Performance Analysis of Hybrid ARQ for Ultra-Reliable Low Latency Communications

João Pedro Battistella Nadas, *Student Member, IEEE*, Oluwakayode Onireti, *Member, IEEE*, Richard Demo Souza, *Senior Member, IEEE*, Hirley Alves, Glauber Brante, *Senior Member, IEEE*, and Muhammad Ali Imran, *Senior Member, IEEE*

***Abstract*— Considering an ultra-reliable low latency communication scenario, we assess the trade-off in terms of energy consumption between achieving time diversity through retransmissions and having to communicate at a higher rate due to latency constraints. Our analysis considers Nakagami-m block-fading channels with Chase combining hybrid automatic repeat request. We derive a fixed-point equation to determine the best number of allowed transmission attempts considering the maximum possible energy spent, which yields insights into the system behavior. Furthermore, we compare the energy consumption of the proposed approach against direct transmission with frequency diversity. Results show substantial energy savings using retransmissions when selecting the maximum number of transmission attempts according to our approach. For instance, considering a Rayleigh channel and smart grid teleprotection applications, our approach uses around 8 times less energy per bit compared with a direct transmission with frequency diversity.**

***Index Terms*— URLLC, energy efficiency, CC-HARQ.**

I. Introduction

THE essence of 5G systems is based on three important use cases: enhanced mobile broadband (eMBB), massive machine-type communication (mMTC), and mission critical MTC (cMTC) [1]. Therefore, 5G systems will not only have to outperform previous generations in terms of requirements such as data rate and capacity [2], but also address new machine-type communication (MTC) usage scenarios [1], including cMTC applications with very high demands for reliability, availability and low latency. Examples of cMTC include applications with ultra-reliable and low latency communication (URLLC) requirements in order to enable real-time automation and control of dynamic processes. Applications such as factory automation through network controlled systems [3], autonomous roads [4], platooning [5], haptic communications [6] and teleprotection in smart grids [7], are predicted cMTC scenarios that can only be enabled by URLLC, which will be supported by 5G networks [8]. Moreover, energy constraints are observed in cMTC, since devices are usually battery-powered. Therefore, energy efficient protocols toward URLLC is a relevant research topic [9]–[12].

There are several strategies to improve energy efficiency in wireless communications, many of which revolve around mitigating the effect of fast-fading through diversity. In systems without latency constraints hybrid automatic repeat request (HARQ) is well known to improve the energy efficiency [13], [14] by providing time diversity. In [13], modulation order, transmit power, number of transmission attempts and code rate are optimized using a realistic energy consumption model for the case of truncated simple and Chase combining HARQ (CC-HARQ), considering fast and block-fading scenarios in a Nakagami-m channel.

On the other hand, adaptive HARQ is studied in [14] and compared to traditional 1-bit feedback HARQ. Modeling the system with a Markov decision process, the authors derive optimal policies for truncated and persistent adaptive HARQ. An analysis using cooperation and simple HARQ is considered in [15], accounting for average delay constraints from a coding and modulation point of view. They propose a solution for power allocation and communication strategy that minimizes the overall power consumption of the system and their results show that the solution can reduce the overall energy consumption. Moreover, Dosti *et al.* [9] and [10] allocate power in order to improve energy efficiency considering truncated simple ARQ and CC-HARQ in a Rayleigh block-fading channel. Furthermore, URLLC and the impact of finite block-length (FBL) in channel capacity are considered, while they present a formal description of the optimization problem and solve it in closed-form using the Karush-Kuhn-Tucker (KKT) conditions; they also show that power allocation in HARQ is a good strategy to improve the system energy efficiency. However, they do not analyze the effect of retransmissions in latency.

Manuscript received November 4, 2018; revised December 21, 2018; accepted December 21, 2018. Date of publication January 7, 2019; date of current version April 5, 2019. This work was supported by EPSRC Global Challenges Research Fund the DARE Project under Grant EP/P028764/1, by CAPES and CNPq, Brazil, and by the Academy of Finland, 6Genesis Flagship (Grant no. 318937), ee-IoT (Grant no. 319008), and Academy Professor (Grant no. 307492). The associate editor coordinating the review of this paper and approving it for publication was Prof. Huang Chen Lee. *(Corresponding author: João Pedro Battistella Nadas.)*

J. P. Battistella Nadas, O. Onireti, and M. A. Imran are with the School of Engineering, University of Glasgow, Glasgow G12 8QQ, U.K. (e-mail: j.battistella-nadas.1@research.gla.ac.uk; oluwakayode.onireti@glasgow.ac.uk; muhammad.imran@glasgow.ac.uk).

R. D. Souza is with the Department of Electrical and Electronics Engineering, Federal University of Santa Catarina, Florianópolis 88040-900, Brazil (e-mail: richard.demo@ufsc.br).

H. Alves is with the Centre for Wireless Communications, University of Oulu, 90014 Oulu, Finland (e-mail: hirley.alves@oulu.fi).

G. Brante is with the Departament of Electrotechnics, Federal Technological University–Paraná, Curitiba 80230-901, Brazil (e-mail: gbrante@utfpr.edu.br).

Digital Object Identifier 10.1109/JSEN.2019.2891221

Sun *et al.* [11] analyzed improving the overall energy efficiency of a 5G URLLC network by considering a resource allocation policy from a queuing perspective. Also considering queue delays, Li *et al.* [16] showed that the policy has great impact on the achievable latency for providing system-wide URLLC. However, they do not consider the effect of fading, as is done here.

In [17] we investigated the energy efficiency of URLLC truncated simple HARQ and a novel optimization strategy is proposed via optimizing the maximum number of allowed transmission attempts, for a block-fading Nakagami-m channel, whilst guaranteeing a maximum latency. We analyze the trade-off between rate and diversity, showing that important energy savings can be obtained. Similarly, in [18] the number of allowed attempts is also optimized, but the focus is to reduce the required bandwidth for the URLLC application.

Unlike [9], [10], [13], [14], in this work we guarantee a maximum latency with a target reliability. The work in [15], by its turn, considers energy consumption using HARQ but only accounts for average delays, which is not suitable for cMTC applications, and in [17] we explored simple HARQ. A power allocation scheme is not considered in this work because it requires longer feedback messages, which can be a problem in URLLC. Moreover, since we consider peak power constraints, the applicability of power allocation strategies is limited.

We investigate the impact of CC-HARQ on the energy consumption of a point to point URLLC system. Furthermore, we derive a fixed point equation to determine the optimal number of attempts considering the maximum possible energy consumption. As the first contribution of this work, we show that despite the added latency for allowing retransmissions, their appropriate use can lead to a better performance when considering URLLC.

The proposed solution is further evaluated in a smart grid teleprotection scenario, as described in the mobile and wireless communications enablers for the twenty-twenty information society (METIS) test case number 5 [7]. It consists of reliably delivering messages within a tight latency constraint between substations for the purpose of triggering protection mechanisms when faults occur, preventing damage to the grid. As a second contribution of this work, results show that using different channels for each HARQ round achieves relevant energy savings compared to using all channels in parallel to achieve frequency diversity, even when accounting for higher data rates required to meet latency constraints in CC-HARQ.

The remainder of this paper is organized as follows. Table I contains a list of symbols used in the paper. Section II presents the system model, Section III contains the optimization problem, Section IV discusses simulation results and Section V concludes the paper.

TABLE I
LIST OF SYMBOLS

Symbol	Definition
Energy and Power	
$\bar{E}_{\mathrm{b}}$	Average energy per successful bit
$\hat{E}_{\mathrm{b}}$	Maximum energy per successful bit
$\bar{E}_{\mathrm{f}}$	Frequency diversity consumption per bit
E_{st}	Circuit start-up energy
η	PA average drain efficiency
$P_{\mathrm{el,rx}}$	Passband circuit consumption
$P_{\mathrm{el,tx}}$	Baseband and RF circuit consumption
P_{el}	Total circuit consumption
P_{PA}	Power amplifier consumption
P_{rf}	Radiated power
$P_{\mathrm{rf,max}}$	Maximum instantaneous radiated power
Fading	
A_0	Attenuation at Reference Distance
α	Path loss exponent
d	Link distance
h	Channel gain
m	Nakagami-m fading parameter
M_{c}	Coding margin
M_{l}	Link margin
N_0	Noise power spectral density
$P_{\mathrm{out},j}$	Probability of failing at the j^{th} attempt
W	Bandwidth
Latency and Reliability	
λ'	Maximum latency
λ	Maximum latency before decoding
R	Rate
$R^\star$	Optimal Rate
R_{max}	Maximum rate
R_{min}	Minimum rate
δ_{fb}	Time to decode feedback signals
δ_{fw}	Time to decode the message
T_{out}	Target outage
Number of Symbols and Bit Lengths	
L_{fb}	Feedback Length
L_{H}	Header Length
L_{D}	Payload Length
L_t	Total message length
n	Total forward symbols ($n_{\mathrm{fw}} + \rho$)
n_{fw}	Number of forward symbols for data
n_{fb}	Number of symbols for feedback
ρ	Number of pilots for channel estimation
SNR	
γ_0	Outage threshold
$\bar{\gamma}$	Average SNR
$\bar{\gamma}_{\mathrm{d}}$	Data transmission average SNR
$\bar{\gamma}_{\mathrm{eff}}$	Effective SNR
$\bar{\gamma}_{\mathrm{p}}$	Pilot average SNR
Symbol Vectors	
$\boldsymbol{p}$	Pilot training sequence
$\boldsymbol{r}$	Received signal
$\boldsymbol{s}$	Encoded message at transmitter
$\boldsymbol{w}$	AWGN
Transmission Attempts	
τ	Number of transmission attempts
$\bar{\tau}$	Average number of transmission attempts
z	Maximum transmission attempts
$z^\star$	Optimal number of attempts
$\hat{z}$	Real relaxation of z
$\hat{z}^\star$	Result of optimizing $\hat{E}_{\mathrm{b}}$

II. SYSTEM MODEL

Consider a point to point communication link, where short messages composed of L_{T} bits are mapped via an encoder to the signal $\boldsymbol{s}$ composed of n symbols to be transmitted over a block-fading channel with gain h. Each symbol period, also denoted channel use, experiences the same channel gain over its n symbols, such that the received signal is $\boldsymbol{r} = h\boldsymbol{s} + \boldsymbol{w}$, where $\boldsymbol{w}$ is the additive white Gaussian noise (AWGN). Additionally, each channel realization h is random and follows a Nakagami-m distribution. Furthermore, in the context of URLLC, the message has to be delivered with very high reliability within a maximum latency λ'.

A. Retransmissions

We investigate the use of time diversity through retransmissions to make the links viable in face of the stringent requirements of URLLC applications, as well as to reduce the energy consumption. Because of the latency constraint,

we consider truncated CC-HARQ, limiting the number of transmission attempts to a maximum of z. In this scheme, if the receiver succeeds in decoding a message, it responds with an acknowledgment (ACK); otherwise, it stores the received signal and responds with a non-acknowledgment (NACK). Upon receiving a NACK, the transmitter resends the same message, which is combined at the receiver using maximum ratio combining (MRC). Then, the receiver tries to decode the combined message and this process is repeated until success or z attempts have been made, after which an error is declared. Moreover, to reduce the communication latency, we assume that the NACK is sent if the accumulated signal-to-noise ratio (SNR) is below a threshold, before trying to decode the entire message. Thus, it only has to be decoded once, when the accumulated SNR is above the threshold.

Based on the error probability of each attempt, we calculate the average number of required transmission trials $\bar{\tau}$ as

$$\bar{\tau}(z) = 1 + \sum_{j=1}^{z-1} P_{\text{out},j}, \tag{1}$$

where $P_{\text{out},j}$ is the probability of failing at the j^{th} attempt.

Note that diversity is not achieved unless the channel varies from one attempt to the other. Thus, slow frequency hopping is employed between consecutive attempts to ensure that each round experiences a different channel realization h_j, where j is the attempt number. In practice, this can be achieved by using a different channel for every attempt.[1] This imposes that h_j must be estimated at the beginning of each attempt.

B. Probability of Outage

It has been shown in [19] that the outage probability P_{out} is a good approximation for the probability of error at high SNR—as in the case of URLLC—even considering a finite block length. For the case of CC-HARQ, where the receiver combines all attempts using maximum ratio combining to increase the chances of successfully decoding the message, the probability of outage after z rounds $P_{\text{out,z}}$ is expressed as [20]

$$P_{\text{out},z} = \frac{\Gamma_{\text{inc}}\left(zm, m\frac{\gamma_0}{\bar{\gamma}}\right)}{\Gamma(zm)}, \tag{2}$$

where $\Gamma_{\text{inc}}(\cdot,\cdot)$ is the lower incomplete gamma function, $\Gamma(\cdot)$ is the complete gamma function, m is the Nakagami fading parameter, $\gamma_0 = 2^R - 1$, R is the data rate and $\bar{\gamma}$ is the average SNR. Meanwhile, $P_{\text{out,z}}$ is well approximated, at high SNR, as [21]

$$P_{\text{out},z} \approx \frac{\left(\frac{m\gamma_0}{\bar{\gamma}}\right)^{mz}}{\Gamma(zm+1)}. \tag{3}$$

C. Channel Estimation

In order to perform coherent detection, the receiver must estimate the channel. Furthermore, as discussed previously, this estimation has to occur before each transmission round in our set-up. This can be done via in-band pilot training, where for each attempt, ρ pilots are used to estimate the channel state information (CSI) at the receiver. Thus, the received signal for the first ρ symbols is given by [22] and [23]

$$\boldsymbol{r} = \sqrt{P_{\text{r}}} h_j \boldsymbol{p} + \boldsymbol{w}, \tag{4}$$

where $\boldsymbol{p} = [p_1\ p_2\ \ldots\ p_\rho]$ is the sequence of pilot symbols and P_{r} is the received signal power, dependent on the transmit power and on the path loss. Using any established channel estimation technique, *e.g.*, minimum mean-square error [24], the receiver obtains an estimate $\tilde{h}_j$ of the channel, which differs from the actual channel realization. Therefore, during the remaining $n_{\text{fw}} = n - \rho$ channel uses, we have

$$\boldsymbol{r} = \sqrt{P_{\text{r}}}(\tilde{h}_j \boldsymbol{s} + (h_j - \tilde{h}_j)\boldsymbol{s}) + \boldsymbol{w}. \tag{5}$$

Thus, the average SNR considering both the channel estimation and the data transmission phases is expressed as

$$\bar{\gamma} = \frac{(n-\rho)\bar{\gamma}_{\text{d}} + \rho\bar{\gamma}_{\text{p}}}{n}, \tag{6}$$

where $\bar{\gamma}_{\text{d}}$ is the average SNR used for data and $\bar{\gamma}_{\text{p}}$ is the average SNR used for pilots. Note that since $\bar{\gamma}$ depends on the transmit power and the large-scale path loss, it is limited by the peak power constraint.

Furthermore, despite depending on the signal and not being Gaussian, the effect of imperfect channel estimation, $(h_j - \tilde{h}_j)\boldsymbol{s}$, can be well modeled as a Gaussian random variable and combined with $\boldsymbol{w}$ into an effective noise perceived by the system, as in [25] and [26]. This effective noise provides a worst case scenario [26] and is modeled as a lower effective average SNR for the purpose of system performance analysis, such that [25]

$$\bar{\gamma}_{\text{eff}} = \frac{\rho\bar{\gamma}_{\text{d}}\bar{\gamma}_{\text{p}}}{1 + \bar{\gamma}_{\text{d}} + \rho\bar{\gamma}_{\text{p}}}. \tag{7}$$

To account for the channel estimation error, we use the effective average SNR $\bar{\gamma}_{\text{eff}}$ obtained via (7) for the purpose of evaluating the outage probability.

When there are peak power limitations, as in most practical applications, more than one pilot must be used to obtain optimal performance [26], as optimally allocating power to one pilot might violate the peak power constraints. In this case, the optimal value of ρ is obtained numerically, with each pilot using the maximum allowed power [26].

In Fig. 1 we have plotted the SNR loss due to imperfect channel estimation considering a peak power constraints versus the number of pilots used and $\bar{\gamma}_{\text{d}}$. In this example, the peak power limitations alongside the path loss yields $\bar{\gamma}_{\text{d}} \leq 40$dB. As discussed before, the pilot transmit power is the maximum allowed such that $\bar{\gamma}_{\text{p}} = 40$dB. We can observe that the effect of imperfect channel estimation is more pronounced when trying to obtain an estimate for the channel with $\bar{\gamma}_{\text{d}}$ close to the pilot SNR $\bar{\gamma}_{\text{p}}$ and fewer pilots are used. This relates to the fact that the power used in estimation is too small in comparison to the signal power. Therefore, we conclude that when choosing the number of pilots, it is important to consider how far from the peak power limitation is the data going to be transmitted. In

[1] In general a communication standard divides the whole available bandwidth into several channels.

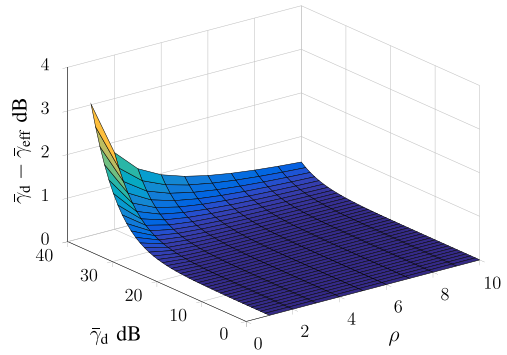


Fig. 1. Effect of imperfect channel estimation on the average SNR for different values of ρ and $\bar{\gamma}_d$. A peak power constraint is considered such that the maximum $\bar{\gamma}_d$ is 40dB, which is the same value for $\bar{\gamma}_p$.

general, it is beneficial to use more pilots when the average SNR of the transmission is close to the maximum allowed by the system. On the other hand, the SNR loss is relatively small and a well designed link margin can be enough to account it.

D. Energy Consumption Model

In this work, we model the energy consumed in the transmission of one message considering the radio startup energy, the pre-transmission processing energy, the energy involved in powering the passband receiver elements, and the electromagnetic radiation, similarly to [13]. However, we consider the exchange of short messages, typical of URLLC, with rates relatively low which cause the arithmetic processing unit clock speed to also be low [13]. This, in turn, causes the energy required for encoding and decoding messages to be small, especially when compared to other consumptions. Thus, we disregard the baseband coding/decoding energy from our model simplifying the presentation without harming the analysis, similarly to [27]. However, since the energy consumed to transmit pilot symbols becomes relevant when the information packets are shorter, we explicit their contribution to the energy consumption in the following analysis.

1) Energy Used by the Transmitter: We assume that, in order to save energy, the transmitter is in idle mode before initiating a transmission, such that it uses a certain startup energy (E_{st}) to wake-up before the first attempt. Both baseband and radio-frequency (RF) circuits, as well as the power amplifier (PA), are used for n channel uses for each attempt.

Next, at the data transmission phase, the remaining $n_{\text{fw}} = n-\rho$ symbols are sent with $\bar{\gamma}_d$ average SNR. The value of n_{fw} is determined based on the rate R (in bits per channel use) and the payload L_{D} and header L_{H} lengths (in bits), such that

$$n_{\text{fw}} = \frac{L_{\text{H}} + L_{\text{D}}}{R}. \tag{8}$$

As in [13], the consumption of baseband and RF circuits is assumed to be constant and equal to $P_{\text{el,tx}}$. Also, the power used to energize passband receiver elements $P_{\text{el,rx}}$ is assumed to be invariant. However, the electromagnetic radiation energy depends on the PA's consumption P_{PA}, which is a function of its average drain efficiency η and of the radiated power P_{rf} [13]

$$P_{\text{PA}} = P_{\text{rf}}\eta^{-1}. \tag{9}$$

Next, P_{rf} is expressed as a function of the path loss and $\bar{\gamma}$,

$$P_{\text{rf}} = N_0 W M_{\text{l}} M_{\text{c}} A_0 d^{\alpha}\bar{\gamma}, \tag{10}$$

where d is the link distance, α is the path loss exponent, N_0 is the noise power spectral density, W is the bandwidth in Hz, the link margin is M_{l}—which includes the noise figure and other unforeseen losses—, M_{c} is the coding margin (further explained in Section II-E) and A_0 is the attenuation at a reference distance. Combining (9) and (10), we have

$$P_{\text{PA}} = A d^{\alpha}\bar{\gamma}/\eta, \tag{11}$$

where $A = N_0 W M_{\text{l}} M_{\text{c}} A_0$. Therefore, to obtain the PA power consumption for the feedback $P_{\text{PA,fb}}$, data transmission $P_{\text{PA,d}}$ and channel estimation $P_{\text{PA,p}}$ phases we use (11) with the respective average SNR for each phase, $\bar{\gamma}_{\text{fb}}$ for the feedback, $\bar{\gamma}_{\text{d}}$ for the data transmission and $\bar{\gamma}_{\text{p}}$ for the channel estimation. Here, the transmit power used for estimation and feedback is always the maximum, such that $\bar{\gamma}_{\text{fb}} = \bar{\gamma}_{\text{p}}$ and $P_{\text{PA,fb}} = P_{\text{PA,p}}$.

When the first attempt fails, the receiver requests a retransmission. Then the transmitter receives an L_{fb} bits long feedback message at each attempt for n_{fb} channel uses, as

$$n_{\text{fb}} = \frac{L_{\text{fb}}}{R}. \tag{12}$$

Therefore, assuming a bandwidth of W, the energy used at the transmitter for τ forward transmission attempts is

$$E_{\text{tx}} = E_{\text{st}} + \frac{\tau}{W}\left[n_{\text{fw}}(P_{\text{el,tx}} + P_{\text{PA,d}}) + E_{\text{p,tx}} + n_{\text{fb}}P_{\text{el,rx}}\right], \tag{13}$$

where $E_{\text{p,tx}} = \rho(P_{\text{el,tx}} + P_{\text{PA,p}})$ denotes the energy used by the transmitter for sending the pilots.

2) Energy Used by the Receiver: Assuming receiver and transmitter use identical radios, the energy used by the former is similar to the one used by the latter. Following the same steps,[2] the energy used by the receiver for τ attempts is

$$E_{\text{rx}} = E_{\text{st}} + \frac{\tau}{W}\left[n_{\text{fb}}(P_{\text{el,tx}} + P_{\text{PA,p}}) + (n_{\text{fw}} + \rho)P_{\text{el,rx}}\right]. \tag{14}$$

3) Average Energy per Successful Bit: The average energy $\bar{E}$ is obtained by considering the average number of transmissions $\bar{\tau}(z)$ and adding (13) with (14), yielding

$$\bar{E} = 2E_{\text{st}} + \frac{\bar{\tau}(z)}{W}\left[n_{\text{fw}}(P_{\text{el}} + P_{\text{PA,d}}) + (\rho + n_{\text{fb}})(P_{\text{el}} + P_{\text{PA,p}})\right] \tag{15}$$

where $P_{\text{el}} = P_{\text{el,tx}} + P_{\text{el,rx}}$.

In order to obtain $\bar{E}_{\text{b}}$, the average energy per successful bit, we normalize the result in (15) by the payload length times the probability of success after z attempts, yielding

$$\bar{E}_{\text{b}}(z) = \frac{\bar{E}}{L_{\text{D}}(1 - P_{\text{out},z})}. \tag{16}$$

[2]Note that we consider the wake-up energy both at transmitter and receiver, thus we assume scheduled-rendezvous [28].

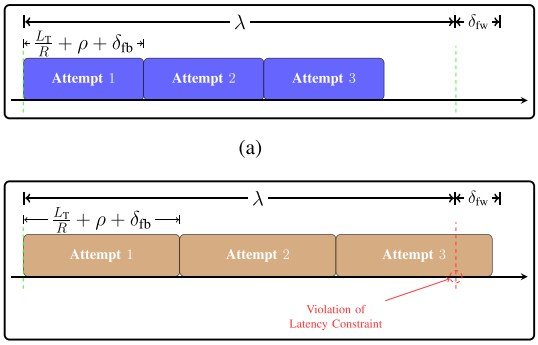


Fig. 2. Illustration of the reason why a maximum latency constraint imposes a minimum rate. In this example, $z = 3$. (a) Latency constraint is respected. (b) Latency constraint is violated.

E. Latency Constraint

The goal of this paper is to investigate the impact of CC-HARQ on the energy consumption of a point to point URLLC system. Traditionally, it is well understood that HARQ improves energy efficiency at the cost of higher latencies [29]. On the other hand, when considering URLLC, a strict maximum latency is imposed [30]. Therefore it is not obvious that CC-HARQ improves the energy efficiency in this scenario.

Here the transmitter must fit all z transmission attempts—and their associated acknowledgments, decoded—within a maximum latency λ' seconds using a bandwidth W. Note that, as mentioned in Section II-A, the receiver estimates the channel at each attempt and if the accumulated SNR is below a threshold, it decodes the header and sends back a NACK immediately after all the symbols have been received and stored. Therefore, the entire message only has to be decoded once, saving latency. The time to decode it δ_{fw} is deducted λ', such that all transmission attempts have to fit within $\lambda = \lambda' - \delta_{\text{fw}}$ seconds. This can be viewed as a constraint on the minimum communication rate R_{min} in bits per channel use. Fig. 2 illustrates this idea in a case where $z = 3$ and with normalized bandwidth. Note that in Fig. 2a, the communication rate is higher than the minimum and it is possible to transmit L_{T} bits z times in less than λ seconds. Conversely, in Fig. 2b, the rate is lower than R_{min} and attempting to communicate L_{T} bits using up to z transmission attempts violates the constraint.

Therefore, R_{min} is determined by calculating the rate at which all z attempts would take $W\lambda$ channel uses, yielding

$$W\lambda = z(n_{\text{fw}} + n_{\text{fb}} + \rho + W\delta_{\text{fb}}), \tag{17}$$

where δ_{fb} is the time it takes for the transmitter to decode the feedback packet. Next, we substitute n_{fw} and n_{fb} defined in (8) and (12), respectively, in (17) while using $R = R_{\text{min}}$. Then using $L_{\text{T}} = L_{\text{H}} + L_{\text{D}} + L_{\text{fb}}$, and solving for R_{min}, we arrive at

$$R_{\text{min}} = z\frac{L_{\text{T}}}{W(\lambda - z\delta_{\text{fb}}) - z\rho}. \tag{18}$$

Note from (18) that using a larger z imposes a higher R_{min}, resulting in a trade-off between diversity and rate. Moreover, despite having excellent performance in terms of error rates, turbo codes, the most commonly used in LTE, have a significant complexity [31] and thus may not be suitable for CC-HARQ considering URLLC applications. Instead, polar codes are good candidates for encoding feedback signals, as they are simple to implement and can be used to encode and decode short feedback messages within negligible time [31], such that $\delta_{\text{fb}} \ll \lambda$. The trade-off is a slight loss in terms of error rate [32], which we have added to the path loss model as M_{c}.

III. Optimization

Our goal is to analyze the trade-off between gaining diversity, by increasing the maximum number of allowed retransmissions, and increasing the data rate, in order to communicate in less than λ seconds. We formally establish the optimization problem in order to minimize the average energy per successful forward bit whilst meeting constraints for maximum instantaneous transmit power $P_{\text{rf,max}}$, latency and reliability (expressed in the form of a target outage T_{out}), as

$$\begin{aligned} \underset{z\in\mathbb{N}^*}{\text{minimize}} \quad & \bar{E}_{\text{b}}(z) && \text{(19a)}\\ \text{subject to} \quad & P_{\text{out},z} \leq T_{\text{out}} && \text{(19b)}\\ & P_{\text{rf}} \leq P_{\text{rf,max}} && \text{(19c)}\\ & R \geq R_{\text{min}}. && \text{(19d)} \end{aligned}$$

A. Optimizing Maximum Energy

Although it can be numerically verified that the objective function is convex with respect to z, and thus has one unique global optimal solution, to the best of our knowledge, it is not possible to prove it analytically due to the shape of $\bar{\tau}$ and where it appears in (19a).

Thus, we propose an alternative approach, where we optimize the maximum energy consumption $\hat{E}_{\text{b}}$ in the same setup. The value of $\hat{E}_{\text{b}}$ is obtained by replacing $\bar{\tau}(z)$ by z in (16), such that

$$\hat{E}_{\text{b}}(z) = \frac{2E_{\text{st}} + \frac{z}{WR}\Phi}{L_{\text{D}}(1 - P_{\text{out},z})}, \tag{20}$$

where $\Phi = L_{\text{fw}}(P_{\text{el}} + P_{\text{PA,d}}) + (L_{\text{fb}} + \rho R)(P_{\text{el}} + P_{\text{PA,p}})$.

This allows us to design a protocol with the worst case scenario in mind, which is a sensible approach in URLLC, and we also show numerically that this result yields almost the same performance as a solution obtained numerically via the problem in (19). Moreover, we show that in this case, the optimal rate $R^\star$ and when R_{min} exceeds this optimal, the optimal number of transmission attempts $\hat{z}^\star$ can be obtained via a floating point equation. Furthermore, using the obtained result, we are able to show that when the link budget is more stringent, as is the case of URLLC, $R^\star$ becomes smaller and using the obtained $\hat{z}^\star$ becomes advantageous.

Theorem 1: The optimal rate $R^\star$ to minimize the maximum energy consumption $\hat{E}_b$ considering CC-HARQ and disregarding the effect of imperfect channel estimation is expressed as

$$R^\star = \min\left(\frac{W_0\left(\frac{\Omega}{\text{e}}\right) + 1}{\ln(2)}, R_{max}\right), \tag{21}$$

where R_{max} is the rate which guarantees the target outage at the maximum possible SNR, according to peak power limitations, e *is Euler's constant, W_0 is the upper branch of the main Lambert-W function,*

$$\Omega = \frac{\frac{L_{fb}}{L_{fw}}(P_{el} + Ad^{\alpha}\bar{\gamma}_p) - \Delta + P_{el}}{\Delta} \tag{22}$$

and

$$\Delta = \frac{Ad^{\alpha} m\Gamma(mz)}{\Gamma_{inc}^{-1}(mz, T_{out})}, \tag{23}$$

given that Γ_{inc}^{-1} is the inverse incomplete gamma function and $L_{fw} = L_H + L_D$.

Proof: Considering an URLLC scenario,

$$P_{\text{out},z} \leq T_{\text{out}} \ll 1, \tag{24}$$

and thus $\hat{E}_{\text{b}}$ approximated

$$\hat{E}_{\text{b}}(z) \approx \frac{2E_{\text{st}} + \frac{z}{WR}\left[L_{\text{fw}}(P_{\text{el}} + P_{\text{PA,d}}) + L_{\text{fb}}(P_{\text{el}} + P_{\text{PA,p}})\right]}{L_{\text{D}}}, \tag{25}$$

considering perfect channel state information at the receiver, for tractability. Using (11) and computing the derivative with respect to $\bar{\gamma}_{\text{d}}$,

$$\frac{\partial \hat{E}_{\text{b}}}{\partial \bar{\gamma}_{\text{d}}} = \frac{z}{WR} L_{\text{fw}} A d^{\alpha} > 0, \tag{26}$$

thus, increasing the transmit power to obtain a better (larger) SNR results in a larger $\hat{E}_{\text{b}}$. Therefore, we assume that the transmit power used is the one that guarantees the target outage, such that solving (2) for $\bar{\gamma}$ results in

$$\bar{\gamma}_{\text{d}} = \frac{m\left(2^R - 1\right)\Gamma(mz)}{\Gamma_{\text{inc}}^{-1}(mz, T_{\text{out}})}. \tag{27}$$

Next, we obtain $P_{\text{PA,d}}(z)$ using (11) and (27) and replace it onto (25). Finally, we solve $\partial \hat{E}_{\text{b}}/\partial R = 0$ yielding

$$\Delta 2^{R^{\star}}(\ln(2)R^{\star} - 1) = \Omega, \tag{28}$$

after simple algebraic manipulations. Lastly, we use the upper part of the main branch of the Lambert-W function to solve (28) for $R^{\star}$, arriving at

$$R^{\star} = \frac{W_0\left(\frac{\Omega}{\text{e}}\right) + 1}{\ln(2)}. \tag{29}$$

However, when accounting for peak power limitations, the SNR in (27) is limited at $\bar{\gamma}_{\max}$, such that setting $P_{\text{rf}} = P_{\text{rf,max}}$ in (11) and solving for $\bar{\gamma}$ yields

$$\bar{\gamma}_{\max} = \frac{P_{\text{rf,max}}}{Ad^{\alpha}}. \tag{30}$$

In other words, $R^{\star}$ is limited by $R_{\max}$, which is obtained using $\bar{\gamma}_{\max}$ in (2) and solving for R with $P_{\text{out},z} = T_{\text{out}}$, as

$$R_{\max} = \log_2\left(1 + \bar{\gamma}_{\max}\frac{\Gamma_{\text{inc}}^{-1}(mz, T_{\text{out}})}{m\Gamma(mz)}\right). \tag{31}$$

Lastly, combining (29) and (31) yields (21). □

Corollary 1: When the optimum rate $R^{\star}$ is smaller than the minimum required rate R_{min}, the optimal number of transmission attempts $z^{\star}$ which optimizes the maximum energy $\hat{E}_b$ in CC-HARQ can be well approximated by $\hat{z}^{\star}$, the solution of

$$2m\hat{z}\left(\ln(2)\frac{L_T}{W\lambda}\hat{z} - 1\right) + \ln(m\hat{z}) = 1 - \ln(T_{out}^2 2\pi) \tag{32}$$

with respect to $\hat{z}$, where $\hat{z}$ is a real relaxation of z.

Proof: The optimization problem is defined as

$$\underset{z\in\mathbb{N}^*}{\text{minimize}} \ \hat{E}_{\text{b}}(z) \tag{33a}$$

$$\text{subject to } P_{\text{out},z} \leq T_{\text{out}} \tag{33b}$$

$$P_{\text{rf}} \leq P_{\text{rf,max}} \tag{33c}$$

$$R \geq R_{\text{min}}. \tag{33d}$$

Next we impose $R^{\star} < R_{\text{min}}$ and consider $\delta_{\text{fb}} \ll \lambda$, thus

$$R = R_{\text{min}} \approx z\frac{L_{\text{T}}}{W\lambda}. \tag{34}$$

Then, obtaining $P_{\text{PA,d}}(z)$ as in the proof of Theorem 1 and replacing (34) into (25) we arrive at

$$\hat{E}_{\text{b}} \approx \frac{2E_{\text{st}} + \frac{\lambda}{L_{\text{T}}}\left[L_{\text{fw}}P_{\text{PA,d}}(z) + L_{\text{fb}}P_{\text{PA,p}} + L_{\text{T}}P_{\text{el}}\right]}{L_{\text{D}}}. \tag{35}$$

Then, considering that $z^{\star}$ can assume real values, we derive $\hat{E}_{\text{b}}$ with respect to $\hat{z}$ and equate to zero. Assuming high spectral efficiency, $2^R \gg 1$, which is true in many practical applications even for small R, it is possible to rewrite $\partial \hat{E}_{\text{b}}/\partial \hat{z} = 0$ as (32). □

The steps to prove the convexity of $\hat{E}_{\text{b}}$ are presented in Appendix A, while the steps used to derive (32) from $\partial \hat{E}_{\text{b}}/\partial \hat{z} = 0$ are outlined in Appendix B. Since it is not possible to have fractions of attempts, in practice, we must use the closest integer to $\hat{z}^{\star}$.

The solution presented in this section can be utilized to obtain insights into the behavior of the optimization problem. For instance, because in URLLC the values of T_{out} are always positive and much smaller than one, the left-hand side of (32) always yields a positive value. Therefore, fixing the values of T_{out}, m, L_{T} and W in (32) and choosing a smaller value for λ—considering a more stringent latency—causes $\hat{z}^{\star}$ to be smaller. In other words, having a more severe constraint in terms of latency imposes that less maximum attempts are optimal, which can be explained by having less time for more attempts. With the same rationale, but now fixing λ and decreasing the number of bits to communicate (L_{T}), yields a larger $\hat{z}^{\star}$. This is because the duration of each attempt is shorter, due to fewer bits being conveyed. Similarly, decreasing m results in a larger $\hat{z}^{\star}$, which is due to the diversity gains being more relevant in a worse channel condition. In addition, considering a communication channel with smaller bandwidth has a similar effect as considering a more strict latency in the solution of (32). Lastly, keeping the parameters on the right-hand side of (32) fixed and increasing the value of T_{out}, (*i.e.* considering a less reliable communication) yields a smaller $\hat{z}^{\star}$, which is explained by the fact that the gains in diversity are less important for a more relaxed reliability.

TABLE II
SIMULATION PARAMETERS

Parameter	Value
Target Outage (T_{out})	10^{-5} [7]
Typical Link Latency (λ)	6.48 ms†
Time to decode feedback (δ_{fb})	0.0213 ms [32]
Maximum Transmit Power ($P_{\text{rf,max}}$)	-16 dB [33]
Bandwidth (W)	180 KHz [33]
Distance (d)	100 m
Spectral Noise Power Density (N_0)	-204 dB
Link Margin (M_{l})	15 dB
Coding Margin (M_{c})	3 dB [32]
Path Loss Exponent (α)	2.5 [34]
Attenuation at reference (A_0)	38.5 dB [34]‡
Header Length (L_{H})	16 bits [13]
Feedback Length (L_{fb})	17 bits††
Payload Length (L_{D})	1216 bits [7]
Radio Startup Energy (ε_{st})	0.125 nJ [27]
Average PA Efficiency (η)	50% [35]
Electronic Power (P_{el})	-15.69 dB [13]

† $\lambda' = 8$ms [7] and $\delta_{\text{fw}} = 1.5$ms [32].
‡ We consider a reference distance of 1 m, unit gain on the antennas and a carrier frequency centered at 2 GHz.
†† We consider 1 bit feedback, such that $L_{\text{fb}} = L_{\text{H}} + 1$.

IV. NUMERICAL RESULTS

In this section, we evaluate the proposed approach using parameters from the METIS test case #5 [7]: smart grid communications. The parameters are summarized in Table II. This example fits within the block-fading model because there is almost no mobility and consecutive transmissions are assumed to be performed in uncorrelated frequency channels. In addition, the Nakagami-m channel model describes the scenario well since it correctly depicts the variability of situations encountered in smart grids, with varied line-of-sight (LOS) conditions [33]. Additionally, to perform slow frequency hopping, we assume that the transmitter uses sub-carriers with independent and identically distributed channel gains and bandwidth W for each transmission attempt, such that the average consumed bandwidth is $\bar{\tau}W$ Hz per message.

First, we show in Fig. 3 and Fig. 4 how $\bar{E}_{\text{b}}$ and $\hat{E}_{\text{b}}$ vary with respect to z for various strategies regarding SNR and using the rate according to[3]

$$R = \max\left(R^{\star}, R_{\min}\right). \tag{36}$$

Moreover, the SNR strategies are characterized by: 1) A benchmark obtained numerically, by optimizing the average SNR to determine its optimal value $\bar{\gamma}_{\text{d}}^{\star}$, 2) using the average SNR that guarantees the target outage $\bar{\gamma}_{\text{d,min}}$ and 3) using the maximum average SNR according to the maximum transmit power $\bar{\gamma}_{\text{d,max}}$. Furthermore, we also show $\hat{E}_{\text{b}}$ using $\bar{\gamma}_{\text{d}}^{\star}$ and $R = R_{\min}$ to illustrate how using a higher rate impacts the value of $\hat{E}_{\text{b}}$. Note that as the curve of $\bar{E}_{\text{b}}$ is a benchmark, it has been generated using a numerically obtained optimal number of pilots for channel estimation,

[3]If $R_{\max} < R_{\min}$ the link cannot be closed for λ and T_{out}.

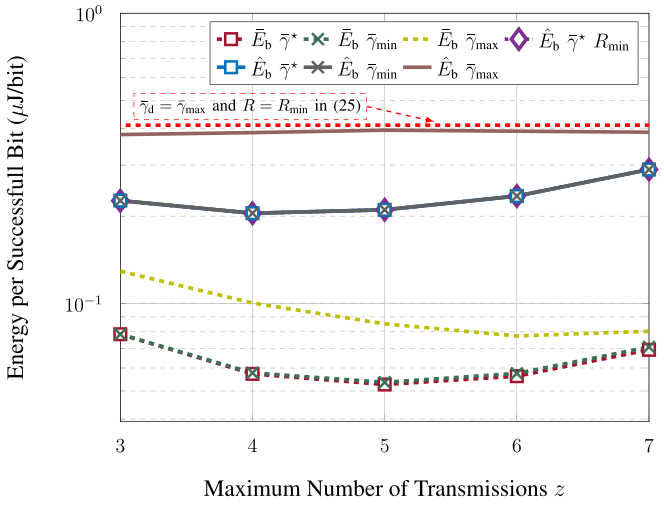


Fig. 3. $\bar{E}_{\text{b}}$ and $\hat{E}_{\text{b}}$ versus z for various strategies and $m = 1$.

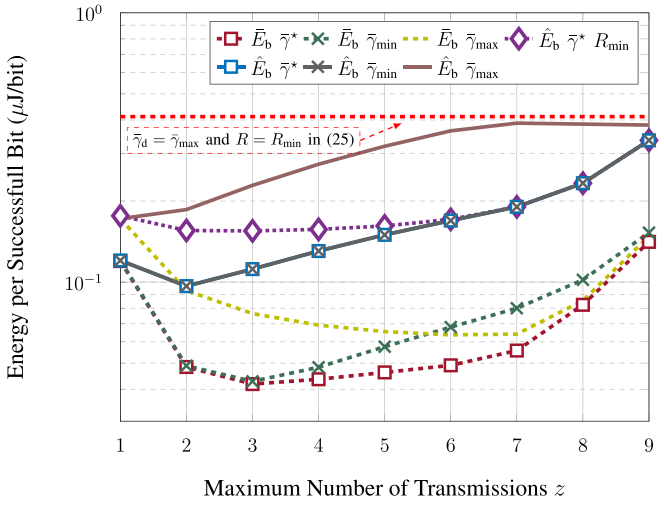


Fig. 4. $\bar{E}_{\text{b}}$ and $\hat{E}_{\text{b}}$ versus z for various strategies and $m = 3$.

while the other curves represent real implementations and therefore use a fixed $\rho = 6$.

In Fig. 3, we show the performance for the case with a Nakagami parameter of $m = 1$. We can notice that because there is no LOS, the diversity gains are not as impactful and increasing the average SNR has little impact on $\bar{E}_{\text{b}}$, such that the curve with $\bar{\gamma}_{\text{d,min}}$ performs very close to the benchmark, more so for the smaller values of z. This means that using the results of Corollary 1, which assumes that using the lowest possible SNR will yield a good performance in such scenarios. Additionally, we observe that optimizing the rate has little to no effect on $\hat{E}_{\text{b}}$, as due to the stringent link budget characteristics, $R^{\star}$ will be close to $R_{\min}$. Moreover, the curve for $\hat{E}_{\text{b}}$ with $\bar{\gamma}_{\text{d}}^{\star}$ matches exactly with the one with $\bar{\gamma}_{\text{d,min}}$ (both for Figs. 3 and 4), as predicted analytically in (26).

On the other hand, in Fig. 4, $m = 3$ is shown, and thus the link is less stringent. Here, the diversity gains of using higher z benefit more from using $\bar{\gamma}_{\text{d}}^{\star}$ in terms of $\hat{E}_{\text{b}}$. However, when z is relatively small (≤ 4), using $\bar{\gamma}_{\text{d,min}}$ still performs very close to the benchmark. Also note that, because of the more

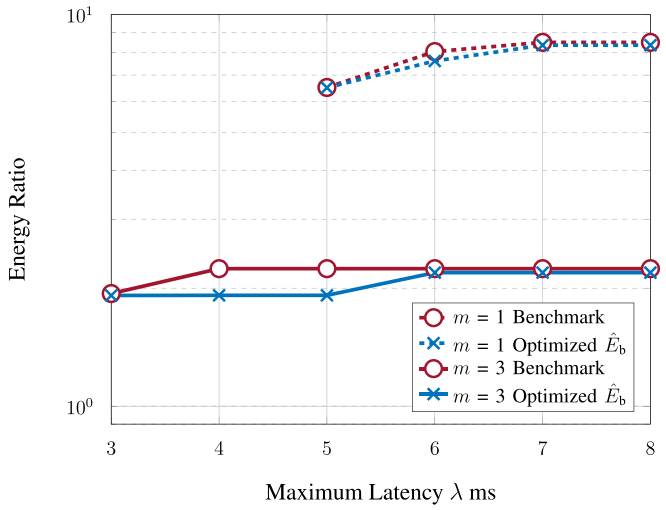


Fig. 5. Performance for different λ with $T_{out} = 10^{-5}$ and comparison between proposed scheme and the benchmark.

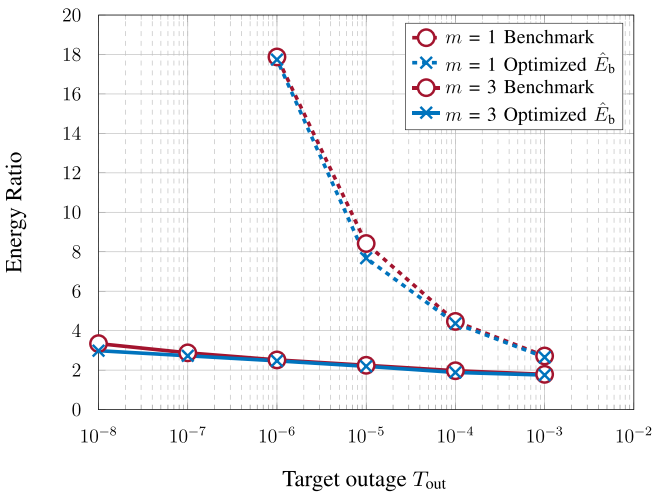


Fig. 6. Performance for different T_{out} with $\lambda = 6.48$ ms and comparison between proposed scheme and the benchmark.

relaxed link budget, there is more freedom to increase the rate and using the results from Theorem 1 results in performance gains, in particular where z is smaller. The red dotted line in both Fig. 3 and Fig. 4 is the result of using $\bar{\gamma}_d = \bar{\gamma}_{max}$ and $R = R_{min}$ in (25). It represents a ceiling in the maximum consumption when $R^\star < R_{min}$ and we operate at the maximum transmit power.

Further, in order to evaluate the performance of the proposed solution, we compare it with the case of a single transmission. To make the comparison fair, we allow the direct transmission to use $\bar{\tau}$ channels to send copies of the message and perform MRC at the receiver, exploiting frequency diversity. However, because it is not possible to use fractions of a sub-carrier and $\bar{\tau}$ is not always an integer, we use the next closest integer for consistency. All channels are estimated separately and this cost is taken into account when computing the energy used to send the message with frequency diversity $\bar{E}_f$, presented in Appendix C. Moreover, the optimal rate considering frequency diversity is similar to the one obtained in (21), and the value of R is determined according to (36).

Fig. 5 shows the ratio between the energy consumption considering frequency diversity and that of our proposed solution. The energy ratio considering the benchmark (obtained numerically) is determined by calculating $\bar{E}_f(\lceil\bar{\tau}\rceil)/\bar{E}_b(z^\star)$, while the ratio which considers the results in Theorem 1 and Corollary 1 is calculated as $\bar{E}_f(\lceil\bar{\tau}\rceil)/\bar{E}_b(\hat{z}^\star)$. As we can observe, for the target latency of 6.48 ms and $m = 3$, the proposed solution outperforms the frequency diversity by a factor of more than 2. Note that here, since $R^\star > R_{min}$, considering a more stringent latency has little effect in the performance as the higher rate was already guaranteeing a more stringent latency. On the other hand, when $m = 1$, the link is more stringent and $R^\star < R_{min}$, such that changes in λ incur in changes to the performance. Also regarding $m = 1$, we can see that for the target latency of 6.48 ms, the proposed scheme is about 8 times better. Because we use the next closest integer to $\bar{\tau}$ when choosing the number of channels to use for frequency diversity, our approach uses less bandwidth on average, in other words, our approach constitutes a better way of using the spectrum. Despite using less bandwidth on average, we can often achieve higher orders of diversity, resulting in energy savings. For example, for the case where $z^\star = 2$, the diversity orders are the same but because $\bar{\tau}$ is close to one, the frequency diversity approach uses almost twice the bandwidth. This result shows the strength of HARQ, achieving high diversity orders at low cost, which is possible because on average it requires few attempts and thus uses a small number of channels. Conversely, when considering frequency diversity, the cost of obtaining a large diversity is to use all channels in every transmission, which is less energy and spectral efficient.

In Fig. 6, we show the performance when considering different levels of reliability, considering the same ratios as before with $\lambda = 6.48$ ms. As we can observe, when we have good channel conditions ($m = 3$), increasing the reliability has less impact in the frequency diversity strategy compared to the case where $m = 1$, because the need for diversity is not as pressing and thus only a few channels are used each time. However, when there is no LOS ($m = 1$), increasing the reliability comes at a high cost for a direct transmission because of the need for diversity to reach the stringent reliability requirement. Since the CC-HARQ strategy does not need to use all of those resources in every attempt, it yields a better performance when considering more reliable specifications. For instance, if we consider no LOS and $T_{out} = 10^{-6}$, our scheme outperforms the direct transmission with frequency diversity by more than 18 times. Furthermore, we can observe from this numerical example, both in Figs. 5 and 6, that optimizing $\hat{E}_b$ has very similar performance compared to the benchmark solution which uses numerically obtained optimal number of pilots, SNR and z.

V. Conclusion

We analyzed the use of truncated CC-HARQ in order to meet the stringent reliability and latency requirements of URLLC whilst increasing energy efficiency in a block-fading Nakagami-m channel. We demonstrated that the energy

consumption depends on the choice of allowed transmission attempts and can be minimized by optimally tuning this parameter. Further, we proposed solving an alternative problem which allows us to arrive at a fixed point equation to determine the optimum number of maximum transmission attempts, which can be used to obtain several insights.

We evaluated the results in a smart grid communication scenario and showed interesting savings in energy when compared to a frequency diversity strategy while using less bandwidth on average, in particular the scheme is better for more strict reliability scenarios. It is clear from the results that truncated CC-HARQ is one viable strategy to reduce energy consumption while meeting the requirements of URLLC when adequately designed, even when accounting for higher data rates required to meet stringent latency requirements.

APPENDIX A
PROOF OF $\hat{E}_{\mathrm{B}}$ CONVEXITY

It is explicit that proving the convexity of $\bar{\gamma}_{\mathrm{d}}$ with respect to $\hat{z}$ is sufficient proof that $\hat{E}_{\mathrm{b}}$ is convex with respect to $\hat{z}$.

Considering high SNR, we can solve (3) for $\bar{\gamma}$ and using the Stirling approximation of factorials to approximate the complete Gamma function we write

$$\bar{\gamma}_{\mathrm{d}} \approx \frac{m 2^{\hat{z}\frac{L_{\mathrm{T}}}{W\lambda}}}{\left(T_{\mathrm{out}}\sqrt{2\pi m\hat{z}}\right)^{\frac{1}{m\hat{z}}}\frac{m\hat{z}}{\mathrm{e}}}, \tag{37}$$

considering high spectral efficiency such that $2^{\hat{z}\frac{L_{\mathrm{T}}}{W\lambda}} \gg 1$.

Equation (37) can be written in the form of $f(\hat{z})g(\hat{z})$, with

$$f(\hat{z}) = m 2^{\hat{z}\frac{L_{\mathrm{T}}}{W\lambda}} \tag{38}$$

and

$$g(\hat{z}) = \frac{1}{\left(T_{\mathrm{out}}\sqrt{2\pi m\hat{z}}\right)^{\frac{1}{m\hat{z}}}\frac{m\hat{z}}{e}}. \tag{39}$$

In turn, $g(\hat{z})$ can be written as $1/(g_1(\hat{z})g_2(\hat{z}))$, with

$$g_1(\hat{z}) = \left(T_{\mathrm{out}}\sqrt{2\pi m\hat{z}}\right)^{\frac{1}{mz}} \tag{40}$$

and

$$g_2(\hat{z}) = m\hat{z}/\mathrm{e}. \tag{41}$$

First we show that $g_1(\hat{z})$ is log-concave, such that

$$\frac{\partial^2 \ln(g_1(\hat{z}))}{\partial \hat{z}^2} \leq 0. \tag{42}$$

Since

$$\frac{\partial^2 \ln(g_1(\hat{z}))}{\partial \hat{z}^2} = \frac{4\ln\left(T_{\mathrm{out}}\sqrt{2\pi m\hat{z}}\right) - 3}{2\left(m\hat{z}\right)^3}, \tag{43}$$

and $m\hat{z} \neq 0$, (42) becomes

$$0 < \hat{z} \leq \frac{\mathrm{e}^{3/2}}{2\pi T_{\mathrm{out}}^2}, \tag{44}$$

after some algebraic manipulations. Because T_{out} is always positive and very small and $\hat{z} \geq 1$, (44) holds in any practical scenario. Thus, $g_1(\hat{z})$ is log-concave.

Moreover, because $g_2(\hat{z})$ is both log-convex and log-concave [36], the product of $g_1(\hat{z})$ and $g_2(\hat{z})$ is also log-concave [36] and therefore its inverse $g(\hat{z})$ is log-convex [36].

Finally, since $f(\hat{z})$ is an exponential function, it is log-convex [36]. The product of two log-convex functions is also log-convex [36], thus $\bar{\gamma} = f(\hat{z})g(\hat{z})$ is also log-convex and therefore convex, concluding the proof. □

APPENDIX B
DERIVATION OF (32)

Calculating $\frac{\partial \hat{E}_{\mathrm{b}}}{\partial \hat{z}} = 0$ yields

$$\frac{\partial \frac{2E_{\mathrm{st}} + \frac{\lambda}{L_{\mathrm{T}}}[L_{\mathrm{fw}} A d^{\alpha} \bar{\gamma}_{\mathrm{d}} + L_{\mathrm{fb}} P_{\mathrm{PA,p}} + L_{\mathrm{T}} P_{\mathrm{el}}]}{L_{\mathrm{D}}}}{\partial \hat{z}} = 0$$

$$\partial \bar{\gamma}_{\mathrm{d}} / \partial \hat{z} = 0. \tag{45}$$

Computing the derivative of the approximated average SNR expressed in (37) results in

$$\Xi\left(2\ln(2)\frac{L_{\mathrm{T}}}{W\lambda} m\hat{z}^2 - 2m\hat{z} + \ln(\hat{z}) + \ln\left(\frac{2\pi T_{\mathrm{out}}^2 m}{\mathrm{e}}\right)\right), \tag{46}$$

where $\Xi = \frac{\mathrm{e} 2^{\hat{z}\frac{L_{\mathrm{T}}}{W\lambda} - \frac{1}{2}\frac{1}{m\hat{z}} - 1}}{m\left(T_{\mathrm{out}}\sqrt{\pi m\hat{z}}\right)^{\frac{1}{m\hat{z}}}}$.

Since $\Xi \neq 0$, replacing (46) in (45) and diving both sides by Ξ results in (32). □

APPENDIX C
ENERGY CONSUMPTION OF FREQUENCY DIVERSITY

Consider that we have $\lceil \tau \rceil$ channels of bandwidth W to send copies of the message, which are combined by the receiver using MRC. In this case, to achieve the target error probability, the effective average SNR of each message is given by

$$\bar{\gamma}_{\mathrm{eff}}^{\mathrm{freq}} = \frac{m\left(2^R - 1\right)}{\left(T_{\mathrm{out}}\Gamma(\lceil \tau \rceil m + 1)\right)^{1/\lceil \tau \rceil m}}. \tag{47}$$

The maximum average SNR for each channel $\bar{\gamma}_{\mathrm{max}}^{\mathrm{freq}}$ is reduced accounting for the radiated power in all channels as $\bar{\gamma}_{\mathrm{max}}^{\mathrm{freq}} = \bar{\gamma}_{\mathrm{max}}/\lceil \tau \rceil$, to account for the peak power constraint.

Making $\bar{\gamma}_{\mathrm{p}} = \bar{\gamma}_{\mathrm{max}}^{\mathrm{freq}}$ and calculating the effective average SNR using (47), we determine the average SNR of the data. Next, following similar steps as for the CC-HARQ case, the energy per successful bit for the case of frequency diversity is

$$\bar{E}_{\mathrm{f}}(\lceil \tau \rceil) = \frac{2E_{\mathrm{st}} + \frac{\lceil \tau \rceil}{W}\left[n_{\mathrm{fw}}(P_{\mathrm{el}} + P_{\mathrm{PA,d}}) + \rho(P_{\mathrm{el}} + P_{\mathrm{PA,p}})\right]}{L_{\mathrm{D}}(1 - P_{\mathrm{out},\lceil \tau \rceil})}. \tag{48}$$

The consumption due to larger bandwidth is accounted for by multiplying the power consumption of the PA by $\lceil \tau \rceil$.

REFERENCES

[1] H. Tullberg *et al.*, "The METIS 5G system concept: Meeting the 5G requirements," *IEEE Commun. Mag.*, vol. 54, no. 12, pp. 132–139, Dec. 2016.

[2] J. G. Andrews *et al.*, "What will 5G be?" *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1065–1082, Jun. 2014.

[3] B. Holfeld *et al.*, "Wireless communication for factory automation: An opportunity for LTE and 5G systems," *IEEE Commun. Mag.*, vol. 54, no. 6, pp. 36–43, Jun. 2016.

[4] M. Gharba *et al.*, "5G enabled cooperative collision avoidance: System design and field test," in *Proc. IEEE Int. Symp. World Wireless, Mobile Multimedia Netw. (WoWMoM)*, Jun. 2017, pp. 1–6.

[5] P. Fernandes and U. Nunes, "Platooning with IVC-enabled autonomous vehicles: Strategies to mitigate communication delays, improve safety and traffic flow," *IEEE Trans. Intell. Transp. Syst.*, vol. 13, no. 1, pp. 91–106, Mar. 2012.

[6] A. Aijaz, M. Dohler, A. H. Aghvami, V. Friderikos, and M. Frodigh, "Realizing the Tactile Internet: Haptic communications over next generation 5G cellular networks," *IEEE Wireless Commun.*, vol. 24, no. 2, pp. 82–89, Apr. 2017.

[7] P. Popovski *et al.*, *Scenarios, Requirements and KPIs for 5G Mobile and Wireless System*, ICT-317669-METIS/D1.1, ICT-317669 METIS project, 2013.

[8] G. Durisi, T. Koch, and P. Popovski, "Toward massive, ultrareliable, and low-latency wireless communication with short packets," *Proc. IEEE*, vol. 104, no. 9, pp. 1711–1726, Aug. 2016.

[9] E. Dosti, U. L. Wijewardhana, H. Alves, and M. Latva-Aho, "Ultra reliable communication via optimum power allocation for type-I ARQ in finite block-length," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2017, pp. 1–6.

[10] E. Dosti, M. Shehab, H. Alves, and M. Latva-Aho, "Ultra reliable communication via CC-HARQ in finite block-length," in *Proc. Eur. Conf. Netw. Commun. (EuCNC)*, Jun. 2017, pp. 1–5.

[11] C. Sun, C. She, and C. Yang, "Energy-efficient resource allocation for ultra-reliable and low-latency communications," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2017, pp. 1–6.

[12] A. Mukherjee, "Energy efficiency and delay in 5G ultra-reliable low-latency communications system architectures," *IEEE Netw.*, vol. 32, no. 2, pp. 55–61, Mar./Apr. 2018.

[13] F. Rosas *et al.*, "Optimizing the code rate of energy-constrained wireless communications with HARQ," *IEEE Trans. Wireless Commun.*, vol. 15, no. 1, pp. 191–205, Jan. 2016.

[14] M. Jabi, M. Benjillali, L. Szczecinski, and F. Labeau, "Energy efficiency of adaptive HARQ," *IEEE Trans. Commun.*, vol. 64, no. 2, pp. 818–831, Feb. 2016.

[15] M. Maaz, P. Mary, and M. Hélard, "Energy minimization in HARQ-I relay-assisted networks with delay-limited users," *IEEE Trans. Veh. Technol.*, vol. 66, no. 8, pp. 6887–6898, Aug. 2017.

[16] C.-P. Li, J. Jiang, W. Chen, T. Ji, and J. Smee, "5G ultra-reliable and low-latency systems design," in *Proc. Eur. Conf. Netw. Commun. (EuCNC)*, Jun. 2017, pp. 1–5.

[17] J. P. B. Nadas, M. A. Imran, G. Brante, and R. D. Souza, "Optimizing the energy efficiency of short term ultra reliable communications in vehicular networks," in *Proc. IEEE Int. Symp. Modeling Optim. Mobile, Ad Hoc, Wireless Netw. (WiOpt)*, May 2017, pp. 1–6.

[18] A. Anand and G. de Veciana, "Resource allocation and HARQ optimization for URLLC traffic in 5G wireless networks," *CoRR*, vol. abs/1804.09201, pp. 1–10, May 2018. [Online]. Available: http://arxiv.org/abs/1804.09201

[19] P. Mary, J.-M. Gorce, A. Unsal, and H. V. Poor, "Finite blocklength information theory: What is the practical impact on wireless communications?" in *Proc. IEEE Globecom Workshops*, Dec. 2016, pp. 1–6.

[20] V. A. Aalo, "Performance of maximal-ratio diversity systems in a correlated Nakagami-fading environment," *IEEE Trans. Commun.*, vol. 43, no. 8, pp. 2360–2369, Aug. 1995.

[21] H. AlQuwaiee and M.-S. Alouini, "New exact and asymptotic results of dual-branch MRC over correlated Nakagami-m fading channels," in *Proc. IEEE Veh. Technol. Conf. (VTC Spring)*, May 2015, pp. 1–5.

[22] S. Kim and H. Yu, "Energy-efficient HARQ-IR for massive MIMO systems," *IEEE Trans. Commun.*, vol. 66, no. 9, pp. 3892–3901, Sep. 2018.

[23] X. Leturc, P. Ciblat, and C. J. Le Martret, "Energy-efficient resource allocation for HARQ with statistical CSI," *IEEE Trans. Veh. Technol.*, vol. 67, no. 12, pp. 11936–11949, Dec. 2018.

[24] M. Hsieh and C. Wei, "Channel estimation for OFDM systems based on comb-type pilot arrangement in frequency selective fading channels," *IEEE Trans. Consum. Electron.*, vol. 44, no. 1, pp. 217–225, Aug. 1998.

[25] B. Hassibi and B. M. Hochwald, "How much training is needed in multiple-antenna wireless links?" *IEEE Trans. Inf. Theory*, vol. 49, no. 4, pp. 951–963, Apr. 2003.

[26] M. C. Gursoy, "On the capacity and energy efficiency of training-based transmissions over fading channels," *IEEE Trans. Inf. Theory*, vol. 55, no. 10, pp. 4543–4567, Oct. 2009.

[27] S. Cui, A. J. Goldsmith, and A. Bahai, "Energy-constrained modulation optimization," *IEEE Trans. Wireless Commun.*, vol. 4, no. 5, pp. 2349–2360, Sep. 2005.

[28] A. Keshavarzian, H. Lee, and L. Venkatraman, "Wakeup scheduling in wireless sensor networks," in *Proc. ACM Int. Symp. Mobile Ad Hoc Netw. Comput.*, 2006, pp. 322–333.

[29] J. Choi, "Energy-delay tradeoff comparison of transmission schemes with limited CSI feedback," *IEEE Trans. Wireless Commun.*, vol. 12, no. 4, pp. 1762–1773, Apr. 2013.

[30] G. Pocovi, B. Soret, K. I. Pedersen, and P. Mogensen, "MAC layer enhancements for ultra-reliable low-latency communications in cellular networks," in *Proc. IEEE Int. Conf. Commun. Workshops (ICC Workshops)*, May 2017, pp. 1005–1010.

[31] M. Sybis, K. Wesolowski, K. Jayasinghe, V. Venkatasubramanian, and V. Vukadinovic, "Channel coding for ultra-reliable low-latency communication in 5G systems," in *Proc. IEEE 84th Veh. Technol. Conf. (VTC-Fall)*, Sep. 2016, pp. 1–5.

[32] M. Shirvanimoghaddam *et al.* (2018). "Short block-length codes for ultra-reliable low-latency communications." [Online]. Available: https://arxiv.org/abs/1802.09166

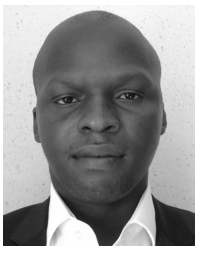
[33] Y. Cao, T. Jiang, M. He, and J. Zhang, "Device-to-device communications for energy management: A smart grid case," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 1, pp. 190–201, Jan. 2016.

[34] T. S. Rappaport, *Wireless Communications: Principles and Practice*, 2nd ed. Upper Saddle River, NJ, USA: Prentice-Hall, 2002.

[35] A. He *et al.*, "System power consumption minimization for multichannel communications using cognitive radio," in *Proc. IEEE Int. Conf. Microw., Commun., Antennas Electron. Syst.*, Nov. 2009, pp. 1–5.

[36] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

João Pedro Battistella Nadas (S'17) was born in Curitiba, Brazil. He received the B.Sc. degree in electronics and telecommunications engineering and the master's degree in telecommunications engineering from the Federal Technological University of Paraná, Brazil, in 2014 and 2016, respectively. During both his degree's, he has simultaneously worked in the industry in various positions, ranging from placements to project management. He is currently pursuing the Ph.D. degree with the University of Glasgow. His research interests include ultra-reliable and low latency communications, HARQ, machine learning applied to communications, cyber-physical systems, and wireless sensor networks.



Oluwakayode Onireti (S'11–M'13) received the B.Eng. (Hons.) degree in electrical engineering from the University of Ilorin, Ilorin, Nigeria, in 2005, and the M.Sc. (Hons.) degree in mobile and satellite communications and the Ph.D. degree in electronics engineering from the University of Surrey, Guildford, U.K., in 2009 and 2012, respectively. From 2013 to 2016, he was a Research Fellow with the Institute for Communication Systems, University of Surrey. He has been actively involved in projects, such as ROCKET, EARTH, Greencom, QSON, and Energy proportional EnodeB for LTE-Advanced and Beyond. He is currently a Research Associate with the School of Engineering, University of Glasgow, Glasgow, U.K. He is currently involved in the DARE project, and an ESPRC funded project on distributed autonomous and resilient emergency management systems. His main research interests include self-organizing cellular networks, energy efficiency, multiple-input multiple-output, and cooperative communications.

Richard Demo Souza (S'01–M'04–SM'12) was born in Florianópolis, Brazil. He received the B.Sc. and D.Sc. degrees in electrical engineering from the Federal University of Santa Catarina (UFSC), Brazil, in 1999 and 2003, respectively. In 2003, he was a Visiting Researcher with the Department of Electrical and Computer Engineering, University of Delaware, USA. From 2004 to 2016, he was with the Federal University of Technology–Paraná, Brazil. Since 2017, he has been with UFSC, where he is currently an Associate Professor. His research interests include wireless communications and signal processing. He is a Senior Member of the Brazilian Telecommunications Society (SBrT). He was a co-recipient of the 2014 IEEE/IFIP Wireless Days Conference Best Paper Award and the 2016 Research Award from the Cuban Academy of Sciences. He has served as an Editor-in-Chief of the SBrT *Journal of Communication and Information Systems*; and as an Associate Editor for the IEEE COMMUNICATIONS LETTERS, the *EURASIP Journal on Wireless Communications and Networking*, and the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY. He was the Supervisor of the Best Ph.D. Thesis Award in electrical engineering in Brazil in 2014.

Hirley Alves received the B.Sc. and M.Sc. degrees in electrical engineering from the Federal University of Technology–Paraná (UTFPR), Brazil, in 2010 and 2011, respectively, and the dual D.Sc. degrees from the University of Oulu and UTFPR in 2015. He was an Adjunct Professor in machine-type wireless communications with the Centre for Wireless Communications (CWC), University of Oulu, Oulu, Finland, in 2017, where he joined CWC as an Assistant Professor in machine-type wireless communications for future networks in 2019. His research interests are wireless and cooperative communications, wireless full-duplex communications, PHY-security, and ultra-reliable communications mechanisms for future machine type wireless networks. He has acted as an organizer, a chair, and serves as a TPC and a tutorial lecturer to several renowned international conferences. He was a co-recipient of the 2017 IEEE International Symposium on Wireless Communications and Systems (ISWCS) Best Student Paper Award and of the 2016 Research Award from the Cuban Academy of Sciences. He is the General Chair of ISWCS 2019. He is actively working on massive connectivity and ultra-reliable low latency communications.

Glauber Brante (S'10–M'14–SM'18) was born in Arapongas, Brazil, in 1983. He received the D.Sc. degree in electrical engineering from the Federal University of Technology–Paraná (UTFPR), Curitiba, Brazil, in 2013. In 2012, he was a Visiting Researcher with the Institute of Information and Communication Technologies, Electronics, and Applied Mathematics, Catholic University of Louvain, Belgium. He is currently an Assistant Professor with UTFPR. His research interests include cooperative communications, HARQ, energy efficiency, and physical layer security. He received the Best Ph.D. Thesis Award in electrical engineering in Brazil in 2014. He was a co-recipient of the 2016 Research Award from the Cuban Academy of Sciences. Since 2018, he has been serving as an Associate Editor for the IEEE COMMUNICATIONS LETTERS and as a Co-Editor-in-Chief for the *Journal of Communication and Information Systems*.

Muhammad Ali Imran (M'03–SM'12) is currently a Professor of Wireless Communication Systems with research interests in self-organized networks, wireless networked control systems, and the wireless sensor systems. He also heads the Communications, Sensing and Imaging CSI Research Group, University of Glasgow. He is also an Affiliate Professor with The University of Oklahoma, USA, and a Visiting Professor with the 5G Innovation Centre, University of Surrey, U.K. He has over 18 years of combined academic and industry experience with several leading roles in multi-million pounds funded projects. He holds 15 patents; has authored or co-authored over 400 journal and conference publications; was editor of two books and authored more than 15 book chapters; and has successfully supervised over 40 postgraduate students at Ph.D. level. He has been a consultant to international projects and local companies in the area of self-organized networks. He is a Fellow IET and a Senior Fellow of HEA.