement achieved by enhanced power control (i.e., power boosting) is only 13% over the same baseline. As explained above, the key reason that the capacity gain of power boosting is limited is because the URLLC capacity is dominated by users with high path loss (i.e., users at the cell edge). Power boosting is not useful for those users as they are already reaching their maximum transmit power.

It is also observed in [13] that the resources consumed by URLLC with uplink PI is almost half of that with power boosting for the same packet arrival rate. This is because in the power-control based approach, the URLLC transmissions must be scheduled with low code rate due to the presence of eMBB

intra-cell interference. This shows that uplink PI not only benefits URLLC but also eMBB users, as more interference-free resources will be left for eMBB transmissions.

Finally, we remark that in the simulation results in [13], the impact of interference made by power boosting to users in the neighbouring cell is not modeled. Power boosting by URLLC users may cause undesirable interference to neighboring cell URLLC users, which needs to be carefully considered for real-world network deployment.

V. CONCLUSIONS

In this paper, we have introduced two techniques developed in 5G NR Release 16 to support dynamic URLLC and eMBB multiplexing in the uplink: uplink preemption indication and enhanced uplink power control. We have explained the main design principles, use cases, implementation challenges, and the pros and cons of these techniques. Through system- and link-level simulations, we have illustrated that uplink PI is more effective than enhanced power control in terms of improving URLLC capacity and resource utilization efficiency. This does come at the cost of higher processing requirements for eMBB users.

ACKNOWLEDGEMENT

The authors would like to thank the systems team working on the 5G project at Qualcomm Technologies, Inc. for the fruitful discussions.

REFERENCES

- [1] C. Li, C.-P. Li, K. Hosseini, S. B. Lee, J. Jing, W. Chen, G. Horn, T. Ji, J. E. Smee, "5G-based systems design for tactile Internet," *Proc. IEEE*, vol. 107, no. 2, pp. 307–324, Feb. 2019.
- [2] 5G America, "New services and applications with 5G ultra-reliable low latency communications," white paper, Nov. 2018. [Online] Available: <http://www.5gamericas.org/>
- [3] E. Dahlman, S. Parkvall, J. Sköld, *5G NR: The next generation wireless access technology*. Academic press, London, UK, 2018.
- [4] M. Khoshnevisan, V. Joseph, P. Gupta, F. Meshkati, R. Prakash, P. Tinnakornsrisuphap, "5G industrial networks with CoMP for URLLC and time sensitive network architecture," *IEEE J. Sele. Areas Commun.*, vol. 37, no. 4, Apr. 2019.
- [5] 3GPP TR 38.824, "Study on physical layer enhancements for NR ultra-reliable and low latency case (URLLC)," Release 16, v16.0.0, Mar. 2019.
- [6] C. Li, J. Jiang, W. Chen, T. Ji, and J. Smee, "5G ultra-reliable and low-latency systems design," *European Conf. on Netw. and Commun. (EuCNC)*, Oulu, Finland, Jun. 2017.
- [7] K. I. Pedersen, G. Pocovi, J. Steiner, and S. R. Khosravirad, "Punctured scheduling for critical low latency data on a shared channel with mobile broadband," in *Proc. Vehic. Technol. Conf. (VTC-Fall)*, Toronto, Canada, Sep. 2017.
- [8] P. Popovski, K. F. Trillingsgaard, O. Simeone, and G. Durisi, "5G wireless network slicing for eMBB, URLLC, and mMTC: A communication-theoretic view," *IEEE ACCESS*, vol. 6, Sep. 2018.
- [9] M. Alsenwi, N. H. Tran, M. Bennis, A. K. Bairagi, and C. S. Hong, "eMBB-URLLC resource slicing: A risk-sensitive approach," *IEEE Commun. Letters*, vol. 23, no. 4, pp. 740–743, Apr. 2019.
- [10] R1-1813929, "UL inter UE Tx prioritization/multiplexing," Qualcomm Incorporated, 3GPP RAN1 meeting #95, Spokane, WA, USA, Nov. 2018.
- [11] 3GPP TS 38.213, "Physical layer procedures for control," Release 15, v15.5.0, Mar. 2019.
- [12] 3GPP TS 38.214, "Physical layer procedures for data," Release 15, v15.5.0, Mar. 2019.
- [13] R1-1903008, "UL inter UE Tx prioritization/multiplexing," Qualcomm Incorporated, 3GPP RAN1 meeting #96, Athens, Greece, Feb. 2019.

⁵At the time of writing, the decision of which power control option (open-versus closed-loop) to adopt is not made in the 3GPP standard body yet.

⁶The uplink URLLC capacity in [13] is defined as the maximum packet arrival rate at which at least 95% of URLLC users can meet the target 10^{-5} BLER within 1 ms latency, i.e. up to 5% of UEs in outage.