Joint Optimization of Link Adaptation and HARQ Retransmissions for URLLC Services

Matha Deghel*, Salah Eddine Elayoubi◇, Ana Galindo-Serrano*, Raphael Visoz*

*Orange Labs, Châtillon, France
◇Laboratoire de Signaux et Systèmes (L2S, UMR8506) CentraleSupélec-CNRS-Université Paris-Sud, Gif-sur-Yvette, France
{matha.deghel, anamaria.galindoserrano, raphael.visoz}@orange.com, salaheddine.elayoubi@centralesupelec.fr

***Abstract*—This work proposes a new joint link adaptation and HARQ (Hybrid Automatic Repeat Request) scheme for URLLC (Ultra-Reliable Low-Latency Communication) services. We consider the case where the transmitter knows only the average SNR (Signal-to-Noise Ratio) and not the instantaneous one. In the proposed scheme, the optimal maximum number of allowable HARQ transmissions and the optimal MCS (Modulation and Coding Scheme) level are determined for each packet to maximize the spectral efficiency. Our adopted approach exploits the channel diversity and increases the flexibility of the scheduling mechanism. Simulation results show that the proposed retransmission policy and link adaptation scheme increases the system performance in terms of spectral efficiency, while satisfying the latency and reliability constraints.**

***Keywords*—5G, URLLC services, link adaption, hybrid automatic repeat request.**

I. Introduction

Ultra-Reliable Low-Latency Communication (URLLC) services are one of the target usage scenarios for 5G communication systems. Such services require short latency and very high reliability [1], and enable real-time control and automation of dynamic processes for vertical applications. The most stringent reliability requirement on URLLC services (currently being standardized) for one transmission of a packet is $1-10^{-5}$ with a user-plane latency of 1 ms [1]. Among others, decreasing the Transmission Time Interval (TTI) and the Round Trip Time (RTT) is suggested as an efficient way to improve the latency performance [2]. On the other hand, one of the key techniques to enhancing the reliability for transmission over fading channels is *link adaptation* [3], also known as Adaptive Modulation and Coding (AMC). Another technique is to rely on the *Hybrid Automatic Repeat Request (HARQ)* protocol at the data link layer.

The HARQ mechanism consists in a combination of ARQ mechanism with channel coding. In ARQ schemes, the receiver examines a packet error using an error detection code such as cyclic redundancy check, and sends a positive Acknowledgment (ACK) or a negative Acknowledgment (NACK) to the transmitter. HARQ is a variation of the ARQ error control scheme, with the objective of reducing the number of transmissions by adding Forward Error Correction (FEC) bits to the existing error detection bits. In HARQ schemes with soft combining, the receiver decodes a retransmitted packet in conjunction with previously transmitted erroneous packets. Chase Combining (CC) [4] and Incremental Redundancy (IR) are two possible methods for soft combining in HARQ. Due to its simplicity and tractability, CC is widely used in the literature and is thus considered in this work. In case of errors with CC HARQ, the same packet is retransmitted and the receiver uses Maximum Ratio Combining (MRC) to combine the received bits with the same bits from previous transmissions. One could think of every packet retransmission as adding extra energy to the received transmission.

Instead of considering link adaptation and HARQ as separate design entities, they can be combined to maximize the system performance. For instance, in [5] these schemes are combined to maximize the spectral efficiency under delay and error performance constraints. The authors in [6] optimize the thresholds of link adaptation for HARQ with IR, where the analysis is based on the instantaneous throughput. The work in [7] also considers HARQ with link adaptation, where the focus is more on scheduling. The work in [8] provides a closed formula of the packet loss probability and the throughput for truncated CC HARQ with link adaptation. In [9], the combination of the two schemes is investigated and analyzed from an information theory perspective.

In this work, we consider both link adaptation and HARQ schemes, under the assumption that the transmitter knows only the average Signal-to-Noise Ratio (SNR) and not the instantaneous one. Such an assumption holds true, for instance, in a highly-mobile scenario where the channel is rapidly varying. A short TTI (i.e. mini-slot) is considered so that URLLC services can be supported by the system. Our aim is to maximize the spectral efficiency of the adopted system given latency and reliability constraints. To this end, we propose a joint HARQ retransmission and link adaptation scheme, where the optimal maximum number of HARQ transmissions and the optimal Modulation and Coding Scheme (MCS) level are determined for each URLLC service and average SNR. It is worth mentioning that URLLC service requirements range from very tight latency constraints, e.g. a 1 or 2 ms budget, to more relaxed ones, e.g. 10 ms budget.

Our work is different from previous works that combine link adaptation and HARQ in the sense that (i) the HARQ optimization, and unlike previous works, is done by deter-

mining the optimal maximum number of transmissions per packet, (ii) our proposed scheme increases the flexibility of the scheduling mechanism. Having such a high flexibility is especially important for scheduling URLLC services with very tight latency budgets, since more scheduling opportunities will be available for these services. It is worth noting that when the latency budget becomes very tight, the benefits of applying our approach is less noticeable.

The rest of this paper is organized as follows. The system model is given in Section II. In Section III, we provide the joint HARQ retransmission and link adaption scheme, and we analyze the corresponding system performance. Section IV is dedicated to numerical results and relevant discussions. We finally conclude the paper in Section V.

II. System Model

In this section, we first present the channel model under consideration. Then, we explain the adopted HARQ protocol and the associated packet error model. Finally, we describe the proposed transmission scheme.

We assume a system working under a Frequency Division Duplex (FDD) mode. It is also assumed that data are sent in fixed-size packets by the transmitter and in general each packet is acknowledged by the receiver.

A. Channel Model

We consider a block fading channel with Rayleigh fading and a coherence time T_{c}. The received SNR then remains constant during a packet transmission, and is independent and identically distributed (i.i.d.) between different transmissions. We denote by γ the instantaneous SNR at the receiver. Let $\bar{\gamma}$ be the average received SNR. The probability distribution function of γ, which we denote by $p_{\gamma}(\gamma)$, can be given as

$$p_{\gamma}(\gamma) = \frac{1}{\bar{\gamma}} \exp\left(-\frac{\gamma}{\bar{\gamma}}\right). \tag{1}$$

Note that we work under the assumption of an Additive White Gaussian Noise (AWGN) channel. Also, a fast-varying channel is assumed, i.e. T_{c} is sufficiently low, so the transmitter does not know the instantaneous Channel State Information (CSI), but only knows $\bar{\gamma}$. Nevertheless, we still consider link adaptation, where several MCS levels are used. Let m represent the MCS level (i.e. mode) with $1 \leq m \leq M$, where M is the number of MCS levels. The approach for link adaptation will be discussed in a subsequent section.

B. HARQ Protocol and Packet Error Model

The considered HARQ protocol uses CC. Under CC type of HARQ, a packet is repeated with the same MCS until successful reception at the receiver or until the maximum number of allowable transmissions is reached. At the receiver, previous erroneous packets are stored in a buffer, so that a current packet retransmission can be combined with previously received erroneous packets before they are passed to decoder. We suppose that MRC technique is used to combine the packets. Let γ_k denote the received SNR at the kth transmission. Then, after the combination of the packets, the resulting SNR at the kth transmission, denoted $\gamma_{\mathrm{G},k}$, can be written as

$$\gamma_{\mathrm{G},k} = \sum_{i=1}^{k} \gamma_i. \tag{2}$$

For the adopted system, an approximation of the packet error probability as a function of the instantaneous SNR for an AWGN channel was found in [5] as

$$f_m(\gamma) = \begin{cases} 1, & \text{if } 0 < \gamma < \gamma_{\mathrm{S},m} \\ a_m \exp\left(-g_m \gamma\right), & \text{if } \gamma \geq \gamma_{\mathrm{S},m} \end{cases} \tag{3}$$

where a_m, g_m and $\gamma_{\mathrm{S},m}$ are parameters that depend on the MCS level, m, and that are found through simulation and curve fitting.

Recalling that under the adopted HARQ protocol the receiver combines all the received packets using the MRC technique, and supposing that MCS level m is used, it can be seen that the packet error probability at the kth transmission is $f_m(\gamma_{\mathrm{G},k})$.

C. Proposed Transmission Policy

Suppose we have a latency budget of T_{lat} (in ms). We consider a short TTI (shorter than the 1 ms in LTE), which is denoted by T_{TTI} (in ms). Let T_{RTT} (in ms) represent the RTT, which is defined in this work as the duration of time between the transmission of a packet and the reception of the corresponding ACK/NACK, including the processing times at the transmitter and the receiver.

Unlike classical approaches where a retransmission is done immediately after the reception of a NACK, here we suppose that the transmitter is allowed to *wait* a duration of time equal to x (in ms) before retransmitting the packet; clearly, x should correspond to an integer number of TTIs. In other words, the waiting time between two transmissions of the same packet (if any) can be given by $T = T_{\mathrm{RTT}} + x$; as presented in Figure 1. Accounting for the latency budget, the *maximum number of transmissions* that can be allowed for a packet, which we denote by $K(x)$, can then be written as

$$K(x) = \left\lceil \frac{T_{\mathrm{lat}}}{T_{\mathrm{RTT}} + x} \right\rceil, \tag{4}$$

where $\lceil z \rceil$ gives as output the least integer that is greater than or equal to z. Note that the maximum number of allowable *retransmissions* is $K(x) - 1$. It should also be noted that, under the latency constraint, the longest allowable waiting time before retransmission (after the reception of a NACK), denoted $x_{\max}$, can be given as follows

$$x_{\max} = T_{\mathrm{lat}} - T_{\mathrm{RTT}} - \frac{1}{2} T_{\mathrm{RTT}}, \tag{5}$$

which corresponds to a minimum number of allowable transmissions (of a packet) equal to 2, i.e. only one allowable retransmission. The expression in (5) results from the fact that if for the first transmission there is a NACK, then:

- The time between the first transmission and the reception of the NACK is T_{RTT}.

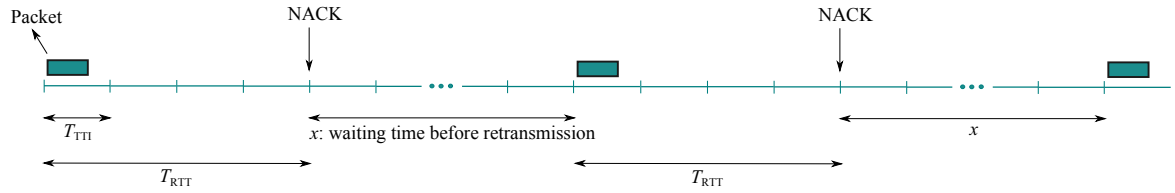

Figure 1: Example illustrating the proposed transmission policy.

– After the second transmission the transmitter waits only for the packet to arrive to the receiver without waiting for the reception of the corresponding ACK/NACK, because anyway the packet cannot be transmitted again. For simplicity, we assume that this waiting time is half the RTT, i.e. $\frac{1}{2}T_{\text{RTT}}$.

From the above considerations we have

$$0 \leq x \leq x_{\max}. \tag{6}$$

Based on all the above, the *transmission policy* we adopt can be described as follows:

- If an ACK is received, the next packet is sent in the next TTI.
- If a NACK is received, the same packet is retransmitted again but after a waiting time equal to x.
- A packet can be transmitted a maximum number of times equal to $K(x)$; i.e. the maximum number of allowable retransmissions is $K(x) - 1$.

The approach to find x, and consequently $K(x)$, will be presented in the subsequent section.

Finally, some remarks on the choice and the importance of the above transmission policy are in order.

1) Recall that we work under the assumption of a block fading channel where the channel varies independently between two transmissions. Under such an assumption, it can be noticed that a system where a retransmission is always done immediately after the reception of a NACK can yield performance similar to the system we adopt, if equal max number of transmissions are always considered for both systems. However, the importance of our proposed transmission policy is the flexibility it provides in terms of scheduling, due to the waiting period (x) after the reception of a NACK and before retransmitting the packet. This flexibility is for instance important for scheduling URLLC services with very tight latency budgets. As a simple example, suppose there is a URLLC service with a latency budget of 10 ms. For such a service, and under some specific settings, we can afford to wait before retransmitting a packet in case of a NACK. If meanwhile there is a URLLC service with a more tight latency budget (e.g. 1 ms), then this service can be scheduled during the waiting period.

2) From another point of view, suppose that the coherence time is greater than the RTT, in which case the channel between two transmissions may not change. Since the reception of a NACK generally implies a bad channel quality, waiting for the channel variation before retransmitting the packet can be beneficial in this case, because we may have a better channel quality (i.e. diversity). Modeling and analyzing such a system is a very difficult task. This explains the choice of our adopted system, which can be seen as a special case of the system described before where we always wait for the channel variation before a retransmission.

III. Proposed Scheme and Performance Analysis

In this section, we first provide an explicit formula for the packet loss probability. Then, the average number of transmissions is derived. Finally, joint retransmission policy and link adaptation scheme is proposed.

A. Packet Loss Probability

Define $q(k, m)$ to be the probability that the k first transmissions of the same packet fail given that MCS level m is used. This probability can be expressed as follows

$$q(k, m) = \mathbb{P}\left\{\text{NACK}_1, \ldots, \text{NACK}_{k-1}, \text{NACK}_k \mid m\right\}. \tag{7}$$

As shown earlier, given the instantaneous SNRs, the packet error probability for the ith transmission is $f_m(\sum_{j=1}^{i} \gamma_j)$ if mode m is used. Thus, probability $q(k, m)$ is the average of $\prod_{i=1}^{k} f_m(\sum_{j=1}^{i} \gamma_j)$ over all the SNR values

$$q(k, m) = \int_0^\infty \ldots \int_0^\infty \Bigg[f_m(\gamma_1) \ldots f_m\Bigg(\sum_{j=1}^{k} \gamma_j\Bigg) \times p(\gamma_1) \ldots p(\gamma_k) \Bigg] d\gamma_1 \ldots d\gamma_k. \tag{8}$$

Using the result in [8], we have

$$q(k, m) = \Gamma_1(k, \eta) + \exp(-\eta) \sum_{i=0}^{k-1} \frac{\eta^i}{i!} \prod_{j=1}^{k-i} \frac{1}{1 + j g_m \bar{\gamma}}, \tag{9}$$

where $\eta = \gamma_{\text{S},m}/\bar{\gamma}$ and $\Gamma_1(k, \eta)$ is the regularized lower incomplete Gamma function defined as

$$\Gamma_1(k, \eta) = \frac{1}{(k-1)!} \int_0^\eta t^{k-1} e^{-t} dt. \tag{10}$$

Define $p(k, m)$ to be the probability to successfully receive a packet in exactly k transmissions given that mode m is used. For $k = 1, \ldots, K(x) - 1$, we have

$$p(k, m) = \mathbb{P}\left\{\text{NACK}_1, \ldots, \text{NACK}_{k-1}, \text{ACK}_k \mid m\right\}. \tag{11}$$

Using the fact that

$$\begin{aligned} \mathbb{P}\left\{\text{NACK}_1, \ldots, \text{NACK}_{k-1}\right\} = \\ \mathbb{P}\left\{\text{NACK}_1, \ldots, \text{NACK}_{k-1}, \text{NACK}_k\right\} + \\ \mathbb{P}\left\{\text{NACK}_1, \ldots, \text{NACK}_{k-1}, \text{ACK}_k\right\}, \end{aligned} \tag{12}$$

for $k = 1, \ldots, K(x) - 1$ we can write

$$p(k, m) = q(k-1, m) - q(k, m). \tag{13}$$

At the final round $K(x)$, even in the case of a NACK, the transmitter moves to the next packet, so we have

$$p(K(x), m) = 1 - \sum_{k=1}^{K(x)-1} p(k, m). \tag{14}$$

Under mode m, the Packet Error Rate (PER) is defined as the probability that after $K(x)$ transmissions the packet is still not received correctly given that mode m is used. This probability is nothing but $q(K(x), m)$, and it can be written as follows

$$\begin{aligned} &q(K(x), m) = \\ &\mathbb{P}\left\{\text{NACK}_1, \ldots, \text{NACK}_{K(x)} \mid m\right\} = \\ &\Gamma_{\text{l}}(K(x), \eta) + \exp(-\eta) \sum_{i=0}^{K(x)-1} \frac{\eta^i}{i!} \prod_{j=1}^{K(x)-i} \frac{1}{1 + j g_m \bar{\gamma}}, \end{aligned} \tag{15}$$

where the second equality results from (9).

It is worth noting that the PER can be expressed as a function of $p(k, m)$ as $q(K(x), m) = 1 - \sum_{k=1}^{K(x)} p(k, m)$.

B. Average Number of Transmissions

Let $\bar{K}(x, m)$ denote the average number of transmissions of the same packet given that mode m is chosen and that x is the waiting time (defined earlier). Recalling that the maximum number of transmissions is $K(x)$, in this case we have

$$\begin{aligned} &\bar{K}(x, m) = \\ &\sum_{k=1}^{K(x)} k\, \mathbb{P}\left\{\text{packet succesfully received in } k \text{ transmissions} \mid m\right\} \\ &= \sum_{k=1}^{K(x)} k\, p(k, m). \end{aligned} \tag{16}$$

Combining (13), (14) and (16), we can express $\bar{K}(x, m)$ as a function of $q(k, m)$ as

$$\bar{K}(x, m) = 1 + \sum_{k=1}^{K(x)-1} q(k, m). \tag{17}$$

Using (9) and (17), we can re-write $\bar{K}(x, m)$ as

$$\begin{aligned} &\bar{K}(x, m) = \\ &1 + \sum_{k=1}^{K(x)-1} \Gamma_{\text{l}}(k, \eta) + \exp(-\eta) \sum_{i=0}^{k-1} \frac{\eta^i}{i!} \prod_{j=1}^{k-i} \frac{1}{1 + j g_m \bar{\gamma}}. \end{aligned} \tag{18}$$

C. Optimal Joint Retransmission Policy and Link Adaptation

As alluded earlier, the main goal of this work is to minimize the resources occupied by every packet given a latency budget, T_{lat}, and a PER upper bound, θ; the PER bound ensures a certain level of reliability for the system. Expressed differently, our objective is to maximize the spectral efficiency for a given URLLC service. This will be done by finding the optimal waiting time x and MCS level m. The considered optimization problem can then be seen as a joint retransmission policy and link adaptation scheme.

Let $S_{\text{e}}(x, m)$ be the *spectral efficiency* metric (in units of bits/symbol), which is defined as follows

$$S_{\text{e}}(x, m) = (1 - q(K(x), m)) \frac{\log_2(\alpha(m))\beta(m)}{\bar{K}(x, m)}, \tag{19}$$

where $\alpha(m)$ and $\beta(m)$ are the constellation size and coding rate, respectively, as functions of mode m. Recall that $\bar{K}(x, m)$ is the average number of (HARQ) transmissions per packet and $q(K(x), m)$ is the PER (given m and $K(x)$).

Our objective is to optimize $S_{\text{e}}(x, m)$ under the considered constraints of latency and reliability (measured using PER). The corresponding optimization problem can be written as follows

$$\begin{aligned} \underset{x,\, m}{\text{maximize}} \quad & S_{\text{e}}(x, m) && \text{(20a)} \\ \text{subject to} \quad & q(K(x), m) \leq \theta, && \text{(20b)} \\ & 0 \leq x \leq x_{\max}, && \text{(20c)} \end{aligned}$$

where (20b) and (20c) represent the reliability and latency constraints, respectively. We next provide a simple but important result that will help us solving the optimization problem.

Proposition 1. *With respect to x, the average number of transmissions, $\bar{K}(x, m)$, is a decreasing function whereas the PER, $q(K(x), m)$, is an increasing function.*

Proof. From the expressions of $q(K(x), m)$ and $\bar{K}(x, m)$ given in (15) and (17), respectively, it is plain to see that the above statement holds. Indeed, for both expressions the only dependence on x is in the upper limit of the summation. □

Solving the Optimization Problem: Recall that x and m are discrete variables since x corresponds to an integer number of TTIs and m represents the MCS level. Obviously, finding an analytical solution for the optimization problem is a very difficult task. However, since x and m are discrete variables, and based on the result of Proposition 1, we can solve this problem using simple numerical computations. Specifically, the following simple procedure can be used:

1) We first find all the pairs (x, m) that satisfy both the constraints in (20b) and (20c). The search space here can be reduced by using the fact that $q(K(x), m)$ is an increasing function with respect to x; see Proposition 1.
2) Then, among the above pairs, we find the one that yields the maximum spectral efficiency, $S_{\text{e}}(x, m)$; ties are broken by selecting the pair with the smallest x so that a minimum latency is ensured. The resulting pair is then the solution for the optimization problem.

It is worth mentioning that for the optimization problem to have a (feasible) solution, we must have $q(K(0), m) \leq \theta$ for at least one MCS level, m. This condition results from the fact that $q(K(x), m)$ increases with x.

Table I: Parameters for PER approximation ([5])

	Mode 1	Mode 2	Mode 3	Mode 4	Mode 5	Mode 6
Modulation	BPSK	QPSK	QPSK	16-QAM	16-QAM	64-QAM
Coding Rate	1/2	1/2	3/4	9/16	3/4	3/4
a_m	274.7229	90.2514	67.6181	50.1222	53.3987	35.3508
g_m	7.9932	3.4998	1.6883	0.6644	0.3756	0.0900
$\gamma_{\mathrm{S},m}$(dB)	−1.5331	1.0942	3.9722	7.7021	10.2488	15.9784

IV. Numerical Results and Discussions

In this section we present numerical results. We set the packet length to 1080 bits, and we consider a convolutionally coded modulation [5]. Under this transmission scheme, the PER approximation parameters of (3) are given in Table I, and are the same as in [5]. Regarding the RTT and TTI, we consider $T_{\mathrm{RTT}} = 0.5$ ms and $T_{\mathrm{TTI}} = 0.125$ ms, which is one among many possible timing attributes related to downlink HARQ for a flexible timing approach in 5G [2].

We first consider a URLLC service with latency budget $T_{\mathrm{lat}} = 5$ ms and reliability $\theta = 10^{-4}$, and we set $\bar{\gamma} = 10$ dB. Note that here the latency budget yields a maximum waiting time $x_{\max} = 4.25$ ms, which can be computed using (5).
In Figures 2 and 3 we depict the spectral efficiency (in bits/symbol) and the PER, respectively, for different combinations of waiting time and MCS level (i.e. mode). We point out that a base-10 log scale is used for the PER axis (in Figure 3). It can be easily seen that, for a fixed waiting time x the PER (strictly) increases when using a higher MCS level (m). Also, for a given MCS level, this function increases with x. More specifically, for $x_1 \leq x_2$ we have $q(K(x_1), m) \leq q(K(x_2), m)$; note that $q(K(x_1), m) = q(K(x_2), m)$ if $K(x_1) = K(x_2)$, where we recall that $K(x)$ is given in (4).

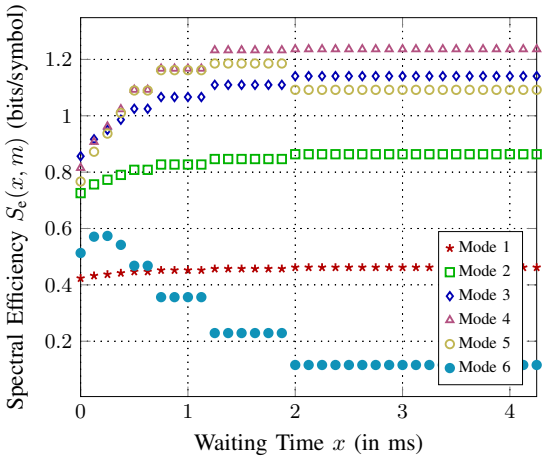

Figure 2: Spectral Efficiency vs Waiting Time x for various MCS levels, with $\bar{\gamma} = 10$ dB and $T_{\mathrm{lat}} = 5$ ms.

In Figure 4, the optimal maximum number of transmissions (resulting from the optimal waiting time), denoted $K_{\mathrm{opt}}(\bar{\gamma})$, and the optimal MCS level, denoted $m_{\mathrm{opt}}(\bar{\gamma})$, are calculated for different values of the average SNR, $\bar{\gamma}$. The same URLLC service as before is considered. Note that the optimal waiting

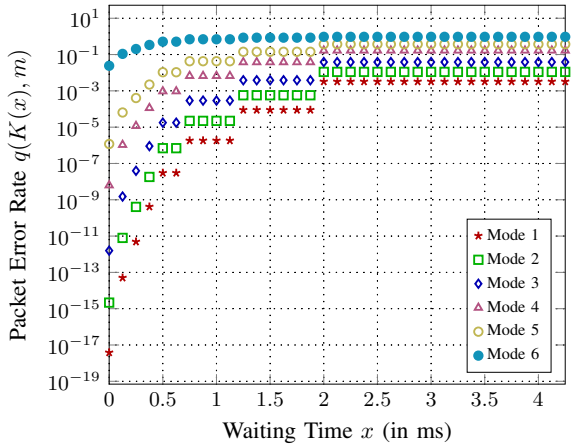

Figure 3: Packet Error Rate (in base-10 log scale) vs Waiting Time x for various MCS levels, with $\bar{\gamma} = 10$ dB and $T_{\mathrm{lat}} = 5$ ms.

time and the optimal MCS level are outputs of the optimization problem. It can be noticed that when $m_{\mathrm{opt}}(\bar{\gamma})$ is the same, $K_{\mathrm{opt}}(\bar{\gamma})$ decreases when $\bar{\gamma}$ increases, which is something expected since the channel quality is, in average, better for greater $\bar{\gamma}$. However, when $m_{\mathrm{opt}}(\bar{\gamma})$ is not the same for instance for two different values of $\bar{\gamma}$, $K_{\mathrm{opt}}(\bar{\gamma})$ does not necessarily decrease with $\bar{\gamma}$. This can be seen for example for $\bar{\gamma} = 10$ dB and 11 dB, where we get $m_{\mathrm{opt}}(10) = 3$, $m_{\mathrm{opt}}(11) = 4$ and $K_{\mathrm{opt}}(10) < K_{\mathrm{opt}}(11)$.

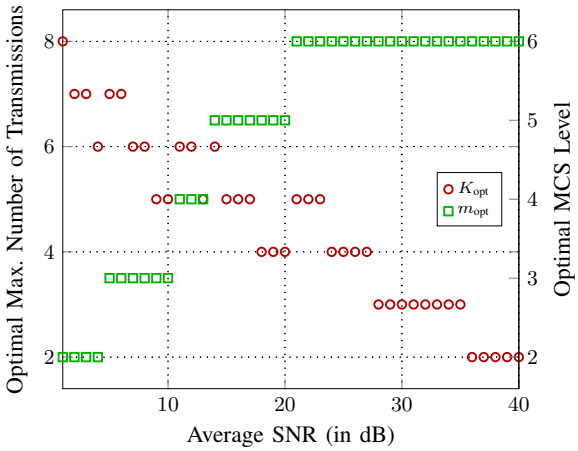

Figure 4: Optimal Max Number of Transmissions and Optimal MCS Level vs Average SNR $\bar{\gamma}$, with $T_{\mathrm{lat}} = 5$ ms and $\theta = 10^{-4}$.

In Figure 5, we compare the optimal spectral efficiency $S_{\mathrm{e}}(x_{\mathrm{opt}}(\bar{\gamma}), m_{\mathrm{opt}}(\bar{\gamma}))$, which is an output of the optimization

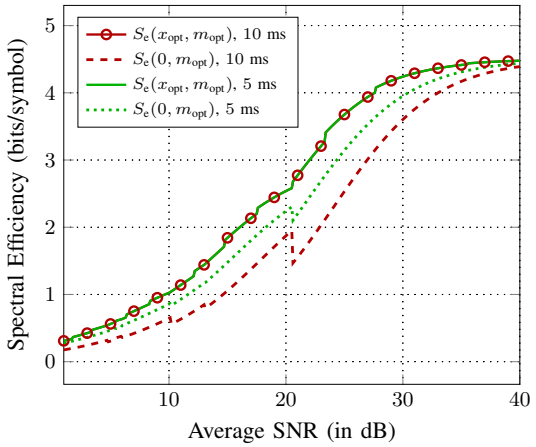

Figure 5: Optimal Spectral Efficiency $S_e(x_{opt}, m_{opt})$ and $S_e(0, m_{opt})$ vs Average SNR for various latency budgets, with $\theta = 10^{-4}$.

problem, with the spectral efficiency of the case where we consider $m = m_{opt}(\bar{\gamma})$ (i.e. with link adaptation) and $x = 0$ as MCS level and waiting time, respectively. We point out that $x = 0$ means that the maximum number of transmissions is equal to $K(0)$, which is the largest possible value of $K(x)$ given a certain latency budget. The spectral efficiencies $S_e(x_{opt}(\bar{\gamma}), m_{opt}(\bar{\gamma}))$ and $S_e(0, m_{opt}(\bar{\gamma}))$ are illustrated for different values of average SNR, $\bar{\gamma}$, and latency budget, T_{lat}. The first thing to note in Figure 5 is that $S_e(x_{opt}(\bar{\gamma}), m_{opt}(\bar{\gamma}))$ increases with $\bar{\gamma}$, which is expected since for each $\bar{\gamma}$ we find the maximum $S_e(x, m)$. Also, it can be noticed that $S_e(x_{opt}(\bar{\gamma}), m_{opt}(\bar{\gamma}))$ is almost the same independently of T_{lat}. Furthermore, unlike $S_e(x_{opt}(\bar{\gamma}), m_{opt}(\bar{\gamma}))$, $S_e(0, m_{opt}(\bar{\gamma}))$ does not always increase with $\bar{\gamma}$. This results from the fact that $S_e(0, m_{opt}(\bar{\gamma}))$ is not the maximum spectral efficiency for each value of $\bar{\gamma}$. Moreover, the gain in terms of spectral efficiency is higher for greater T_{lat}; for instance, at $\bar{\gamma} \approx 21$ dB we have $S_e(x_{opt}(\bar{\gamma}), m_{opt}(\bar{\gamma})) \approx 2.5$ bits/symbol, and $S_e(0, m_{opt}(\bar{\gamma})) = 1.5$ and 2.1 bits/symbol for $T_{lat} = 10$ and 5 ms, respectively.

V. Conclusions

In this paper, we propose a new joint link adaptation and HARQ scheme for URLLC services. A fast-varying channel where the transmitter knows only the average SNR (and not the instantaneous one) is assumed. In order to provide flexibility and transmission diversity, we introduce a waiting time before the retransmission after a NACK is received. We find the optimal waiting time, and consequently the optimal maximum number of transmissions, and the optimal MCS level for each packet to maximize the spectral efficiency while satisfying latency and reliability constraints. We show that the proposed scheme increases the system performance in terms of spectral efficiency, especially when the latency budget is not very tight.

Finally, it is important to note that in this work we are only considering time diversity as a means to improve the system performance. A possible extension of this work would be to combine this diversity with, for instance, frequency diversity or spatial diversity, and to investigate the performance that can be achieved in this case.

References

[1] 3GPP, "Study on scenarios and requirements for next generation access technologies," 3GPP TR 38.913 v14.2.0, Tech. Rep., March 2017.

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

[3] M.-S. Alouini and A. J. Goldsmith, "Adaptive modulation over nakagami fading channels," *Wireless Personal Communications*, vol. 13, no. 1, pp. 119–143, May 2000.

[4] D. Chase, "Code combining - a maximum-likelihood decoding approach for combining an arbitrary number of noisy packets," *IEEE Transactions on Communications*, vol. 33, no. 5, pp. 385–393, May 1985.

[5] Q. Liu, S. Zhou, and G. B. Giannakis, "Cross-layer combining of adaptive modulation and coding with truncated ARQ over wireless links," *IEEE Transactions on Wireless Communications*, vol. 3, no. 5, pp. 1746–1755, September 2004.

[6] C. G. Kang, S. H. Park, and J. W. Kim, "Design of adaptive modulation and coding scheme for truncated hybrid ARQ," *Wireless Personal Communications*, vol. 53, no. 2, pp. 269–280, April 2010.

[7] H. Zheng and H. Viswanathan, "Optimizing the ARQ performance in downlink packet data systems with scheduling," *IEEE Transactions on Wireless Communications*, vol. 4, no. 2, pp. 495–506, March 2005.

[8] X. Lagrange, "Performance analysis of HARQ protocols with link adaptation on fading channels," *Annales des Télécommunications*, vol. 66, pp. 695–705, 2011.

[9] P. Wu and N. Jindal, "Performance of hybrid-ARQ in block-fading channels: A fixed outage probability analysis," *IEEE Transactions on Communications*, vol. 58, no. 4, pp. 1129–1141, April 2010.