Aalborg Universitet

A Blind Retransmission Scheme for Ultra-Reliable and Low Latency Communications

Abreu, Renato Barbosa; Berardinelli, Gilberto; Jacobsen, Thomas; Pedersen, Klaus I.; Mogensen, Preben Elgaard

A Blind Retransmission Scheme for Ultra-Reliable and Low Latency Communications

Renato Abreu*, Gilberto Berardinelli*, Thomas Jacobsen*, Klaus Pedersen*†, Preben Mogensen*†

*Dept. of Electronic Systems, Aalborg University; †Nokia Bell Labs, Aalborg, Denmark

Email: rba@es.aau.dk

Abstract—This work is related to 5G new radio concept design, with focus on ultra-reliable and low latency communication (URLLC) use cases. We mainly target to achieve the stringent latency and reliability requirements for transmissions over the air interface, such as 99.999% success probability within 1 ms. Meeting these requirements in an efficient way, that is, without draining the network capacity is one of the main challenges for the new radio standardization. In this work, we propose a scheme to perform blind retransmissions on shared radio resources together with the application of successive interference cancellation to receive remaining non-decoded data with low delay penalty. The method avoids control errors and extra delays existent on feedback-based retransmission schemes. The investigations also show that blind retransmission on shared resources is more resource efficient than a conservative single shot transmission, depending on the number of users sharing the resources.

I. INTRODUCTION

The advent of ultra-reliable and low latency communication (URLLC) for mission critical applications in cellular networks brings new challenges due to specific characteristics of these systems, such as tight delay and reliability tolerances (e.g. $1 - 10^{-5}$ within 1 ms) and in some cases, infrequent small data traffic [1]. URLLC requires a careful redesign of technology components such as radio numerology, frame structure, scheduling and transmission protocols [2]. Acknowledged transmission mechanisms suffer from inherent delays due to the round trip time (RTT) of the feedback signaling, impacting negatively the latency distribution and potentially jeopardizing the possibility of coping with the URLLC target. Besides that, errors can occur either in the decoding of the feedback or grant signaling messages, affecting the reliability of system [3].

Semi-persistent scheduling (SPS) was extended in LTE Release-14 for faster uplink (UL) access reducing the overhead caused by the request/grant procedures. For unpredictable data traffic, pre-scheduled allocation could result in wasting of radio resources in case user equipment (UE) has no data available for transmission. So, it was proposed that SPS resources could be shared by multiple UEs [4]. In the case that more than one UE transmit at the same time in the shared resources, a collision happens and the base station (BS) may not decode the data. So, the collision should be detected in order to arrange a retransmission of the data of each UE. This can result in an extended latency and compromise the application in URLLC use cases. It should be noted that retransmissions in shared resources are not supported for SPS in LTE, meaning that they should only be scheduled in dedicated resources.

The usage of a shared channel for retransmissions was considered in [5]. In that case, a shared retransmission resource is pre-scheduled to a group of UEs. If more than one UE fails on their initial transmissions, they need to content for the pre-scheduled resource. The procedure relies on a feedback signal to solve the contention for the resource.

Different multi-user detection (MUD) approaches exist to combat interference at the receiver. In conventional successive interference cancellation (SIC), the signal with large signal-to-interference-plus-noise-ratio (SINR) is decoded, reconstructed and subtracted from the aggregated signal. Subsequently, the signal with low SINR is decoded from the other signal [6]. Recently, coded random access schemes using SIC receivers have being proposed in [7]. Such techniques have the potential of boosting cell throughput but the increased average delay does not cope with URLLC requirements.

In 5G New Radio (NR), it is expected that URLLC exploits the usage of blind repetitions, in order to increase the success probability of transmitting a message with low delay penalty [8], [9]. The node just proactively retransmits for a predetermined number of attempts or until a positive acknowledgment is received, rather than stop and wait for a feedback upon each transmission. However, this method can also lead to poor resource utilization and excessive interference, since further retransmissions might not be needed if the message is already detected on the initial transmission.

In summary, retransmissions are beneficial to improve reliability but the problems are:

- blind retransmissions can drain capacity
- stop-and-wait protocols lead to a delay penalty

Hence, in this work we evaluate a scheme that permits the nodes to perform blind retransmissions with low delay penalty and improved resource utilization. A receiver that performs the cancellation of initially decoded transmissions is considered for recovering retransmissions on a shared resource pool. We describe a simple analytical model used to evaluate its success probability achieved with different configurations. We also compare its performance in terms of resource utilization and latency with other schemes.

The paper is organized as follows. Section II describes the proposed scheme. Section III formulates the system model. The performance evaluation is presented in Section IV, and the conclusions are drawn in Section V.
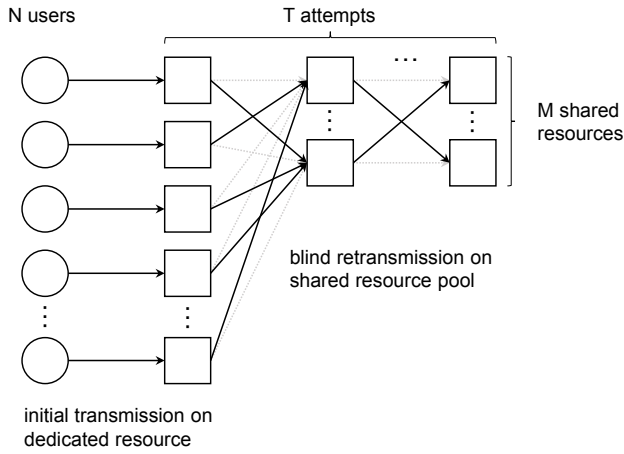
Fig. 1. Example of M shared resources and T transmissions by N UEs.

Fig. 2. Example of reception process with shared retransmission resource.

II. BLIND RETRANSMISSION OVER SHARED RESOURCES

Fig. 1 illustrates a group of N UEs performing the initial transmission on dedicated resources, and the principle of sharing M resources to perform blind retransmissions in a total of T transmission attempts.

Thought the principle could be used for both downlink and uplink, it should be more relevant in uplink, where each transmitter node might not be interested on the data decoding of the other nodes. It is important to mention that the UEs should be time and frequency synchronized in the uplink. In the proposed scheme a group of UEs perform their initial transmission on dedicated resources that can be granted or semi-statically assigned. Subsequently, the devices transmit again the same information $T-1$ times without waiting for a feedback, aiming low latency and reliability. However, instead using dedicated resources, the UEs in the group perform their repetitions using a shared resource pool, for better resource utilization. The shared resource pool can be pre-reserved and its size should be smaller than the amount of resources utilized for the dedicated transmissions ($M < N$). If the pool contains multiple resources, the one to be used for each retransmission can be predefined or randomly selected to avoid extra control signaling.

Since the UEs in the group can perform the same procedure, collision will occur during the retransmissions. Then, a successive interference cancellation (SIC) receiver is used to recover a payload that was possibly not decoded on the initial transmission. Since the initial transmission occurs in "safer" resources, most of them should be early decoded for a low initial block error rate (BLER) target. The already decoded signals can be then subtracted from the received signal in the shared resources, therefore increasing the chances of correctly retrieving the payloads whose detection had failed earlier.

Fig. 2 illustrates the reception process. The received signal $y_{m,j}$ on a shared resource $j \in \{1, ..., M\}$ at a certain retransmission attempt is a combination of the signals from all the UEs retransmitting in there, considering also the channel
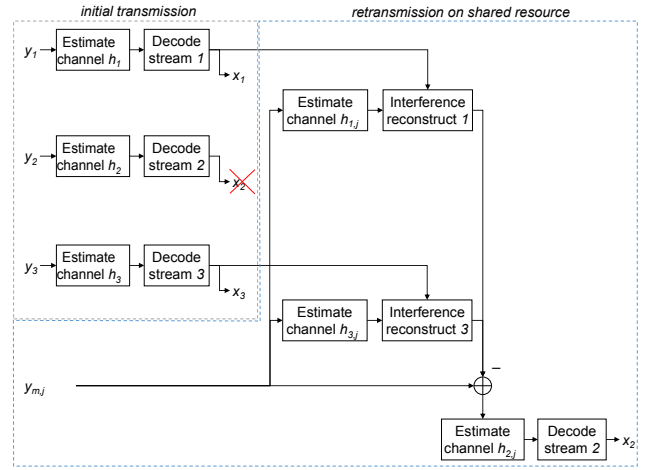
effect over each transmission stream. This can be written as

$$y_{m,j} = \sum_{i \in \Psi} h_{i,j} x_i + \sum_{i \in \Omega} h_{i,j} x_i + w \ , \tag{1}$$

where x_i are the signals transmitted by the UEs, $h_{i,j}$ are the channel fading coefficients of the i-th UE transmitting over the j-th resource, w denotes the Gaussian noise, Ψ is the set of indexes of the UEs whose payload was not yet decoded, and Ω is the set of the ones whose payload was decoded, and are being retransmitted over the same shared resource j. So, the receiver should be able to detect the UEs and estimate their channel responses (for instance, by assuming orthogonal reference sequences used by the different UEs) and reconstruct the signal from the previously decoded ones. After that, it cancels their interference over the non-decoded signals. That is part of the SIC decoding process. Ideally, each successfully decoded replica should permit to remove its interference in the other replicas, at each retransmission. Therefore, the successive decoding process on the shared channel resolves fast the remaining non-decoded transmissions.

The scheme can be summarized as follows (e.g. for an uplink transmission implementation):

1) The BS configures semi-persistent or dynamically granted resources for the UEs initial UL transmission.
2) The BS also configures the UEs to perform blind retransmissions on shared retransmission channels.
3) The UE performs the initial transmission in dedicated channel and blind retransmissions on a shared channel according to the configuration from steps 1 and 2.
4) The BS attempts to decode the initial transmissions from the UEs in their dedicated resources and store the successfully decoded signals.
5) The BS attempts decoding the shared channel after subtracting the already decoded signals from the combined received signal.

III. Success Probability Model

To investigate the reliability achieved with the described transmission procedure we model the probability to successfully deliver a data packet. The following assumptions are considered in this study:

- Same error probability on the initial transmission P_1 for the grouped UEs.
- The decoded transmissions can be fully canceled from the shared resource.
- For a predefined pool with $M > 1$, the retransmission occurs in one randomly selected resource from the pool.
- A transmission can be decoded on shared resource in case it does not collide with other non-decoded transmission.

The probability of u UEs to fail on the initial transmission and contending on the shared resource pool with a UE of interest is given by

$$P_f(u) = \binom{N-1}{u-1}(P_1)^{u-1}(1-P_1)^{N-u}. \tag{2}$$

For u UEs failing on the first transmission, the probability of a UE of interest to be the only failing UE transmitting in a certain resource from the pool is

$$P_g(u) = \left(\frac{M-1}{M}\right)^{u-1}. \tag{3}$$

For one retransmission attempt ($T = 2$), the probability of UE transmission to be singleton, that is, the only transmission that was not yet decoded in a certain shared resource is given by

$$P_s = \sum_{n=1}^{N} P_f(n)P_g(n) =$$
$$= \sum_{n=1}^{N} \binom{N-1}{n-1}(P_1)^{n-1}(1-P_1)^{N-n}\left(\frac{M-1}{M}\right)^{n-1}. \tag{4}$$

And the final probability that a packet transmission to be successfully received can be given by

$$P_r = (1 - P_1) + P_1 P_s (1 - P_2), \tag{5}$$

where P_2 is the error probability in the retransmission.

In a typical feedback-based retransmission scheme, the error probability of the control signaling should be taken into account [3]. However, in the studied scheme the signaling errors do not appear. Instead, equation (5) considers the contention when using the shared retransmission resources, which is the probability of being singleton P_s.

IV. Performance Evaluation

In this section, we present first the reliability and resource utilization analysis, and then a case study with latency evaluation. We compare the described scheme with an aggressive single shot transmission. We also consider for the sake of comparison, the feedback-based scheme in which an UL grant is needed for the retransmissions, as was recently agreed for NR in 3GPP [10].
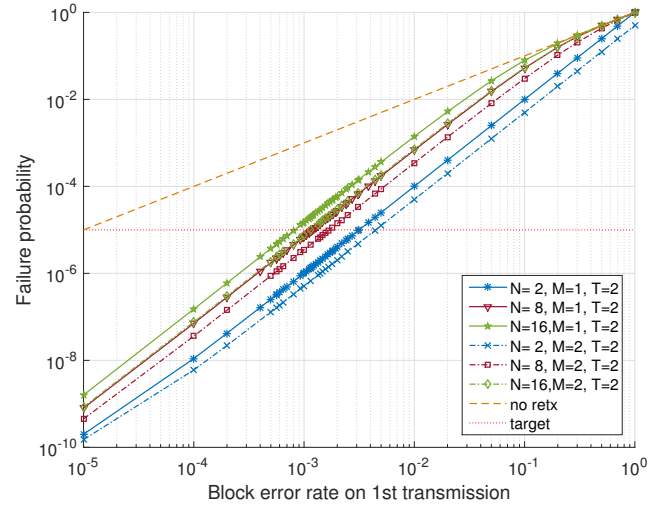

Fig. 3. Reliability for N UEs sharing M retransmission resources and $T = 2$.

A. Reliability and resource efficiency

Employing the model presented in the previous section, we first analyze the resulting failure probability for different number of UEs grouped to share the retransmission resource pool. Fig. 3 shows the final failure probability $(1 - P_r)$ achieved. As in [3] and [5] the failure probability for any retransmission (which should be singleton in our case) is assumed to be 10^{-5} after the detection and soft-combining with the initial transmission. It is obvious that the failure probability reduces with the lower block error rate on the initial transmission. In any case, for the assumed failure probability on the retransmission, the final failure probability is lower than for a baseline case without retransmission. The initial BLER for achieving the target success probability of $1 - 10^{-5}$ is in the order of $\sim 10^{-3}$. For instance, for 16 UEs sharing 2 resources and for 8 UEs sharing 1 resource the initial BLER should be at most 1.2×10^{-3} to meet the target.

The relation between the maximum number of UEs that can be grouped and the initial BLER to achieve the target success probability for different sharing settings is illustrated in Fig. 4. The curve for $T = 3$ transmission attempts was derived through simulation. It is obvious that the higher the number of resources in the shared pool, the higher is the number of UEs that can be supported in the group for the same initial BLER. For instance, from $M = 1$ to $M = 3$ and initial BLER of $\sim 10^{-3}$, the number of UEs sharing the pool can increase from 10 to 30.

To account for the resource utilization, it was applied the same procedure as in [5]. For the single shot transmission the used resources per bit can be calculated as

$$\phi_c = \frac{1}{r_c(1 - P_c)}, \tag{6}$$

where r_c is the rate of a robust modulation and coding scheme, considering ideal link adaptation, which gives a failure probability P_c that in this case should be equal to 10^{-5}. The rates
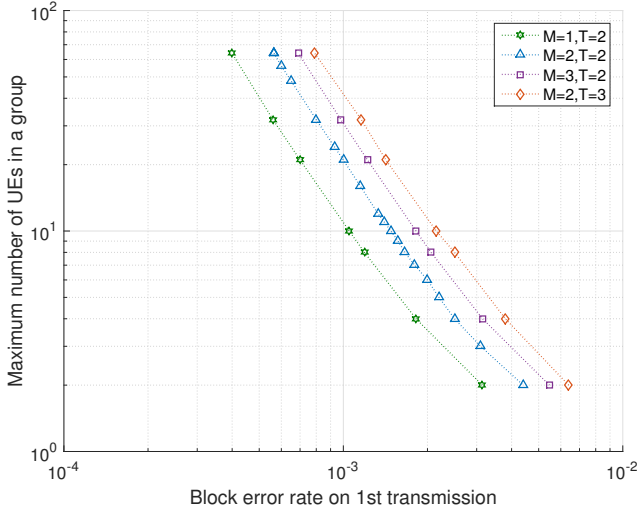
Fig. 4. Maximum number of UEs in the group versus initial BLER.

Fig. 5. Gain on resource efficiency compared to a single shot aggressive transmission.

are obtained considering the link performance with turbo codes for the transmission of a small packet of 32 bytes.

For blind retransmissions over shared resources we calculate resource utilization as

$$\phi_s = \frac{1}{r_1(1-P_1)} + \frac{M}{r_2(1-P_2)N} \ , \tag{7}$$

including the resources occupied for the dedicated initial transmission and the M resources shared by N grouped UEs for the case of one retransmission attempt. The rate for the initial transmission r_1 and for the retransmission r_2 are assumed equal here.

We can calculate the resources utilized in the case of a feedback-based retransmission scheme with the following equation

$$\phi_f = \frac{1}{r_1(1-P_1)} + \frac{P_1}{r_2(1-P_2)(1-\xi)} \ , \tag{8}$$

where ξ is the failure probability of the feedback signal which carries the retransmission grant.

The resource efficiency of two shared retransmission configurations ($M = 1$ and $M = 2$, for $T = 2$) is compared against a transmission that targets 10^{-5} BLER in a single shot. To compare with a feedback-based retransmission scheme we assume a fixed failure probability of $\xi = 10^{-3}$ for the feedback signal. The used resources for the initial transmission and its failure probability is set to be the same as for the blind retransmission scheme with $M = 1$ for the shared pool.

Fig. 5 shows the obtained gain in terms of bits per symbol as function of the number of grouped UEs sharing the resource pool. It can be observed that the gain for $M = 1$ is generally higher, though it requires a lower initial BLER as shown on previous figures. Also for $M = 1$, in case there are only 2 UEs sharing the resource, no gain is achieved. For $M = 2$, a gain on resource efficiency is achieved when the number of UEs sharing the pool is higher than 5. In both cases, the gain
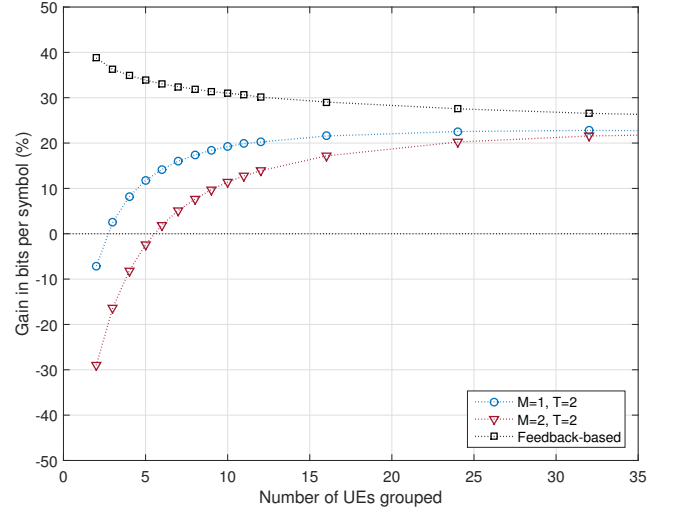
saturates at $\sim 23\%$, since a high number of UEs sharing the pool requires higher initial BLER targets which translates in lower code rates. In practice, such groups with high number of UEs can be formed, for instance, by machine-type communications devices with similar traffic characteristics and located in the same area. In a high mobility scenario, the grouping may require a more complex coordination. It can be also noticed in Fig. 5 that the feedback-based retransmission scheme has, in general, a better resource efficiency. The difference tends to decrease when comparing to the cases where more UEs can be grouped to share the retransmission resources. And, as mentioned previously, the feedback-based scheme comes with the cost of the extra signaling. This can translate to higher latencies as will be discussed next.

B. Case study

Here we consider a frame-based system alike LTE where the resources are arranged in a time-frequency grid and the transmissions occur in mini-slots of a few OFDM symbols (2 to 7) as considered for NR [9]. For uplink transmissions without grant, like in SPS, the latency of a packet transmission is composed by the frame alignment time, transmitter processing, propagation and receiver processing time. The alignment time is a random value from the moment a packet arrives in the transmission buffer until the beginning of the next transmission time interval (TTI). As in [11] and [12] we assume a fast processing time of 1 TTI for transmitting/receiving and also 1 TTI for processing, both in the UE and in the BS side. For the feedback-based retransmission, the HARQ round trip time should also be accounted. The value of it is scaled with the TTI duration and is considered to take 4 TTIs, matching with the time between the beginning of a transmission attempt until the end of its feedback processing. Queuing delays in the UEs transmission buffers are not considered.

Fig. 6 shows the complementary cumulative distribution function (CCDF) of the achievable latencies in terms of TTIs,
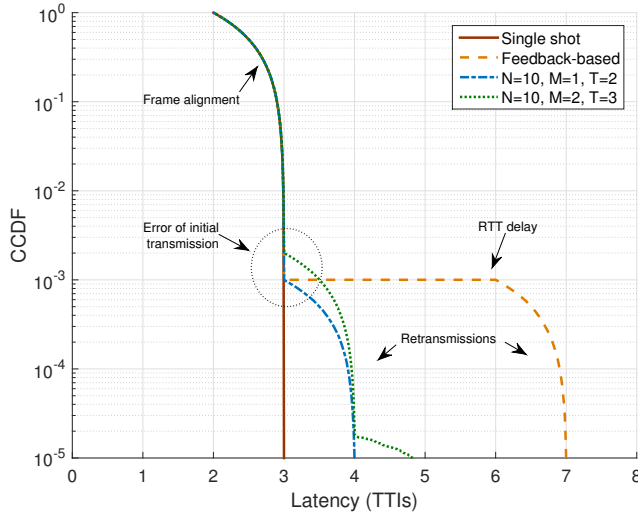
Fig. 6. Example of latency CCDFs for different settings.

TABLE I
ACHIEVABLE LATENCY AT 10^{-5} FAILURE PROBABILITY

Example of numerology configuration	TTI size (ms)	Single shot	Feedback -based	Shared pool
60 KHz SCS, 7 symbols	0.125	0.375	0.875	0.5
15 KHz SCS, 2 symbols	0.143	0.429	1.0	0.572
30 KHz SCS, 7 symbols	0.25	0.75	1.75	1.0
15 KHz SCS, 7 symbols	0.5	1.5	3.5	2.0

from the time a packet arrives in the transmission buffer until it is received and decoded. We can observe that the single shot transmission obviously achieves the lower latency of 3 TTIs at the 10^{-5} percentile, with the cost of low resource efficiency as discussed previously. The blind retransmissions using shared pools with $M = 1$ and $M = 2$ for 2 and 3 transmission attempts respectively, take 4 to 5 TTIs. While the feedback-based option takes 7 TTIs due to the impact of the RTT on the retransmissions.

Considering in particular the baseline URLLC target of $1 - 10^{-5}$ success probability within 1 ms, we show some cases on Table I for different mini-slot configurations, highlighting the options that do not meet the requirement. Mini-slot durations will depend on the subcarrier spacing (SCS) and on the number of OFDM symbols for a given SCS, adopted according to the type of deployment and carrier frequency [13]. It is important to note that the assumed processing times and RTT can be optimistic for the practical NR implementation. If the RTT takes longer time (for instance 8 TTIs like is typically in LTE), then the feedback-based option would not to meet the latency constraints even for very short TTIs.

V. CONCLUSIONS

In this paper we have proposed a scheme for URLLC in which groups of UEs can use a shared resource pool to perform blind retransmissions. The scheme avoids possible errors and delays caused by feedback signaling and re-scheduling

procedures for retransmission. One or more retransmission opportunities can be provided on the shared resources.

The scheme can be more resource efficient than single shot transmissions, especially when more UEs share the retransmission resources. While if the number of UEs is too large the efficiency gain saturates since the BLER for the initial transmission needs to be low. Feedback-based retransmissions have generally better resource utilization than the studied scheme, but might not be able to achieve strict URLLC targets, depending on the numerology and processing times.

The studied solution does not require extra control signaling to allow the UE to perform retransmissions. It counts with an interference cancellation receiver that should be able to reconstruct retransmissions that were previously decoded and subtract them from the received signal in the shared resources. Further, it can be beneficial to consider the performance with multi-user detection receivers which have the potential to capture multiple non-decoded retransmissions on shared resources.

REFERENCES

[1] 3GPP TR 38.913, "Study on Scenarios and Requirements for Next Generation Access Technologies," 3GPP Tech. Rep., V14.0.0, Oct., 2016.
[2] K. I. Pedersen, G. Berardinelli, F. Frederiksen and A. Szufarska, "A flexible 5G frame structure design for frequency-division duplex cases", IEEE Communications Magazine, vol. 54, no. 3, pp. 53-59, March, 2016.
[3] H. Shariatmadari, Z. Li, S. Iraji, M. A. Uusitalo, R. Jäntti, "Control channel enhancements for ultra-reliable low-latency communications", IEEE ICC Workshops, May, 2017.
[4] 3GPP TR 36.881, "Study on latency reduction techniques for LTE," 3GPP Tech. Rep., V14.0.0, Jun., 2016.
[5] R. Abreu, P. Mogensen and K. I. Pedersen,"Pre-Scheduled Resources for Retransmissions in Ultra-Reliable and Low Latency Communications", IEEE Wireless Communications and Networking Conference (WCNC), March, 2017.
[6] Miridakis et. al., "A Survey on the Successive Interference Cancellation Performance for Single-Antenna and Multiple-Antenna OFDM Systems", vol. 15, no. 1, IEEE Comm. Surveys & Tutorials, 2013
[7] E. Paolini, C. Stefanovic, G. Liva, P. Popovski, "Coded Random Access: Applying Codes on Graphs to Design Random Access Protocols", IEEE Comm. Magazine, June, 2015.
[8] 3GPP TR R1-1612246, "Discussion on HARQ support for URLLC", RAN1 #87, Reno, Nevada, 2016.
[9] 3GPP TR 38.802, "Study on New Radio Access Technology," 3GPP Tech. Rep., v14.2.0, Sept., 2017.
[10] 3GPP TSG RAN WG1 Meeting #90, "RAN1 ChairmanâĂŹs Notes", Prague, Czech Republic, Aug., 2017.
[11] G. Pocovi, B. Soret, K. I. Pedersen, and P. Mogensen, "MAC Layer Enhancements for Ultra-Reliable Low-Latency Communications in Cellular Networks," in IEEE International Conference on Communications (ICC) 2017 Workshop, May, 2017.
[12] T. Jacobsen, R. B. Abreu, G. Berardinelli, K. I. Pedersen, P. E. Mogensen, I. Kovcs, and T. Kozlova, "System level analysis of uplink grant-free transmission for URLLC," in IEEE Globecom Workshops, Dec., 2017.
[13] A. A. Zaidi et al., "Waveform and Numerology to Support 5G Services and Requirements," in IEEE Communications Magazine, vol. 54, no. 11, pp. 90-98, Nov., 2016.