System Level Analysis of K-Repetition for Uplink Grant-Free URLLC in 5G NR

Jacobsen, Thomas Haaning; Abreu, Renato Barbosa; Berardinelli, Gilberto; Pedersen, Klaus Ingemann; Kovács, István; E. Mogensen, Preben

Published in:
European Wireless 2019 Conference, EW 2019

Publication date:
2019

Document Version
Early version, also known as pre-print

System Level Analysis of K-Repetition for Uplink Grant-Free URLLC in 5G NR

Thomas Jacobsen*†, Renato Abreu*, Gilberto Berardinelli*, Klaus Pedersen*†,
István Z. Kovács†, Preben Mogensen*†
*Dept. of Electronic Systems, Aalborg University; †Nokia Bell Labs, Aalborg, Denmark
Email: {tj, rba}@es.aau.dk

***Abstract*—Ultra-reliable low-latency communications (URLLC) sets high service requirements for the fifth generation (5G) new radio (NR) standard. Grant-free (GF) transmissions is considered a promising technique for reducing the latency in the uplink. To achieve efficient radio resources utilization, sharing of resources is required for sporadic uplink traffic. Repetitions based transmission schemes aims to enhance the reliability of GF transmissions. However, repetitions may also generate excessive interference and cause additional queuing, harming the reliability and latency. In this work, we explore radio resource management (RRM) configurations for repetition based transmission schemes. That includes the number of repetitions, the allocation size per transmission (sub-band), sub-band hopping and uplink power control. Evaluations are conducted in a 5G NR compliant multi-user multi-cell simulation scenario with sporadic uplink GF URLLC transmissions. Our findings suggest that repetitions based schemes can, with a careful selection of the sub-band size and uplink power control parameters, achieve comparable URLLC performance with retransmission based schemes when the effect of queuing is disregarded.**

I. Introduction

The fifth generation (5G) new radio (NR) standard target to support the challenging Ultra-Reliable Low-Latency Communication (URLLC) service requirements [1]. The third generation partnership project (3GPP) has adopted the baseline URLLC requirement which is 1 ms one-way latency deadline for transmitting a packet with a reliability of 99.999% [2]. Grant-free (GF) is a recognized approach to reduce the latency in uplink transmissions, by skipping the scheduling request procedure. With unpredictable URLLC traffic, GF transmissions over orthogonal preallocated resources becomes resource inefficient as resources can be left unused. Sharing of preallocated resources between URLLC sources, can enhance the resource efficiency [3]. The price to pay, is that GF transmissions become subject to intra-cell interference. Retransmission schemes such as hybrid automatic repeat request (HARQ) are known for improving the transmission reliability. However, it comes at the expense of an increased latency as the terminal needs to wait for the feedback before performing a retransmission, being affected by the feedback round-trip-time (RTT) [4].

Different transmission schemes have been considered for enabling GF URLLC. The use of repetitions is one simple way of enhancing the reliability, by transmitting consecutive replicas of the packet without waiting for feedback prior to transmitting the next one. The 3GPP NR Release-15 standard has established the configuration of GF transmissions, known as configured grant, through radio resource control (RRC) with possible activation via downlink control channel [5]. The framework allows the configuration of the physical layer parameters including the settings of K-repetitions, i.e., K consecutive transmissions of the same packet. Our recent work [6] evaluated three schemes for sporadic GF URLLC transmissions in uplink; K-repetitions, Reactive HARQ and Proactive (repetitions with early termination), along with a grant-based reference. Results strongly indicated that the K-repetitions scheme was subject to high interference from the excessive channel use. Full-band transmission repetitions was used, hence not considering the use of higher order modulation and coding scheme (MCS) and hopping between sub-bands. Contention-based transmission schemes using repetitions are studied in [7], where the optimum number of consecutive transmissions is found. A simplified scenario and reception model are considered. In [8] deterministic access patterns based on combinatorial code design are utilized and shows promising gains compared to transmission in random chosen access slots, when ideal interference cancellation of decoded replicas is assumed. Recent work [9] evaluates a repetition based scheme along with two feedback based schemes using analytical tools in a single-cell scenario. The contribution does not consider the effect of inter-cell interference, NR system settings for evaluation and the possibility of transmission repetitions to finish earlier than the feedback based schemes.

This work conducts a thorough evaluation of the transmission repetition parameters; number of repetitions, the chosen MCS and resource allocation in multiple sub-bands, hopping through the allocated sub-bands, along with optimized uplink power control settings. A feedback stop-and-wait retransmission scheme referred to as Reactive HARQ is included as baseline. The evaluation is done using detailed system level simulations capturing the major performance influencing factors in both, the multiple-access protocol layer and physical layer in the radio access network stack, with commonly agreed models in 3GPP. The simulator is also used e.g. in [10], [11].

The reminder of the paper is structured as follows. Section II describes the network and traffic model. Section III presents the K-repetition transmission scheme with intra-slot frequency hopping. The simulation assumptions and methodology are described in Section IV. Section V presents the performance evaluation, followed by Section VI, which concludes the work.

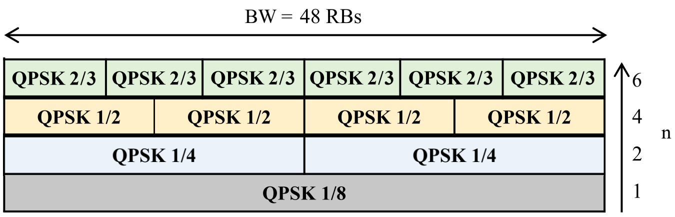

Fig. 1. Examples of radio resource allocations of n sub-bands and corresponding MCS over BW RBs [15].

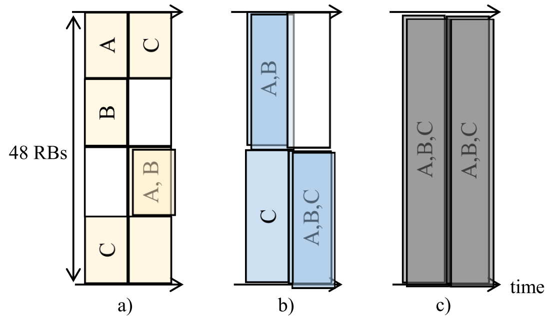

Fig. 2. Realizations of GF transmissions with a) $n = 4$, b) $n = 2$ and c) $n = 1$ sub-bands over K repetitions using sub-band hopping for UE A, B and C.

II. Setting the Scene

We consider a multi-user multi-cell synchronous network consisting of C cells and N URLLC user equipments (UE) uniformly distributed per cell. We assume that the UE connect to the strongest cell, and acquires full synchronization with the network in both time and frequency. Each URLLC UE generates a small packet of size B according to a Poisson arrival process with average rate λ. The aggregated URLLC offered load per cell is therefore given by $L = \lambda \cdot N \cdot B$.

The URLLC UEs are configured for GF transmission over a set of preallocated radio resources. These resources can span multiple sub-bands and are available in every transmission time interval (TTI). We consider an OFDM uplink channel with a bandwidth composed of BW resource blocks (RB) available in the frequency domain. The BW RBs are divided into n sub-bands. Short TTI of duration T are used for GF transmissions. The base station configures the UEs to transmit K consecutive replicas of the packet, hopping to a randomly selected sub-band at each transmission attempt. Note that the same sub-band can be selected with a certain probability, limiting the gain in terms of frequency diversity. However, the potential of interference diversity is kept in this case. It is also important to observe that, this approach is different from the hopping mechanism specified in 3GPP Release-15 [12], which only allows alternate hopping between two sub-bands. Besides, the support of intra-slot repetition within the 14 symbols slot is still under discussion in 3GPP for Release-16 [13].

With the fixed packet size B and bandwidth BW, increasing n also mean that the size of each sub-band is reduced, which implies that the transmission MCS needs to be increased, as illustrated in Fig. 1 for different options of n and for $BW = 48$ RBs. Open loop power control is utilized to regulate the target receive power density at each cell as defined in [14].

III. K-repetitions scheme

Upon arrival of a URLLC packet for immediate transmission at the UE, the packet is prepared for transmission and when ready, the data transmission is performed in the next TTI. For $K > 1$ the repetitions are assumed to be carried out in consecutive TTIs. Upon the end of each transmission, the receiving cell needs to process the received packet and for $K > 1$, combine the received repetitions. A maximum of one transmission can be carried out per TTI per UE. Therefore, ongoing transmissions may force a new packet arrival to wait until its completion, hence being subject to queuing. The latency of a packet that is decoded after $1 < k < K$ replicas is therefore given by

$$t_k = t_{queue} + t_{prep} + t_{align} + k \cdot t_{TTI} + t_{proc}. \quad (1)$$

While the latency contributions t_{prep}, t_{proc}, total transmission time $k \cdot t_{TTI}$ and t_{align} are either known or its upper bound are given, t_{queue} upper bound is not straight forward to determine as it depends on the UE load subject to λ and the number of repetitions K. It should be noted from (1) that, the latency is counted from the moment that the packet is generated, until the moment that any replica is successfully received. The latency of packets that are not received after K-repetitions is accounted as infinite.

Different realizations of GF transmissions are shown in Fig. 2 where GF transmissions are carried out with $K = 2$ and for different number of sub-bands n using sub-band hopping. Increasing the number of sub-bands means that, for unchanged L and the number of transmission repetitions K, the probability of overlaying transmissions is reduced. Further, with $K > 1$ and multiple sub-bands ($n > 1$), frequency hopping can be applied to randomize and reduce systematic transmission overlaying. The total collision probability, i.e., that all K repetitions from a UE have an overlaying transmission, as a function of K and n is shown in Fig. 3 using (9) from [7]. The load in this case is generated by $N = 100$ UEs and $\lambda = 10$ packets per second (PPS). From Fig. 3 we observe that the collision probability is reduced when $K > 1$ and $n > 1$.

Though the total collision probability tends to decrease with K and n, this does not necessarily lead to a reliability improvement. Increasing n and the corresponding MCS, also implies that a higher energy per bit is needed to sustain a transmission reliability target. This can be obtained either by increasing K or increasing the receive power density target through uplink power control, which both implies an increase in channel usage or interference power. Further, the choice of K is bounded by the URLLC latency requirement. And the received power density target is bounded by the UE maximum transmission power. It is therefore not a trivial optimization problem to maximize the URLLC performance, while accounting the diversity gains of using repetitions on sub-bands, the additional interference generated by the repetitions and the uplink power control.

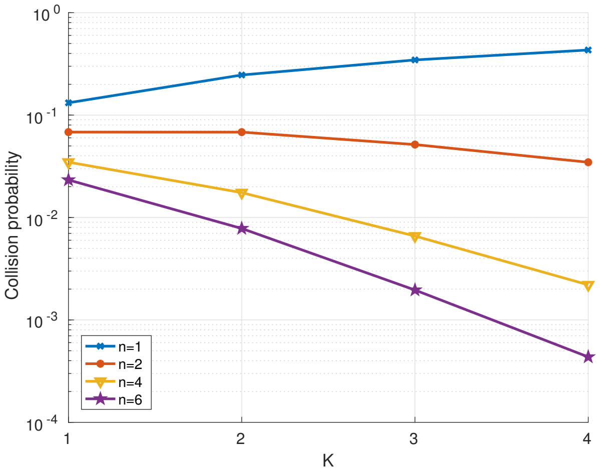

Fig. 3. Collision probability as a function of the number of sub-bands n and repetitions K using (9) from [7]. The load is given by $N = 100$ UEs with $\lambda = 10$ PPS.

IV. Evaluation Methodology

For the performance evaluation we use system level simulations. The evaluation assumptions are in line with URLLC evaluations for 5G NR defined in [16], and are summarized in Table I. A network consisting of $C = 21$ cells is used. The cells are distributed at 7 sites with 3 sectors each, resulting in a regular hexagonal urban macro layout with an inter-site distance of $500\,\mathrm{m}$. URLLC UEs are uniformly distributed outdoors. The uplink bandwidth is $10\,\mathrm{MHz}$, spanning $BW = 48$ RBs. Each RB has 12 sub-carriers with a spacing of $15\,\mathrm{kHz}$. A mini-slot of 2 OFDM symbols is used giving a TTI length of $T = 0.143\,\mathrm{ms}$. The 3D Urban Macro (UMa) channel model is used.

Traffic is generated with a Poisson arrival rate $\lambda = 10$ PPS per UE and $B = 32$ bytes. The packet generation rate was chosen as a trade-off between queuing, number of deployed UEs and simulation time. The offered load is varied by changing the number of UEs per cell. It is assumed that each generated replicas is transmitted using the same redundancy version, and that the receiver combines them using chase combining.

A minimum-mean square error with interference rejection combining (MMSE-IRC) receiver with 2 antennas is assumed. The successful reception of a transmission sample depends on the SINR after the receiver combining. The post-processing SINR values for all sub-carrier including inter- and intra-cell interference are calculated and converted, according to the modulation, to a symbol-level mutual information metric as described in [17]. This metric is mapped through a link-to-system table, depending on the coding rate, to a block error probability value. This value is used for determining if the packet was successful or not. The latency of the packet is then registered, counting from the moment the packet arrived in transmitter buffer until the moment it was successfully received.

The key performance indicator is the achieved outage probability, i.e., the complement of the reliability, which the target

TABLE I
Simulation Assumptions

Parameters	Assumption
Layout	Hexagonal grid composed of 7 sites with 3 sectors/site (21 cells), 500 meters of inter-site distance, wrap-around enabled
Channel model	3D Urban Macro (UMa)
Carrier frequency	$4\,\mathrm{GHz}$
UE distribution	100% uniformly distributed outdoor, $3\,\mathrm{km/h}$ for modeling fading channel
Base station receiver	MMSE-IRC with 2 antennas
Receiver noise figure	$5\,\mathrm{dB}$
Thermal noise	$-174\,\mathrm{dBm/Hz}$
UE transmitter	1 antenna, max. transmit power of $23\,\mathrm{dBm}$
Bandwidth	$10\,\mathrm{MHz}$
Frame numerology	$15\,\mathrm{kHz}$ sub-carrier spacing, $t_{TTI} = 0.143\,\mathrm{ms}$ short-TTI (2 symbols mini-slot), 12 sub-carriers/RB
Latency contributions	$t_{prep} = t_{TTI}$, $t_{proc} = t_{TTI}$ and $t_{align} = [0, t_{TTI}]$,
Configured grant	2-symbols periodicity (every TTI), $n = 1$ use 48 RBs (QPSK1/8), $n = 2$ use 24 RBs (QPSK1/4), $n = 4$ use 12 RBs (QPSK1/2), $n = 6$ use 8 RBs (QPSK3/4). Random sub-band hopping is allowed.
URLLC traffic model	FTP Model 3 with Poisson arrival rate of $\lambda = 10$ packets/sec per UE and $B = 32$ bytes payload

for URLLC is 10^{-5} before $1\,\mathrm{ms}$. The evaluation methodology is conducted in two steps. Firstly, a sensitivity study on the achieved outage probability according the number of sub-bands n relative to the receive power density target P_0, is conducted. This is made for both, $K = 2$ and $K = 4$, as they fit with 1 ms latency requirement given the adopted numerology. Secondly the maximum load L, of which the reliability requirement can be met is found for $K = 2$, $K = 4$ when the best choices of n and P_0 found in the first step are applied. The sensitivity study is conducted using a similar methodology as the one presented in [11], where it is applied on the reactive HARQ baseline scheme.

V. Performance Evaluation

Firstly, we search empirically for the optimal power control setting that leads to the lowest outage probability for each scheme. Four different numbers of sub-bands are considered with $n = \{1, 2, 4, 6\}$. This means sub-bands size of 48, 24, 12 and 8 RBs using MCSs QPSK$1/8$, QPSK$1/4$, QPSK$1/2$ and QPSK$3/4$ respectively. The offered load is $L = 0.256\,\mathrm{Mbps}$ per cell, equivalent to $N = 100$ UEs per cell transmitting $B = 32$ bytes packets with $\lambda = 10$ PPS each. This load was observed to be the highest URLLC load achievable with the baseline reactive HARQ scheme in this scenario [11].

Fig. 4 shows the obtained outage probability after K-repetitions for $K = 2$. It possible to note that the lowest outage probability obtained are comparable for QPSK$1/8$ with $P_0 = -107\,\mathrm{dBm}$, QPSK$1/4$ with $P_0 = -104\,\mathrm{dBm}$ and QPSK$1/2$ with $P_0 = -98\,\mathrm{dBm}$. The optimal P_0 value naturally increases with the MCS given the higher SINR requirement for reliable decoding. The outage probability

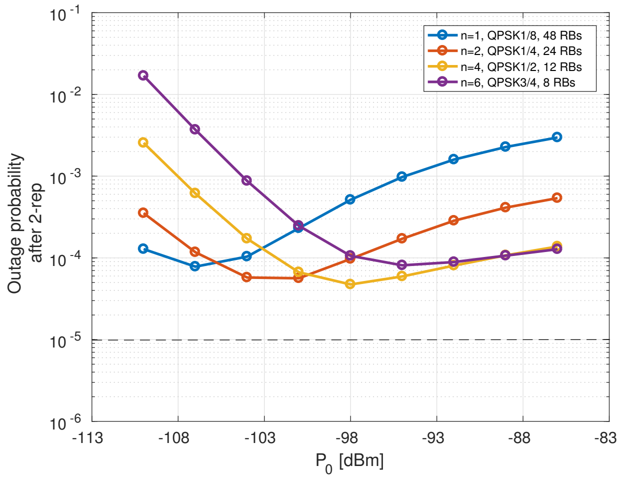

Fig. 4. Sensitivity of outage probability in relation to P_0 and n for $K = 2$.

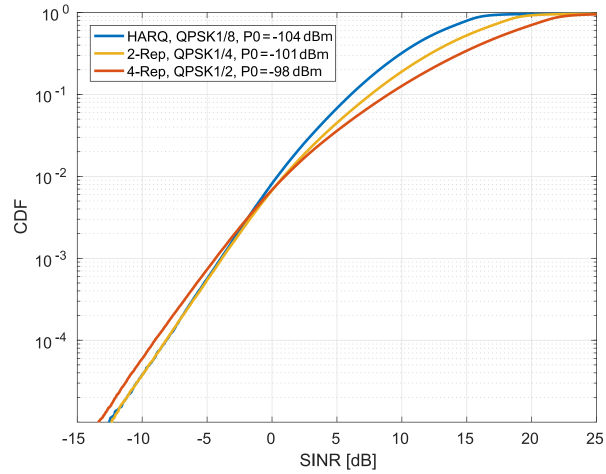

Fig. 6. CDF of the SINR for the different schemes.

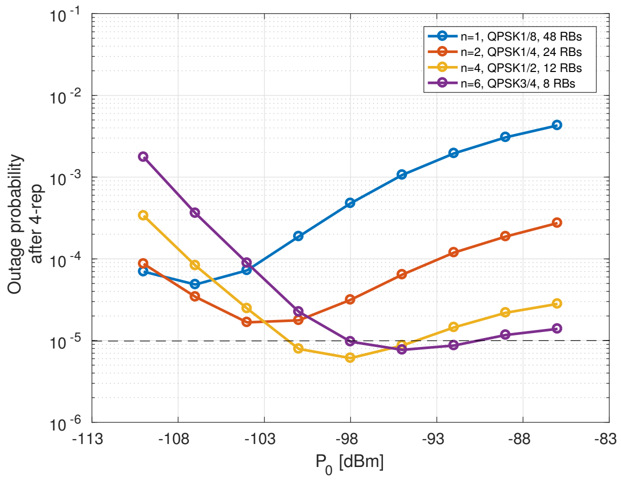

Fig. 5. Sensitivity of outage probability in relation to P_0 and n for $K = 4$.

value in the order of 10^{-4} indicates that the URLLC reliability target can not be met with any of the settings for the applied load. This means that the gain from applying more sub-bands does not sufficiently compensate for the extra interference caused with the repeated transmission.

The same analysis is carried for K-repetitions with $K = 4$ in Fig. 5. In this case we can note an considerable improvement in the outage probability, when comparing the best performance obtained with $\mathrm{QPSK}1/8$ and the performance with a higher order MCS such as $\mathrm{QPSK}1/2$. The achieved outage probability using $\mathrm{QPSK}1/2$ with $P_0 = -98\,\mathrm{dBm}$ gets down to the order of 10^{-5} after the 4 repetitions. The better performance is due to the higher diversity and combining gain obtained with the repetitions in detriment of the higher interference caused by the replicas. With $K = 4$ more energy per bit can be accumulated in time improving the robustness.

The cumulative distribution function (CDF) of the SINR for each scheme, using the configuration that allows the lowest outage probability, is shown in Fig. 6. The increase on 50th percentile SINR between HARQ, $K = 2$ (2-rep) and $K = 4$ (4-rep) corresponds respectively to the increase in optimum P_0 value. 2-Rep has similar SINR tail as HARQ, however due to higher MCS the achieved reliability tends to degrade. It important to note that both, HARQ and 2-rep permit two transmission attempts. 4-rep shows an SINR degradation of $\approx 1\,\mathrm{dB}$ on the low quantiles $< 10^{-4}$, but the combination of the 4 repetitions increases the resultant reliability.

Fig. 7 shows the complementary cumulative distribution function (CCDF) of the latency for the baseline Reactive HARQ and for the K-repetition schemes. For the considered load and packet arrival rate, it can be noted that target latency of $1\,\mathrm{ms}$ and reliability of $1 - 10^{-5}$ can only be reached with the HARQ scheme. Though with 4 repetitions a low outage can be achieved, queuing delays caused by the replicas in the transmission buffer prolong the tail of the latency distribution. As for the illustrated example, considering an average of $\lambda = 10$ PPS generated by the higher layers, it rises to $\lambda = K \cdot 10$ PPS with K repetitions. This can causes an increased queuing such that the latency deadline is exceeded if an early replica is not promptly received. For HARQ, it is important to mention that a retransmission has priority over the initial transmission. So it is very unlikely that a packet retransmission is queued.

The bar plot in Fig. 8 summarizes the maximum URLLC load which can be achieved with each transmission scheme while meeting the $1 - 10^{-5}$ reliability target, disregarding queuing delays. K-repetitions with $K = 2$ supports the lowest load of $0.051\,\mathrm{Mbps}$, while with $K = 4$ a load of $0.307\,\mathrm{Mbps}$, 20% higher than with reactive HARQ, can be supported. It is important to highlight that, satisfying the latency constraint such as $1\,\mathrm{ms}$ will depend on the traffic. Transmissions from UEs with higher packet arrival rates are more susceptible to queuing delays for higher values of K.

VI. Conclusion

In this work we have studied the performance of K-repetitions with intra-slot frequency hopping schemes for URLLC. An extensive exploration of the parameter space involved in GF transmissions with K-repetitions was conducted.

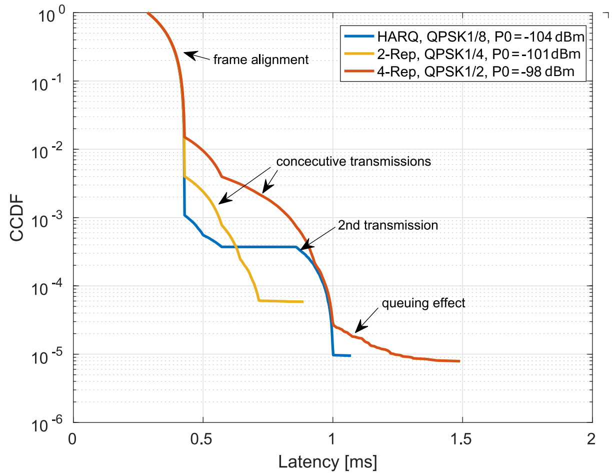

Fig. 7. Complementary cumulative distribution function of the latency for K-repetitions with $K = 2$, $K = 4$ and the HARQ baseline ($L = 0.256$ Mbps).

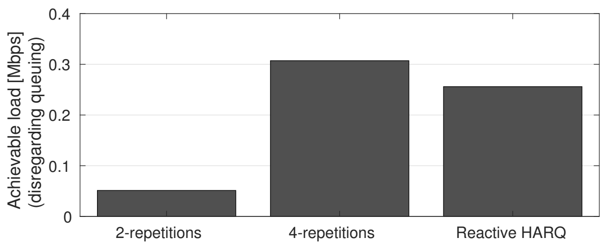

Fig. 8. Maximum loads supported with $K = 2$, $K = 4$ and reactive HARQ, neglecting queuing delays.

That involves the number of transmission repetitions, the sub-band allocation size per transmission, the usage of sub-band hopping and uplink power control RRM mechanism.

By increasing the number of sub-bands, and the number of transmission repetitions, gains can be harvested from a reduced interference probability and with frequency diversity through sub-band hopping. However, when a larger number of sub-bands is used, a higher receive power density or number of repetitions is also needed, which also increase the generated interference.

Our evaluations are conducted in a multi-user multi-cell network to include the effects of intra-cell and inter-cell interference within a 5G NR compliant scenario with sporadic uplink GF URLLC transmissions. Our findings show that K-repetitions can, with a similar latency budget, reach lower outage probabilities than a GF HARQ baseline, with optimized power control settings, number of repetitions and number of sub-bands. However, the queuing effect, potentially cause K-repetitions to violate the latency requirement.

Acknowledgments

This research is partially supported by the EU H2020-ICT-2016-2 project ONE5G. The views expressed in this paper are those of the authors and do not necessarily represent the project views.

References

[1] 3GPP TS 38.300 V15.2.0, "NR; NR and NG-RAN Overall Description," Jun. 2018.

[2] 3GPP TR 38.913 v14.1.0, "Study on Scenarios and Requirements for Next Generation Access Technologies," Mar. 2017.

[3] 3GPP TR 36.881 v14.0.0, "Study on latency reduction techniques for LTE," Jul. 2016.

[4] S. Sesia, I. Toufik, and M. Baker, *LTE - The UMTS Long Term Evolution: From Theory to Practice*, 2nd ed. Wiley, 2011, pp. 108–120.

[5] 3GPP TS 38.331 V15.2.1, "NR; Radio Resource Control (RRC) protocol specification," Jun. 2018.

[6] T. Jacobsen, R. Abreu, G. Berardinelli, K. Pedersen, P. Mogensen, I. Z. Kovács, and T. K. Madsen, "System Level Analysis of Uplink Grant-Free Transmission for URLLC," in *2017 IEEE Globecom Workshops*, Dec. 2017.

[7] B. Singh, O. Tirkkonen, Z. Li, and M. A. Uusitalo, "Contention-Based Access for Ultra-Reliable Low Latency Uplink Transmissions," *IEEE Wireless Communications Letters*, Apr. 2017.

[8] C. Boyd, R. Vehkalahti, and O. Tirkkonen, "Combinatorial code designs for ultra-reliable IoT random access," in *2017 IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, Oct 2017, pp. 1–5.

[9] G. Berardinelli, N. H. Mahmood, R. Abreu, T. Jacobsen, K. Pedersen, I. Z. Kovács, and P. Mogensen, "Reliability Analysis of Uplink Grant-Free Transmission Over Shared Resources," *IEEE Access*, vol. 6, pp. 23 602–23 611, Apr. 2018.

[10] G. Pocovi, K. I. Pedersen, and P. Mogensen, "Joint Link Adaptation and Scheduling for 5G Ultra-Reliable Low-Latency Communications," *IEEE Access*, vol. 6, pp. 28 912–28 922, May 2018.

[11] R. Abreu, T. Jacobsen, G. Berardinelli, K. Pedersen, I. Z. Kovács, and P. Mogensen, "Power control optimization for uplink grant-free URLLC," in *2018 IEEE Wireless Communications and Networking Conference (WCNC)*, Apr. 2018.

[12] 3GPP TS 38.214 v15.3.0, "NR; Physical layer procedures for data," Sep. 2018.

[13] R1-1813118, "On Configured Grant enhancements for NR URLLC," Nov. 2018.

[14] 3GPP TS 38.213 v15.3.0, "Physical layer procedures for control (Release 15)," Sep. 2018.

[15] T. Jacobsen, R. B. Abreu, G. Berardinelli, K. I. Pedersen, I. Kovács, and P. E. Mogensen, "Joint Resource Configuration and MCS Selection Scheme for Uplink Grant-Free URLLC," in *2018 IEEE Globecom Workshops*, 2018, (Accepted/in press).

[16] 3GPP TR 38.802 v14.0.0, "Study on New Radio Access Technology," Mar. 2017.

[17] R. Srinivasan, J. Zhuang, L. Jalloul, R. Novak, and J. Park, "IEEE 802.16m Evaluation Methodology Document (EMD)," IEEE 802.16 Broadband Wireless Access Working Group, Tech. Rep. IEEE 802.16m-08/004r2, Jul. 2008.