Radio Resource Allocation and Retransmission Schemes for URLLC over 5G networks

Salah Eddine Elayoubi, Patrick Brown, Matha Deghel, Ana Galindo-Serrano, Salah Elayoubi

Radio Resource Allocation and Retransmission Schemes for URLLC over 5G networks

Salah Eddine Elayoubi[1], Patrick Brown[2], Matha Deghel[3], Ana Galindo-Serrano[2]

[1]Laboratoire de Signaux et Systèmes (L2S, UMR8506) CentraleSupélec, Gif-sur-Yvette, France
[2]Orange Labs, Chatillon, France
[3] Nokia Bell Labs, Nozay, France
salaheddine.elayoubi@centralesupelec.fr, {patrick.brown, anamaria.galindoserrano}@orange.com, matha.deghel@nokia-bell-labs.com

***Abstract*—Low latency targets for Ultra-Reliable Low Latency Communications (URLLC) may be conflicting with their stringent reliability requirements due to the need for re-transmissions. We explore in this paper the different resource allocation schemes for transmissions and re-transmissions depending on the requirements of the underlying service and on the traffic characteristics, focusing on Industrial Internet of Things (IIoT). We namely consider schemes with individual reservation versus a pool of contention-based reserved resources. We provide novel resource allocation schemes for initial transmissions and re-transmissions and derive corresponding analytical models for loss rates. We then show how to set the system parameters that allow meeting the URLLC requirements with low resource consumption.**

Keywords— URLLC, grant-free transmissions, contention-based access, packet replicas, collision

I. Introduction

In 5G networks, *Ultra-Reliable Low-Latency Communication (URLLC)* is the class of services with the most stringent latency and reliability requirements [1]. This class of services is arguably the most challenging and intriguing because, generally speaking, guaranteeing low latency is conflicting with achieving ultra-high reliability. In the 3GPP (3rd Generation Partnership Project) standard, a general URLLC requirement is $99.999\,\%$ target reliability with 1 ms (two-way) user-plane latency [2]. The reliability here is defined as the percentage of packets that are correctly received within the delay budget. Decreasing the Transmission Time Interval (TTI) length is one efficient way to shorten the latency in the system [3], [4]. Another way to reduce latency is to use *grant-free* scheduling, instead of the Long-Term Evolution (LTE) grant-based scheduling approach, as waiting for the grant penalizes the latency [5]. In this grant-free fast uplink access, neither issuing a scheduling request nor waiting a scheduling grant are required [6]. On the other hand, retransmission is a key enabler for improving the reliability performance [5], but again, using classical Hybrid Automatic Repeat Request (HARQ) retransmission procedures introduces additional latency [5] and other re-transmission schemes are needed for URLLC.

This paper focuses on the grant-free approach and explores two related access schemes. Specifically, if the packet arrivals are periodic, a cyclic reservation (also known as semi-persistent scheduling) is the most suitable scheme. Under this scheme, each user has preallocated resources that repeat according to a predefined periodicity [7]. If, however, the packet arrivals are sporadic and/or the number of users exceeds the amount of resources, then contention-based access is the appropriate scheme to be exploited. In this case, the users contend to access some shared time and frequency resources which are preallocated for the contention procedures [4].

For use cases characterized by periodic packet generation, we propose a scheme where sufficient resources are reserved for the packet transmissions of each User Equipment (UE). However, as some of the packets may be lost due to bad radio conditions, retransmissions may be needed and a pool of common resources is also periodically reserved for re-transmissions. If the size of this pool is equal to the amount of resources reserved for first transmissions, all lost packets can be resent. However, we propose to minimize the size of this pool so that the overall resource consumption is reduced while satisfying the target reliability. We also consider a joint optimization of the link level and the resource allocation. Note that this scheme supposes that the latency target allows that at least one acknowledgment (ACK) can be received for retransmissions to occur. If the latency constraint is so tight that no ACK can be received, all users have to retransmit automatically their packets.

For the sporadic packet case, we combine grant-free contention-based scheme with packet repetitions. Indeed, the UE cannot wait for the ACK before retransmitting its (erroneous) packet as in classical HARQ, as the base station may not realize that it attempted a transmission. In this regard, the approach that consists in sending multiple copies of the same packet without waiting for the acknowledgments was introduced as an efficient way to improve the reliability performance. This approach is already adopted as a solution in the 3GPP standard [8]. Such an approach will result in collisions between some of the (re)transmitted packets, which will impact the reliability level that can be achieved. Hence, it is important to carefully design the contention-based scheme, which will determine the resource allocation policy an active user will follow to send the replicas of each of its packets. In cases where the latency constraint allows for receiving an ACK, we exploit this additional information about the packets that have been correctly received in order to provide a second retransmission opportunity that reduces further the loss rate.

We then derive the minimal amount of resource reservation so that the performance targets are achieved.

Even if the two schemes (individual allocation versus contention-based) correspond in general to use cases with different traffic characteristics, there are some use cases where both schemes are possible. For instance, when there is a limited number of users generating sporadic traffic and with a latency budget that allows for receiving an ACK, both schemes can be used and we show the parameter regions where each of the proposed schemes is better.

The original contributions of this paper are the following:

1) We provide a general framework for resource allocation for URLLC services in 5G and guidelines on the optimal choices.
2) We propose novel resource allocation schemes for transmissions and retransmissions that meet the performance targets with low resource consumption.
3) We derive closed form expressions for reliability performance under the different schemes that fit very well with simulation results.
4) We propose a cross-layer scheme where both the link level and the resource allocation are considered for meeting the targets.

The rest of the paper is structured as follows. In Section II, we provide some details about the URLLC use cases. Section III derives and illustrates the optimal resource allocation for the individual reservation case. Section IV deals with the sporadic traffic case. Section V applies both individual reservation and contention-based schemes to use cases that allow that and derives the best one based on the system parameters. We finally draw conclusions in Section VI.

II. URLLC use cases and requirements

From all URLLC use cases, the most challenging ones arise in the industrial sector (IIoT), where latency requirements are of 1 ms Round Trip Time (RTT) and reliability of 99.99999%. Depending on the type of production or activity developed in the industrial site the communication pattern between machines or controllers and machines may vary. In [9], there are two main low-latency groups of use cases classified according to the communication pattern, i.e., the motion control and the discrete automation. Representative examples of the former use cases are motion control of robots, machine tools, as well as packaging and printing machines. Discrete automation encompasses all types of production that result in discrete products: cars, chocolate bars, etc.

In motion control applications, a controller interacts with a large number of sensors and actuators. The controller periodically submits instructions to the devices, which return a response within a cycle time. The messages are typically small, e.g., 56 bytes. The cycle time can be as low as 2 ms, setting stringent end-to-end latency constraints on message forwarding (1 ms). Additional constraints on isochronous message delivery add tight constraints on jitter (1 μs), and the communication service has also to be highly available (99.9999%). The message transmission in this type of application will therefore follow a deterministic behavior, that is why in what follows we will refer to them as periodic traffic.

For the discrete automation applications, a large number of sensors distributed over the plant forward measurement data to process controllers on a periodic or event-driven base. This use case requires a high communication service availability (99.99%), an end-to-end latency ranging between 1 ms and 100 ms and data rates rather low since each transaction typically comprises less than 256 bytes.

We are interested in this paper in these two families of industrial use cases, and will refer to them in the following using the generic terms of deterministic and sporadic traffics.

III. Deterministic packet arrivals

We start by use cases with deterministic packet arrivals. We consider a system with N UEs, indexed by i and show how the resource allocation can be performed for first transmissions and for retransmissions.

A. Resource allocation

Radio resources are allocated into the time/frequency domain. In particular, in the time domain, they are allocated every TTI. In 4G, a TTI lasts for 1 ms, while different TTI sizes are being defined for 5G. In the frequency domain, instead, the total bandwidth is divided in sub-channels whose size depends on the numerology. A combination of a TTI and a subchannel is called Resource Block (RB) and corresponds to the smallest radio resource unit assigned to a UE for data transmission.

To guarantee deterministic scheduling, we propose that a periodic resource reservation be performed. In order to satisfy reliability targets for URLLC, users are assigned a robust Modulation and Coding Scheme (MCS) that ensures a low Block Error Rate (BLER). For a size of an application packet of b bit, a spectral efficiency of the used MCS of η bit/s/Hz, a bandwidth per RB of ω and a TTI τ, the number of physical RBs, R, for transmitting an application packet is $R = \lceil b/(\eta\tau\omega) \rceil$, where $\lceil x \rceil$ (resp. $\lfloor x \rfloor$) denotes the smallest integer larger than x (resp. the largest integer smaller than x).

However, some packets will be lost with a packet error rate that depends on the chosen MCS. An additional amount of resources should thus be reserved for retransmissions. This amount is less or equal to the amount of resources reserved initially. As the services are delay-constrained, it is reasonable to allow only one retransmission, but our model can be easily extended to a larger number of retransmissions.

B. Optimal resource allocation

It is worth noting that the latency constraint has a large impact on resource allocation. Indeed, the total time between the packet generation and the termination of its retransmission has to be less than the latency target. This introduces constraints on the amount of TTIs consumed for the transmissions and the retransmissions and thus on the amount of needed spectrum knowing the required amount of resources. Let us now study the impact of this latency constraint on the feasibility of

the resource allocation. For the ease of reading and without any loss of generality, we define a "resource unit"(RU) equal to R RBs, so that each packet occupies 1 unit. Let M be the amount of RUs per TTI; it is obtained by dividing the amount of available spectrum W by the available amount of spectral resources per unit: $M = \lfloor W/(R\omega) \rfloor$. The amount of resource units that have to be reserved for first transmissions being equal to the number of UEs N, the resources for first transmissions are spanned over a number of TTIs equal to $\lceil \frac{N}{M} \rceil$. Let the delay before receiving an ACK be equal to t_a and the delay constraint of the service be equal to T, the amount of resources allocated to retransmissions, K, has to verify the following constraint:

$$\left\lceil \frac{K}{M} \right\rceil + \left\lceil \frac{N}{M} \right\rceil \leq \frac{T - t_a}{\tau} \tag{1}$$

The feasibility of this constraint depends on the service and system parameters (latency constraint, ACK response time, number of users, amount of available spectrum). We now suppose that (1) is feasible and derive the optimal value K^* for satisfying the reliability constraint. We start by observing that the number of lost packets follows a binomial distribution of parameters (N, δ_1), where δ_1 is the error probability of the first transmissions, as for all $i \leq N$, user's i transmission process is a Bernoulli random variable ϵ_i that is equal to 1 with probability δ_1 and to 0 otherwise.

To evaluate the reliability of our resource allocation mechanism, we have to consider two possible events for loss as follows. First, if the number of needed resources for retransmissions is larger than K, some of the lost packets cannot be retransmitted, leading to a definite loss. Second, even if there is enough space for a retransmissions, the retransmission may fail again. Note also that, for the first event, i.e., when there is no enough space to accommodate all retransmissions, we select randomly K packets among the lost ones for retransmission.

Proposition 1. *The final error rate for a UE is:*

$$e(K, \delta_1, \delta_2) = \sum_{n=0}^{N-1} C_{N-1}^{n} \delta_1^{n+1} (1-\delta_1)^{N-1-n} \times \left(\delta_2 I_{n+1 \leq K} + \frac{\delta_2 K + n + 1 - K}{n+1} I_{n+1>K} \right) \tag{2}$$

where C_N^n is the binomial coefficient and I_A is an indicator function equal to 1 if condition A is verified and to 0 otherwise. δ_2 denotes the error rate for a second transmission.

Proof. The probability of having n packets lost in the first round is given by the binomial law as errors are independent. The sum represents the events of having n lost packets among the first transmissions of the $N-1$ other users. The first term within the sum is the binomial law, multiplied by δ to consider that the user of interest has lost its packet; and the second term characterizes the retransmissions. Here, if there is enough space for all lost packets to be retransmitted, the error probability is δ_2, giving the term $\delta_2 I_{n+1 \leq K}$. Otherwise, it is equal to 1 if the packet is not selected for retransmissions (with probability $\frac{n+1-K}{n+1}$, giving the term $\frac{n+1-K}{n+1} I_{n+1>K}$) and to δ_2 if it is selected (giving the term $\delta_2 \frac{K}{n+1} I_{n+1>K}$). Note that the error probability reduces to $\delta_1 \delta_2$ for $K = N$. □

The optimal reservation of resources for retransmissions (denoted by K^*) is the smallest K so that:

$$e(K^*) \leq \Theta \tag{3}$$

Θ denotes the reliability target and this reliability constraint comes in addition to the latency constraint (1).

C. Further repetitions for increasing reliability

In the previous sections, we explored the optimal resource reservation scheme for first transmissions and retransmissions, after ACK reception. Even if it is not realistic to consider waiting for another acknowledgment cycle for retransmitting lost packets, due to the stringent delay constraints, blind retransmissions can be envisaged as follows. If the amount of reserved resources for retransmissions, K, is larger than the number of lost packets, some users who received a NACK for their first transmissions can attempt for a second retransmission in the unused resources. For this, the users have to be aware of the remaining free resources, but this can be incorporated in the ACK sent by the base station (e.g. a grouped ACK with resource allocation).

Let n be the number of lost packets from the first transmissions. If $n \geq K$, no second retransmissions are possible. However, if $n \leq \lfloor \frac{K}{2} \rfloor$, all users can retransmit their packets without collisions (following for example some scheme that repeats the initial allocation for retransmissions in the remaining resources). However, if $\lfloor \frac{K}{2} \rfloor < n < K$, not all re-transmissions can be ensured. In this case, the base station selects a subset of users for second re-transmissions (e.g. at random, as it is done for first retransmissions if $n > K$).

Proposition 2. *The error probability in the case of several repetitions can be computed as follows:*

$$e(K, \delta_1, \delta_2, \delta_3) = \sum_{n=0}^{N-1} C_{N-1}^{n} \delta_1^{n+1} (1-\delta_1)^{N-1-n} \times \left(\delta_2 \delta_3 I_{n+1 \leq \lfloor \frac{K}{2} \rfloor} + \frac{\delta_2 K + n + 1 - K}{n+1} I_{n+1 \geq K} + \delta_2 \frac{\delta_3 (K - (n+1)) + 2(n+1) - K}{n+1} I_{\lfloor \frac{K}{2} \rfloor < n+1 < K} \right) \tag{4}$$

Where δ_3 is the probability that a second retransmission fails.

Proof. Equation (4) differs from equation (2) in the case where there is enough space for making a second retransmission for all of the lost packets ($\delta_2 \delta_3 I_{n+1 \leq \lfloor \frac{K}{2} \rfloor}$ or for a subset of them ($\delta_2 \frac{\delta_3 (K-(n+1)) + 2(n+1) - K}{n+1} I_{\lfloor \frac{K}{2} \rfloor < n+1 < K}$). Retransmitted packets for the second time are lost with probability δ_3 [1]. □

[1]Note that if the transmissions are well distributed in the frequency dimension, the events of loss are independent for the different packets, leading to $\delta_1 = \delta_2 = \delta_3$. If, in addition packets are combined following an HARQ-like scheme, $\delta_3 < \delta_2 < \delta_1$

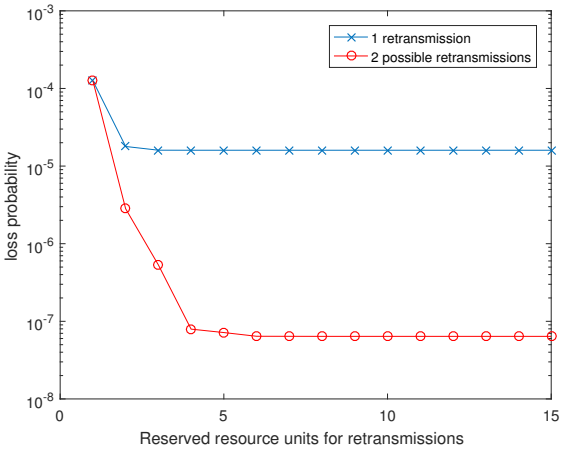

Figure 1: Probability of packet loss.

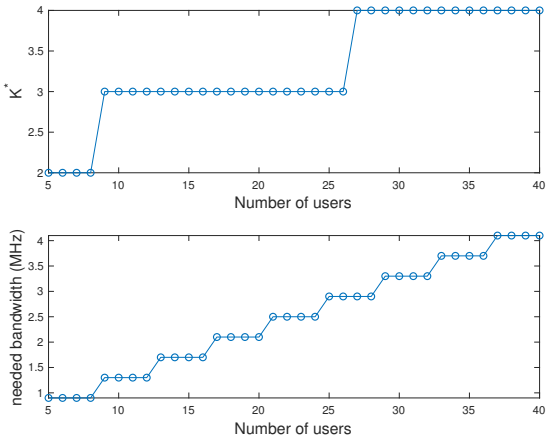

Figure 2: Impact of the number of users on the resource reservation (the amount of resource units K^* and the amount of spectrum).

We illustrate in Figure 1 the impact of further repetitions on the performance, for a system whose parameters are expressed in Table I. In addition, we considered a packet loss rate of $\delta_1 = \delta_2 = \delta_3 = 0.005$. The figure illustrates that the scheme with further repetitions achieves much lower final loss rates for the same amount of reserved resource units. We then evaluate in Figure 2 the needed resources when the number of users increases, for a latency constraint of 1 ms and a target reliability of 10^{-6}. When the number of users increases, not only the needed amount of resources for first transmissions increases, but also the needed resources of retransmissions (K^*). This makes the latency constraint (1) difficult to meet, unless the amount of spectrum used for URLLC (W) is increased. Note that this increase is not linear but K^* is a step function, because of the rounding operator in equation (1). K^* also increases slowly with the number of users, while the required spectrum increases more rapidly. This is because the addition of one resource unit for retransmissions allows serving the replicas for a large number of additional users, while the need for resources for the initial transmission increases the required spectrum.

Table I: System and service parameters

Applicative packet size, b	100 bits
Number of UEs, N	15
Reserved bandwidth for URLLC service, W	2.5 MHz
Subcarrier spacing, ω	15 KHz
Smallest time scheduling unit (TTI), τ	0.144 ms (2 symbols per TTI)
Spectral efficiency of the selected MCS, η	1 bit/s/Hz
Reliability target, θ	$1 - 10^{-5}$
Acknowledgment time (t_a)	0.288 ms

D. Joint optimization of link level and resource allocation

In the previous section, we have developed an optimization framework of the resource allocation, supposing that the MCS is a priori chosen so that it minimizes the loss rate δ_1. However, this may be only a local optimum considering a cross-layer view that also considers the link level. Here we do not consider tight link adaptation as it needs training and may be inconsistent with the stringent latency requirements, but limit ourselves to a choice of the MCS based on the average radio conditions (i.e. slow fading). The problem now becomes a two dimensional optimization problem, where both the MCS choice and the resource reservation for retransmissions have to be jointly optimized.

There is clearly a trade-off between the number of resources reserved for the first transmission and for retransmissions. Indeed, a more robust MCS ensures fewer initial losses, and then less reservation of resources for retransmissions, but has a lower spectral efficiency leading to more reserved resources for the first transmission. There is an optimal trade-off to seek. Let μ be the selected MCS for transmissions and $\delta_j(\mu)$ be the corresponding loss probability for transmission $j \geq 1$. The loss probability becomes (in the case of only one retransmission), equal to $e(K, \delta_1(\mu), \delta_2(\mu))$, where the function $e(.)$ is defined in equation (2). The MCS μ can also be introduced in the expression of loss in the extended case (4).

On the other hand, while the reliability constraint (3) remains unchanged, the latency constraint (1) depends on the used MCS, as a more robust MCS requires more RBs per packet, leading to the following constraint:

$$\left\lceil \frac{K}{\left\lfloor \frac{W}{\omega \left\lceil \frac{b}{\eta(\mu)\tau\omega} \right\rceil} \right\rfloor} \right\rceil + \left\lceil \frac{N}{\left\lfloor \frac{W}{\omega \left\lceil \frac{b}{\eta(\mu)\tau\omega} \right\rceil} \right\rfloor} \right\rceil \leq \frac{T - t_a}{\tau} \quad (5)$$

The objective is to minimize the total resource consumption (for first transmissions and retransmissions):

$$\min_{\mu,K} (N + K) \left\lceil \frac{b}{\eta(\mu)\tau\omega} \right\rceil \quad (6)$$

subject to (3) and (5). This optimization problem is easily solved by an exhaustive search as follows. Note that, for a fixed MCS, the reliability increases when K increases but the

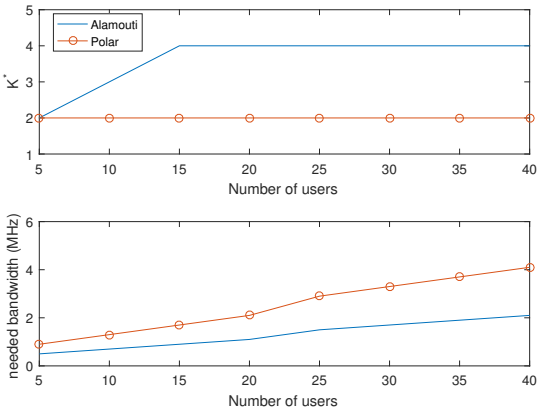

Figure 3: Impact of MCS on the optimal resource reservation.

delay increases too. $K^*(\mu)$ is the minimal K such that the loss probability is less than the target Θ; $K^*(\mu)$ exists only if $e(N, \delta_1(\mu), \delta_2(\mu)) < \Theta$ and the latency constraint (1) is verified for the computed $K^*(\mu)$. The optimal MCS is the one that leads to the minimal value of $(N + K^*(\mu)) \left\lceil \frac{b}{\eta(\mu)\tau\omega} \right\rceil$.

Figure 3 shows the reserved resources for retransmissions and the overall spectral resource consumption for two MCSs: Alamouti 2*2 OFDM (efficiency of 4 bit/s/Hz) and Polar-Alamouti 2*2 OFDM (efficiency of 2 bit/s/Hz). Loss rates, computed using link level simulations, are equal to 0.005 and 0.001 respectively (average user SNR of 15 dB). While polar coding reduces the initial packet loss and requires less reservation for retransmissions (with a target reliability of 10^{-6}), it reduces the spectral efficiency, leading thus to higher overall resource consumption when accounting for both latency and reliability constraints.

IV. Sporadic packet arrivals

We now turn to another set of interesting use cases, where users do not have always packets to send. In each cycle, packet arrivals are thus sporadic and reserving resources for each user is clearly under optimal, as the number of users, N, may be very large and the probability that a user generates a packet during a cycle, p, may be low. Our proposal is to deal with this traffic in a contention-based manner, i.e. to reserve a pool of resources where users who have packets to transmit contend. Packets are thus subject to collisions, in addition to the losses introduced by the wireless channel. In order to increase the probability of success, each packet may be sent $\beta \geq 1$ times. We call these replicas.

In a context similar to what we consider here, the authors in [10] propose to send these replicas in consecutive TTIs, where the resources used by each replica are randomly selected from the set of available RUs in each TTI; our scheme provides more flexibility in the resource allocation process of these replicas, which results in less collisions between the (re)transmitted packets. In [11], an uplink transmission scheme is proposed in which the resources are split into shared and dedicated parts. It relies on advanced receiver processing in order to satisfy the URLLC constraints.

A set of K transmission units are reserved for uplink transmissions in each cycle. Each packet occupies one unit, as computed in section III-A.

A. Computation of the loss probability

We now provide the loss probability for the contention-based scheme. Note that when β copies of a packet are sent, a collision occurs if all these copies collide with other transmissions. The collision rate is measured from a predefined-user perspective, given that this user has data to transmit. Even if a packet is collision-free, it may be lost due to bad radio conditions. Let $e_c(N, K, \beta, p)$ denote this loss probability.

Proposition 3. *The loss probability under the contention-based approach with replicas can be expressed as follows*

$$e_c(N, K, \beta, p) = 1 - \sum_{l=1}^{\beta} (-1)^{l+1} C_\beta^l \left((1-p) + p \frac{C_{K-l}^\beta}{C_K^\beta} \right)^{N-1} (1-\delta_1)^l. \quad (7)$$

Proof. Define $\mathcal{A}_i$ to be the event that the i-th resource is free, i.e. no (other) active user chooses this resource for its packet transmissions and this resource is not subject to an error. We would like to express the probability that one of the β resources is free, i.e. $\mathbb{P}\{\mathcal{A}_1 \cup \ldots \cup \mathcal{A}_\beta\}$. To this end, we determine the probability that a subset of l resources is free. Note that in a set containing β resources there are C_β^l subsets of size l. All l resources will be collision-free if all other users are either not transmitting or non of their β RUs fall in the l slots. For a given user, this happens with probability

$$1 - p + p \frac{C_{K-l}^\beta}{C_K^\beta}, \quad (8)$$

where p represents the probability that a user is active. Since there are $N-1$ other users and errors are independent, the probability that all l slots of this subset are collision-free and error-free is:

$$\mathbb{P}\{\mathcal{A}_1 \cap \ldots \cap \mathcal{A}_l\} = \left(1 - p + p \frac{C_{K-l}^\beta}{C_K^\beta} \right)^{N-1} (1-\delta_1)^l.$$

We here use the initial error probability δ_1 for all replicas as combination of packets is not easy to achieve, as the base station does not know in advance the position of packets to combine. Using the above, we conclude that

$$\mathbb{P}\{\mathcal{A}_1 \cup \ldots \cup \mathcal{A}_\beta\} = \sum_{l=1}^{\beta} (-1)^{l+1} C_\beta^l \, \mathbb{P}\{\mathcal{A}_1 \cap \ldots \cap \mathcal{A}_l\}$$
$$= \sum_{l=1}^{\beta} (-1)^{l+1} C_\beta^l \left(1 - p + p \frac{C_{K-l}^\beta}{C_K^\beta} \right)^{N-1} (1-\delta_1)^l.$$

Leading to the loss probability (7), which concludes the proof. □

Note that a slightly more general expression may be derived. In (8) the probability that a user is active may be specific to each user leading to the following expression for the loss probability, which we will use in the next section:

$$e_c(N,K,\beta,\mathbf{p}) = 1- \sum_{l=1}^{\beta}(-1)^{l+1}C_\beta^l \prod_{i=1}^{N-1}\left((1-p_i)+p_i\frac{C_{K-l}^{\beta}}{C_K^{\beta}}\right)(1-\delta_1)^l, \quad (9)$$

where $\mathbf{p} = (p_1, ..., p_{N-1})$ is the vector of activity probabilities of the other users.

Before moving to the performance evaluation using the analytical formula (7), we proceed to its validation with respect to simulations. We construct a discrete event simulator where a RU consists in a certain amount of subcarriers reserved on one slot; this amount of subcarriers for constituting one resource unit is called a frequency unit. In particular, there are K_f frequency units that are continuously reserved. At each cycle, users that have packets to send choose randomly β resource units by selecting at random one frequency unit on one of the time slots available for transmissions (there are K_t time slots due to the delay budget). Note that $K = K_f K_t$.

We plot in Figure 4 the packet loss probability when varying the activity ratio p using (7) and the simulator for $N = 30$, $K = 12$, $\delta_1 = 10^{-3}$, $\beta = 4$. The figure shows a perfect match.

We also exploit the simulator for testing the impact of our assumption on independent errors. In fact, the reduction of the length of the TTI may result in correlated errors for channels with small coherence time. This is especially relevant for the case of transmissions with replicas as these replicas may be transmitted in adjacent TTIs. Figure 4 illustrates the impact of correlated errors on the loss rate, considering a system whose parameters are defined in Table I and where the $K = 12$ resources are reserved on a basis of 3 frequency units with 4 time units (i.e. a delay budget of 0.6 ms within which all the replicas have to be sent). We consider a channel whose coherence time is larger than one time unit, so that errors arrive in burst, i.e. when an error occurs, it lasts for several time units. We consider both medium correlation of errors, where error bursts are spread over two scheduling units and 30 KHz, and high correlation where an error is followed by a burst that covers the whole remaining time slots, on all reserved frequency units. It can be observed that correlation between errors increases the loss rate, but this increase remains limited as collisions remain the main cause for losses.

We also compare by simulation the proposed scheme with the state of the art. We take as baseline the Aloha-like scheme where each packet is transmitted once in one of the available resources (called URT of Unique Random Transmission). We also simulate the scheme of [10] where the replicas are sent in consecutive TTIs (called OT for One Transmission per TTI). In order to have comparable results between OT and RT, we consider the same number of replicas ($\beta = 2$) in both (UT corresponds to one replica). The results show that schemes with several replicas have significant gains with respect to baseline Aloha, and that our scheme (RT) outperforms the OT

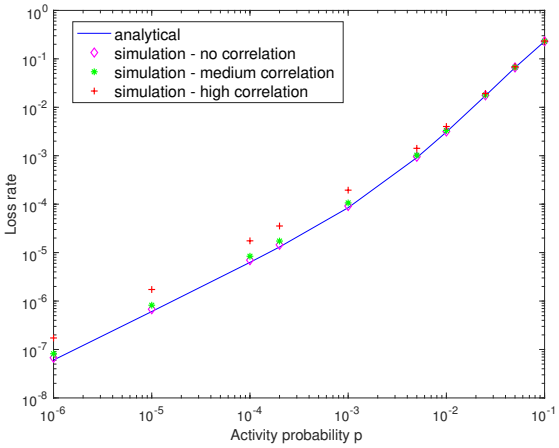

Figure 4: Validation of the analytical expression for the probability of collisions and impact of correlated errors. $N = 30$, $K = 12$, $\delta_1 = 10^{-3}$, $\beta = 4$.

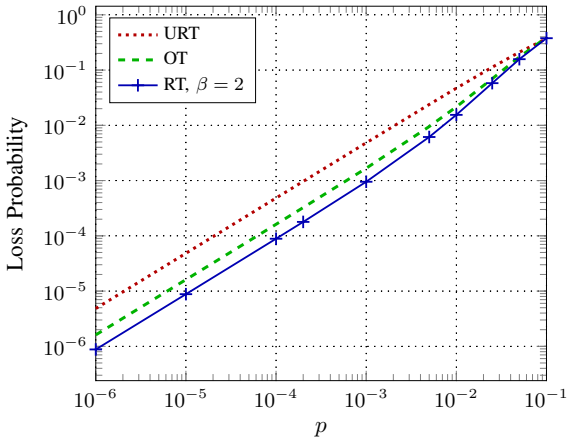

Figure 5: Loss probability for URT, OT and RT ($N = 30$, $K = 12$).

scheme. It is to note that our scheme has also the advantage of flexibility, as the number of replicas can be adapted for minimizing the probability of loss, as will be shown next.

B. Optimal retransmissions and resource allocation

Equation (7) gives the packet loss probability for a given number of replicas β and a given set of reserved resources K. These parameters, β and K, have to be chosen so that the latency and reliability constraints are satisfied with the lowest possible resource reservation. Using the same notation as in section III-A, the latency constraint can be expressed as:

$$\left\lceil \frac{K}{M} \right\rceil \leq \frac{T}{\tau} \quad (10)$$

where $M = \left\lfloor \frac{W}{R\omega} \right\rfloor$ is the amount of resource units per TTI. The number of replicas and the amount of resources are to be chosen so that K is minimized while satisfying latency constraint (10) and the reliability constraint:

$$e_c(N,K,\beta,p) \leq \Theta \quad (11)$$

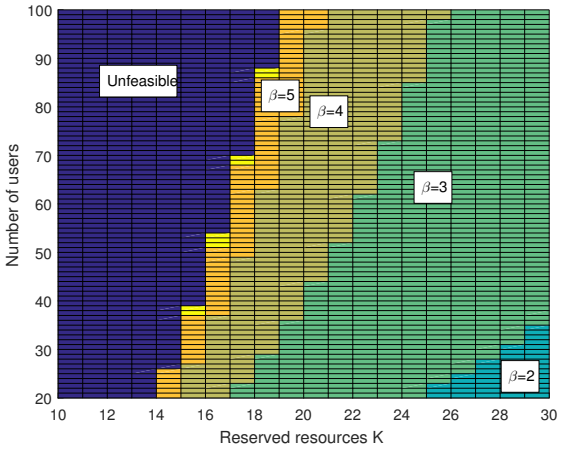

Figure 6: Number of needed replicas for the target reliability. This number ranges for $\beta^* = 2$ (low right corner of the figure) to $\beta^* = 6$ (the small yellow areas in the left). For a small K, the stringent reliability target is not achievable (left part of the figure).

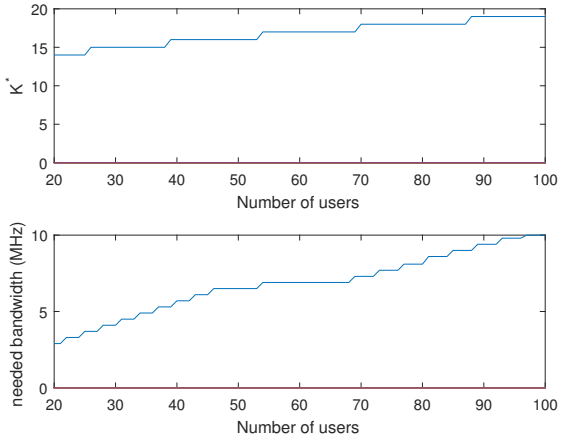

Figure 7: Reserved spectrum for the target reliability.

Figure 6 shows the number of replicas needed for obtaining a reliability target of 10^{-6}, for different numbers of users and different sizes of the reserved pool, while keeping $p = 10^{-4}$. The need for replicas increases when N increases, but also when resources become scarcer. Also, the reliability target becomes unfeasible when N is too large or K is too small.

We now move to the computation of the amount of resources to be reserved for ensuring both the target reliability and the target latency constraint of 1 ms (equation (10)). Figure 7 shows the minimal amount of resources and the corresponding amount of spectrum, based on the parameters of Table I, and on the number of replicas represented in Figure 6. It can be observed that the amount of required spectrum increases with the number of users, reaching up to 10 MHz for N=100.

C. Exploiting acknowledgments for limiting retransmissions

We now explore how to take advantage of acknowledgments received from the base station, as done in the deterministic traffic case. Indeed, there might be room for receiving an acknowledgment for the correctly received packets (e.g. a shared ACK). In this case, only users who have a packet to transmit and who did not receive an ACK will send the replicas. Two sets of resources are then reserved: A first set of K_1 resources where active users send β_1 replicas, and another set, after a time t_a, of size K_2 where only users who did not receive a ACK retry sending with β_2 replicas.

Proposition 4. *For small transmission probabilities p the loss probability after the two rounds of replicas can be approximated as follows*

$$e_f(N, K_1, K_2, \beta_1, \beta_2, p) \approx e_c(N, K_1, \beta_1, p) e_c(N, K_2, \beta_2, \boldsymbol{p}_1); \quad (12)$$

$$\boldsymbol{p}_1 = (1, pe_c(N, K_1, \beta_1, p), ..., pe_c(N, K_1, \beta_1, p)) \quad (13)$$

Proof. The probability of error after the first phase being computed as in equation (7), the error probability of this scheme is computed by:

$$e_f(N, K_1, K_2, \beta_1, \beta_2) = e_c(N, K_1, \beta_1, p) \times e_c^2(N, K_1, K_2, \beta_1, \beta_2)) \quad (14)$$

where $e_c^2(N, K_1, K_2, \beta_1, \beta_2)$ is the loss probability for the second set of replicas, knowing that all first replicas of the target user have been lost. If the two stages were independent, the error probability at the second stage would have been the error probability after sending the first set of replicas, $e_c(N, K_2, \beta_2, pe_c(N, K_1, \beta_1, p))$. Indeed, the probability of a user being active during the second phase is the probability of him being active in the cycle (p), multiplied by the probability that all his replicas have been lost in the first set ($e_c(N, K_1, \beta_1, p)$), and the loss probability in the second stage can be obtained as for equation (7) replacing p by $pe_c(N, K_1, \beta_1, p)$.

We have noted that the main cause for an unsuccessful transmission is the presence of collisions. So in the second stage another users should be present with high probability. We then modify the probability of activity, accounting for the presence of at least one active user (with probability 1) in addition to random users. This is of course an approximation as there may be losses in cases where there are collisions with two different users, but the simulation results that we will present next show that this is a good approximation. □

Proposition 5. *The following gives an upper bound for the loss probability after the two rounds of replicas*

$$e_f(N, K_1, K_2, \beta_1, \beta_2, p) < C_{N-1}^n p^n (1-p)^{N-1-n} \times \sum_{n=1}^{N-1} e_c(n+1, K_2, \beta_2, 1) e_c(n+1, K_1, \beta_1, 1) \quad (15)$$

Proof. Let $\mathcal{F}_n$ be the event that n users were active during the first stage. Let $\mathcal{S}_m$ be the event that m users are active during

the second stage. Let $\mathcal{C}_1$ (resp. $\mathcal{C}_2$) be the event the first (resp. the second) round was unsuccessful for the user considered:

$$e_f(N,K_1,K_2,\beta_1,\beta_2,p) = \sum_{n=1}^{N-1}\sum_{m=1}^{n} \mathbb{P}\{\mathcal{C}_2 \cap \mathcal{S}_m \cap \mathcal{F}_n \cap \mathcal{C}_1\}$$
$$= \sum_{n=1}^{N-1}\sum_{m=1}^{n} \mathbb{P}\{\mathcal{C}_2 \mid \mathcal{S}_m \cap \mathcal{F}_n \cap \mathcal{C}_1\}\mathbb{P}\{\mathcal{S}_m \mid \mathcal{F}_n \cap \mathcal{C}_1\}\times$$
$$\times \mathbb{P}\{\mathcal{F}_n \mid \mathcal{C}_1\}\mathbb{P}\{\mathcal{C}_1\} =$$
$$\sum_{n=1}^{N-1}\sum_{m=1}^{n} \mathbb{P}\{\mathcal{C}_2 \mid \mathcal{S}_m\}\mathbb{P}\{\mathcal{S}_m \mid \mathcal{F}_n \cap \mathcal{C}_1\}\mathbb{P}\{\mathcal{F}_n \mid \mathcal{C}_1\}\mathbb{P}\{\mathcal{C}_1\} \leq$$
$$\sum_{n=1}^{N-1} \mathbb{P}\{\mathcal{C}_2 \mid \mathcal{S}_n\}\mathbb{P}\{\mathcal{F}_n \mid \mathcal{C}_1\}\mathbb{P}\{\mathcal{C}_1\} \sum_{m=1}^{n} \mathbb{P}\{\mathcal{S}_m \mid \mathcal{F}_n \cap \mathcal{C}_1\} =$$
$$\sum_{n=1}^{N-1} \mathbb{P}\{\mathcal{C}_2 \mid \mathcal{S}_n\}\mathbb{P}\{\mathcal{F}_n \mid \mathcal{C}_1\}\mathbb{P}\{\mathcal{C}_1\}.$$

The first equality derives from the law of total probability while the second equality derives from the chain rule (or general product rule). The third equality results from the fact that, knowing $\mathcal{S}_m$, $\mathcal{C}_2$ is independent of $\mathcal{F}_n$ and $\mathcal{C}_1$. The inequality is due to the fact $\mathbb{P}\{\mathcal{C}_2 \mid \mathcal{S}_m\}$ is an increasing function of m so that $\mathbb{P}\{\mathcal{C}_2 \mid \mathcal{S}_m\} \leq \mathbb{P}\{\mathcal{C}_2 \mid \mathcal{S}_n\}$. The fact that $\sum_{m=1}^{n} \mathbb{P}\{\mathcal{S}_m \mid \mathcal{F}_n \cap \mathcal{C}_1\} = 1$ leads to the last line of this equation. Note that $\mathbb{P}\{\mathcal{C}_2 \mid \mathcal{S}_n\} = e_c(n+1, K_2, \beta_2, 1)$. Finally

$$\mathbb{P}\{\mathcal{F}_n \mid \mathcal{C}_1\}\mathbb{P}\{\mathcal{C}_1\} = \mathbb{P}\{\mathcal{C}_1 \mid \mathcal{F}_n\}\mathbb{P}\{\mathcal{F}_n\} = \quad (16)$$
$$e_c(n+1, K_1, \beta_1, 1) C_{N-1}^{n} p^n (1-p)^{N-1-n}. \quad \square$$

The expression of proposition 2 is an approximation that needs to be validated. We validate it using a discrete event simulator for $N = 20$, $K_1 = 12$, $K_2 = 8$, $\delta_1 = 10^{-3}$, $\beta_1 = \beta_2 = 2$. Figure 8 shows an almost perfect fit of the simulation and analytical results, both for the error after the first phase as well as the error after the two phases. In addition, Figure 8 shows the upper bound derived in proposition 3. It can be observed that this upper bound is very tight for small p, while the approximation fits very well for all values of p.

We now use the approximation of proposition 2 to evaluate the performance of the two-stage scheme. Figure 9 shows the amount of reserved resources for the case with further retransmissions, in comparison with the case where only one stage of replicas is allowed (for $\delta_1 = 10^{-3}$, $\beta_1 = 4$, $\beta_2 = 2$ and $p = 10^{-3}$). We can observe that the two-stage scheme allows reserving less resources in total. However, this does not mean that this 2-stage scheme implies necessarily lower resource reservation, as the total latency budget for retransmissions is reduced by the acknowledgement time. Figure 9 shows also the amount of spectrum to be reserved for the one-stage vs. two-stage schemes, for the parameters of Table I. It is observed that the one-stage scheme outperforms the two-stage scheme for a low number of users, as the acknowledgment time ($t_a = 2$ TTI) will have a large impact in this case, but the two schemes are almost equivalent for larger loads. It is worth noting that,

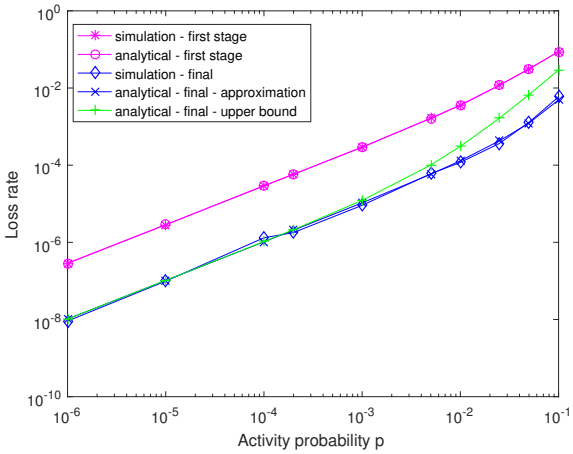

Figure 8: Validation of the analytical expression for a two-phase contention-based scheme.

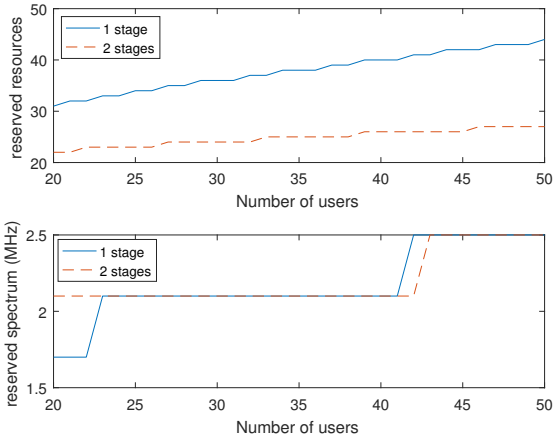

Figure 9: Performance of the 2-stage contention-based scheme.

even if the two-stage scheme does not outperform the one-stage scheme for the amount of needed spectrum, the unused RUs (as $K_1 + K_2 < K$) may be used for other services like eMBB (enhanced Mobile Broadband).

V. Individual Reservation versus Contention-Based Schemes

In the previous sections of this paper, we considered two resource reservation schemes that correspond to two use cases:

1) Individual reservation for the first transmissions, and a pool for retransmissions, after ACKs are received.
2) A contention-based scheme with a common pool for replicas. If the delay budget allows for receiving an ACK, a two-stage contention-based scheme is possible.

While the former scheme is natural for use cases with deterministic traffic arrivals and the latter is suitable for sporadic traffic use cases, there are use cases where both schemes are possible. A typical example is the use case where a limited number of users generates packets randomly, but the

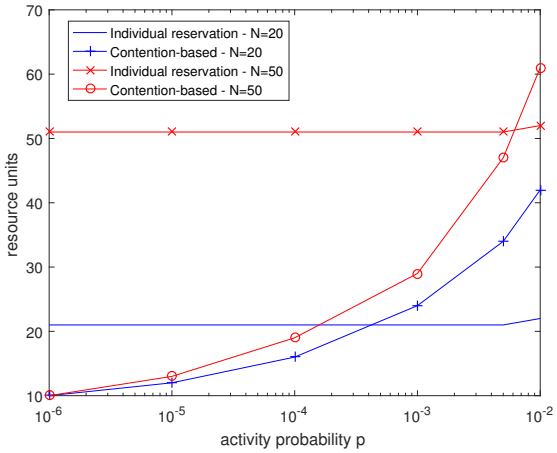

Figure 10: contention-based versus individual reservation.

delay budget allows for receiving acknowledgements. While the two-stage contention-based scheme can be applied as it has been defined in the previous section, the individual reservation scheme has to be adapted to the sporadic traffic. In this case, resources are reserved for users at each cycle, even if they may not have packets to send. The base station requests retransmissions in the common pool for packets that are lost. The error probability of equation (4) has to integrate the fact that users are not always active, but only with probability p:

$$e(K) = \sum_{n=0}^{N-1} C_{N-1}^{n} (p\delta_1)^{n+1} (1-p\delta_1)^{N-1-n} \times \qquad (17)$$
$$(\delta_2\delta_3 I_{n+1\leq \lfloor \frac{K}{2} \rfloor} + \frac{\delta_2 K + n + 1 - K}{n+1} I_{n+1\geq K} +$$
$$\delta_2 \frac{\delta_3(K-(n+1)) + 2(n+1) - K}{n+1} I_{\lfloor \frac{K}{2} \rfloor < n+1 < K})$$

Figure 10 illustrates the performances for the two resource allocation schemes in the sporadic traffic case, for parameters of table I, while varying the activity probability p and for two cases for the number of users ($N = 20$ and $N = 50$). The first observation is that the amount of needed resources in the individual reservation scheme is relatively stable as the amount of resources in the retransmissions pool remains limited compared to the number of users, while the contention based scheme needs much more resources when users become more active. The figure suggests the existence of an optimal choice, depending on the activity profile of users. For low activity profiles, it is better to use the contention-based scheme as this reduces the amount of resources, but starting from a certain activity level, individual reservation is needed as collisions become too frequent.

VI. Conclusion

In this paper, we developed a framework for radio resource allocation for URLLC traffic in 5G use cases. We considered two classes of use cases, depending on the traffic generation profile. For deterministic packet generation, individual reservation of resources is needed for the first transmissions, while a pool of resources is reserved for retransmissions, scheduled by the base station. When traffic is sporadic, a contention-based scheme is adequate, where several replicas of each packet are randomly placed at different resources in order to increase the probability of success, despite possible collisions. In both cases, we derived analytical expressions for the reliability and used them to estimate the amount of resources needed for satisfying the reliability and latency targets. For use cases that allow the usage of both schemes, we showed how to choose between individual reservation and contention-based schemes for the lowest possible resource consumption.

References

[1] P. Popovski, "Ultra-reliable communication in 5G wireless systems," in *1st International Conference on 5G for Ubiquitous Connectivity*, Nov 2014, pp. 146–151.

[2] 3GPP, "Study on scenarios and requirements for next generation access technologies," 3GPP TR 38.913 v14.2.0, Tech. Rep., June 2017.

[3] K. I. Pedersen, S. R. Khosravirad, G. Berardinelli, and F. Frederiksen, "Rethink hybrid automatic repeat request design for 5G: Five configurable enhancements," *IEEE Wireless Communications*, vol. PP, no. 99, pp. 2–8, 2017.

[4] 3GPP, "Study on latency reduction techniques for LTE," 3GPP TR 36.881 v14.0.0, Tech. Rep., June 2016.

[5] S. Sesia, I. Toufik, and M. Baker, *LTE, The UMTS Long Term Evolution: From Theory to Practice, 2nd Edition*. Wiley Publishing, 2011.

[6] P. Schulz, M. Matthe, H. Klessig, M. Simsek, G. Fettweis, J. Ansari, S. A. Ashraf, B. Almeroth, J. Voigt, I. Riedel, A. Puschmann, A. Mitschele-Thiel, M. Muller, T. Elste, and M. Windisch, "Latency critical IoT applications in 5G: Perspective on the design of radio interface and network architecture," *IEEE Communications Magazine*, vol. 55, no. 2, pp. 70–78, February 2017.

[7] D. Jiang, H. Wang, E. Malkamaki, and E. Tuomaala, "Principle and performance of semi-persistent scheduling for VoIP in LTE system," in *2007 International Conference on Wireless Communications, Networking and Mobile Computing*, Sept 2007, pp. 2861–2864.

[8] 3GPP, "Physical layer procedures for data," 3GPP TR 38.214 v15.1.0, Tech. Rep., March 2018.

[9] *Service requirements for next generation new services and markets*, 3GPP, 3 2018, v16.3.0.

[10] B. Singh, O. Tirkkonen, Z. Li, and M. A. Uusitalo, "Contention-based access for ultra-reliable low latency uplink transmissions," *IEEE Wireless Communications Letters*, vol. 7, no. 2, pp. 182–185, April 2018.

[11] R. Kotaba, C. N. Manchón, T. Balercia, and P. Popovski, "Uplink transmissions in URLLC systems with shared diversity resources," *IEEE Wireless Communications Letters*, pp. 1–1, 2018.