Ultra-Reliable Communication for Services with Heterogeneous Latency Requirements

Kalør, Anders Ellersgaard; Popovski, Petar

Published in:
2019 IEEE Globecom Workshops (GC Wkshps)

DOI (link to publication from Publisher):
10.1109/GCWkshps45667.2019.9024507

Publication date:
2020

Document Version
Accepted author manuscript, peer reviewed version

Link to publication from Aalborg University

Citation for published version (APA):
Ultra-Reliable Communication for Services with Heterogeneous Latency Requirements

Anders E. Kalør
Department of Electronic Systems
Aalborg University, Denmark
Email: aek@es.aau.dk

Petar Popovski
Department of Electronic Systems
Aalborg University, Denmark
Email: petarp@es.aau.dk

Abstract—Ultra-reliable communication (URC) is often studied with very strict and homogeneous latency requirements, commonly referred to as ultra-reliable low-latency communication (URLLC). However, in many scenarios the tolerated latencies may vary across users, and treating all users equally may lead to unnecessary over-provisioning of resources. In this paper, we study URC with orthogonal and non-orthogonal access in uplink scenarios where users have heterogeneous latency requirements. Users with strict latency requirements are given resources that are localized in time, while users with less strict latency are given resources that are spread across time and with intermediate feedback. We show that exploiting differences in the tolerated latency can lead to both a significant increase in reliability, and to more efficient use of resources.

I. INTRODUCTION

Ultra-reliable communication (URC) plays a central role in the support of emerging wireless applications, such as industrial automation, smart grids, and virtual reality, where required packet error rates are in the range of $10^{-9}$–$10^{-5}$ [1]. In many cases, URC is blended with strict latency requirements in the order of a few milliseconds, which has led to the introduction of Ultra-Reliable Low-Latency Communication (URLLC), one of the three defining pillars in 5G along with enhanced Mobile Broadband (eMBB) and massive Machine-Type Communication (mMTC) [2]. As a result, URLLC has received much attention during the past years as part of the research activities towards 5G and is expensive in terms of spectral efficiency [4].

However, several URC use cases have less stringent latency requirements than those usually studied under URLLC, such as remote health monitoring and disaster-and-rescue scenarios (see e.g. [1], [5], [6]). This opens the possibilities for using more degrees of diversity and more coordination. Examples could be the ability to spread transmissions across time, acquire channel state information (CSI) to allow for precoding, or to provide intermediate feedback (e.g. stop-feedback) to the transmitter to schedule resources with higher granularity, thus reducing the resource overhead. Common to all of these methods is that they introduce a delay in the communication, and hence cannot be considered for the traditional URLLC use case. Furthermore, it is likely that a single base station will serve applications with heterogeneous latency requirements. Architectures that support heterogeneous services in the same network have been widely studied in the literature through the concept of network slicing [7]. However, most focus has been on the co-existence of eMBB, URLLC and mMTC [8], [9], while the case with diverse latency requirements within the ultra-reliable regime has received less attention. Nevertheless, due to the high over-provisioning required in URLLC regime, it is generally desirable to exploit the additional delay that can be tolerated by some applications.

Motivated by this observation, in this paper we study the scenario in which a base station serves URC devices with heterogeneous, but still moderate latency requirements. We limit the focus to the feedback aspect, i.e. the case where some of the users can tolerate latencies that allow for (short) intermediate feedback during the transmission, while other users cannot. We study various feedback schemes in settings with orthogonal and non-orthogonal access, and quantify the gains in terms of rate and spectral efficiency that can be achieved by exploiting the feedback.

To illustrate the overall idea, consider a wireless interface between a number of ultra-reliable users and a base station, comprising frequency and time resources as depicted in Fig. 1. The resources colored in gray are occupied by users that require very low latency, and hence must be localized in time, while the hatched resources are users with less strict latency requirements that can span several time slots. Feedback is given after every second time slot. In Fig. 1a both user groups are treated equally as URLLC users and multiplexed orthogonally. There is no use of feedback, and hence each user is allotted sufficient (dedicated) resources to cope with potentially bad channel conditions in order to ensure high reliability. On the other hand, Fig. 1b illustrates the idea of multiplexing the users that can tolerate higher latency non-orthogonally with the low-latency users. Furthermore, the base station transmits a stop-feedback signal as soon as the transmission has completed, so as to limit the resource overhead and the interference to the low-latency users. In addition, due to the non-orthogonal multiplexing, the total number of users that can be supported is larger, and the time diversity allows for spreading the
interference across multiple low-latency users, thereby gaining diversity with respect to interference. In summary, the scenario in Fig. 1b has both better utilization of resources and more diversity, thus facilitating the ability to provide high reliability.

The remainder of the paper is organized as follows. Section II presents the system model, and in Section III we present the different scheduling and feedback schemes that we consider. The schemes are evaluated in Section IV and finally the paper is concluded in Section V.

II. SYSTEM MODEL

We consider the uplink in a system comprising $N$ users and a single base station. The air interface is divided into time slots, and each time slot is further divided into $S$ minislots and $F$ frequency channels, as illustrated in Fig. 2. One frequency channel and one time slot constitute a radio resource, which represents the minimum granularity of the scheduler and is assumed to be within the time/frequency coherence interval of the channel. Due to the latency requirements, which are relatively strict for all applications that we consider, we assume that the transmitters have no information about the channel, while the base station has full CSI, acquired through an idealized estimation process during the URC transmission. To satisfy the strict reliability requirements, we assume that the users are pre-assigned radio resources, so that the interference can be controlled. Each user accesses its assigned resources in a grant-free manner and is active with probability $p$.

We consider two groups of users which have identical reliability requirements but different latency requirements. The common reliability requirement is given in terms of a minimal probability, $\epsilon$, of successful transmission within their tolerated latency. Regarding the latency requirements, the users in the first group, which we refer to as low-latency users (URC-LL), have very strict latency requirements and must transmit within a single minislot. On the other hand, the users in the second group have less stringent latency requirements, and need to finish within two time slots ($2S$ minislots). We refer to the second group as high-latency users (URC-HL). The fundamental difference between the two user groups is that the base station can provide feedback to the users in second group between the two time slots. The feedback schemes, which are assumed instantaneous and error free, are outlined in detail in the next section.

We study both orthogonal and non-orthogonal transmissions, and assume a Rayleigh block fading channel. We further assume that the users employ frequency hopping between the minislots so that they experience independent channel coefficients. Denoting the symbols transmitted by user $n$ in frequency channel $f$ and minislot $s$ as $X_{s,f}^{(n)}$ the signal received at the base station in slot $s$, $f$ reads

$$Y_{s,f} = \sum_{n=1}^{N} \delta_{s,f}^{(n)} H_{s,f}^{(n)} X_{s,f}^{(n)} + Z_{s,f},$$

where $\delta_{s,f}^{(n)}$ is a Bernoulli random variable indicating whether user $n$ is active in the slot, $H_{s,f}^{(n)} \sim \mathcal{CN}(0, \Gamma)$ is the channel coefficient of user $n$ in the slot, and $Z_{s,f} \sim \mathcal{CN}(0, I)$ is additive noise.

We remark that it may be beneficial to multiplex the URC users non-orthogonally with eMBB as discussed in [8]. However, since we are only concerned about the URC use
case, we assume that the resources are dedicated to URC
and note that superimposing eMBB traffic would merely add
uncorrelated interference to the URC users.

To quantify the performance of the schemes we define the
following metrics. First, we consider the maximum per-user
rate, denoted by \(r_{LL}\) and \(r_{HL}\), for URC-LL and URC-HL,
respectively, that provide a certain reliability \(\epsilon\). To indicate
the utilization of the resources, we introduce the ratio between
the average rate supported by the channel, \(\mathbb{E}[R]\), and the maximum
rate, \(C_r\), required to satisfy the reliability requirement \(\epsilon\) for
the respective scheme. Mathematically, this is expressed as

\[
\frac{\mathbb{E}[R]}{C_r} = \sup \{ r \mid \Pr(E) \leq 1 - \epsilon \},
\]

where \(\Pr(E)\) is the probability of error. In general, it is
desirable for the number to be small, as this reflects a small
resource overhead. Notice that the ratio can be less than one if
the rate distribution is asymmetric. Although the data packets
are small, the finite blocklength effects are known to have
little impact on the outage capacity [10], and thus we study
the scenario in the infinite blocklength regime.

III. SCHEDULING AND FEEDBACK SCHEMES

We study a total of three transmission policies as illustrated
for two URC-LL and two URC-HL users in Fig. 3. The
first is orthogonal access without feedback, which reflects the
situation of treating both user groups equivalently according to
the most strict requirements. The remaining two policies are
based on non-orthogonal access; one without feedback, and
one with stop-feedback.

Since we consider the per-user error probability, and the
users are assumed to use frequency hopping so that they experience independent channel realizations, we omit the
dependency on the slot and user in the channel coefficients.
Instead, we denote them by \(H_k\) where \(k\) indicates the resource index (frequency and time). When necessary, we distinguish
between URC-LL and URC-HL users using the subscripts LL
and HL, e.g. \(H_{LL,k}\) and \(H_{HL,k}\).

A. Orthogonal access without feedback

The orthogonal access scheme reflects the standard grant-
free transmission, in which users are assigned dedicated resources that they access if they have data to transmit.
Furthermore, in line with the majority of the current research,
no distinction is made between the low-latency and high-
latency users. The situation is illustrated in Fig. 3a, where both URC-LL and URC-HL users are assigned 3 frequency
slots within the same minislot.

We denote the number of frequency resources assigned to
each user by \(K\). The reliability of each user is then given by

\[
\Pr(E) = \Pr \left[ \frac{1}{K} \sum_{k=1}^{K} \log_2 \left( 1 + |H_{k}|^2 \right) < r \right],
\]

where \(r\) is the transmission rate and \(|H_{k}|^2\) are independent exponentially distributed random variables with mean \(\Gamma\).

For a given reliability requirement \(\epsilon\), the rate \(r\) can be
determined using Monte Carlo simulation by setting Eq. (3)
equal to \(\epsilon\). However, this can be difficult to compute for small
\(\epsilon\), which are usually of interest for URC. As an alternative, the
rate may instead be selected conservatively using the bound
derived in Appendix A as

\[
r = \max_{\epsilon > 0} \frac{1}{tK} \log_2 (1 + \mathbb{E}[(1 + |H_{k}|^2)^{-\epsilon}]),
\]

where the expectation can be calculated using Monte Carlo
simulation and \(t > 0\) can be optimized to maximize the rate.

B. Non-orthogonal without feedback

We now turn our attention to non-orthogonal allocations.
The motivation for this is twofold. First, non-orthogonal access
is beneficial when the access probability \(p\) is relatively low,
since the probability of unused resources is lower. Secondly,
the non-orthogonality allows for higher frequency and time
diversity gain, as each resource can serve multiple users. We
first consider the case without feedback. For simplicity, we
assume that each resource is allocated to one URC-LL and one
URC-HL user, as shown in Fig. 3b. However, the resources
could as well be shared among users of the same group, which
is likely to be beneficial if the activation probability is low.
Although Fig. 3b illustrates a diagonal frequency hopping pattern for the URC-HL users, the analysis is valid as long
as each channel resource is not shared by multiple URC-HL
users, and each frequency channel is used at most once by the
same URC-HL user within a time slot.

We assume that the base station initially attempts to de-
code the URC-LL users while treating the URC-HL users as
noise, since URC-LL users receive all their channel resources
within one minislot. The URC-LL users that are successfully
decoded are subsequently cancelled and the URC-HL users are 
decoded. Notice that even if the decoding of a URC-LL user 
fails it may still be possible for the base station to decode the 
URC-HL users by treating the URC-LL interference as noise. 
Denoting by $\delta_{HL,k}$ the Bernoulli random variable indicating 
whether the URC-HL assigned to resource $k$ in the current 
minislot is active, the error probability of the URC-LL user is 
\[
\Pr(E_{LL}) = \Pr\left[ \frac{1}{K_{LL}} \sum_{k=1}^{K_{LL}} \log_2 \left( 1 + \frac{|H_{LL,k}|^2}{1 + \delta_{HL,k}|H_{HL,k}|^2} \right) < r_{LL} \right], 
\]
where the subscripts LL and HL are used to distinguish 
between the URC-LL and URC-HL users, respectively. As in 
the orthogonal case, the rate can be selected according to the 
bound in Appendix A as 
\[
r_{LL} = \max_{t>0} \frac{1}{tK_{LL}} \log_2 (e) - \frac{1}{t} \log_2 \left( \mathbb{E} \left[ \left( 1 + \frac{|H_{LL,k}|^2}{1 + \delta_{HL,k}|H_{HL,k}|^2} \right)^{-1} \right] \right). 
\]
Due to the interference cancellation procedure, the URC-HL 
users can experience three scenarios in each resource: (i) The 
URC-LL user is not active, (ii) the URC-LL user is active, 
successfully decoded and cancelled, and (iii) the URC-LL user 
is active but not decoded. The resulting error probability can 
be written 
\[
\Pr(E_{HL}) = \Pr\left[ \frac{1}{K_{HL}} \sum_{k=1}^{K_{HL}} \log_2 \left( 1 + \frac{|H_{HL,k}|^2}{1 + \delta_{LL,k}(1 - \gamma_{LL,k})|H_{LL,k}|^2} \right) < r_{HL} \right], 
\]
where $\delta_{LL,k}$ is the binary variable indicating whether the 
URC-LL user is active in resource $k$, and $\gamma_{LL,k} = 1$ if the 
URC-LL user is successfully decoded, otherwise $\gamma_{LL,k} = 0$. 
Notice that even though the URC-LL and URC-HL users 
have the same reliability requirements, due to the interference 
cancellation procedure it is not optimal to select $r_{LL}$ = $r_{HL}$ 
and $K_{LL}$ = $K_{HL}$. Instead, the rates depend on both the 
activation probability of the interfering users, as well as the 
number of resources assigned to URC-LL and URC-HL. 
For this reason, deriving bounds on the rate for URC-HL 
is challenging. However, a valid bound can be obtained by 
assuming that the URC-HL user is decoded first, while treating 
the URC-LL transmissions as interference. This gives the rate 
\[
r_{HL} = \max_{t>0} \frac{1}{tK_{HL}} \log_2 (e) - \frac{1}{t} \log_2 \left( \mathbb{E} \left[ \left( 1 + \frac{|H_{HL,k}|^2}{1 + \delta_{LL,k}|H_{LL,k}|^2} \right)^{-1} \right] \right). 
\]

C. Non-orthogonal with stop-feedback

The non-orthogonal scenario with stop-feedback is equivalent 
to the previous case without feedback, with the addition 
of a feedback signal from the base station after the initial 
time slot, that indicates to the URC-HL users whether their 
transmission has completed successfully. Consequently, the 
URC-LL users may experience less interference if a URC-HL 
user succeeds already within the first time slot (see Fig. 3c). 
This in turn slightly increases the reliability of URC-LL. 
We denote the number of resources given in the first and 
second time slots by $K_{HL}^{(1)}$ and $K_{HL}^{(2)}$, respectively, so that 
$K_{HL}^{(1)} + K_{HL}^{(2)} = K_{HL}$. The probability that a URC-HL user 
succeeds after the first time slot is 
\[
\Pr(E_{HL}^{(1)}) = \Pr\left[ \frac{1}{K_{HL}} \sum_{k=1}^{K_{HL}^{(1)}} \log_2 \left( 1 + \frac{|H_{HL,k}|^2}{1 + \delta_{LL,k}(1 - \gamma_{LL,k})|H_{LL,k}|^2} \right) < r_{HL} \right]. 
\]
As a result, the expected number of resources allocated to 
a URC-HL user is $K_{HL}^{(1)} + \Pr(E_{HL}^{(1)}|K_{HL}^{(2)})$. This reflects the 
central advantage of feedback, namely a higher granularity in 
the resource assignment, which in turn results in less resource 
overhead.

While the rates for URC-HL users are the same as in the 
case without feedback, the URC-LL users can support slightly 
higher rate. However, including this into the calculation of 
the bounds is challenging due to the dependence between the 
interference experienced by the URC-LL users and the rate of 
the URC-LL users. Hence, we will resort to using the same 
bounds as in the case without feedback for both URC-HL and 
URC-LL users.

IV. NUMERICAL EVALUATION

In this section we present numerical results to illustrate 
the reliabilities under the schemes described in the previous 
section. Since the dependency between successful decoding 
of URC-LL and URC-HL render the error probability calculations 
difficult, we approximate the results using Monte Carlo 
simulations.

We consider a scenario of a total of $N = 20$ users, divided 
into 10 URC-LL and 10 URC-HL users. The air interface 
contains $F = 10$ frequency channels and each time slot is 
comprised of $S = 5$ minislots. The activation probability is 
$p = 0.5$ and the channel gains are normalized to $\Gamma = 1$. In 
the case of orthogonal access, each user is assigned 5 frequency 
resources so that the users occupy a total of two time slots. 
In the non-orthogonal schemes, each user is given 10 resources. 
More specifically, the URC-LL users are each allocated a 
dedicated time slot, i.e. $K_{UL} = 10$, while each URC-HL user 
is assigned one frequency resource in each minislot, so that 
their resources are equally divided between the two time slots 
i.e. $K_{HL}^{(1)} = K_{HL}^{(2)} = 5$.

The error probabilities for the various schemes and the rate 
bounds are shown for URC-HL and URC-LL users in Figs. 4
and 5, respectively. In both cases, the non-orthogonal schemes support higher rates than the orthogonal in the high-reliability region. This indicates that despite the interference caused by non-orthogonality, the fact that twice as many resources can be assigned to each user results in higher reliability. For URC-HL, the scheme without feedback and the scheme with stop-feedback result in the same error probabilities, as stop-feedback only impacts the URC-HL users that have successful transmissions. However, in the case with URC-LL users the stop-feedback scheme results in higher reliability than the other schemes due to the reduced interference experienced in the second timeslot.

The utilization of the schemes are shown for URC-HL in Fig. 6. Again, the two non-orthogonal schemes outperform the orthogonal due to the higher number of resources and hence improved average channel conditions. For large target error probabilities, \( \epsilon_{HL} \), the non-orthogonal schemes perform equivalently and the utilization ratio tends towards zero. However, as the target reliability increases, the stop-feedback scheme becomes more efficient as less users are given excess resources.

The case with URC-LL is shown in Fig. 7, where it can be seen that due to the lack of feedback, all schemes exhibit high overhead in the high reliability region. While the overhead for the orthogonal scheme is largest, the stop-feedback scheme is slightly higher than the one without feedback. This indicates that, despite users in this scheme can transmit with higher rate to achieve a certain \( \epsilon_{LL} \) (see Fig. 5), the average rate overhead is even larger.

V. CONCLUSION

In this paper we have investigated ultra-reliable communication in a scenario where some users require very low latency, while others can tolerate higher latencies. We have studied how the increased time diversity and the use of intermediate feedback to the high-latency users can help supporting ultra-reliable communication with both orthogonal and non-orthogonal access. More specifically, we show that the ability to use stop-feedback can lead to higher reliability and better resource utilization. Furthermore, even without feedback non-orthogonal access outperforms orthogonal access in the ultra-reliable regime both in terms of reliability and resource.
efficiency due to increased time, frequency and interference diversity gains. This suggests that adapting the use of the radio resources to the latency requirements is highly beneficial in the ultra-reliable regime.

Future research can be in the direction of studying the use of power control as well as more sophisticated feedback schemes that exploit the information that the base station has obtained during the initial time slot, such as channel estimations. Furthermore, the model could be generalized e.g. to allow for more general resource allocations, heterogeneous reliability requirements, and eMBB traffic.

APPENDIX A
DERIVATION OF LOWER BOUND ON RATE

We derive the lower bounds on the rate for the non-orthogonal case in Eq. (5), from which the orthogonal access can be obtained by setting $\delta_{LL,k} = 0$. By fixing the error probability as $\Pr(E_{LL}) = \epsilon_{LL}$ and by following the same procedure as in [8] we obtain

$$\epsilon_{LL} = \Pr\left[ \sum_{k=1}^{K_{LL}} \log_2 \left( 1 + \frac{|H_{LL,k}|^2}{1 + \delta_{HL,k}|H_{HL,k}|^2} \right) > K_{LL}r_{LL} \right] \leq t \log_2 \left( \prod_{k=1}^{K_{LL}} \left( 1 + \frac{|H_{LL,k}|^2}{1 + \delta_{HL,k}|H_{HL,k}|^2} \right)^{-t} \right)$$

(10)

$$= \Pr\left[ -t \log_2 \left( \prod_{k=1}^{K_{LL}} \left( 1 + \frac{|H_{LL,k}|^2}{1 + \delta_{HL,k}|H_{HL,k}|^2} \right) \right) < -K_{LL}r_{LL}t \right]$$

(11)

$$= \Pr\left[ \prod_{k=1}^{K_{LL}} \left( 1 + \frac{|H_{LL,k}|^2}{1 + \delta_{HL,k}|H_{HL,k}|^2} \right)^{-t} < 2^{-K_{LL}r_{LL}t} \right]$$

(12)

$$\leq \frac{1}{2^{-K_{LL}r_{LL}t}} \sum_{k=1}^{K_{LL}} \left( \frac{|H_{LL,k}|^2}{1 + \delta_{HL,k}|H_{HL,k}|^2} \right)^{-t}$$

(13)

$$= \frac{1}{2^{-K_{LL}r_{LL}t}} \sum_{k=1}^{K_{LL}} \left( 1 + \frac{|H_{LL,k}|^2}{1 + \delta_{HL,k}|H_{HL,k}|^2} \right)^{-t}$$

(14)

Here, (11) is obtained by moving the summands inside the logarithm, and then multiplying both sides of the inequality by $-t$. In (12) we have raised both sides to the power of two, and (13) follows from the Markov inequality. Using the fact that the terms inside the expectation are independent and identically distributed we arrive at the expression in (14). The lower bound on the rate for a given $\epsilon_{LL}$ can then be obtained by rewriting the expression as

$$r_{LL} \geq \frac{1}{tK_{LL}} \log_2 \left( \epsilon_{LL} \right) \geq \frac{1}{tK_{LL}} \log_2 \left( \frac{1 + \Delta_{HL,k}|H_{HL,k}|^2}{1 + \Delta_{HL,k}|H_{HL,k}|^2} \right)^{-t} \right) \right)$$

(15)

where the expectation can be approximated using Monte Carlo simulation and $t > 0$ can be optimized so as to maximize the rate.

ACKNOWLEDGMENT

This work has been in part supported the European Research Council (ERC) under the European Union Horizon 2020 research and innovation program (ERC Consolidator Grant Nr. 648382 WILLOW) and Danish Council for Independent Research (Grant Nr. 8022-00284B SEMIOTIC)

REFERENCES


