Energy-Latency Tradeoff in Ultra-Reliable Low-Latency Communication with Short Packets

Apostolos Avranas*, Marios Kountouris*, and Philippe Ciblat†
*Mathematical and Algorithmic Sciences Lab, Paris Research Center, Huawei France
†Télécom ParisTech, Université Paris-Saclay, F-75013 Paris, France
Emails: {apostolos.avranas,marios.kountouris}@huawei.com, philippe.ciblat@telecom-paristech.fr

***Abstract*—We consider an ultra-reliable low-latency communication (URLLC) system with short packets employing hybrid automatic repeat request (HARQ). Depending on the delay of HARQ feedback and retransmissions, the latency constraint can be either violated or fulfilled at the expense of power consumption. We focus on the energy-latency tradeoff and examine whether it is better to do one-shot transmission or use HARQ. We analyze the energy consumption for incremental redundancy (IR) HARQ and compare it with the no HARQ case. The analysis relies on closed-form expressions for the outage probability of IR-HARQ with variables both the blocklength and the power. Our results show that for a wide range of blocklength, when the feedback delay is more than half the latency constraint, it is beneficial in terms of energy to use one-shot transmission (i.e. no HARQ).**

I. Introduction

Future evolution of mobile communication systems (5G new radio) is giving rise to new uses of wireless communications in areas such as augmented and virtual reality (AR/VR), industrial control, automated transportation and robotics. 5G is envisaged to support mission-critical Internet-of-Things (IoT) applications and ultra-reliable low-latency communication (URLLC) scenarios with strict requirements in terms of latency (ranging from 1 ms and below to few milliseconds) and reliability (higher than 99.999%). This entails a fundamental paradigm shift from throughput-oriented system design towards an holistic design for guaranteed and reliable end-to-end latency.

Guaranteeing URLLC requirements is a challenging task even in simple settings as URLLC drives the system to new, unexplored operating regimes. The performance is constrained by challenging fundamental tradeoffs between delay, throughput, energy and error probability. The predominance of short messages, together with the need to reduce the packet duration, implies that small blocklength channel codes are also used. This results in a rate penalty term and transmission rates with non-zero error probability, revisiting key insights obtained via asymptotic information theoretic results. Recent progress has quantified the effect of finite blocklength, providing tight bounds and accurate normal approximation for the maximum coding rate to sustain the desired packet error probability (PEP) for a given packet size [1].

In order to compensate for the reliability loss introduced by short packets, highly reliable communications mechanisms creating diversity have to be carried out, such as hybrid automatic repeat request (HARQ). However the benefits of time diversity could be rather limited under stringent latency constraints. Moreover, the benefit of feedback-based retransmissions (even with error-free but delayed feedback) is questionable since each transmit packet is much smaller due to energy and latency constraints, thus more prone to errors. Additionally, energy considerations, in particular power consumption, are of cardinal importance in the design of URLLC systems, and there is an inherent power-latency tradeoff. A transmission can be successful (or its PEP may be kept unaltered) with minimum delay at the expense of additional or high power usage. In the short-packet regime, this interplay is more pronounced as latency is minimized when all packets are jointly encoded, whereas power is minimized when each packet is encoded separately. Note that power is the energy consumed over symbol period.

In this paper, we analyze the fundamental tradeoff between latency (in terms of feedback/retransmission delay) and average consumed energy in URLLC with incremental redundancy (IR)-HARQ. Considering that short packets have to be decoded with a certain PEP and latency, we give an answer whether it is beneficial to do one-shot transmission or split the packet into sub-codewords and use IR-HARQ. Prior work has considered the problem of throughput maximization by either adjusting the blocklength of each IR-HARQ round using the same power [2] or via rate refinement over retransmissions of equal-sized and constant energy packets [3]. Equal-sized and constant energy packets and rate maximization under a reliability constraint is considered in [4]. In [5], sphere packing is used for optimizing the blocklength of every transmission with equal power. In contrast to prior work, here we study the problem of average energy consumed minimization to guarantee both PEP and latency (URLLC) constraints by properly adapting both the blocklength and the power of each transmission. A key result of our paper is that one-shot transmission (no HARQ) should be used when the feedback delay is more than half the latency constraint for low and moderate blocklength.

II. System Model

We consider a transmitter having to send B information bits within a certain predefined latency, which can be expressed with a certain predefined maximum number of channel uses, denoted by N. If no ARQ/HARQ mechanism is utilized, the

packet of B bits is sent at once (one-shot transmission) and its maximum length is N. When HARQ is employed, we consider hereafter IR-HARQ with two transmissions, i.e. one retransmission without loss of generality since our objective is to highlight the potential gain (or loss) from packet splitting. Our results can be easily extended to multiple retransmissions without altering the conclusions. Let n_1 and n_2 be the number of channel uses (equivalently the symbols) for the first and second transmission, respectively. In order to fulfill the URLLC latency constraint, we have $n_1 + n_2 \leq N - D$ where D is a penalty due to delay for the receiver to process/decode the first packet and send acknowledgment (ACK/NACK). The IR-HARQ mechanism operates as follows: B information bits are encoded into a parent codeword of length n_1+n_2 symbols. Then, the parent codeword is split into two fragments of codeword (sub-codewords), the first with length n_1 and the second with length n_2. The receiver requests transmission of the second sub-codeword only if it is unable to correctly decode the message using the first fragment of the codeword. In that case, the receiver attempts to jointly decode the entire codeword, i.e. the concatenation of the first and second fragment. We assume that the receiver knows perfectly whether or not the message is correctly decoded (through CRC) and ACK/NACK is received error free but with delay.

The channel is considered to be quasi-static along with the whole HARQ mechanism, i.e. the channel coefficients remain constant during the packet retransmissions. This is a relevant model for URLLC applications and short-length packet transmissions. For a system operating at carrier frequency $f_c = 2.5$ GHz and coherence time $T_c = 1$ ms (latency constraint), the receiver speed is $v = cB_d/f_c \approx 180$ km/h, where $B_d = 0.423/T_c$ [6, (8.20)], is the Doppler spread c is the speed of light; this is a relatively high speed for most mission-critical IoT or tactile Internet applications. Therefore, our communication scenario consists of a point-to-point link with additive white Gaussian noise (AWGN). Specifically, in m-th round ($m \in \{1, 2\}$), the fragment (sub-codeword) $c_m \in \mathbb{C}^{n_m}$ is received with power $P_m = \frac{||c_m||^2}{n_m}$ and distorted by an additive white circularly-symmetric complex Gaussian random process with zero mean and unit variance. As the channel is static along with the transmission, the channel gains are constant and also the noise variance is assumed equal to one without loss of generality.

III. Problem Statement and Preliminaries

Our objective is to derive the best HARQ mechanism, by optimally tuning n_1, n_2 as well as P_1 and P_2 (power assigned to each fragment of the codeword), which minimizes the average consumed energy for a given packet error probability and latency constraint (URLLC requirements). For that, we first need to characterize the error probability of the associated HARQ mechanism as a function of the quadruple (n_1, P_1, n_2, P_2). To derive the packet error probability for short packets (finite blocklength), we cannot resort to the standard asymptotic regime related to large blocklengths.

In IR-HARQ with one retransmission, the packet error probability or equivalently the outage probability, denoted by ε, can be expressed as

$$\varepsilon = \mathbb{P}(\Omega_1 \cap \Omega_2) \tag{1}$$

where Ω_1 is the event "the first fragment of length n_1 and energy per symbol P_1 is not correctly decoded" and Ω_2 is the event "the concatenation of the first fragment and the second one of length n_2 and energy per symbol P_2 is not correctly decoded".

When *infinite* blocklength regime is assumed, the error occurs when the mutual information is less than a threshold and for IR-HARQ, it can be easy to see that $\Omega_2 \subseteq \Omega_1$ [7], [8] leading to $\varepsilon = \mathbb{P}(\Omega_2)$. In contrast, when a real coding scheme (and so *finite* blocklength) is used, this is not true anymore [8], and a closed-form expression for ε is intractable. Therefore in the majority of prior work on HARQ (see [8] and references therein), the exact outage probability ε is replaced with the simplified $\bar{\varepsilon}$ defined as

$$\bar{\varepsilon} = \mathbb{P}(\Omega_2),$$

since ε and $\bar{\varepsilon}$ perform quite closely numerically. In the remainder of the paper, we assume that this approximation applies also for Polyanski's framework [1]. Then, $\bar{\varepsilon}$ can be upper bounded [1, Lemma 14] and also lower bounded as in [9] by employing the $\kappa\beta$-bounds proposed in [1]. Both bounds have the same first two dominant terms and lead to the following form:

$$\bar{\varepsilon} = Q\left(\frac{n_1 \ln(1+P_1) + n_2 \ln(1+P_2) - B\ln 2}{\sqrt{\frac{n_1 P_1(P_1+2)}{(P_1+1)^2} + \frac{n_2 P_2(P_2+2)}{(P_2+1)^2}}}\right) \tag{2}$$

where $Q(x)$ is the complementary Gaussian cumulative distribution function. Setting $n_2 = 0$ we obtain the common formula describing the error probability of the first fragment:

$$\varepsilon_1 = \mathbb{P}(\Omega_1) = Q\left(\frac{n_1 \ln(1+P_1) - B\ln 2}{\sqrt{\frac{n_1 P_1(P_1+2)}{(P_1+1)^2}}}\right). \tag{3}$$

For the sake of clarity, we may mention the dependency on the variables, i.e. $\bar{\varepsilon}(n_1, P_1, n_2, P_2)$ instead of $\bar{\varepsilon}$ and $\varepsilon(n_1, P_1)$ instead of ε.

IV. Optimization Problems

We focus on the minimization of the average energy consumed to achieve a target reliability T_{rel} ($T_{\text{rel}} = 99.999\%$ in 3GPP URLLC or equivalently an outage probability $P_{out} = 1 - T_{\text{rel}} = 10^{-5}$) without violating the latency constraint $n_1 + n_2 \leq N - D$ by properly setting n_1, n_2, P_1, and P_2. At first, we will assume no feedback penalty ($D = 0$) since the extension is straightforward by setting $N' = N - D$ in the latency constraint.

In the rest of the paper, we address (i) optimized IR-HARQ, (ii) partially optimized IR-HARQ when $n_1 = n_2 = N/2$, and (iii) no HARQ ($n_1 = N$ and $n_2 = 0$).

A. Optimized IR-HARQ

The problem is stated as follows:

Problem 1:

$$\min_{n_1,P_1,n_2,P_2} \quad n_1P_1 + n_2P_2\varepsilon_1 \tag{4}$$

$$\text{s.t.} \quad n_1 + n_2 \leq N \tag{5}$$

$$\bar{\varepsilon} \leq 1 - T_{\text{rel}} \tag{6}$$

Our objective is to show that inequalities in the constraints (5)-(6) can be replaced with equalities. For doing that, we need Lemma 1.

Lemma 1: The optimal solution $(n_1^\star, P_1^\star, n_2^\star, P_2^\star)$ satisfies $\varepsilon_1 \geq \bar{\varepsilon}$.

The proof is quite simple since if $\varepsilon_1 < \bar{\varepsilon}$ at the optimal solution, then the quadruple $(n_1^\star, P_1^\star, n_2^\star, 0)$ offers a lower consumed average energy which leads to a contradiction.

Then we have the following results.

Proposition 1: If a quadruple (n_1, P_1, n_2, P_2) satisfies $\varepsilon_1 > \bar{\varepsilon}$ then $\bar{\varepsilon}$ is decreasing with respect to P_2 at that point.

Proof: See Appendix A. ■

Proposition 1 enables us to force the constraint (6) to be an equality. Indeed, assuming that the optimal point $(n_1^\star, P_1^\star, n_2^\star, P_2^\star)$ satisfies $\bar{\varepsilon} < 1 - T_{\text{rel}}$ (we also know that the optimal point satisfies the condition of Proposition 1 according to Lemma 1), then $P_2^\star$ can be decreased to P_2' such that $\bar{\varepsilon} = 1 - T_{\text{rel}}$. This implies that $(n_1^\star, P_1^\star, n_2^\star, P_2')$ is a better solution than the optimal one which leads to contradiction coming from the assumption $\bar{\varepsilon} < 1 - T_{\text{rel}}$ at the optimal solution.

Proposition 2: Let $\mathcal{D} = \{(n_1, P_1, n_2, P_2) \in \mathbb{R}_+^4 : 1/2 > \varepsilon_1(n_1, P_1) > \bar{\varepsilon}(n_1, P_1, n_2, P_2) > Q(\sqrt{2B\ln 2}/3)\}$. As long as $(an_1, P_1/a, n_2, P_2) \in \mathcal{D}$, $\varepsilon_1(an_1, P_1/a)$ and $\bar{\varepsilon}(an_1, P_1/a, n_2, P_2)$ are decreasing with respect to a.

Proof: See Appendix B. ■

The above result implies that *given an energy budget, it is preferable to spread it into many symbols with low power than to few ones with high power.* Consequently, if we search for an optimal point $(n_1^\star, P_1^\star, n_2^\star, P_2^\star)$ satisfying $0.5 > \varepsilon_1 > \bar{\varepsilon} = 1 - T_{\text{rel}} > Q(\sqrt{2B\ln 2}/3)$, then the constraint (5) also becomes an equality. Indeed, assuming that for the optimal point $n_1^\star + n_2^\star < N$, then any $a > 1$[1] such that $(an_1^\star, P_1^\star/a, n_2^\star, P_2^\star) \in \mathcal{D}$ and $an_1^\star + n_2^\star \leq N$ yields a better solution. Actually $an_1 \in \mathbb{R}_+$ whereas the blocklength can only be a natural number. To overcome this issue, we assume the scheme with $a = (n_1^\star + 1)/n_1^\star$ is still in $\mathcal{D}$, i.e. increasing the blocklength of the first fragment by one symbol is possible in $\mathcal{D}$.

Propositions 1 and 2 allow us to assert that the constraints are satisfied with equality. Therefore the four optimization variables in Problem 1 are reduced to only two. In Section VI, we will see that belonging to the set $\mathcal{D}$ especially for the optimal solution is not restrictive at all and so Proposition 2 applies in practice. In Fig. 1, we plot the cost function given by (4) with respect to (n_1, P_1) in the feasible domain given by constraints (5)-(6). We observe that the cost function is

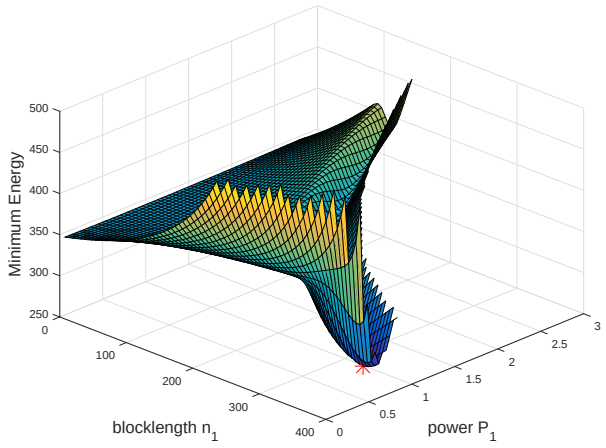

Fig. 1. Average consumed energy versus (n_1, P_1) for $N = 400$, $B = 32$ bytes, and $T_{\text{rel}} = 99.999\%$. The red asterisk marks the minimum.

neither convex nor quasi-convex, thus standard optimization tool cannot be used. As the optimization problem is reduced to find a two-dimensional (2D) bounded parameter, we resort to exhaustive search. More precisely, we use a 2D search over (n_1, P_1) with $n_2 = N - n_1$ and a bisection method to find P_2. The bisection method is efficient since the outage probability ε is a decreasing function with respect to P_2. The complexity is $\mathcal{O}(\theta^{-2}\log(1/\theta))$, where θ is the approximation error.

B. Partially optimized IR-HARQ

We consider here the case where the retransmission packet has the same blocklength as the first packet ($n_1 = n_2$). That case is referred to as partially optimized IR-HARQ, where the sole parameters to optimize are the power P_1 and P_2.

Problem 2:

$$\min_{P_1,P_2} \quad \frac{N}{2}(P_1 + P_2\varepsilon_1)$$

$$\text{s.t.} \quad \bar{\varepsilon} = 1 - T_{\text{rel}}$$

Only one-dimensional exhaustive search over P_1 is needed since, once again, P_2 can be found through a bisection method solving $\varepsilon = 1 - T_{\text{rel}}$. The complexity is $\mathcal{O}(\theta^{-1}\log(1/\theta))$.

C. No HARQ

We assume that $n_1 = N$ and $n_2 = 0$ (one-shot transmission). As the outage is a decreasing function with P_1 (see (2)), we just have to find the root in P of the equation

$$\varepsilon_1(N, P) = 1 - T_{\text{rel}}. \tag{7}$$

A bisection method can be used, whose complexity is $\mathcal{O}(\log(1/\theta))$.

[1]There exists at least one $a > 1$ in $\mathcal{D}$ by continuity of ε_1 and $\bar{\varepsilon}$ with respect to a.

V. Asymptotic Regime

The consumed energy for sending a fixed number of B information bits is a non-increasing function with respect to the latency N since, as seen in Problem 1, the optimal solution for a given N is feasible solution for $(N+1)$. But, as seen in Fig.2, an asymptotic value occurs when $N \to \infty$.

Proposition 3: Let $(n_1^\star, P_1^\star, n_2^\star, P_2^\star)$ be the optimal point of Problem 1, $E_i = n_i^\star P_i^\star$ the energy spent on the i-th fragment and $\beta = n_1^\star/N \in (0,1)$. The minimum average consumed energy under the constraints given by Problem 1 is independent of β when $N \to \infty$ and equals to the solution of the following optimization problem:

$$\min_{E_1,E_2} \quad E_1 + Q\left(\frac{E_1 - B\ln 2}{\sqrt{2E_1}}\right)E_2$$
$$\text{s.t.} \quad E_1 + E_2 = E^{\infty}_{\text{No}-\text{HARQ}}$$

with $E^{\infty}_{\text{No}-\text{HARQ}} = \frac{(Q^{-1}(1-T_{\text{rel}}))^2}{2}\left(1+\sqrt{1+\frac{2B\ln 2}{(Q^{-1}(1-T_{\text{rel}}))^2}}\right)^2$.

Proof: See Appendix C. ■

Notice that $E^{\infty}_{\text{No}-\text{HARQ}}$ corresponds to the required average energy when $N \to \infty$ for the case of no HARQ.

VI. Numerical Results and Discussion

In this section, we provide numerical results based on our analysis as a means to shed light to whether or not it is beneficial from an energy point of view to split the packet transmission in URLLC systems. Except otherwise stated, we set $B = 32$ bytes and $T_{\text{rel}} = 99.999\%$. According to these values, we have $1 - T_{\text{rel}} \gg Q(\frac{\sqrt{2B\ln 2}}{3}) \approx 1.7 \cdot 10^{-10}$ and it is reasonable to consider design parameters n_1 and P_1 such that $\varepsilon_1 < 0.5$. Thus forcing the parameters (n_1, P_1, n_2, P_2) to be in $\mathcal{D}$ is not restrictive at all; hence we consider the constraints of the optimization problems as equalities.

In Fig. 2, we plot the minimum average consumed energy versus N (with $D = 0$) for the two HARQ and no HARQ schemes. As stated in Proposition 2, the consumed energy for sending a packet of B information bits decreases for any configuration when N increases. Nevertheless, the gain is less substantial when N is large enough since an asymptotic floor occurs. In the asymptotic regime, we have $E^{\infty}_{\text{noHARQ}} = 278$ for no HARQ and as anticipated both other configurations converge to the same smaller value $E^{\infty}_{\text{HARQ}} = 210$. Clearly, for $D = 0$, IR-HARQ always performs not worse than no HARQ. This is expected since the feedback of the IR-HARQ mechanism enables, at times, to use only n_1 channel uses (thus, saving energy by not using the remaining n_2).

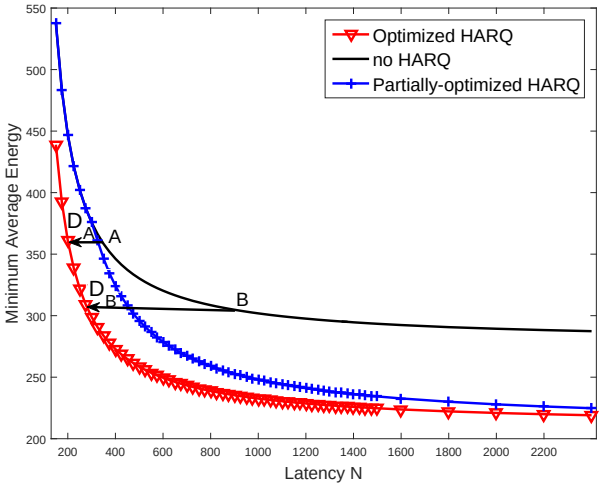

Fig. 2. Minimum average energy versus latency N (with $D = 0$).

The effect of feedback delay D on the performance can be observed in Fig. 2 (see points A and B). Let N_A (resp. N_B) be the minimum satisfied latency for a given consumed energy E_A (resp. E_B) when no HARQ is used. Using optimized IR-HARQ leads to lower latency $N'_A = N_A - D_A$ (resp. $N'_B = N_B - D_B$) for the same amount of energy. Consequently, D_A (resp. D_B) is the latency gain of optimized IR-HARQ against no HARQ. In other words, optimized IR-HARQ can support a feedback delay $D < D_A$ (resp. $D < D_B$) while offering gain in terms of energy consumption when this energy is upper bounded by E_A (resp. E_B). In other words, under reasonable feedback delay values, it is preferable to split the packet into two fragments using IR-HARQ.

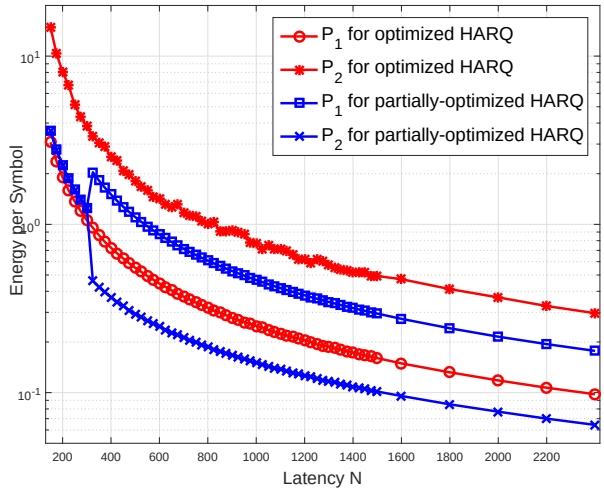

Fig. 3. Power allocation for minimum average energy in both optimized and partially optimized IR-HARQ.

In Fig. 3, we show the optimal power allocation $(P_1^\star, P_2^\star)$ versus the latency N for both IR-HARQ configurations. We see that in optimized IR-HARQ we always have $P_1^\star < P_2^\star$, and on the contrary in partially optimized IR-HARQ, we have $P_1^\star > P_2^\star$ for large N but $P_1^\star = P_2^\star$ for small N, i.e. same performance as no HARQ. The explanation is based on the fact that HARQ has the benefit of *early termination* offering the possibility of no retransmission, thus saving power-blocklength resources. The advantage of HARQ, as compared to no HARQ, is more pronounced when early termination occurs very frequently but without sacrificing a large amount of energy for lowering ε_1, i.e. when both ε_1 and $n_1 P_1$ are small. In the no HARQ case, a smaller error can be achieved even by decreasing the energy (increasing the

available blocklength leads to even less energy - see the curve for no HARQ in Fig. 2 as N grows). Therefore, both n_1P_1 and ε_1 can be kept small by increasing n_1 and decreasing P_1 in the optimized IR-HARQ. That's why we get that the optimal $(n_1^\star, n_2^\star)$ leads to $n_1^\star > n_2^\star$ (specifically $n_1^\star \approx 0.89N$ for almost any value of N), and that P_1^* is small compared to P_2^*. In contrast, in the partially optimized IR-HARQ, one cannot adapt n_1. Therefore, decreasing ε_1 depends on the available N. If N is inadequate, then decreasing ε_1 requires excessively high P_1, which yields to an inefficient solution. That is why for small N, ε_1 is almost 1 (i.e. retransmission should always be employed) and the behavior of IR-HARQ is similar to no HARQ. When N becomes sufficiently large, then the only solution for decreasing ε_1 is increasing P_1. That is why $P_1^\star > P_2^\star$ in the partially optimized IR-HARQ case.

In Fig. 4, we plot the difference (in percentage) between the energy consumed in no HARQ and the optimal average energy consumed in IR-HARQ versus D. Positive gains mean that an IR-HARQ mechanism performs better than no HARQ. We observe that the splitting approach (optimized IR-HARQ)

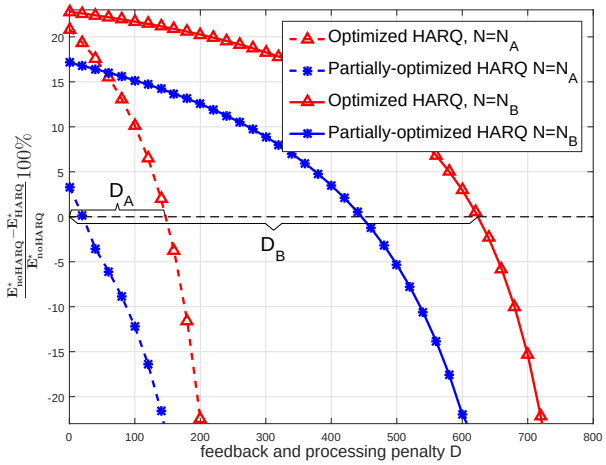

Fig. 4. $(E^\star_{\text{no HARQ}} - E^\star_{\text{HARQ}})/E^\star_{\text{no HARQ}}$ in % versus D for $N_A = 350$ and $N_B = 900$.

is better than one-shot transmission (no splitting) for a large amount of feedback delay D (around $N/2$ or even more). As N increases, the amount of feedback delay that optimized IR-HARQ can support while being more energy efficient also increases. For example, when $N = N_A = 330$, we have $D_A = 0.42N_A$, whereas when we increase to $N = N_B = 900$ we have $D_B = 0.69N$. Therefore, as N grows, IR-HARQ becomes a more robust solution with respect to feedback delay. Note also that an unoptimized or partially optimized IR-HARQ does not necessarily provide better performance than no HARQ, even with almost zero feedback delay.

VII. Conclusion

In this paper, we investigated the energy-latency tradeoff in URLLC with short packets and analyzed the energy consumption of IR-HARQ in the finite-blocklength regime. The main takeaway of this paper is that a properly optimized IR-HARQ scheme can be beneficial in terms of energy as long as the feedback delay is reasonable compared to the packet size.

Appendix A
Proof of Proposition 1

We want to prove that $\partial\bar{\varepsilon}/\partial P_2 < 0$ if $\bar{\varepsilon} < \varepsilon_1$. Assuming $y = 1/(P_2+1)^2$, it is easy to show that

$$\frac{\partial\bar{\varepsilon}}{\partial P_2} < 0 \Leftrightarrow \frac{\partial\bar{\varepsilon}}{\partial y} > 0 \Leftrightarrow h(y) > 0 \tag{8}$$

where $h(y) = k_2 - yk_1 + n_2(1 - y + y\ln(y)/2)$ with $k_1 = n_1\ln(1+P_1) - B\ln 2$, $k_2 = n_1\left(1 - 1/(1+P_1)^2\right)$. It is also easy to prove that $h(y)$ is a monotonically decreasing function. If $h(1) \geq 0$, then (8) is straightforwardly satisfied. If $h(1) < 0$, then it exists $y_0 \in (0,1)$ such that $h(y_0) = 0$. So for $y \in [y_0, 1]$, we get $h(y) \leq 0$, which implies that at that interval $\bar{\varepsilon}$ is decreasing with respect to y. As a consequence, for $y \in [y_0, 1]$, we have $\bar{\varepsilon} \geq \bar{\varepsilon}|_{y=1} = \bar{\varepsilon}|_{P_2=0} \Leftrightarrow \bar{\varepsilon} \geq \varepsilon_1$ which is prevented according to the assumption $\bar{\varepsilon} < \varepsilon_1$. Consequently, y does not belong to $[y_0, 1]$, and belongs to $(0, y_0)$ where (8) holds again.

Appendix B
Proof of Proposition 2

We have $\varepsilon_1 = Q(F_1(a))$ and $\bar{\varepsilon} = Q(F(a))$ where

$$F_1(a) = \frac{g_1(a) - c}{\sqrt{g_2(a)}} \text{ and } F(a) = \frac{g_1(a) + c_1 - c}{\sqrt{g_2(a) + c_2}},$$

with $g_1(a) = an_1\ln(1+\frac{P_1}{a})$, $g_2(a) = an_1(1-1/(1+P_1/a)^2)$, $c_1 = n_2\ln(1+P_2)$, $c_2 = n_2(1-1/(1+P_2)^2)$, and $c = B\ln 2$. As we consider a point in $\mathcal{D}$, we get

$$\varepsilon_1 < 0.5 \Leftrightarrow an_1\ln(1+P_1/a) > c \Rightarrow E_1 > B\ln 2 \tag{9}$$

where $E_1 = n_1P_1$. To prove (9), we used the following inequality $\ln(1+x) \leq x$ when $x \geq 0$. Once again, belonging to $\mathcal{D}$ leads to

$$F_1(a) \leq F(a) \leq \sqrt{2B\ln 2}/3. \tag{10}$$

We want to show that ε_1 and $\bar{\varepsilon}$ are decreasing functions with respect to a, i.e. $F_1'(a) \geq 0$ and $F'(a) \geq 0$ where $f'(a)$ stands for $\mathrm{d}f/\mathrm{d}a$ for any mapping f. As $g_1(a), g_2(a), g_1'(a)$ and $g_2'(a)$ are strictly positive, we have

$$F_1'(a) \geq 0 \quad \Leftrightarrow \quad 2g_1'(a)g_2(a) \geq g_2'(a)(g_1(a) - c) \tag{11}$$
$$\Leftrightarrow \quad c \geq E_1H(P_1/a) \tag{12}$$

and

$$F'(a) \geq 0 \Leftrightarrow 2g_1'(a)(g_2(a) + c_2) \geq g_2'(a)(g_1(a) + c_1 - c) \tag{13}$$
$$\Leftrightarrow c \geq E_1H(P_1/a) + (c_1 - K(P_1/a)c_2) \tag{14}$$

where

$$x \mapsto H(x) = \frac{2x + 4 - \ln(1+x)(\frac{4}{x} + x + 3)}{x(x+3)},$$

and

$$x \mapsto K(x) = \frac{2(x+1)^3\left(\ln(1+x) - \frac{x}{x+1}\right)}{x^2(x+3)}.$$

After some algebraic manipulation, (11) and (13) are equivalent to

$$F_1(a) \leq \frac{2g_1'(a)\sqrt{g_2(a)}}{g_2'(a)} = \sqrt{E_1}W(P_1/a, 0) \quad (15)$$

$$F(a) \leq \frac{2g_1'(a)\sqrt{g_2(a)+c_2}}{g_2'(a)} = \sqrt{E_1}W(P_1/a, \frac{c_2}{E_1}) \quad (16)$$

with

$$(x,y) \mapsto W(x,y) = K(x)\sqrt{y + \frac{x+2}{(1+x)^2}}. \quad (17)$$

Therefore, we now want to prove that either (12) or (15) holds for any $x > 0$, and either (14) or (16) holds for any $x > 0$. For doing that, we split the analysis into two intervals on x.

- If $x \in (0, 484)$: the function $x \mapsto W(x,0)$ is a positive unimodal function converging to zero when x goes to ∞. For $x \in (0,484)$, it is easy to check that $W(x,0) \geq W(0,0) = \sqrt{2}/3$. As $W(x,y) > W(x,0)$ for any $y \geq 0$, we obtain that $\sqrt{E_1}W(x,y) \geq \sqrt{E_1}W(x,0) \geq \sqrt{2E_1}/3$. Due to (9), we have $\sqrt{E_1}W(x,y) \geq \sqrt{E_1}W(x,0) \geq \sqrt{2B\ln 2}/3$. According to (10), we check that $\sqrt{E_1}W(x,y) \geq \sqrt{E_1}W(x,0) \geq F(a) \geq F_1(a)$. Therefore, (15) and (16) hold.
- If $x \in [484, \infty)$: in that interval, we can see that $H(x) \leq 0$ which implies that (12) holds.
  Now it remains to check that either (14) or (16) holds. For doing that, we need to distinguish two cases:
  - If $c_1 \leq 10.37c_2$: one can check that $K(x)$ is an increasing function. Therefore for $x \geq 484$, we get $K(x) \geq K(484) > 10.37$. Consequently, $c_1 - K(x)c_2 < 0$. As $H(x) \leq 0$ too for $x \geq 484$, it is easy to show that (14) holds.
  - If $c_1 > 10.37c_2$: this inequality leads to $P_2 > 31866$ which implies that $c_2 \approx n_2 (>1)$. Consequently, according to (17), $\sqrt{E_1}W(x, c_2/E_1) \geq K(484)\sqrt{n_2} > 10.37$. If (16) does not hold, one can see that $\bar{\varepsilon} < Q(10.37) \approx 1.7{\cdot}10^{-25}$. As this error does not correspond to any reasonable operating point, we can consider that (16) holds.

Appendix C
Proof of Proposition 3

We remind that the value of the optimal point $(n_1^\star, P_1^\star, n_2^\star, P_2^\star)$ depends on the blocklength N. Assume that $\lim_{N\to\infty} E_i = \infty$ for either $i = 1$ or $i = 2$, then the average energy $E_1 + \varepsilon_1 E_2 \to \infty$ since $\varepsilon_1 \geq \bar{\varepsilon} = 1 - T_{\text{rel}} > 0$. For a finite N, say N_f, the optimal point spends a finite amount of average energy. So considering $N \geq N_f$ cannot increase the optimal average energy since the previous solution with $N = N_f$ remains a feasible point. Thus, growing N to infinity necessary results in a smaller average energy. Therefore, $\lim_{N\to\infty} E_i < \infty, i \in \{1,2\}$. According to Proposition 2, the optimal solution uses the whole blocklength, so $\lim_{N\to\infty} n_1^\star = \beta N = \infty$ and $\lim_{N\to\infty} n_2^\star = (1-\beta)N = \infty$. Consequently, $\lim_{N\to\infty} P_i^\star = 0, i \in \{1,2\}$. We prove below that $\lim_{N\to\infty} n_i^\star \ln(1 + P_i^\star) = E_i$:

$$\begin{aligned} \frac{P_i^\star}{P_i^\star+1} \leq \quad & \ln(1+P_i^\star) \quad && \leq P_i^\star \\ \Rightarrow \lim_{N\to\infty} \frac{n_i^\star P_i^\star}{P_i^\star+1} \leq \lim_{N\to\infty} & n_i \ln(1+P_i^\star) && \leq \lim_{N\to\infty} n_i^\star P_i^\star \\ \Rightarrow \frac{E_i}{0+1} \leq \lim_{N\to\infty} & n_i \ln(1+P_i^\star) && \leq E_i. \end{aligned}$$

According to (6), we have

$$1 - T_{\text{rel}} = \lim_{N\to\infty} Q\left(\frac{n_1^\star \ln(1+P_1^\star) + n_2^\star \ln(1+P_2^\star) - B\ln 2}{\sqrt{n_1^\star P_1^\star \frac{P_1^\star+2}{(P_1^\star+1)^2} + n_2^\star P_2^\star \frac{P_2^\star+2}{(P_2^\star+1)^2}}}\right)$$

$$\Leftrightarrow 1 - T_{\text{rel}} = Q\left(\frac{E_1 + E_2 - B\ln 2}{\sqrt{E_1 \frac{0+2}{(0+1)^2} + E_2 \frac{0+2}{(0+1)^2}}}\right)$$

$$\Leftrightarrow B\ln 2 = E_1 + E_2 - Q^{-1}(1-T_{\text{rel}})\sqrt{2(E_1+E_2)}. \quad (18)$$

In (18), we have to solve a second-order polynomial to exhibit $E_1 + E_2$. We obtain that

$$E_1 + E_2 = \frac{(Q^{-1}(1-T_{\text{rel}}))^2}{2}\left(1 + \sqrt{1 + \frac{2B\ln 2}{(Q^{-1}(1-T_{\text{rel}}))^2}}\right)^2. \quad (19)$$

Using (3), $\lim_{N\to\infty} n_1^\star \ln(1+P_1^\star) = E_1$, and $\lim_{N\to\infty} P_1^\star = 0$, we get

$$\varepsilon_1 = Q\left(\frac{E_1 - B\ln 2}{\sqrt{2E_1}}\right).$$

The right-hand side (RHS) of (19) corresponds to the energy of no HARQ and so is denoted by $E_{\text{no-HARQ}}^{\infty}$. Indeed, the RHS of (19) can also be obtained by neglecting the third term of the RHS of equation [1, (4.309)], where this equation plays the same role as (18).

References

[1] Y. Polyanskiy, "Channel coding: Non-asymptotic fundamental limits," Ph.D. dissertation, Princeton University, Nov. 2010.

[2] B. Makki, T. Svensson, and M. Zorzi, "Finite block-length analysis of the incremental redundancy HARQ," *IEEE Wireless Commun. Lett.*, vol. 3, no. 5, pp. 529–532, Oct. 2014.

[3] P. Wu and N. Jindal, "Coding versus ARQ in fading channels: How reliable should the PHY be?" *IEEE Trans. on Commun.*, vol. 59, no. 12, pp. 3363–3374, Dec. 2011.

[4] S. H. Kim, D. K. Sung, and T. Le-Ngoc, "Performance analysis of incremental redundancy type hybrid ARQ for finite-length packets in AWGN channel," in *Proc. IEEE Global Commun. Conf. (Globecom)*, Atlanta, GA, USA, Dec. 2013.

[5] A. R. Williamson, T. Chen, and R. D. Wesel, "A rate-compatible sphere-packing analysis of feedback coding with limited retransmissions," in *Proc. IEEE ISIT*, Cambridge, MA, USA, July 2012.

[6] J. Gibson, *The Communications Handbook*. CRC press, 2002.

[7] G. Caire and D. Tuninetti, "The throughput of hybrid ARQ protocols for the Gaussian collision channel," *IEEE Trans. on Inf. Theory*, vol. 47, no. 5, pp. 1971–1988, July 2001.

[8] C. L. Martret, A. Leduc, S. Marcille, and P. Ciblat, "Analytical performance derivation of hybrid ARQ schemes at IP layer," *IEEE Trans. on Commun.*, vol. 60, no. 5, pp. 1305–1314, May 2012.

[9] J. Park and D. Park, "A new power allocation method for parallel AWGN channels in the finite block length regime," *IEEE Wireless Commun. Lett.*, vol. 16, no. 9, pp. 1392–1395, Sept. 2012.