POLITECNICO DI MILANO

Master of Science in Telecommunication Engineering

Department of Electronics, Information and Bioengineering

Hybrid Retransmission Scheme for QoS-defined 5G Ultra-Reliable Low-Latency Communications

Supervisor: Prof. Maurizio MAGARINI

Co - supervisors: Dr. Silvio MANDELLI

Prof. Luca REGGIANI

Thesis of:

Luca Buccheri, ID 850817

Academic Year 2016-2017

Alla mia famiglia,
grazie di tutto.

Contents

Sommario

Una delle principali sfide nelle reti 5G di nuova generazione é quella di fornire comunicazioni ultra affidabili a bassa latenza (in inglese URLLC).
Grazie ai recenti progressi nella teoria dell'informazione sui principi che governano le trasmissioni con pacchetti brevi, é stato evidenziato che, per la dimensione del pacchetto corto tipico di comunicazioni URLLC, il raggiungimento di una maggiore affidabilitá si paga con un ridotto limite massimo di rate di trasmissione, che avrá cosí una minore efficienza spettrale.
Quindi le ritrasmissioni sono utilizzate in LTE e programmate per l'utilizzo in 5G, in modo da raggiungere l'affidabilitá con un migliore consumo di risorse, al prezzo di una maggiore latenza di pacchetto.
In questa tesi viene proposto il design di un link URLLC per testare le strategie di ritrasmissione considerate in letteratura. L'analisi dei compromessi e limitazioni risiede nella richiesta troppo aggressiva di risorse wireless per ottenere prestazioni URLLC, o la non conformitá dei requisiti URLLC in un regime conservativo di allocazione risorse.
Viene quindi proposta una nuova strategia il cui scopo é di soddisfare i requisiti URLLC e contemporaneamente ridurre al minimo il consumo di risorse.
Gli schemi di ritrasmissione sono valutati con simulazioni per evidenziare i vantaggi dello schema proposto, fornendo al contempo approfondimenti sulle prestazioni delle tecniche HARQ negli scenari URLLC.

Abstract

One of the key challenges in next generation 5G networks is to deliver Ultra-Reliable Low-Latency Communications (URLLC).

Recent advances in information theory about principles that govern short packet transmissions pointed out that, for the URLLC typical short packet dimension, achieving higher reliabilities comes at the price of a lower maximum achievable rate, thus lower spectral efficiency.

Hence retransmissions are used in LTE and planned for 5G, in order to achieve reliability with a better resource consumption, at the price of increased packet latency.

In this thesis it is proposed the design of a URLLC link to test retransmission strategies considered in the literature. The analysis of their tradeoffs and limitations are the too aggressive demand of wireless resource to achieve URLLC performances, or the non compliance of URLLC requirements in a conservative regime of resource delivery.

Then it is proposed a novel strategy whose purpose is matching the URLLC requirements and concurrently minimizing the resource consumption. The retransmission schemes are evaluated with simulations to enlighten the advantages of the proposed scheme, while providing insights into the performance of HARQ techniques in URLLC scenarios.

Chapter 1

Introduction

1.1 5G: more than the evolution of LTE

Wireless communication is becoming a commodity, like electricity or water [1], and the fifth generation (5G) of mobile networks will enable a large class of new services, radically transforming the information and communication technology (ICT) field.

At present, the wireless environment is dominated by wireless technologies for local high-speed use, such as Wi-Fi and Bluetooth, and cellular technologies for wide-area use, such as fourth generation Long Term Evolution (4G LTE).

The principal contribution of LTE was the enhancement of the mobile internet access, established with 3G with the Mobile Broad Band (MBB) operating mode. Video, voice and other applications are delivered with peaks in the data rate of several hundreds of MB/s.

The 5G infrastructure is not planned to be just a release of current network generations, but it is conceived as the nervous system of a digital society, where the 5G operating system (OS) is the distributed software that runs on

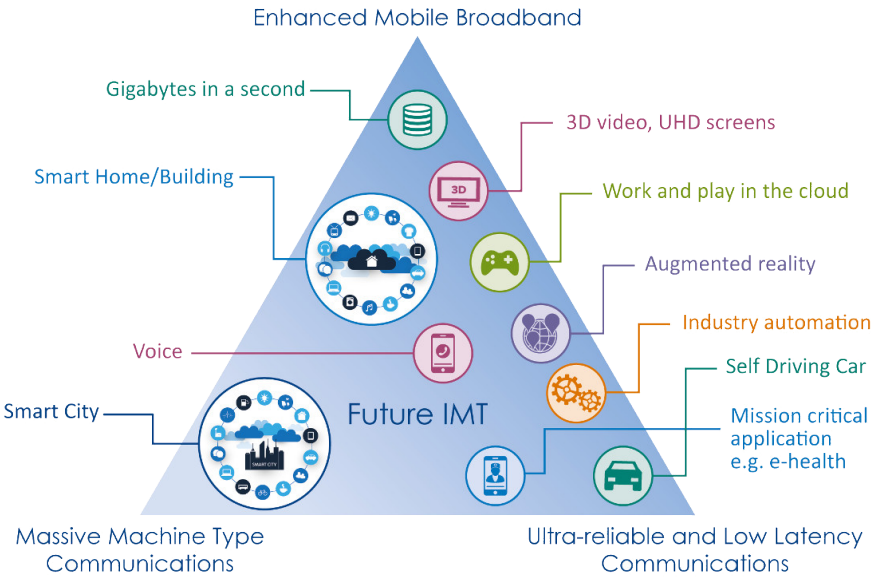

Figure 1.1: Three dimensions to performance improvements with usage scenarios for 2020 and beyond.

top and its implementation is oriented to deal with cognitive objects, such as sensors, machines, robots, drones.

1.1.1 Three main directions for 5G

Figure 1.1 shows the three main directions pushed by the European research community for 5G, Horizon 2020 [2]; the improves with respect to 4G are in terms of user data rates and especially latency, reliability, speed, mobility and spectral efficiency.

Here is the list of the three use cases for 5G communications.

- *enhanced Mobile Broadband (eMBB)*: The evolution of MBB is conceived to have peak data rates greater than 10 Gbps and reduced latencies. To achieve these requirements there will be an extension

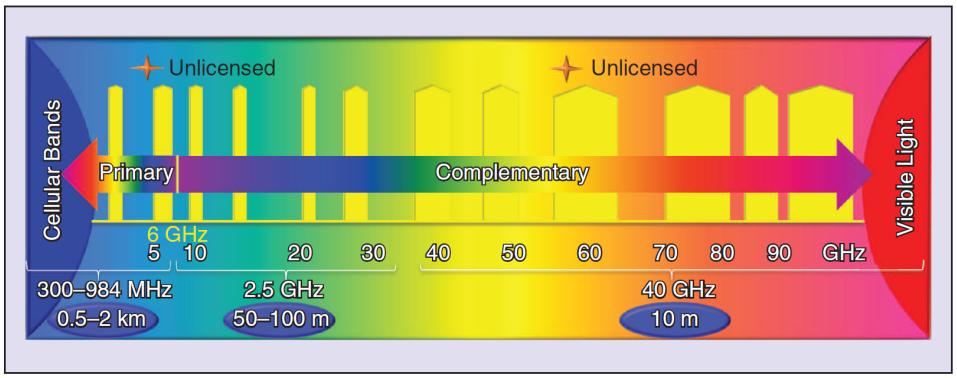

Figure 1.2: Spectra availability for 5G deployment [2].

to the usual cellular spectrum utilization (see Figure 1.2) with carrier frequencies below 6 GHz: 5G will use in addition carriers from 10 to 30 and from 40 to 90 GHz, with a lower coverage than cellular bands, but with an expected available bandwidth 2.5 and 40 GHz respectively.

- *Ultra-Reliable and Low-Latency Communication (URLLC)*: It is the most innovative feature of 5G, as it will be used for *mission critical communication*, like reliable remote action with robots or coordination among vehicles.

  The target is to define for that type of communications a wireless link where the connectivity is guaranteed for more than 99.999 % of the time; an example is Industry 4.0, where different parts of an object or a machine need not to be physically attached, as long as they can use mission-critical ultra-reliable links to work in concert toward accomplishing a production task.

  The ultra-reliability joint with strict latency requirements < 1 ms may allow critical applications that involves danger for the human life; future examples are traffic-safety-related communications among vehicles in autonomous car driving [3] and health applications as remote

surgery and remote monitoring of patients through wearable devices.

- *Massive M2M (Machine-to-Machine) Communication (MM2MC)*: This mode already emerges as an extension of the 4G LTE systems and refers to support of a massive number (tens of thousands) machines in a given area which transmits small data blocks sporadically. This is relevant for large-scale distributed systems, e.g., smart grid and metering sensors, and the network is designed to simultaneously serve the aggregated traffic and to minimize the device energy consumption, so that the battery lifetime is extended up to 10 years.

In the thesis work the focus will be on URLLC, where ultra-reliability and low latency have to be matched together in order to comply with strict requirements of delay-critical services.

1.1.2 5G Operating Regions

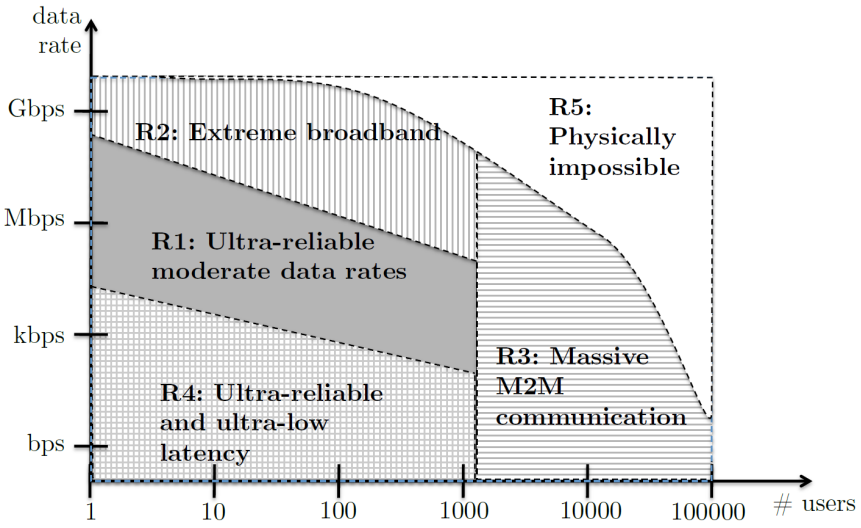

Figure 1.3: Operating regions of the 5G wireless systems [4].

In Figure 1.3 are illustrated the expected operating regions of the 5G

wireless systems in function of the data rate and the number of connected devices in a service area, where we can collocate the operating mode defined in the previous subsection. The numbers are not precise and only depict the order of magnitude.

In this figure, LTE and Wi-Fi are collocated in region **R1**, whose shape outlines that the data rate of each user decreases as the user population increases.

The region **R2** features extreme broadband rates, where we could collocate the eMBB traffic, and region **R1** could be dedicated to moderate rates delivered, for some services, with *ultra-reliability* or with more stringent latency requirements than today's communication systems.

Contrary to the broadband regime, the region **R3** and **R4** feature smaller data rates and the devices send information through short messages.

In region **R3** these short messages are coming from a large number of machines/sensors, supporting the Machine-to-Machine Communication mode.

In region **R4** we collocate the URLLC systems, and short packets are the first step to deal with a stringent latency requirement, e.g. 1 ms, and with a low probability that the latency doesn't exceed it, in the order of $1 - 10^{-x}$, with $x \geq 5$.

The short-packet paradigm is still a research field in Information Theory where the efforts are in the definition of efficient channel coding techniques for such short blocks. In this thesis it is explained a recent information-theoretic result [5] that defines an upper bound on the feasible channel rate in short-packet communication.

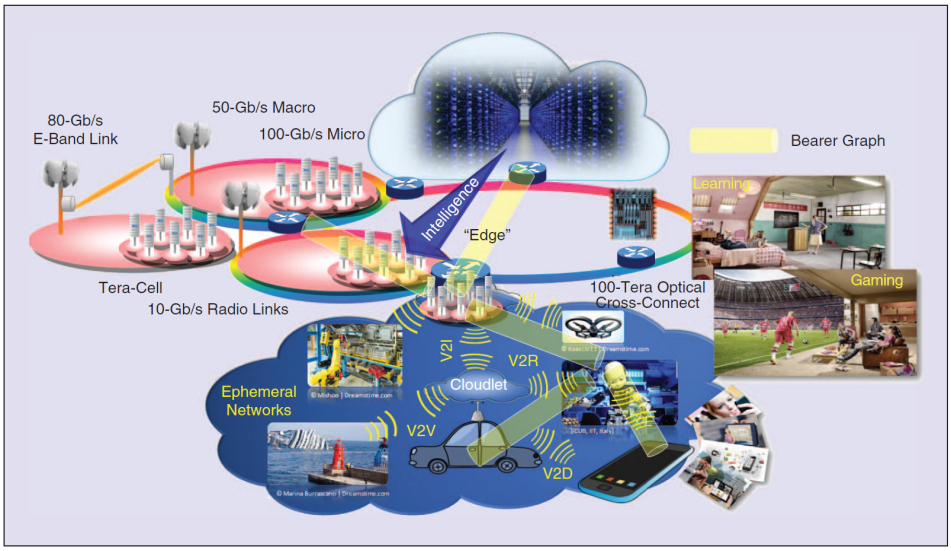

Figure 1.4: Network and services capabilities at horizon 2020 and beyond [2].

1.1.3 An information-oriented society

The vision for 5G [2] is to realize the concept of a bearer graph, where different communication paradigms as vehicle-to-vehicle (V2V) and vehicle-to-robot (V2R) are operating simultaneously with reliable and delay-aware wireless links.

In David Soldani opinion, head of 5G Technology in Nokia, ICT will find application to generate new services at a low cost for improving the quality of life, meaningful to what really matters for the society we are living.

Moreover the socioeconomic and business implications of this vision are huge; new different jobs from today will be created, and its number will be far greater than the jobs driven by human labor that are lost due to automation.

Today the main control variables of our economy are still human intelligence, attention, effort, time, and we are witnessing the migration of industries to regions where labor costs are lower.

The 5G vision can reshape this economy equation by taking over many cognitive tasks that human can or cannot do and improving quality of life. The belief is that 5G infrastructure can be a catalyst of the second machine age [1] and empower intelligent machines to flood the landscape of new jobs.
Socioeconomic benefits can include the reduction of human efforts in jobs subject to computerization and robotization, with lower operating costs and conditions, higher local production, worker safety and product quality.
Moreover the network operator will play a big role in this transformation, leading to a new business model, in a much stronger position in the competition with the Over-The-Top players.

1.2 Retransmissions in URLLC and analysis

In this thesis we focus on reliability and latency of a communication between the base station and a single device, in specific on the physical layer retransmission strategies and their possible enhancement.
The basic concept of retransmission is that, in a communication between a transmitter (TX) and a receiver (RX), whenever RX detects an error in the transmission, sends a feedback signaling to the transmitter and requests a retransmission of the same packet.
If you consider a probability that a block of data is sent with errors, defined as BLock Error Rate (BLER), the scope of retransmission is to lower the probability of needing further retransmissions.
This concept is better explained with an example in Figure 1.5 on the Automatic Repeat reQuest (ARQ) protocol, where we have control messages to confirm the ACKnowledge (ACK) of the information or the Non-ACKnowledge (NACK), the single block of data has a failed transmission probability ϵ, and the RX discards every corrupted message.

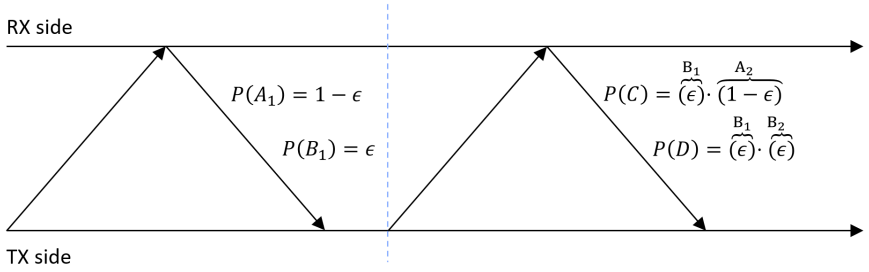

Figure 1.5: ARQ scheme and error probabilities. $A_{1,2}$ and $B_{1,2}$ represent the probability of ACK and NACK respectively for the first and the second transmission.

1.2.1 Automatic Repeat Request for Ultra-Reliability

If we consider a single transmission, there would be a universal set of the possible events

$$U = \{A_1, B_1\}, \tag{1.1}$$

where we have that, in general, A_i and B_i are respectively the reception of an ACK or a NACK for the ith transmission.

Here the BLER coincides with ϵ and the reliability is $1 - \epsilon$, as the successful transmission is the complementary event of the failed one.

We write now the universal set when the first transmission fails and a retransmission is required; U is written as

$$U = \{A_1, A_2 \cap B_1, B_2 \cap B_1\} = \{A_1, C, D\}, \tag{1.2}$$

where C and D are respectively the reception of an ACK and a NACK after a first failure, composed by events A_2 and B_2 joint to event B_1.

A_2 and B_2 have the same probability as A_1 and B_1, and the joint probabilities, given that B_i is independent from A_{i+1} or B_{i+1}, are written as the prod-

uct of the probabilities, $P(C) = P(B_1) \cdot P(A_2)$ and $P(D) = P(B_1) \cdot P(B_2)$. The reliability is given by the sum of the probabilities of all the successful events, in this case A_1 and C, so that $P(A_1) + P(C) = 1 - \epsilon^2$. Equivalently, reliability is the probability of the complementary event of a failure at the second transmission, $1 - P(D) = 1 - \epsilon^2$.
We can extend this concept to n retransmission, obtaining a reliability $1 - e^n$ with n total transmissions.

1.2.2 Contribution to existing PHY strategies

In this thesis we will discuss about the evolution of the ARQ method, called Hybrid ARQ (HARQ); the two main methods studied here, Chase Combining (CC) and Incremental Redundancy (IR), have better performance than ARQ as they store the noisy blocks of data in case of error and combine them with the retransmitted ones.
In this way the receiver exploits the residual information of the corrupted packets instead of depending just on retransmission to decode successfully.
The other topic touched by this work is the study of *retransmission schemes*, that define the type of control messages sent and the frequency of retransmission. The studied techniques are:

- NACK-based scheme: a retransmission is scheduled and sent only if the transmitter receives a NACK from the receiver or has not received an ACK for a given amount of time, due to control channel errors.

- BLIND scheme or GRANT-FREE: the terminal aggressively sends transmissions at every available transmission chance, until it explicitly receives an ACK from the receiver.

- HYBRID scheme or MULTI-MODE: the system uses a NACK-based

scheme for the first transmissions and it is designed to switch to a more aggressive scheme to prevent latency from exceeding a defined *latency budget* of URLLC communication.

While the first two are state of the art techniques, the latter is the novelty we get from this thesis work.

1.2.3 Thesis outline

The thesis will be organized as following.

Chapter 2 will be dedicated to Link Adaptation (LA) or Adaptive Modulation and Coding (AMC) and to the information-theoretic concept of Maximal Coding Rate (MCR).

Chapter 3 will be dedicated to the characterization of the wireless fading channel, by discussing about mobility of the device, coherence time, Multi-Input-Multi-Output (MIMO) transmission and beamforming, time evolution of the channel.

In chapter 4 we enter in the details of CC-HARQ and IR-HARQ techniques and the model adopted to define them, and we discuss about the behaviour of the state of the art retransmission schemes.

The HYBRID scheme and its implementation together with Link Adaptation is discussed in chapter 5.

Finally in Chapter 6 retransmission schemes are compared by means of simulations, highlighting the flexibility of the HYBRID scheme compared to the others in terms of both resource consumption and latency.

The scope, as it will be shown by simulation results, is to achieve the QoS of the transmission even when NACK-based strategy fails, but having at the same time a higher spectral efficiency than the BLIND aggressive scheme.

Chapter 2

Link Adaptation and Maximal Coding Rate

Link Adaptation, or Adaptive Coding and Modulation (ACM), is a procedure in wireless communications to match the Modulation and Coding Scheme (MCS) to the link quality, in order to adapt the rate of transmission and send information with a minimum degree of reliability.
In section 2.1 we discuss about Channel Coding and LA in LTE, while in section 2.2 we define the upper bound on the maximum channel rate, when we are in *short packet* paradigm. This limit will be the principal concept for the definition of a Theoretical LA.

2.1 Channel Coding and Link Adaptation in LTE

Channel coding is one of the most important aspects in digital communication systems, which can be considered as the main difference between analog and digital systems, in order to enable error detection and error correction. In LTE Forward Error Correction (FEC) introduces redundancy bits that

are computed from the information bits, which is done either blockwise (so-called block coding) or convolutional, where the coded bit depends not only on the current data bit, but also on the previous bits.

The combination of a single modulation scheme and increasing code rates r, that are the ratio between the information bits and the bits of a codeword, generates almost parallel BLER curves, as in Figure 2.1, which have increasing efficiency. The ACM feature is to operate with a combination of modulation scheme and code rate in order to guarantee at least a *target* BLER, in the example 10%, according to the Signal to Noise Ratio (SNR) of the channel.

In LTE the SNR measurements is associated to a Channel Quality Indica-

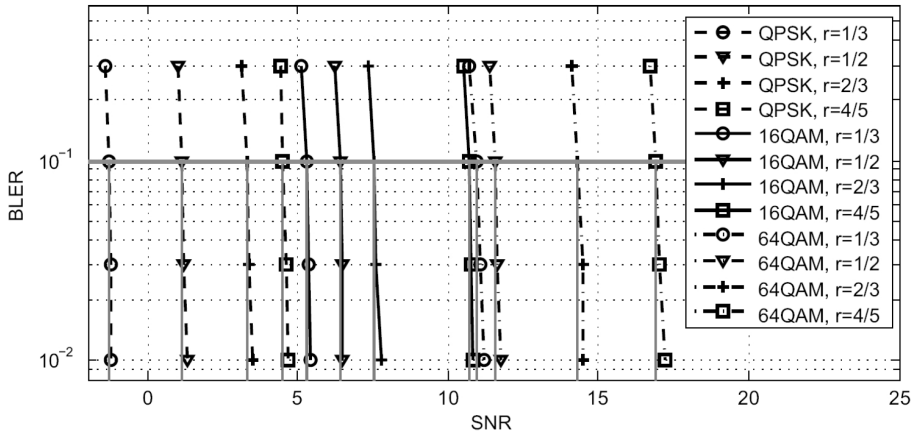

Figure 2.1: BLER versus SNR for various combinations of modulation scheme and code rate [6].

tor (CQI), which considers the receiving capabilities of the UE; for example UE with a receiver of better quality can report better CQI for the same channel quality, and thus can transmit/receive data with higher MCS.

Here in table 2.1 is reported the CQI mapping in a downlink transmission with LTE release 8; it is a look-up table with a connection between the indicator's value reported by the UE device and the selection of a MCS by the

CQI Index	Modulation	Code Rate × 1024	Channel Rate
0	No Transmission		
1	QPSK	78	0.1523
2	QPSK	120	0.2344
3	QPSK	193	0.3770
4	QPSK	308	0.6016
5	QPSK	449	0.8770
6	QPSK	602	1.1758
7	16QAM	378	1.4766
8	16QAM	490	1.9141
9	16QAM	616	2.4063
10	64QAM	466	2.7305
11	64QAM	567	3.3223
12	64QAM	666	3.9023
13	64QAM	772	4.5234
14	64QAM	873	5.1152
15	64QAM	948	5.5547

Table 2.1: Overview of different CQI and the relative MCS choice [7].

eNodeB such that the UE can decode the data with an error rate probability not exceeding 10%.

The choice in the modulation scheme is done between QPSK, 16QAM, 64QAM with three different modulation order, that is the number of bits carried per symbol $[\frac{\text{bits}_{\text{codeword}}}{\text{symbol}}]$, in this case 2, 4 and 6.
The table is arranged in ascending order according to the Channel Rate, expressed as the number of information bits carried per symbol

$$\underbrace{\frac{\text{bit}_{\text{info}}}{\text{symbol}}}_{\text{Channel Rate}} = \underbrace{\frac{\text{bits}_{\text{codeword}}}{\text{symbol}}}_{\text{Modulation Order}} \cdot \underbrace{\frac{\text{bits}_{\text{info}}}{\text{bits}_{\text{codeword}}}}_{\text{Code Rate}}. \tag{2.1}$$

2.2 Maximal Coding Rate

As anticipated also in the Introduction, in URLLC we are dealing with a short packet size, from 32 to 200 bytes [8, 9].

From an information theory point of view, shannon capacity C [10] has

been the usual performance metric to define the largest rate on an AWGN channel with packet error probability ϵ arbitrarily small, by using a pair of encoder/decoder with packet length n arbitrarily large.
URLLC use case will use instead small packets, so that both n and ϵ have to be considered and C become less meaningful to describe the largest channel rate. In the last years, information theory is investigating the performance of these small packets, deriving the concept of Maximal Coding Rate (MCR) R^* [5, 11].
The notation we use is:

1. k as the dimension in bits of the block of information to send,

2. n as the number of symbols, or *channel uses*, to carry the information bits along the channel,

3. R^* as the channel rate determined by the ratio $R^* = k/n$.

In case of AWGN channel with SNR γ, the MCR $R^*(n, \epsilon)$ is expressed as

$$R^*(n, \epsilon) = C(\gamma) - \sqrt{\frac{V(\gamma)}{n}} Q^{-1}(\epsilon) + \mathcal{O}\left(\frac{\log_2 n}{n}\right), \tag{2.2}$$

where $Q(\cdot)$ is the Q-function and V is the so-called *dispersion* of the channel

$$C(\gamma) = \log_2(1 + \gamma), \tag{2.3}$$

$$V(\gamma) = \gamma \frac{(2 + \gamma)}{(1 + \gamma)^2} (\log_2 e)^2. \tag{2.4}$$

What is defined here is a more general metric of traditional information-theoretic metrics, such as capacity C [12] and outage capacity C_ϵ [13].
In facts the outage capacity C_ϵ is defined as the largest rate k/n such that, for every sufficiently large packet length n, there exists an encoder/decoder

pair (f_n, g_n) whose packet error probability does not exceed ϵ. It is possible to obtain this metric from R^* via

$$C_\epsilon = \lim_{n\to\infty} R^*(n, \epsilon). \tag{2.5}$$

C is obtained in similar way from C_ϵ via

$$C = \lim_{\epsilon\to 0} C_\epsilon = \lim_{\epsilon\to 0} \lim_{n\to\infty} R^*(n, \epsilon). \tag{2.6}$$

Both capacity and outage capacity are valid performance metrics for current wireless systems where packet size is typically large, while maximal coding rate R^* can be a better performance metric for URLLC and massive M2M communications. We notice a penalty in Equation (2.2) on C proportional to $\sqrt{V/n}$ and the inverse of the q-function of ϵ, so that a reduction in the BLER target leads to a decrease in the MCR as well.

In [5] the channel dispersion V is defined as the stochastic variability of the channel relative to a deterministic channel with the same capacity, determined by the finite length of the transmission. Also [11] confirms this concept, by defining the communication channel can be thought of as a *bit pipe of randomly varying size*, behaving as a Gaussian random variable with mean C and variance V/n.

In Fig. 2.2 we consider an AWGN channel with SNR $\gamma = 1$, so that $C = 1$, and the MCR is plotted in function of the blocklength n with three values of ϵ.

We modify equation (2.2) by replacing the term $\mathcal{O}\left(\frac{\log_2 n}{n}\right)$ with $\frac{\log_2 n}{2n}$, leading to an approximation, commonly referred as *normal approximation* [11]

$$R^*(n, \epsilon) = C(\gamma) - \sqrt{\frac{V(\gamma)}{n}} Q^{-1}(\epsilon) + \frac{\log_2 n}{2n}. \tag{2.7}$$

We can see that all the curves tend to C for n tending towards infinite, that they incur in an increasing penalty according to a lower ϵ value, and that this last parameter can be made arbitrary small in order to deal with URLLC examples.

The simulation results of this thesis work are obtained using this performance metric, that is optimistic with respect to the utilization of a real pair of encoder/decoder. We believe that this is a good starting point for a future work around the topic of retransmission in URLLC use case.

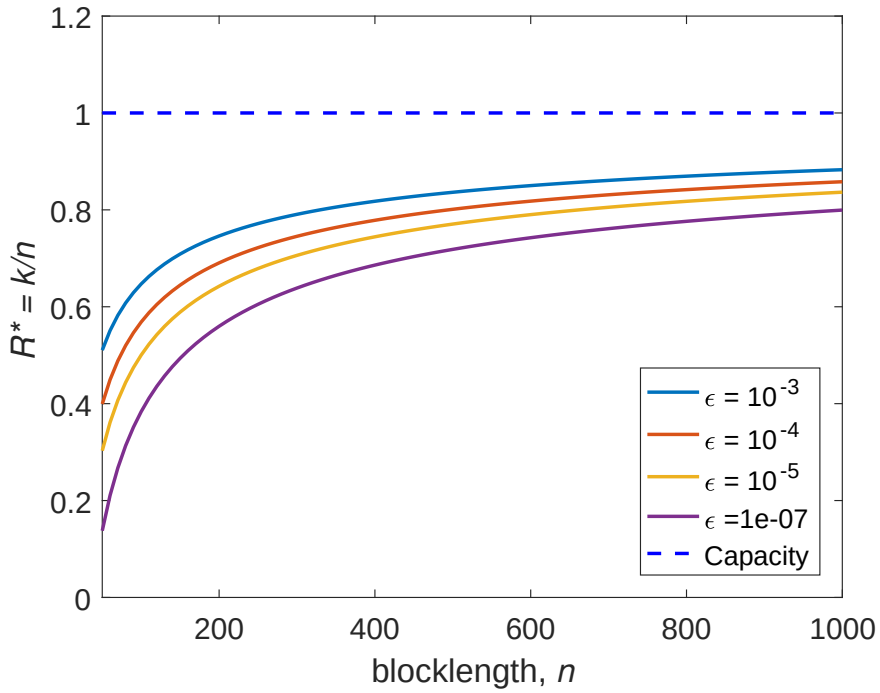

Figure 2.2: Normal approximation on $R^*(n, \epsilon)$ for the AWGN channel with SNR $\gamma = 1$.

Chapter 3

Link Design for URLLC

In this chapter we discuss about the model design of the radio link for URLLC, in order to define the SNR at the receiver and have a mapping between SNR and the channel rate selected by the LA procedure.
Section 3.1 is dedicated to the description of the Rayleigh block-fading model, while in section 3.2 we describe MIMO communication with beamforming at the transmitter and receiver side.
In section 3.3 we discuss about the wireless resource element and the granularity in the allocation of resources to the transmitter.

3.1 Wireless channel: Rayleigh block-fading time-correlated model

In this section a description of the model used to characterize the mobile wireless channel is presented. Some approximations on the frequency selectivity and the structure of the channel are done, however the simulation results based on this model can be considered meaningful.
A defining characteristic of the mobile wireless channel is the variations of

the channel strength over time and frequency. There are roughly two types of variations, the *large-scale fading* and the *small-scale fading*.
The first one is due to path loss of signal, as a function of distance, and shadowing by large objects.
The second one depends on constructive and destructive interference of multiple signal paths between transmitter and receiver.
While large-scale fading is more relevant to issues such as cell-site planning, we focus on the small-scale fading or *fast fading*, more relevant to the design of reliable communication systems.
In the next subsections we recall the necessary contents for the *coherence time* definition, the block-fading model as in [11] and we recall how to define the channel filter taps.

3.1.1 Rayleigh variable and Coherence Time

We start the topic from the discrete-time baseband channel model discussed in [14], such that we have

$$y[m] = \sum_{l} h_l[m]x[m-l] + w[m], \tag{3.1}$$

where $w[m]$ is the low-pass filtered Gaussian noise, $x[m]$ is the transmitted symbol, $h[m]$ is the channel impulse response and $y[m]$ is the output.
This model is the first basis to define the stochastic properties of the single channel filter tap and its autocorrelation in time, by considering also the concept of Coherence Time.
The variable behavior of the channel depends on the constructive and destructive interference of the multiple paths of the signal, due to multiple reflections experienced in a wireless environment.

Rayleigh random variable

In cellular environment the *Rayleigh fading* model is adopted for its simplicity, and it starts with the assumption that there are a large number of statistically independent scattered paths in the delay window corresponding to a single tap.

Each path attenuates the signal in a different way and changes the signal phase such that the latter can be considered as random variable, statistically independent from the others, drawn from a uniform distribution between 0 and π.

Hence, each tap $h_l[m]$ is the sum of a large number of such small independent circular symmetric random variables. It follows that $\Re(h_l[m])$ is the sum of many small independent real random variables, and so by the Central Limit Theorem, it can be modeled as a zero-mean Gaussian random variable.

Similarly, because of the uniform phase, $\Re(h_l[m]\mathrm{e}^{\mathrm{j}\phi})$ is Gaussian with the same variance for any fixed ϕ. This assures us that $h_l[m]$ is a circular Gaussian symmetric variable $\mathcal{CN}(0, \sigma_l^2)$, and its magnitude $|h_l[m]|$ is a *Rayleigh* random variable with density

$$\boxed{\frac{x}{\sigma_l^2}\exp\left\{\frac{-x^2}{2\sigma_l^2}\right\}, \quad x \geq 0.} \tag{3.2}$$

The squared magnitude $|h_l[m]|^2$ is exponentially distributed with density

$$\boxed{\frac{1}{\sigma_l^2}\exp\left\{\frac{-x}{\sigma_l^2}\right\}, \quad x \geq 0.} \tag{3.3}$$

Autocorrelation function and Coherence Time

The following step is to characterize the variation in time of the Rayleigh variable; in order to do that, we introduce the auto-correlation function of

each tap-gain of the filter $R_l[n]$, defined as

$$R_l[n] = \mathbb{E}\left\{h_l^*[m]h_l[m+n]\right\}. \tag{3.4}$$

Then we define the Coherence Time T_c as the time after which the process is decorrelated with itself to $\frac{1}{e}$.

As we have the sample autocorrelation, with an enough small sampling time, we have $\overline{n}$ as the number of samples after which the energy of the single tap is divided by e. If $R_l[0] = \sigma_l^2$, we obtain

$$\frac{|R_l[\overline{n}]|}{R_l[0]} = \frac{1}{e}. \tag{3.5}$$

The variance σ_l^2 in this model is unitary, as we want to consider the channel as a transfer function without any gain or attenuation, so that we may consider separately the multi-path effect and the attenuation due to propagation in a specific wireless environment.

Moreover we want to consider the relation between the decorellation and the coherence time, which is in turn inversely related to the maximum velocity of a mobile user. We expand this concept in the next example with a moving device, a transmitting antenna and a reflecting wall.

Doppler Shift with reflecting wall and moving antenna

Consider Fig. 3.1 with a fixed antenna transmitting the sinusoid $\cos 2\pi ft$, a person with device moving away from the transmitter with velocity v, an indefinitely large wall such to have a perfect reflection of the electromagnetic wave.

Considering the receive antenna at location r_0 at time 0, so that $r(t) = r_0 + vt$, we write the direct and reflected contribution to the electric field at

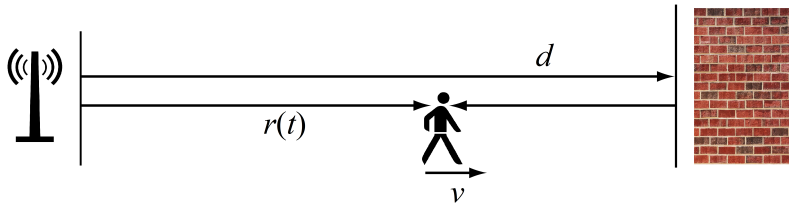

Figure 3.1: Illustration of a direct path and a reflected path, with a transmit antenna in the left and a reflecting wall in the right.

the receiver

$$E_r(f,t) = \frac{\alpha \cos 2\pi f[(1-v/c)t - r_0/c]}{r_0 + vt} - \frac{\alpha cos 2\pi f[(1+v/c)t + (r_0 - 2d)/c]}{2d - r_0 - vt}, \tag{3.6}$$

where we assume the same antenna gain α for both waves.

Both direct and reflected waves are sinusoids at frequency $f(1 \pm v/c)$, experiencing a maximum *doppler shift* $f_v := \pm fv/c$, and since the coherence time is the time-domain dual of the doppler shift and it is also related to the autocorrelation of the Rayleigh process, we have somehow a relationship between the mobility of the user and the correlation between channel samples.

Let us consider the popular rule-of-thumb formula used in practice [15]

$$T_c \simeq \frac{0.423}{f_v}. \tag{3.7}$$

If we have a carrier frequency $f_c = 2$ GHz, a person moving with $v = 5$ Km/h, we have a doppler shift $f_v = f_c v/c = 9.26$ Hz and a coherence time $T_c = 45.7$ ms.

Rayleigh time-correlated variable generation

What we want to do now is to emulate the extraction of 2 samples from a time-correlated Rayleigh process by considering that we want to define the decorrelation after one Time Transmission Interval (TTI), a parameter determining the time slot dedicated for a single block of transmission.
Let's take a generic correlation term ρ and two independent extractions $h_1^{'}$ and $h_2^{'}$ drawn from a zero-mean white circular Gaussian process with unitary variance $\mathcal{CN}(0,1)$. If we define

$$\begin{aligned} h_1 &= h_1^{'} \\ h_2 &= h_1\rho + h_2^{'}\sqrt{1-\rho^2}, \end{aligned}$$

and we evaluate the cross-correlation between h_1 and h_2, we end up with

$$\mathbb{E}\left\{h_1^* h_2\right\} = \mathbb{E}\left\{h_1^{'*} h_1^{'}\rho\right\} + \mathbb{E}\left\{h_1^{'*} h_2^{'}\sqrt{1-\rho^2}\right\} = \rho \tag{3.8}$$

as $h_1^{'}$ and $h_2^{'}$ are independent extractions. If we extract from the same process $h_3^{'}$ and we want to define h_3 correlated with h_2 with the same term ρ, we have

$$h_3 = h_2\rho + h_3^{'}\sqrt{1-\rho^2}.$$

The cross-correlation between h_2 and h_3 is still ρ, moreover the one between h_1 and h_3 is

$$\mathbb{E}\left\{h_1^* h_3\right\} = \mathbb{E}\left\{h_1^{'*} h_2\rho + h_1^{'*} h_3^{'}\sqrt{1-\rho^2}\right\} = \mathbb{E}\left\{h_1^* h_2\right\}\rho = \rho^2 \tag{3.9}$$

as $h_1^{'}$ and $h_3^{'}$ are independent extractions.
Considering an ergodic continuous-time process of the channel variation, ρ is

the correlation between two values separated by a time interval, furthermore doubling that time interval corresponds to raising to the second power the correlation term.

Hence to define the decorrelation after one TTI ρ_{TTI}, with TTI in ms, we define the correlation term after 1 ms ρ_{ms} and we name as ρ_{T_c} the definition of decorrelation after the Coherence Time, as seen in Equation (3.4). We end up with

$$\rho_{T_c} = \frac{1}{e}, \tag{3.10}$$

$$\rho_{\text{ms}} = (\rho_{T_c})^{\frac{1}{T_c}}\,, \tag{3.11}$$

$$\rho_{\text{TTI}} = (\rho_{\text{ms}})^{\text{TTI}} = \left((\rho_{T_c})^{\frac{1}{T_c}}\right)^{\text{TTI}} = \frac{1}{e}^{\frac{\text{TTI}}{T_c}}\,, \tag{3.12}$$

with T_c and TTI expressed in ms.

If we consider the previous example of the moving user, a coherence time $T_c = 45.7$ ms corresponds to $\rho_{\text{TTI}} = 0.9969$ when the TTI length is 0.14 ms. This high value of correlation let us make the assumption that it is possible to consider the single fading coefficient h_l of Equation (3.1) constant over the symbols of a packet.

Furthermore, with the assumption on the bandwidth disclosed in section 3.1.2, we will end up with a channel with an impulsive response, in order to exploit the Rayleigh block-fading model.

3.1.2 Block-fading model

We would like to reduce the impulse response of a channel $\mathbf{h} = [h_1, h_2, \cdots h_N]$ with N coefficient to a single tap function in order to have a flat frequency response of the channel.

Let's consider a general parameter of wireless systems, the delay spread

T_{d}, defined as the difference in propagation time between the longest and shortest path, counting only the paths with significant energy

$$T_{\mathrm{d}} \coloneqq \max_{i,j} |\tau_i(t) - \tau_j(t)|. \tag{3.13}$$

This parameter can be evaluated through measurements in the propagation environment of the system. For a cell with a linear extent of few kilometers or less, it is likely to have one or two microseconds of delay spread, and as the coverage area reduces, T_{d} shrinks too.

The corresponding frequency-domain parameter is the *coherence bandwidth* W_{c}, that is the range of frequencies over which the channel can be considered flat fading

$$W_{\mathrm{c}} = \frac{1}{2T_{\mathrm{d}}}. \tag{3.14}$$

By making the assumption to have a delay spread in the order of nanoseconds, we have a coherence bandwidth enough large to transmit over a 20 MHz bandwidth, the one considered for this work, with a hypothetically flat channel.

Hence, based on a single tap channel which has a variable behavior in time, we assume to have a single fading coefficient h for the overall packet duration and a new realization in case of retransmission, with a correlation degree dependent on the period of retransmission.

This is the block-fading channel model [11], that is rewriting Equation 3.1 as

$$y[m] = hx[m] + w[m]. \tag{3.15}$$

We have thus the basis to the model we considered in this work to derive the SNR at the receiver for every transmission, in a Single-Input-Single-Output (SISO) system, that is the connection between two single antennas

at transmitter and receiver side.

We derive now MIMO transmission with beamforming in transmission and reception.

3.2 MIMO channel: Singular Value Decomposition and beamforming

When we consider multiple antennas at the TX and RX side, the received signal at every RX antenna is a superposition of the signals coming from the different TX; each one of them corresponds to a SISO system with transfer function h_{ij} from the ith TX antenna to the jth RX antenna, as shown in the example in Fig. 3.2 with a 3×3 MIMO system.

The model for the single output y_i of the system is

$$y_i = h_{i1}x_1 + h_{i2}x_2 + \cdots + h_{in}x_n + w_i,$$

where w_i is drawn from a complex white Gaussian noise process. If we have n transmitters and m receivers, we can write the MIMO system in matrix form, in the compact (3.16) or in the extended form (3.17).

$$\mathbf{y} = \mathbf{H}\boldsymbol{x} + \mathbf{w} \tag{3.16}$$

$$\underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_M \end{bmatrix}}_{\mathbf{y}} = \underbrace{\begin{bmatrix} h_{11} & h_{12} & \dots & h_{1n} \\ h_{21} & h_{22} & \dots & h_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ h_{m1} & h_{m2} & \dots & h_{mn} \end{bmatrix}}_{\mathbf{H}} \cdot \underbrace{\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}}_{\boldsymbol{x}} + \underbrace{\begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_m \end{bmatrix}}_{\mathbf{w}} \tag{3.17}$$

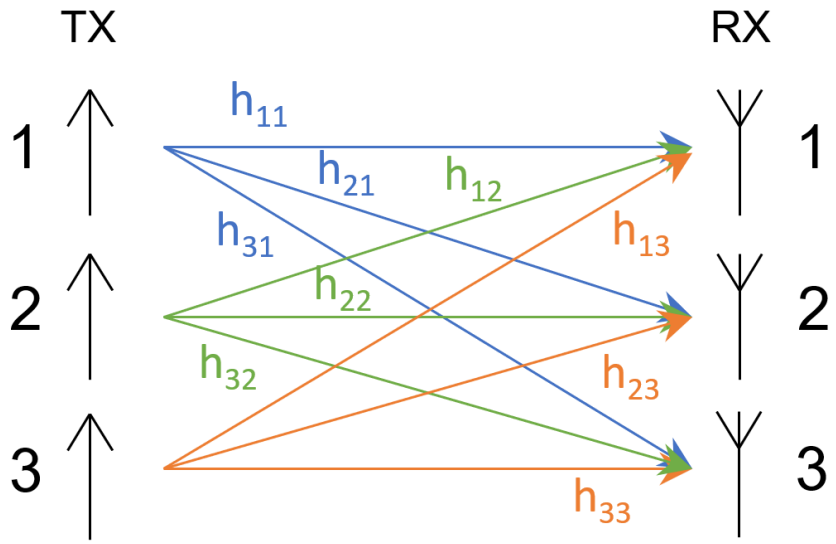

Figure 3.2: MIMO transmission with superposition of TX signals at the RX side.

In this MIMO system $\mathbf{H}$ belongs to $\mathbb{C}^{m\times n}$ and w_i is additive Gaussian noise sample drown from a stationary memoryless process.

Let's consider the Singular Value Decomposition (SVD) of the $m \times n$ matrix $\mathbf{H}$

$$\mathbf{H} = \mathbf{U}_m \mathbf{D} \mathbf{V}_n^H, \tag{3.18}$$

a generalization of the eigendecomposition of a positive semidefinite square matrix to any $m \times n$ matrix. $\mathbf{U}_m$ and $\mathbf{V}_n$ are unitary matrices ($\mathbf{U}_m^H = \mathbf{I}_m$ and $\mathbf{V}_n \mathbf{V}_n^H = \mathbf{I}_n$), more specifically $\mathbf{U}_m$ columns are the eigenvectors of $\mathbf{H}\mathbf{H}^H$, and $\mathbf{V}_n$ columns are the eigenvectors of $\mathbf{H}^H\mathbf{H}$.

The non-negative entries in the diagonal of $\mathbf{D}$ are called *singular values*, and correspond to the square roots of the eigenvalues of $\mathbf{H}\mathbf{H}^H$

$$\mathbf{H}\mathbf{H}^H = \mathbf{U}_m \mathbf{D} \mathbf{V}_n^H \mathbf{V}_n \mathbf{D}^H \mathbf{U}_m^H = \mathbf{U}_m \mathbf{D}\mathbf{D}^H \mathbf{U}_m^H = \mathbf{U}_m \mathbf{\Lambda}_m \mathbf{U}_m^H. \tag{3.19}$$

With the SVD of $\mathbf{H}$ we write

$$\mathbf{y} = \mathbf{U}_m \mathbf{D} \mathbf{V}_n^H \boldsymbol{x} + \mathbf{w}. \tag{3.20}$$

By imposing a rotation of the observation space we obtain

$$\mathbf{U}_m^H \mathbf{y} = \mathbf{y}' = \mathbf{U}_m^H (\mathbf{U}_m \mathbf{D} \mathbf{V}_n^H \boldsymbol{x} + \mathbf{w}) = \mathbf{D} \mathbf{V}_n^H \boldsymbol{x} + \mathbf{w}', \tag{3.21}$$

where $\mathbf{w}$' is still a complex gaussian white noise vector, as its covariance is $E\left\{\mathbf{U}_m^H \mathbf{w} \mathbf{w}^H \mathbf{U}_m\right\} = E\left\{\mathbf{U}_m^H N_0 \mathbf{I}_m \mathbf{U}_m\right\} = N_0 \mathbf{I}_m$.

If we call σ_l the lth singular value and we expand (3.21), we can write

$$\mathbf{y}' = \underbrace{\begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ 0 & 0 & \ddots & \vdots \\ 0 & 0 & 0 & \sigma_n \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 \end{bmatrix}}_{\mathbf{D}} \cdot \underbrace{\begin{bmatrix} \mathbf{v}_1^H \\ \mathbf{v}_2^H \\ \vdots \\ \mathbf{v}_n^H \end{bmatrix}}_{\mathbf{V}^H} \cdot \underbrace{\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}}_{\boldsymbol{x}} + \underbrace{\begin{bmatrix} w_1' \\ w_2' \\ w_3' \\ \vdots \\ w_m' \end{bmatrix}}_{\mathbf{w}'} \tag{3.22}$$

with $\sigma_1 > \sigma_2 > \cdots \sigma_n \geq 0$ and $\mathbf{v}_1^H = [v_{11}^*, v_{21}^*, \cdots v_{n1}^*]$.

The previous operations are useful to define n equations, where the singular values are sorted in descending order and identify a single equivalent channel. If the transmitter has a perfect knowledge of the channel, it may exploit this information to distribute the signal over the transmitting antennas in an optimal way. As the best singular value σ_1 is situated in the first row of the linear system (3.22), we can us the first column of matrix $\mathbf{V}$ to perform this shaping.

Let's consider a symbol a drawn from a memoryless process with zero mean

and variance the mean power of the transmission $\mathrm{P_a}$. Hence we define

$$\boldsymbol{x} \coloneqq \mathbf{v}_1 \cdot a \tag{3.23}$$

and we rewrite the rotated observation (3.22) as

$$\mathbf{y}' = \mathbf{D} \begin{bmatrix} \mathbf{v}_1^H \\ \mathbf{v}_2^H \\ \vdots \\ \mathbf{v}_n^H \end{bmatrix} \mathbf{v}_1 \cdot a + \mathbf{w}' = \mathbf{D} \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \cdot a + \mathbf{w}'. \tag{3.24}$$

The *beamforming* performed on the TX signal allow the transmitter to distribute the power over the antennas in the best way and, with the same procedure at the receiver, the power is concentrated on the best equivalent channel, while the others observations only contain the termal noise.
Hence we consider only the first equation

$$y_1' = \sigma_1 \cdot a + w_1 \tag{3.25}$$

and its equivalent SNR γ is

$$\gamma = \frac{\sigma_1^2 \mathrm{P_a}}{N_0 \mathrm{B}}. \tag{3.26}$$

where B is the transmission bandwidth and $\mathrm{P_a}$ the mean power of the block of symbols.
Of course there is a big enhancement with respect to SISO communication. The procedure above introduces a gain G^M, that we call MIMO gain, to the SNR of the SISO system at the receiver.
In the single-antenna case the channel does not provide a gain to the transmission, but it is a representation of multiple reflectors. So its mean power

is 1, while the mean gain of a 2×8 MIMO is approximately 11.14.

3.3 Wireless Resource Element: granularity in time and frequency

What we want to define in this section is the amount of resources dedicated to the single device according to the quality of the channel, whose characterization has been so far explained.

Given the SNR and number of information bits k willing to send, the Maximal Coding Rate $R^*(n, \epsilon)$ determines the number of channel uses n to deliver the total information with BLER ϵ. In LTE the channel uses n are OFDM symbols, sent over the system bandwidth whose spacing, defined as SubCarrier Spacing (SCS), is 15 KHz.

The sampling frequency f_s of the symbols is 14 KHz, lower than the spacing SCS in order to guarantee a guard time to compensate for the delay spread. So if we define the TTI duration as the time to transmit 14 OFDM symbols, we come up with a TTI = 1 ms, as defined in the LTE standard.

In the following, we consider the 5G multi-carrier system being standardized by 3GPP as New Radio (NR), in particular we considered the definition of mini-slot, that is the time to transmit 2 symbols in one TTI.

So if one OFDM symbol lasts 71.4μs, that is $1/f_s$, one mini-slot has a duration of 143 μs.

We consider also that the resource elements can be grouped in a Physical Resource Block (PRB), grouping together multiple subcarriers and being the unit element for allocation.

If we define with n_t and n_s the number of symbols in a TTI and the number of subcarriers in a PRB respectively, the minimum group of wireless

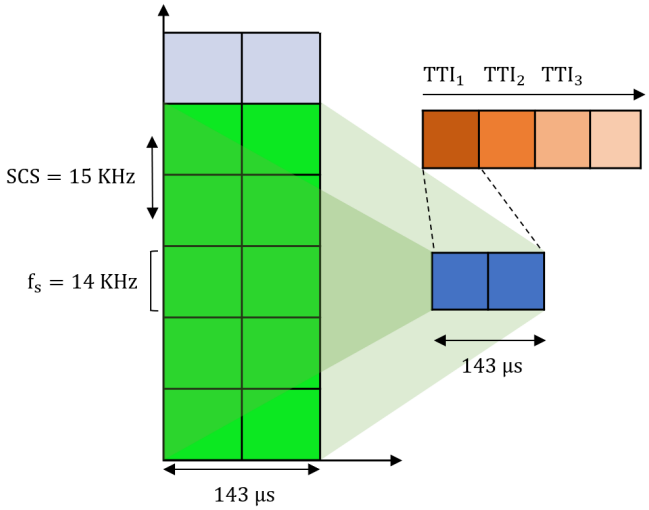

Figure 3.3: Visual description of Time Transmission Interval (TTI), SubCarrier Spacing (SCS) and sampling frequency f_s.

resources contains $n_s \cdot n_t$ symbols. Therefore, given the number n of channel uses required by the transmission, the total amount of resource allocated is

$$(P \cdot n_s) \cdot n_t \geq n, \tag{3.27}$$

where P and $(P \cdot n_s)$ are the number of PRBs and the number of subcarriers respectively.

In Fig. 3.4 we suppose to have $n = 7$ from the MCR, $n_t = 2$ and $n_s = 4$. We have in this case a granularity of 8 symbols, and the algorithm delivers to the TX device 1 PRB of 8 symbols.

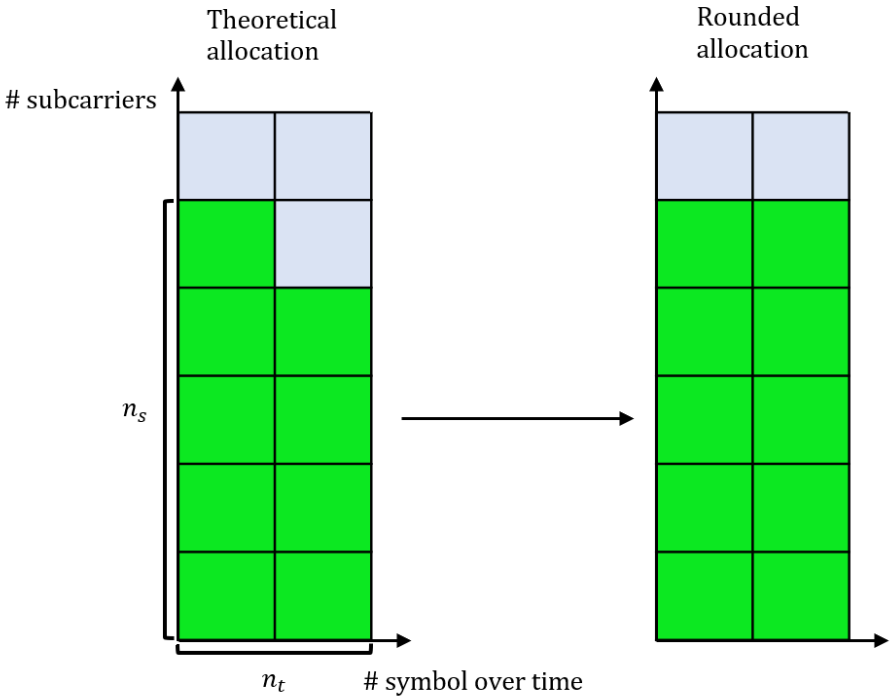

Figure 3.4: Resource allocation with $n = 7$, $n_t = 2$ and $n_s = 4$.

Chapter 4

Retransmission techniques

This chapter is dedicated to the retransmission techniques, which involve error control, soft combining methods and schemes that rule the time interval after which the retransmission is triggered in case of failed transmission, named as Transmission Period (TP).

Retransmission in LTE and 5G is pursued in the time domain by targeting a higher BLER than the one required for Ultra-Reliability, and lowered by exploiting retransmissions, as we have already discussed in section 1.2.1 with the ARQ protocol .

In section 4.1 we present the features of HARQ-CC and HARQ-IR techniques and the model we adopted to integrate the methods in the thesis work.

In section 4.2 we present the state of the art retransmission schemes, NACK and BLIND-based.

4.1 HARQ combining techniques

In LTE HARQ combining techniques are used since they have better performances than ARQ; in the latter case the receiver, whenever it requires

a new transmission, discards the corrupted block, and it relies on the new received block for the decoding.

In HARQ instead the main concept is the *soft combining* of the received packets, corresponding to the storage and the combination of the previously received transmissions, in order to increase consecutively the probability of successful decoding.

There are two main HARQ techniques, namely Chase Combining (CC) and Incremental Redundancy (IR).

In this paper we will derive HARQ performance by following the model in [16], also considered as reference for current simulators.

HARQ-CC performance

Accordingly, all $(k-1)$ retransmissions sent with CC-HARQ consist in the same signal sent during the first transmission, allowing the receiver to combine all of the k received transmission attempts, thus achieving performance equivalent to an AWGN channel with SNR

$$\gamma_k^{eq} = \alpha^{k-1} \sum_{j=1}^{k} \gamma_j \ , \tag{4.1}$$

where $\alpha < 1$ is the combining efficiency, modeling the losses due to the channel measurement errors.

Note that γ_k^{eq} is typically higher than the single transmissions SNR γ_j, due to the coherent combining of the different components with the same useful signal part but independent noise realizations.

In the Appendix A it is derived the relationship between the soft combining and the sum of the SNRs of the transmissions, by presenting the theory behind the Maximum Likelihood Estimators.

HARQ-IR performance

On the other hand, in IR-HARQ technique, a different version of the *redundant* bits is sent, resulting in a decreased coding rate of the combined packet and, most of all, a higher probability of successful decoding.
In order to do this, the rate R of the output of the encoder is adapted by a *rate matching* block (Figure 4.1) that increases the code rate for the first transmission with the *puncturing* procedure, by removing parity bits, and lowers it for the following retransmissions by repeating the parity bits. The code rate after the rate matching is named Effective Code Rate (ECR), that is the ratio between the number of bits feeding the encoder and the number of bits after the rate matching.

IR could allocate less resources than CC in the retransmission by only

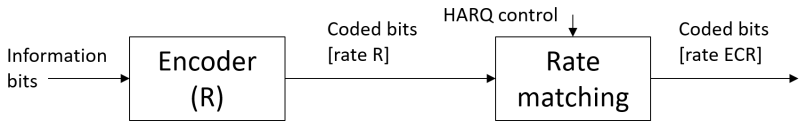

Figure 4.1: Rate-matching procedure. The HARQ control is to choose a different ECR according to the number of the retransmission4.

sending the redundancy in the retransmission, achieving better spectral efficiency at the price of reliability. However, the model gets more complicated and doesn't let the comparison between CC and IR; hence in order to compare their performance in terms of reliability, we use the model in [16], where IR and CC retransmissions have the same allocation.
As it is done for CC-HARQ, the equivalent SNR after IR-HARQ combining is given by

$$\gamma_k^{eq} = \alpha^{k-1} \cdot \eta(\mathrm{MCR}_1)_k \cdot \sum_{j=1}^{k} \gamma_j, \tag{4.2}$$

$$\eta(\mathrm{MCR}_1)_k = \begin{cases} 1 & \text{if IR, } k = 1 \\ \eta(\mathrm{MCR}_1) & \text{if IR, } k \geq 2. \end{cases} \tag{4.3}$$

where $\eta(\mathrm{MCR}_1)$ is the gain of IR with respect to CC, and MCR_1 is the MCR of the first transmission.

The values of $\eta(x)$ are computed in [16] and reported in Table 4.1, where

Table 4.1: Simulated IR gain $\eta(MCR_1)$ table with QPSK and 16-QAM.

Modulation	MCR_1	ECR_1	$\eta\,(\mathrm{MCR}_1)$
QPSK	0.40	0.20	1.03
QPSK	0.60	0.30	1.01
QPSK	0.80	0.10	1.04
QPSK	1	0.50	1.11
QPSK	1.20	0.60	1.20
QPSK	1.40	0.70	1.35
QPSK	1.60	0.80	1.59
16-QAM	2.00	0.50	1.46
16-QAM	2.20	0.55	1.54
16-QAM	2.40	0.60	1.59
16-QAM	2.60	0.65	1.71
16-QAM	2.80	0.70	1.81
16-QAM	3.00	0.75	1.93

we modeled $\mathrm{MCR}_1 = \log_2(M) \cdot \mathrm{ECR}_1$, being M-QAM the modulation and ECR the Effective Code Rate of the first transmission.

From the numerical values of the IR gain $\eta(\mathrm{MCR}_1)$, we can notice that

- the IR gain is generally increasing with the MCR, with the exception of the MCR region where, in practical Modulation and Coding schemes, a modulation switch occurs. Notice that, at MCR = 2, the system switches between QPSK and 16-QAM.
- When MCR is close to zero, the IR gain is negligible.

Since we have only discrete measurements of $\eta(\mathrm{MCR}_1)$, in this work we smoothed the values in Table 4.1, with a moving average filter of length 3,

and interpolated linearly between the resulting points in order to compute $\eta(\mathrm{MCR}_1)$ for all the possible value of $\mathrm{MCR}_1 > 0$.

This is necessary because, in our simulations, the transmission rate is computed with (2.2) and, consequently, it does not assume only discrete values.

4.2 Retransmission schemes: state of the art

In this Section we first analyze how the probability of error is integrated in the retransmission model, then we discuss about NACK and BLIND-based schemes main features.

The inversion of the equation (2.7) leads to the equation below

$$\epsilon^*(k,n) \approx Q\left(\frac{nC - k + (\log n)/2}{\sqrt{nV}}\right) \tag{4.4}$$

where n is already the rounded value of the amount of resources selected, discussed in section 3.3.

This BLER value ϵ^* fits into a random experiment with different channel realizations, where *latency* and *bandwidth allocation* for every transmission/retransmission are collected. It is the threshold that determines the success or failure of the single transmission, generally it assumes a smaller value than the BLER target.

This is due to the fact that the resource elements are sent in groups, as seen in Equation (3.27), so for a greater amount of resources than the one evaluated in (2.2), and for a fixed SNR, we have a lower ϵ^* value.

In case the system is not able to deliver the number of resources evaluated, happening when the bandwidth request by the user is greater than the system bandwidth, the BLER value ϵ^* will be greater than the BLER target.

4.2.1 NACK-based scheme

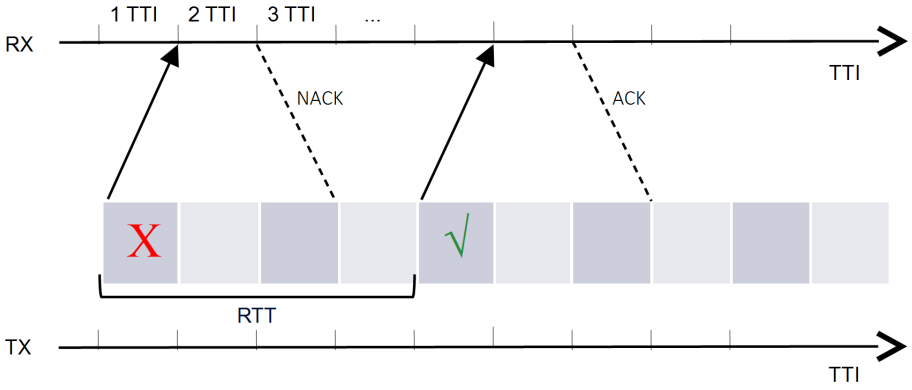

Figure 4.2: NACK-based scheme. The transmission is succesfully received after one retransmission.

In NACK-based approach, the transmitter TX sends to the receiver RX a packet and it waits for the ACK/NACK feedback, as can be seen in Fig. 4.2: in this example the RX decodes successfully the packet after the first retransmission.

In this thesis, we define as Round Trip Time (RTT) the time interval to have a complete ACK/NACK process, including 1 TTI to forward the transmission and having error control at RX, 1 TTI to process the received transmission and the feedback transmission, 1 TTI to forward the feedback and 1 TTI to process the feedback and the next transmission.

This scheme is good for the spectral efficiency because it retransmits only when it is needed. However, the delay introduced at every RTT may let the transmitter fail to comply with URLLC latency requirements.

4.2.2 BLIND scheme

The BLIND strategy differs from the NACK-based because it does not wait for the control message, conversely it continuously transmits the packet,

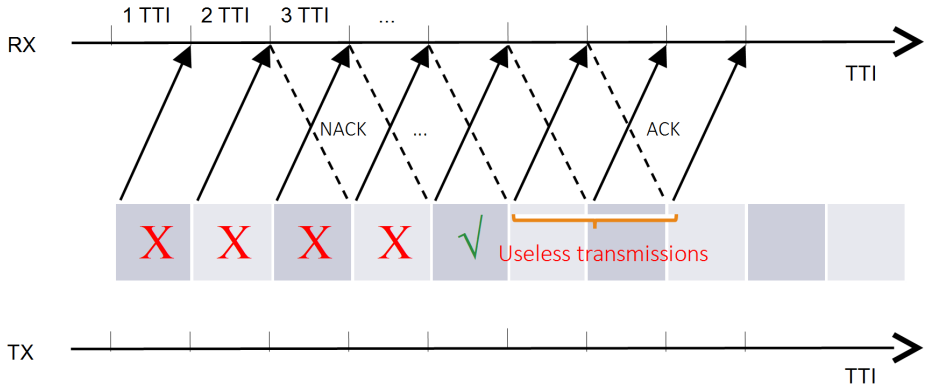

Figure 4.3: BLIND scheme. Here the Tx needs 5 BLIND transmissions to have success. Note that 3 additional transmission are still sent before the processing of the ACK. The Transmission period is $T_p = 1$ TTI.

with a Transmission Period T_p that is shorter than the RTT, until the TX receives an ACK.

BLIND scheme is strong as far as latency is concerned, due to the greater amount of transmission opportunities offered; in contrast we have a poorer spectral efficiency, due to the fact that TX sends more transmission than actually needed. For example in Fig. 4.3 there are 3 useless transmissions if we impose $T_p = 1$ TTI.

Chapter 5

HYBRID scheme and implementation of LA algorithm

The idea of a retransmission scheme that may give a contribute to the actual state of the art is born during my internship in Nokia Bell-Labs, Stuttgart. The HYBRID scheme has brought an answer to the task of the thesis work, to analyze tradeoffs in latency and spectral efficiency of current retransmission methods and to find a new solution.

Its patent's name is *Multi-Mode retransmission scheme for Wireless Networks*, filed on 7 August 2017, and the simulation analysis on the novelty compared to the state of the art techniques are already submitted as a contribution to IEEE WCNC 2018 conference on Wireless Communication and Networking.

In section 5.1 we discuss about the proposed scheme, in section 5.2 we present the implementation of the theoretical Link Adaptation and its expected behavior.

5.1 Proposed Scheme: HYBRID Strategy

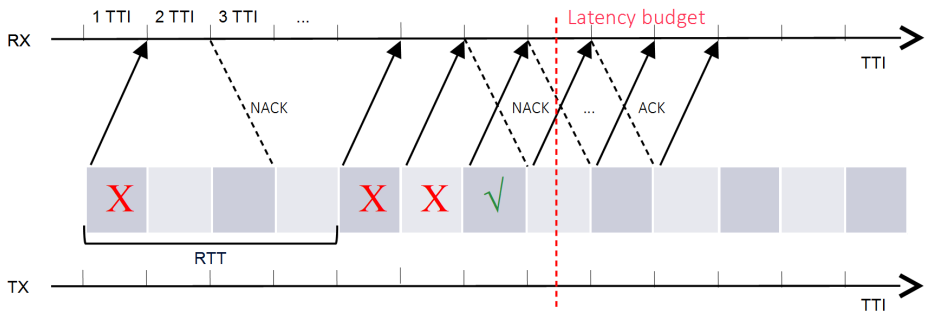

Figure 5.1: HYBRID scheme. In the example Tx is able to comply with the latency requirement by switching scheme and sending three BLIND retransmissions, besides the three additional transmission sent with $T_p = 1$ TTI.

The principal concept of the novelty is to combine the NACK and BLIND-based strategies in order to switch from the first to the second scheme and to take advantage from both. The device minimizes the resource consumption by using NACK-based scheme and changes to a more aggressive retransmission when needed, without being conservative on the bandwidth utilization. This idea can be effective in a multiple-access system, where the devices shares the same resources, that for example cannot be wasted by unnecessary BLIND retransmissions, although it has been designed to compensate for the high latency of NACK-based scheme.
In the HYBRID scheme there could be two ways to switch from NACK to BLIND:

1. after N NACK-based transmissions;

2. at the last useful transmission before a *latency budget*, corresponding to the URLLC requirements; in the example in Fig. 5.1, the device would not meet them if it had to wait the RTT before retransmission.

We stress that, with a BLER target ϵ, the first transmission has approxi-

mately $1 - \epsilon$ probability to be successful, thus leading to a reliability of 90 % for the first transmission in case $\epsilon = 10\%$. Hence for the majority of cases the resource usage will be the same of NACK-based scheme, and for the remaining we exploit the effectiveness of BLIND strategy.
We have found thus a tradeoff between an excellent latency and a small amount of occupied resources.

5.2 Implementation of theoretical LA algorithm

Now we dedicate a section to present the algorithm performing the theoretical Link Adaptation, where "theoretical" means without the use of modulation and coding schemes.
In the first subsection we have a differentiation between UpLink (UL) and DownLink (DL) transmission and the steps of the iterative process, in order to define the correct amount of resources.
The second subsection is dedicated to an analysis on the out-of-band probability, happening when resource requests are exceeding the available bandwidth.

5.2.1 SNR definition for UL and DL and iterative LA procedure

A communication between a user device, with usually low available power, and a base station, which can dispose of a greater amount of power. We consider a limited power P_{tx} for UL transmission and a theoretical infinite power available for the base station, leading to two different behaviors in the resource allocation. The limitations and constraints for the total transmit power P_{tx} are the following.

1. P_{tx} is divided among the allocated subcarriers. Therefore, the single TX antenna to RX antenna SNR before propagation and fading η decreases with the total allocated bandwidth, i.e.

$$\eta = \frac{P_{tx}}{(N_0)B} , \tag{5.1}$$

where $B = P \cdot n_s \cdot \text{SCS}$ turns out to be the allocated bandwidth. P and n_s are respectively the number of PRBs and the number of subcarrier grouped together in a PRB, as seen in Equation (3.27).

2. The received power corresponds to the transmission power after propagation loss and shadowing, that could be defined with reference values of the propagation model assumed for the specific propagation environment. To maintain general validity we define a maximum SNR_{max} at the receiver, corresponding to conveying all the power in a single resource block.

3. An upper limit $\overline{\text{SNR}}$ on the resulting SNR has to be respected. This is adopted in practical systems for saving energy when the transmission achieves the maximum modulation and coding scheme and limiting the interference toward other cells and devices.

We notice that the SNR values SNR_{max} and $\overline{\text{SNR}}$ correspond to the SNR between one transmit and one receive antenna; the final γ will be derived generating the channel fading ρ of each antenna-antenna path in order to evaluate the MIMO gain G^M, as discussed in 3.2.
Given these constraints, we also display the process to compute the total number of allocated subcarriers $n_f = n_s \cdot P$ and the MCR.
We start the allocation process from an optimistic hypothesis, i.e. only 1 PRB is needed for satisfying the BLER target.

Therefore, it is possible to initialize the allocated power in $n_f = n_s$ subcarriers, according to (3.27).

The iterative search of the final MCR and the corresponding BLER to rule the retransmissions is made according to the following steps:

1. $\mathrm{SNR} = \min\{\mathrm{SNR}_{\max}/\mathrm{n}_f, \overline{\mathrm{SNR}}\}$ (power constraints);
2. $\gamma = \mathrm{SNR} \cdot G^M$;
3. evaluation of MCR and number of needed subcarriers n_f from (3.27) and (2.2);
4. if n_f is changed, repeat from 1. Otherwise, continue to next step.
5. evaluation of the number of channel uses delivered n_{eff} and corresponding block error probability ϵ^*.

For DL transmission is different, as the SNR value is always corresponding to $\overline{\mathrm{SNR}}$ due to infinite available power and hence the MCR is evaluated in one iteration. We will find again this behavior in the simulation results.

5.2.2 Behavior of LA procedure

Let's analyze now the possibility of having a resource request exceeding the available bandwidth. For example, we consider a system with bandwidth B = 20 MHz and BLER target $\epsilon = 10^{-1}$, and two systems working with $\epsilon = 10^{-6}$ and system bandwidths B = 20, 1000 MHz.

In Figure 5.2 we have run the algorithm in function of the MIMO gain G^M for UL transmissions, with an $\text{SNR}_{\max} = 15$ dB and an upper limit $\overline{\text{SNR}} = 6$ dB an the settings on B and ϵ above. The vertical axis represents the number of subcarrier composing the system bandwidth B, having a spacing SCS = 15 KHz. Hence we have 1333 subcarriers in 20 MHz and 66666 in 1 GHz. First

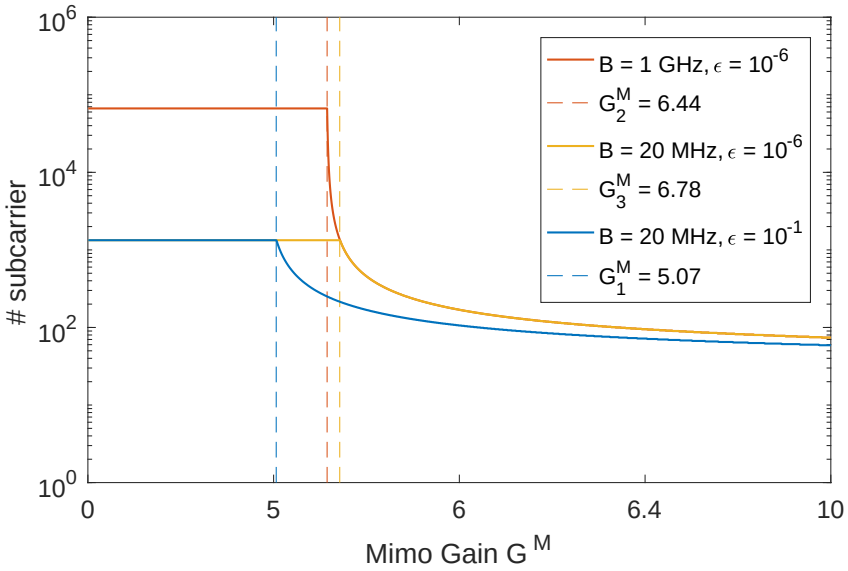

Figure 5.2: Resource allocation with $SNR_{max} = 15$ dB and $\overline{SNR} = 6$ dB.

of all we notice that for B = 20 MHz, $\epsilon = 10^{-1}$ and G^M smaller than 5.07, the system allocates the maximum amount of resources. If $\epsilon = 10^{-6}$, the allocation curve changes and requires a higher gain for the same allocation, to ensure a higher reliability; the curve is limited by the two different system bandwidths.

So it is possible to find the minimum gains G_1^M, G_2^M and G_3^M to not exceed B in allocation and evaluate the out-of-bandwidth probabilities. In figures 5.3, 5.4 and 5.5, probability density functions of the gains in 1×1, 2×4 and 2×8 MIMO systems are plotted; the area delimited by the pdfs and the thresholds is the probability to not fulfill Link Adaptation.

This is due to the fact that, in the out-of-bandwidth condition, we use a higher rate than the maximum allowed on a channel with a given value of SNR, if we want to deliver information with an error rate ϵ. The consequence is to have a higher BLER than ϵ on the transmission with probability P_1, P_2 and P_3 in the figures. We will see this concept further in the simulation results.

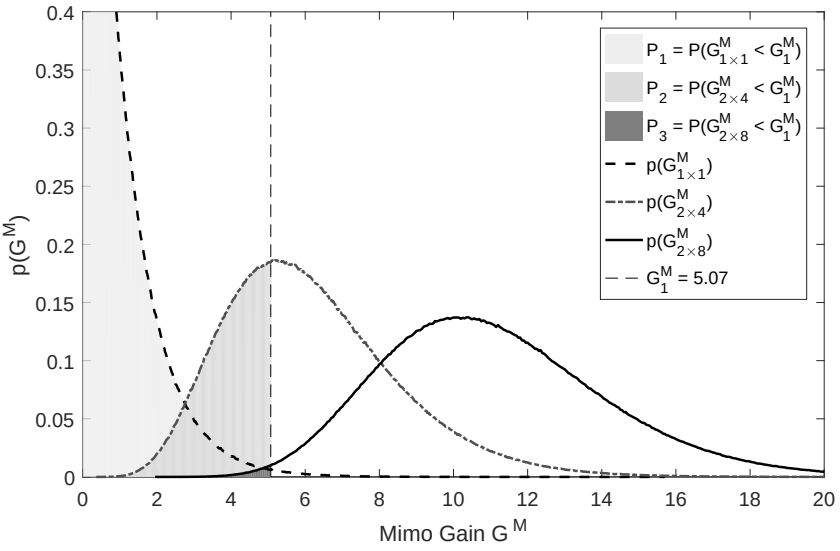

Figure 5.3: Distributions of 1×1, 2×4 *and* 2×8 *MIMO gain with* $B = 20$ *MHz and* $\epsilon = 10^{-1}$. *Out-of-bandwidth probabilities are* $P_1 = 0.994$, $P_2 = 0.35$, $P_3 = 5.6 \cdot 10^{-3}$.

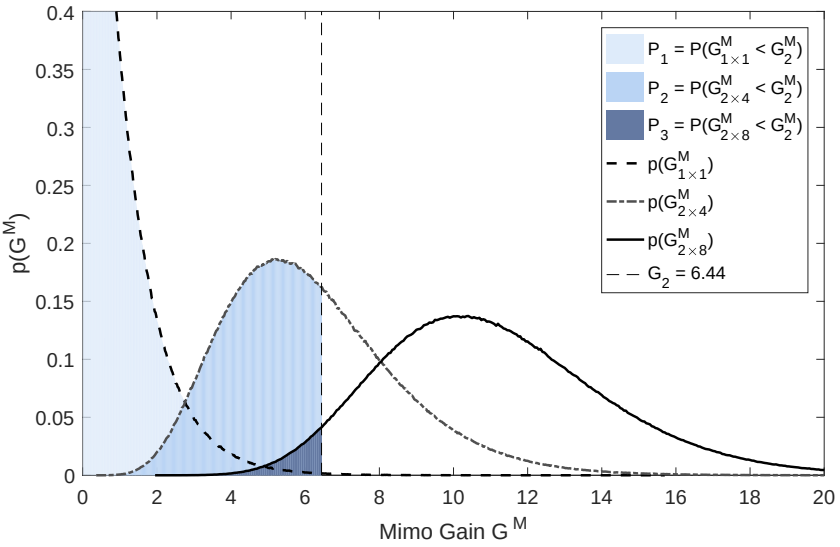

Figure 5.4: Distributions of 1×1, 2×4 *and* 2×8 *MIMO gain with* $B = 1$ *GHz and* $\epsilon = 10^{-6}$. *Out-of-bandwidth probabilities are* $P_1 = 0.998$, $P_2 = 0.6$, $P_3 = 0.037$.

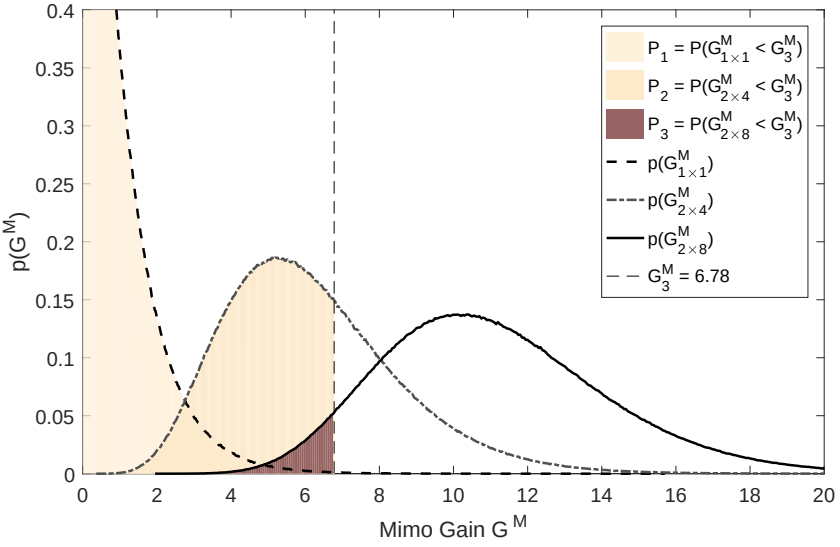

Figure 5.5: Distributions of 1×1, 2×4 *and* 2×8 *MIMO gain with* $B = 20$ *MHz and* $\epsilon = 10^{-6}$. *Out-of-bandwidth probabilities are* $P_1 = 0.999$, $P_2 = 0.648$, $P_3 = 0.052$.

Chapter 6

Simulation results

In order to quantify the performances of the retransmission schemes, we evaluated latency performances by means of Probability Mass Functions (PMF) of the transmissions needed to have a correct transmission, while the resource consumption has been studied with a Cumulative Distribution Function (CDF) of the bandwidth occupation.
In section 6.1 we have some comments on the simulation assumptions, in section 6.2 we analyse the performance of the three retransmission schemes. Also we compare CC-HARQ and IR-HARQ performance in typical URLLC working conditions in section 6.3.
In section 6.4 we propose secondary results, concerning the relationship between deep fades and NACK events and the retransmission performance by tuning the coherence time value.

6.1 Simulation assumptions

In Table 6.1 are reported the simulation parameters. The MIMO scheme addresses a typical URLLC scenario with 8 antennas at the base station and 2 antennas at the device side [17].

Table 6.1: Simulation Parameters

Parameter	**Value**	
Scenario	Single Cell, Single Transmission	
UE Type	UL SPS URLLC	DL SPS URLLC
System Bandwidth	20 MHz	5 MHz
SCS	15 kHz	
TTI	0.143 ms	
RTT	4 TTI	
BLIND TX Period T_p	1 TTI	
HYBRID switching N	1	
Central Frequency	2 GHz	
User Mobility	Slow mobility $\approx$ 5 km/h	
Channel Coherence Time T_c	50 ms	
BLER target	10^{-1} / 10^{-6}	10^{-1}
Combining efficiency α	0.95	
Packet Size	400 bits	
PRB dimension n_s	1 subcarrier	
Starting Average SNR with all power in 1 Subcarrier	15 dB	Inf
Limit on the SNR given by PSD $\rightarrow \overline{\text{SNR}}$	6 dB	-11 dB
BS Antennas	8	
UE Antennas	2	
Link Adaptation	Theoretical, from[5] and (3.27)	
QoS Latency Budget	1 ms	
MonteCarlo iterations	10^7	

Notice that the scheme and the simulated latencies are consistent with Downlink (DL) transmissions, while, for Uplink (UL), we need a Semi-Persistent Scheduled allocation for the first packet transmission [18], otherwise the scheduling request delay should be considered in the resulting packet latency.

Therefore, in our simulations, there is no delay due to the first transmission scheduling request procedure while the retransmission delays are considered.

The URLLC mini-slot with a TTI of 143 μs is adopted [19], a Subcarrier Spacing (SCS) of 15 kHz, as well as a RTT = 4 TTI between the transmission and the reception of ACK/NACK feedbacks.
In section 3.2 we defined the MIMO transmission, performed assuming perfect CSI and linear precoding at the transmitter for selecting and using all the available power on the channel corresponding to the maximum singular value of the MIMO channel matrix.
Frequency selectivity is not considered and, therefore, the fading coefficient ρ is assumed constant in all the band. However, the time selectivity is considered, with a channel coherence time T_c.

6.2 Retransmission schemes performances

We have two sets of results, one for UL transmission, which simulates limited transmit power P_{tx}, and one for DL transmission, where $\overline{SNR}$ is reduced compared whit UL case.
For UL we have Figs. 6.1 and 6.2, which report for each retransmission scheme the Probability Mass Function (PMF) of a successful transmission latency and the Cumulative Density Function (CDF) of the total time-frequency resource consumption, respectively.
Figs. 6.3 and 6.4 refer to the same functions for DL transmission.
Notice that CC-HARQ has been used and 10^7 single packet transmissions have been simulated.

UpLink

The four different scenarios plotted in Figs. 6.1 and 6.2 are commented in the bullet list below.

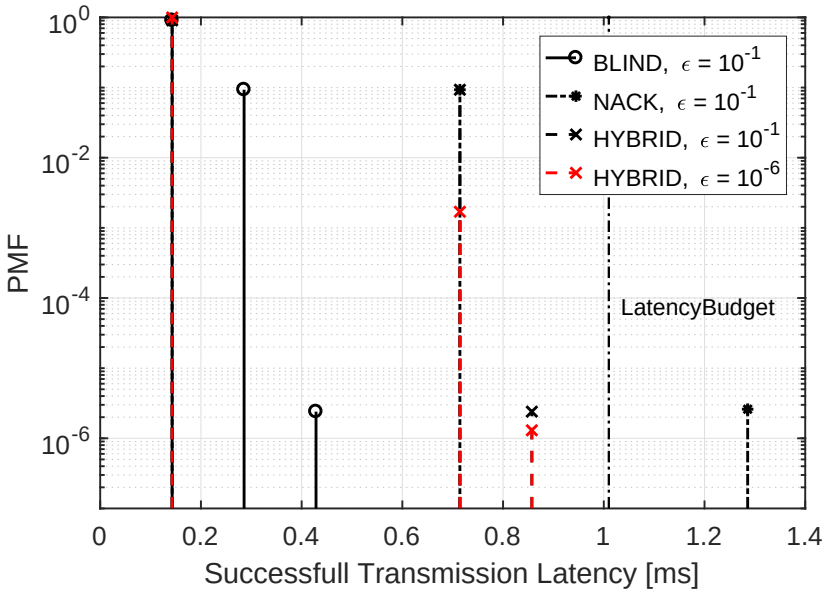

Figure 6.1: PMF of the packet latency with the considered retransmission schemes. Simulation parameters for UL are in Table 6.1 and retransmission performance is modeled with CC HARQ.

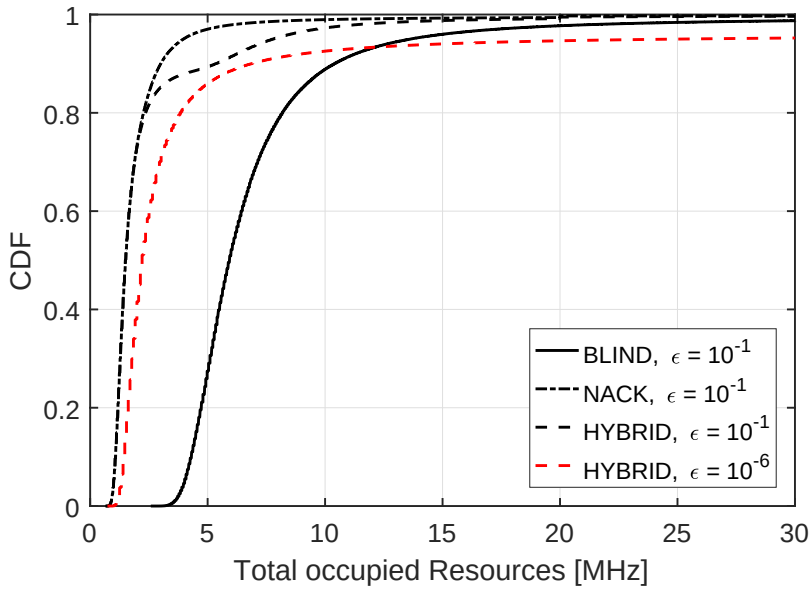

Figure 6.2: CDF of the total time-frequency resources allocated to the transmission with the considered retransmission schemes.
Simulation parameters for UL are in Table 6.1, retransmission performance is modeled with CC HARQ.

1. The NACK-based model is not able to satisfy the latency budget of 1 ms with the reliability target of $1 - 10^{-6}$, as the probability of failure with one retransmission is equal to the probability of success after the latency requirement, with value $2.6 \cdot 10^{-6}$.
   Nevertheless, in terms of spectral efficiency it is the most efficient scheme.

2. The BLIND scheme is the best approach in terms of latency and reliability, due to the Transmission Period of 1 TTI, shorter than the RTT = 4 TTI in the NACK-based scheme.
   However, the resource consumption in the BLIND scheme is almost 4 times that in the NACK-based scheme, due to the 3 unnecessary transmissions before ACK reception.

3. Even if the latency/reliability performance of the HYBRID scheme are worse than the BLIND scheme, the HYBRID scheme is able to match the latency budget in a reliable way, and its spectral efficiency is the same of NACK one in more than 80% of the cases, achieving important savings in terms of occupied time-frequency resources.

4. When the target first transmission BLER is much lower, e.g. $\epsilon = 10^{-6}$, we obtain obviously better performances in terms of latency/reliability. However, the system does not achieve the desired BLER due to sporadic deep fades, leading to poor MCR and resource requirements exceeding the available bandwidth, in this case increased to 1 GHz. The out-of-bandwidth probability P_3 evaluated in Fig. 5.4 is equal to $3.7 \cdot 10^{-3}$, higher than ϵ, strictly related to the experienced BLER at the first transmission, roughly $10^{-3} \gg 10^{-6}$. According to this high resource consumption (Fig. 6.2), we can conclude that a very

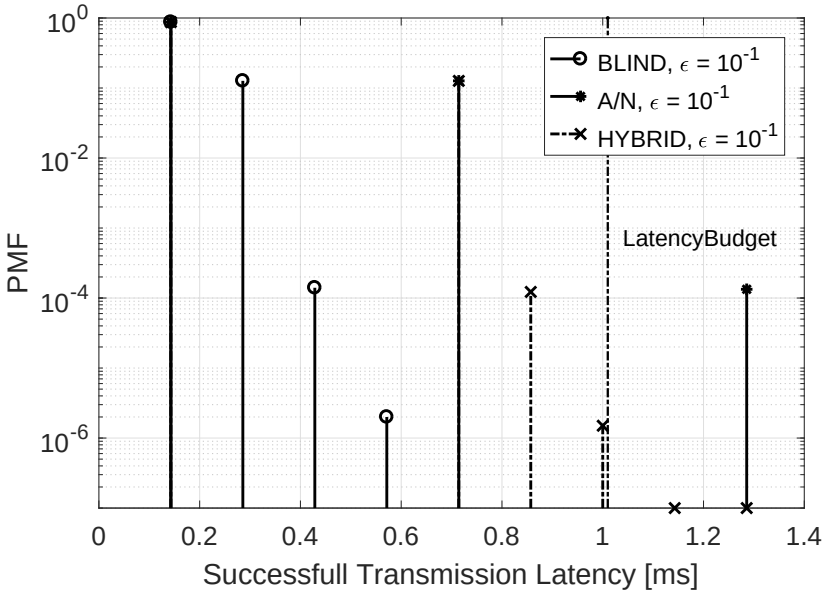

Figure 6.3: Simulation parameters for DL are in Table 6.1, retransmission performance is modeled with CC HARQ.

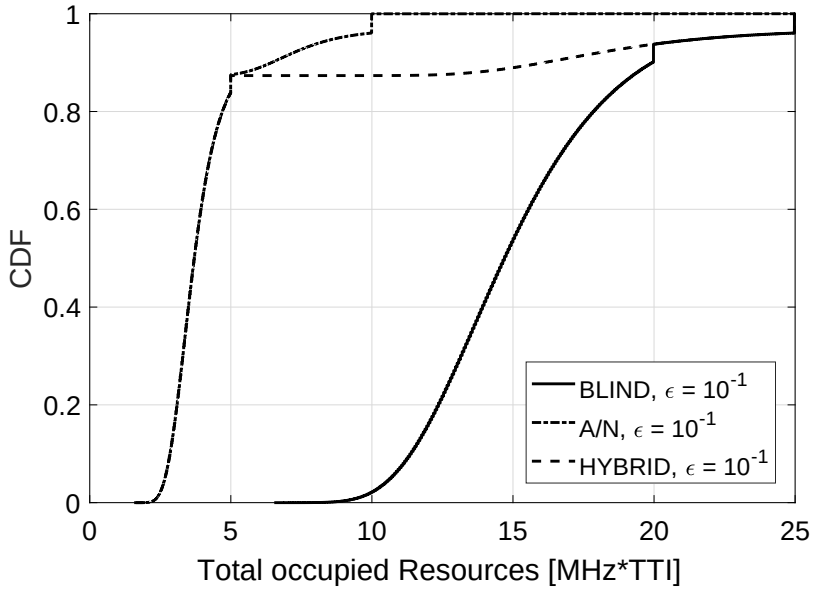

Figure 6.4: Simulation parameters for DL are in Table 6.1, retransmission performance is modeled with CC HARQ.

low target BLER at the first transmission is not an effective strategy, especially when the transmission might experience deep fades or weak link budget SNRs.

It can be noticed that, in all the retransmission schemes, the PMF of the latency starts to exhibit a floor due to very poor performance in the rare occasions that a deep fade condition occurs.
In these situations, ultra-reliability performance, e.g. PMF $\leq 10^{-5}$, are dominated by deep fading conditions, and we should let the channel evolve to leave critical fast fading realizations.

DownLink

For DL case we have Figs. 6.3 and 6.4, where we imposed a low value of $\overline{\text{SNR}}$ to test the situation of poor link budget.
The bandwidth for a single user is limited to 5 MHz to avoid very low MCR when we are in a deep fade condition, that would let the device using all the available bandwidth.
The performance is similar to UL transmission. The difference here is that the HYBRID scheme experiences the switch in 12.6% of the cases, corresponding to the BLER of the first transmission.
The point is that, as discussed in the simulation assumptions, in DL the SNR is fixed to the limit $\overline{\text{SNR}}$ and the MCR is directly related to that value of SNR, while in UL there is an iterative search of both MCR and SNR. So here the switch is more evident and predictable.

6.3 Comparison HARQ-CC and HARQ-IR performances

In the previous discussion, we have applied CC HARQ to the retransmission schemes, although IR HARQ is expected to have always better performance. Indeed, we can notice from Fig. 6.5 that, with a target BLER of 10%, IR achieves a better SNR distribution after the first retransmission.

However, for the "unlucky" packets that arrive to the second retransmissions, we can observe in Fig. 6.6 that the combined SNR is very low and similar for IR and CC.

This effect is due to the fact that these unlucky packets are the ones experiencing a deep fade, hence with a poor channel rate and a weak gain of IR w.r.t. CC (see Table 4.1).

Therefore, CC HARQ becomes an interesting choice for these URLLC configurations since its performance is similar to IR at the reliability target of URLLC and its implementation is typically simpler than IR one.

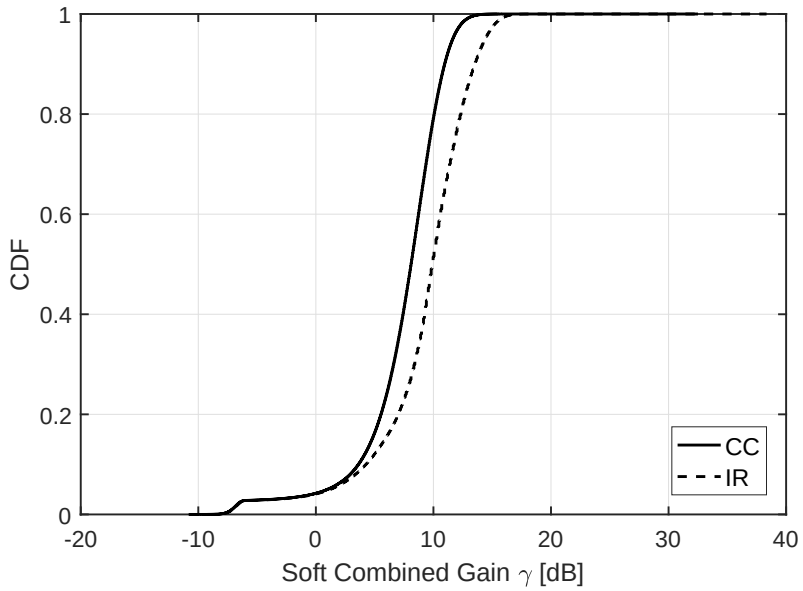

Figure 6.5: CDF of the combined SNR after the first retransmission with CC and IR.

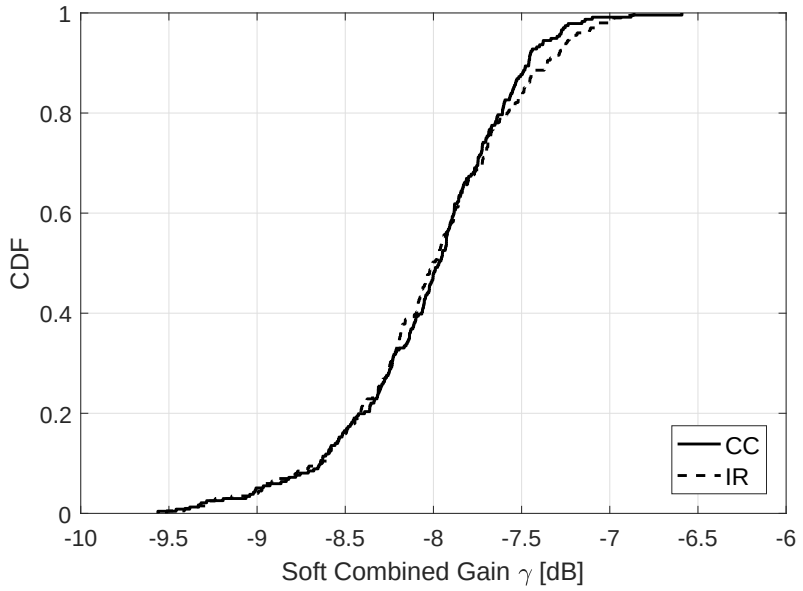

Figure 6.6: CDF of the combined SNR after the second retransmission with CC and IR. $= 10^8$ transmissions were simulated to obtain this Figure.

6.4 Secondary results

6.4.1 Correlation between SNR distributions and NACK events

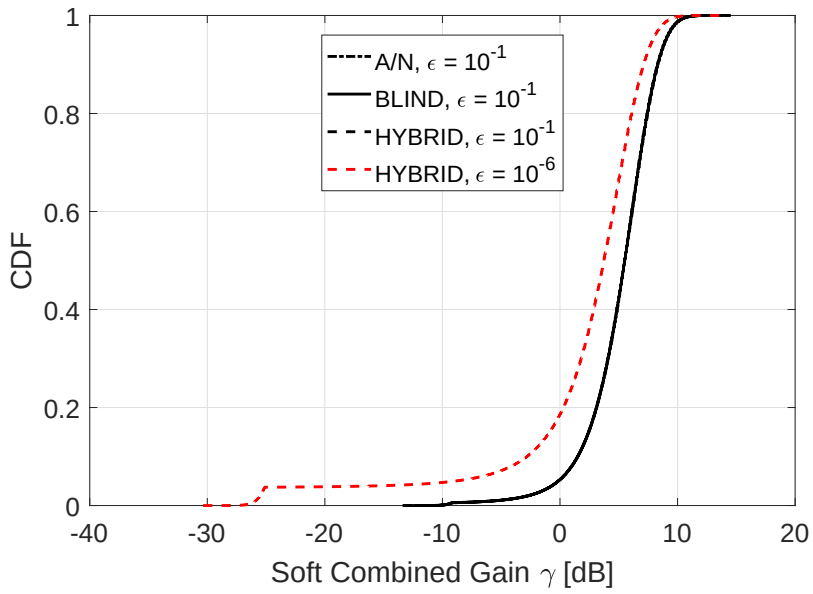

Figure 6.7: CDF of the SNR of the 1st transmission for UL case.

Here we want to consider again the UL example in section 6.2 in order to stress the relationship between the failure of the transmission and a poor channel quality.

In Figure 6.7 the CDF of the SNR of the 1st transmission is plotted, showing a slightly lower distribution for the device which aim to transmit with a BLER $10^{-6} \ll 10^{-1}$. Then we evaluated the distribution of the SNR, or the Soft Combined Gain, that have led to a NACK event. In Figures 6.8 and 6.9 we evaluate the distributions that have caused respectively one NACK event and two consecutively NACK events respectively.

The first consideration is that in Figure 6.7, for the scheme using a BLER target 10^{-6}, there is a probability 0.038 to use all the system bandwidth,

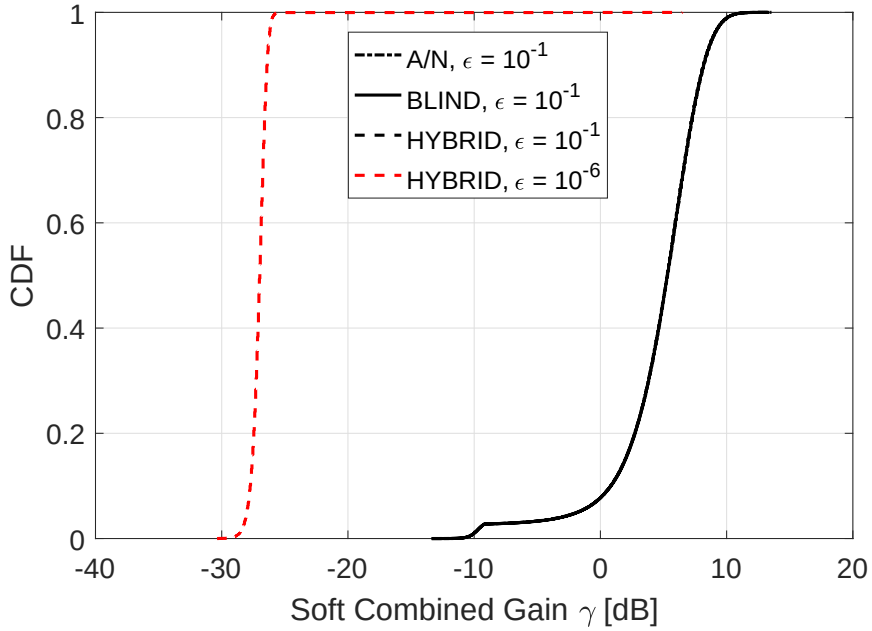

Figure 6.8: CDF of the SNR of the 1st TX that caused 1 NACK.

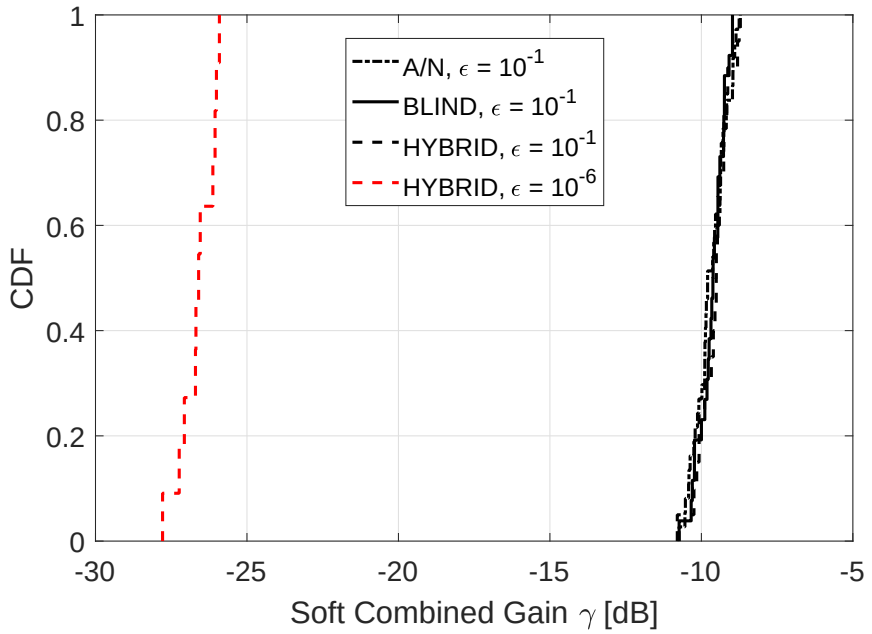

Figure 6.9: CDF of the combined gain γ of the 1st Retransmission that caused a second NACK.

comparable to the result evaluated in section 5.2.2. The $\text{SNR}_{\text{start}}$ is reduced to a lower value than -25 dB.

If we state that in Figure 6.1 the experienced BLER at the first transmission is roughly $10^{-3} \ll 0.038$, it is most likely that the first transmission failures are caused by low SNR values.

Figure 6.8 confirms the statement, as the SNR at the 1st transmission that caused the 1st NACK is below -25 dB with probability $\backsimeq 1$.

The same consideration is valid also for the schemes using a BLER target 10^{-1}, as the step in Figure 6.8 around -10 dB leads to a curve in 6.8 still distributed around that value.

6.4.2 Coherence Time and Time Diversity

The last topic we want to address is how the velocity of the device can affect our retransmission performances.

We reported in Table 6.1 a Coherence Time $T_c = 50$ ms, corresponding to roughly 5 Km/h, now we want to check what happens if the device goes 4 time slower, so that $T_c = 200$ ms.

In Figure 6.10 the Hybrid scheme for UpLink experiences a probability to have 2 retransmission that is higher when T_c is higher, that is when the device has a lower maximum velocity.

As we anticipated in section 3.1, we have a lower *diversity* in time when the device is slow, having more time-correlated channel realizations, leading to a fixed channel when the device is not moving.

We can state that when deep fade occurs the transmission fails and, by comparing the two T_c values, a faster device has a higher gain in diversity, despite the URLLC purpose is matched in both cases.

Nevertheless we have in theory a better gain in retransmission with faster

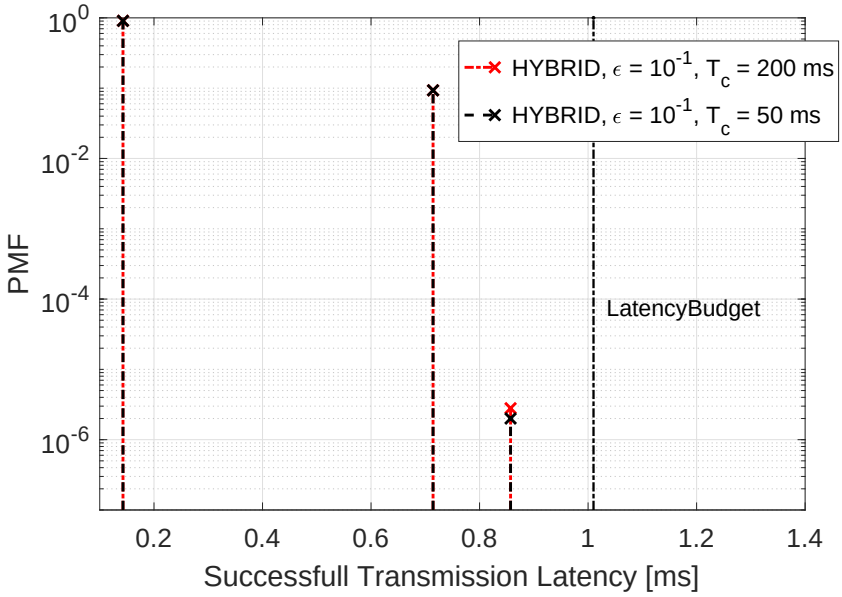

Figure 6.10: Difference in the retransmission performances. A maximum velocity $v = 4.57$ *Km/h leads to* $T_c = 50$ *ms, while* $v = 1.1425$ *Km/h leads to* $T_c = 200$ *ms.*

devices, with an obvious limit given by the degradation of the channel estimation accuracy.

Conclusion

In the thesis work we have analyzed the performance of the state of the art NACK-based and BLIND retransmission schemes and proposed a novel HYBRID scheme, whose purpose it to match with better flexibility the URLLC QoS latency budget requirement.

So the HYBRID scheme is able to achieve the QoS requirements with important savings in terms of resources w.r.t. the BLIND scheme, which is the most aggressive one.

The second contribution of this work is to tackle the differences between CC and IR performance. While in non critical working conditions IR typically outperforms CC in terms of equivalent SNR, when transmission experiences a deep fade their performance becomes almost similar.

Since deep fades dominate performance at the very high URLLC reliability targets, we have observed that CC becomes a strong candidate for its easier implementation.

Appendix A

Maximum Likelihood Estimation

A.1 Maximum Likelihood Estimation with Gaussian Noise

This first section recalls the theory about Maximum Likelihood Estimation (MLE) with Additive Gaussian Noise (AGN); it is based on the maximization of the probability density function (pdf) associated to a observation vector $\mathbf{x}$ with respect to an unknown parameter vector $\boldsymbol{\theta}$:

$$\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}_{ML} = \arg\max_{\boldsymbol{\theta}} p\left[\mathbf{x}|\boldsymbol{\theta}\right]. \tag{A.1}$$

With AGN, the relationship between observation vector $\mathbf{x}$ and parameters can be expressed as:

$$\mathbf{x} = \mathbf{h}\left(\boldsymbol{\theta}\right) + \mathbf{w}, \tag{A.2}$$

where $\mathbf{h}(\cdot)$ is in principle a non-linear function and $\mathbf{w}$ is the noise vector. Assuming that the quantity $\mathbf{h}(\boldsymbol{\theta})$ is purely deterministic and that the noise pdf $p_w[\mathbf{w}]$ is known, it is possible to write for the additive noise model the likelihood function as

$$p[\mathbf{x}|\boldsymbol{\theta}] = p_w[\mathbf{x} - \mathbf{h}(\boldsymbol{\theta})]. \tag{A.3}$$

If we compute $\log(\cdot)$ of the likelihood function, the log-likelihood function is defined:

$$\mathcal{L}(\mathbf{x}|\boldsymbol{\theta}) = \ln p(\mathbf{x}|\boldsymbol{\theta}),$$

and therefore the MLE is given by its maximization [20]

$$\boldsymbol{\theta}_{ML}(\mathbf{x}) = \arg\max_{\boldsymbol{\theta}} \mathcal{L}(\mathbf{x}|\boldsymbol{\theta}).$$

If we define the error $\delta\hat{\boldsymbol{\theta}}$ as the difference between estimation of parameter and parameter itself, the bias of the MLE is defined as

$$\mathbf{b}(\hat{\boldsymbol{\theta}}) = E[\delta\hat{\boldsymbol{\theta}}] = E[\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}]. \tag{A.4}$$

For a set of parameter $\boldsymbol{\theta}$, we can write the Mean Square Error (MSE) of their MLE

$$\mathbf{MSE}(\hat{\boldsymbol{\theta}}) = E[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^H] = \mathrm{cov}[\hat{\boldsymbol{\theta}}] + \mathbf{b}(\hat{\boldsymbol{\theta}})\mathbf{b}(\hat{\boldsymbol{\theta}})^H, \tag{A.5}$$

and in case of single-parameter estimation we can simplify the equation above into

$$\mathrm{MSE}(\hat{\theta}) = E[|\hat{\theta} - \theta|^2] = \mathrm{var}(\hat{\theta}) + |\mathrm{b}(\hat{\theta})|^2, \tag{A.6}$$

where the MLE variance $\text{var}(\hat{\theta})$ is an indicator on the uncertainty of the estimator, and $\text{b}(\hat{\theta})$ its bias.

Well designed ML estimator are asymptotically *unbiased*, i.e. $\lim_{N\to\infty} \mathbf{b}(\hat{\boldsymbol{\theta}}) \to \mathbf{0}$, and $\lim_{N\to\infty} E[\hat{\boldsymbol{\theta}}] \to \boldsymbol{\theta}$. Hence $\mathbf{MSE}(\hat{\boldsymbol{\theta}})$ is composed by the covariance matrix of the estimator, or its variance in case of single-parameter estimation.

A.1.1 Linear Gaussian model $\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w}$ with $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{C}_w)$

By assuming that $\mathbf{h}(\cdot)$ consists in a pre-multiplication by the matrix H, the likelihood can be written as the gaussian pdf of the noise centered in the mean value, as explained in Equation (A.3)

$$p(\mathbf{x}|\boldsymbol{\theta}) = \frac{1}{(2\pi)^N |\mathbf{C}_w|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}\left[\mathbf{x} - \mathbf{H}\boldsymbol{\theta}\right]^H \mathbf{C}_w^{-1}\left[\mathbf{x} - \mathbf{H}\boldsymbol{\theta}\right]\right\},$$

where $|\cdot|$ is the matrix determinant, $(\cdot)^H$ the transposed matrix of conjugate elements, defined as Hermitian transposition, and the log-likelihood function is given by

$$\mathcal{L}\left(\mathbf{x}|\boldsymbol{\theta}\right) = -\frac{N}{2}\ln\left(2\pi\right) - \frac{1}{2}\ln|\mathbf{C}_w| - \frac{1}{2}\left[\mathbf{x} - \mathbf{H}\boldsymbol{\theta}\right]^H \mathbf{C}_w^{-1}\left[\mathbf{x} - \mathbf{H}\boldsymbol{\theta}\right].$$

The MLE can be computed as

$$\arg\max_{\boldsymbol{\theta}} \mathcal{L}\left(\mathbf{x}|\boldsymbol{\theta}\right) = \boldsymbol{\theta}_{ML} = \arg\min_{\boldsymbol{\theta}}\{\left[\mathbf{x} - \mathbf{H}\boldsymbol{\theta}\right]^H \mathbf{C}_w^{-1}\left[\mathbf{x} - \mathbf{H}\boldsymbol{\theta}\right]\}$$

The minimization of the quadratic form leads to the MLE

$$\boldsymbol{\theta}_{ML} = \underbrace{\left(\mathbf{H}^H\mathbf{C}_w^{-1}\mathbf{H}\right)^{-1}\mathbf{H}^H\mathbf{C}_w^{-1}}_{\mathbf{A}}\mathbf{x} = \mathbf{A}\mathbf{x}. \tag{A.7}$$

If the estimator is well designed, the bias (A.4) is null, so the MSE of the estimator (A.5) reduces itself in

$$\begin{aligned}\mathbf{MSE}(\boldsymbol{\theta}_{ML}) = \mathrm{cov}[\boldsymbol{\theta}_{ML}] = E[(\boldsymbol{\theta}_{ML} - \boldsymbol{\mu}_{ML})(\boldsymbol{\theta}_{ML} - \boldsymbol{\mu}_{\mathrm{ML}})^H] = \\ = E[\boldsymbol{\theta}_{ML}{\boldsymbol{\theta}_{ML}}^H] - \boldsymbol{\mu}_{ML}{\boldsymbol{\mu}_{ML}}^H,\end{aligned}$$

having that $\boldsymbol{\mu}_{ML} = E[\boldsymbol{\theta}_{ML}] = E[\mathbf{Ax}] = \mathbf{A}\boldsymbol{\mu}_{\mathbf{x}}$. Hence it is possible to write the MSE as

$$\begin{aligned}\mathbf{MSE}(\boldsymbol{\theta}_{ML}) &= E[\mathbf{A}\mathbf{x}\mathbf{x}^H\mathbf{A}^H - \mathbf{A}\boldsymbol{\mu}_{\mathbf{x}}{\boldsymbol{\mu}_{\mathbf{x}}}^H\mathbf{A}^H] \\ &= \mathbf{A}E[\mathbf{x}\mathbf{x}^H - \boldsymbol{\mu}_{\mathbf{x}}{\boldsymbol{\mu}_{\mathbf{x}}}^H]\mathbf{A}^H = \mathbf{A}\mathbf{C}_x\mathbf{A}^H. \qquad \text{(A.8)}\end{aligned}$$

We exploit the fact that $\mathbf{x}$ and $\mathbf{w}$ have pdfs with the same second order moments, leading to $\mathbf{C}_x = \mathbf{C}_w$. By substituting the value of A in the equation we obtain the following relationship

$$\begin{aligned}\mathbf{MSE}(\hat{\boldsymbol{\theta}}) &= \overbrace{\left(\mathbf{H}^H\mathbf{C}_w^{-1}\mathbf{H}\right)^{-1}\mathbf{H}^H\mathbf{C}_w^{-1}}^{\mathbf{A}} \cdot \mathbf{C}_w \cdot \overbrace{\mathbf{C}_w^{-1}\mathbf{H}\left(\mathbf{H}^H\mathbf{C}_w^{-1}\mathbf{H}\right)^{-1}}^{\mathbf{A}^H} \\ &= \left(\mathbf{H}^H\mathbf{C}_w^{-1}\mathbf{H}\right)^{-1} \qquad \text{(A.9)}\end{aligned}$$

that is the principal concept of the soft combining mechanism of HARQ-CC, as it is discussed in next sections.

A.2 Chase Combining HARQ model

We consider an AWGN channel where we try to correct residual block errors by retransmitting the same block of symbols and applying CC-HARQ. Accordingly, for each symbol we want to estimate the same scalar signal s over N transmissions, stored in a vector Nx1 of independent noisy measure-

ments $\mathbf{x} = \mathbf{h}s + \mathbf{n}$. Here the matrix $\mathbf{H}$ is reduced to a Nx1 vector $\mathbf{h}$ because it is a single-parameter estimation.

The problem by itself is already solved by the previous Section. Nevertheless, we would like to focus and derive the gain provided by CC-HARQ in case of Perfect CSI (P-CSI).

Therefore, the Linear Gaussian Model of Section 2 is suited to cope with the CC-HARQ problem, and we have the expression of the received signal

$$\mathbf{r} = \mathbf{h}s + \mathbf{n} \tag{A.10}$$

where the complex vector $\mathbf{h}$ is the channel vector, whose elements are equivalent complex equivalent channels at the receiver and $\mathbf{n}$ is a zero-mean complex Gaussian vector with diagonal covariance matrix $\mathbf{C_n} = \text{diag}\left(\sigma_1^2, ..., \sigma_N^2\right)$. The receiver, conversely, makes use of an estimation of the channel and the noise figure in the decoding procedure, leading to a different model for the received signal

$$\hat{\mathbf{r}} = \hat{\mathbf{h}}s + \mathbf{n}', \tag{A.11}$$

where the estimated channel vector is $\hat{\mathbf{h}}$ and noise covariance matrix is $\hat{\mathbf{C}}_\mathbf{n} = \text{diag}\left(\hat{\sigma}_1^2, ..., \hat{\sigma}_N^2\right)$.

Hence, when we use the estimator in eq. (A.7), the received signal, following the model in eq. (A.10), is linearly multiplied by the kernel $\mathbf{A}$ of the estimator itself, which exploits the estimations of channel vector and noise covariance matrix extracted from the estimated model in eq. (A.11).

In the next section we will derive the formulae for performance of CC-HARQ and discuss the implications of Perfect (P-CSI).

A.3 Performance Derivation

By exploiting the theory of Section A.1, one can substitute quantities from Equation (A.7), to derive the MLE for the signal s

$$\hat{\mathrm{s}}_{ML} = \left(\hat{\mathbf{h}}^H \hat{\mathbf{C}}_{\mathbf{n}}^{-1} \hat{\mathbf{h}}\right)^{-1} \hat{\mathbf{h}}^H \hat{\mathbf{C}}_{\mathbf{n}}^{-1} \mathbf{r} = \left(\hat{\mathbf{h}}^H \hat{\mathbf{C}}_{\mathbf{n}}^{-1} \hat{\mathbf{h}}\right)^{-1} \hat{\mathbf{h}}^H \hat{\mathbf{C}}_{\mathbf{n}}^{-1} (\mathbf{h}s + \mathbf{n}). \quad \text{(A.12)}$$

One should note that all the elements q given by $\mathbf{x^H A y}$, where x and y are column vectors with N elements and $\mathbf{A} = \mathrm{diag}(A_1, \cdots, A_N)$, are scalar and they can be written as

$$q = \mathbf{x^H A y} = \sum_{i=1}^{N} x_i y_i A_i. \quad \text{(A.13)}$$

Therefore, one can write and simplify the estimation error as follows

$$\begin{aligned}
\epsilon = \hat{\mathrm{s}}_{ML} - s &= \left(\hat{\mathbf{h}}^H \hat{\mathbf{C}}_{\mathbf{n}}^{-1} \hat{\mathbf{h}}\right)^{-1} \hat{\mathbf{h}}^H \hat{\mathbf{C}}_{\mathbf{n}}^{-1} (\mathbf{h}s + \mathbf{n}) - s = \\
&= \left(\left(\hat{\mathbf{h}}^H \hat{\mathbf{C}}_{\mathbf{n}}^{-1} \hat{\mathbf{h}}\right)^{-1} \hat{\mathbf{h}}^H \hat{\mathbf{C}}_{\mathbf{n}}^{-1} \mathbf{h} - 1\right) s + \left(\hat{\mathbf{h}}^H \hat{\mathbf{C}}_{\mathbf{n}}^{-1} \hat{\mathbf{h}}\right)^{-1} \hat{\mathbf{h}}^H \hat{\mathbf{C}}_{\mathbf{n}}^{-1} \mathbf{n} = \\
&= \left(\frac{\sum_{i=1}^{N} (\hat{h}_i^* h_i)/\hat{\sigma}_i^2}{\sum_{i=1}^{N} |\hat{h}_i|^2/\hat{\sigma}_i^2} - 1\right) s + \frac{\sum_{i=1}^{N} (\hat{h}_i^*/\hat{\sigma}_i^2) n_i}{\sum_{i=1}^{N} |\hat{h}_i|^2/\hat{\sigma}_i^2}, \quad \text{(A.14)}
\end{aligned}$$

and compute its bias μ_{CC} and variance σ^2_{CC} to write the MSE of the CC-HARQ estimate

$$\begin{aligned}
\mu_{\mathrm{CC}} = E[\epsilon] &= \left(\frac{\sum_{i=1}^{N} (\hat{h}_i^* h_i)/\hat{\sigma}_i^2}{\sum_{i=1}^{N} |\hat{h}_i|^2/\hat{\sigma}_i^2} - 1\right) s + \frac{\sum_{i=1}^{N} (\hat{h}_i^*/\hat{\sigma}_i^2) 0}{\sum_{i=1}^{N} |\hat{h}_i|^2/\hat{\sigma}_i^2} = \\
&= \left(\frac{\sum_{i=1}^{N} (\hat{h}_i^* h_i)/\hat{\sigma}_i^2}{\sum_{i=1}^{N} |\hat{h}_i|^2/\hat{\sigma}_i^2} - 1\right) s, \quad \text{(A.15)}
\end{aligned}$$

$$\sigma^2_{\mathrm{CC}} = E[(\epsilon - \mu_{\mathrm{CC}})(\epsilon - \mu_{\mathrm{CC}})^*] = E\left[\frac{\sum_{i=1}^{N}(\hat{h}_i^*/\hat{\sigma}_i^2)n_i}{\sum_{i=1}^{N}|\hat{h}_i|^2/\hat{\sigma}_i^2}\frac{\sum_{j=1}^{N}(\hat{h}_j/\hat{\sigma}_j^2)n_j^*}{\sum_{i=1}^{N}|\hat{h}_i|^2/\hat{\sigma}_i^2}\right] =$$

$$= \frac{\sum_{i=1}^{N}\sum_{j=1}^{N}(\hat{h}_i^*/\hat{\sigma}_i^2)(\hat{h}_j/\hat{\sigma}_j^2)E[n_i n_j^*]}{\left(\sum_{i=1}^{N}|\hat{h}_i|^2/\hat{\sigma}_i^2\right)^2} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{N}(\hat{h}_i^*\hat{h}_j)/(\hat{\sigma}_i^2\hat{\sigma}_j^2)\sigma_i^2\delta(i=j)}{\left(\sum_{i=1}^{N}|\hat{h}_i|^2/\hat{\sigma}_i^2\right)^2} =$$

$$= \frac{\sum_{i=1}^{N}(|\hat{h}_i|^2\sigma_i^2)/\left(\hat{\sigma}_i^2\right)^2}{\left(\sum_{i=1}^{N}|\hat{h}_i|^2/\hat{\sigma}_i^2\right)^2} = \frac{\sum_{i=1}^{N}(|\hat{h}_i|^2/\hat{\sigma}_i^2)\left(\sigma_i^2/\hat{\sigma}_i^2\right)}{\left(\sum_{i=1}^{N}|\hat{h}_i|^2/\hat{\sigma}_i^2\right)^2}, \tag{A.16}$$

$$\mathrm{MSE} = |\mu_{\mathrm{CC}}|^2 + \sigma^2_{\mathrm{CC}}. \tag{A.17}$$

You can notice that this definition of the MSE coincides with Equation (A.6), since the variance of the estimator is equivalent to the variance of the error.

A.3.1 Perfect CSI impact on CC-HARQ MSE

When we have perfect CSi at the receiver, the estimations of the channel and the noise variances corresponds to the real ones, such that $\hat{\mathbf{h}} = \mathbf{h}$ and $\hat{\sigma}_i^2 = \sigma_i^2$. In this case, the MLE bias and variance can be expressed as

$$\mu_{\mathrm{CC}}^{(P)} = \left(\frac{\sum_{i=1}^{N}(\hat{h}_i^*\hat{h}_i)/\hat{\sigma}_i^2}{\sum_{i=1}^{N}|\hat{h}_i|^2/\hat{\sigma}_i^2} - 1\right)s = 0, \tag{A.18}$$

$$\sigma_{\mathrm{CC}}^{2,(P)} = \frac{\sum_{i=1}^{N}(|\hat{h}_i|^2/\hat{\sigma}_i^2)\left(\hat{\sigma}_i^2/\hat{\sigma}_i^2\right)}{\left(\sum_{i=1}^{N}|\hat{h}_i|^2/\hat{\sigma}_i^2\right)^2} = \frac{1}{\left(\sum_{i=1}^{N}|\hat{h}_i|^2/\hat{\sigma}_i^2\right)} = \frac{1}{\left(\sum_{i=1}^{N}|h_i|^2/\sigma_i^2\right)}, \tag{A.19}$$

$$\mathrm{MSE} = \sigma_{\mathrm{CC}}^{2,(P)}. \tag{A.20}$$

If we evaluate the inverse of the MSE as performance metric of the CC estimate, we come back the HARQ-CC equivalent SNR in case of P-CSI

[21]:

$$\lambda^{(P)} = (\mathrm{MSE})^{-1} = \sigma_{\mathrm{CC}}^{-2,(P)} = \sum_{i=1}^{N} |h_i|^2/\sigma_i^2 = \sum_{i=1}^{N} \mathrm{SNR}_i, \tag{A.21}$$

where $\mathrm{SNR}_i = |h_i|^2/\sigma_i^2$ is the i-th transmission true SNR at the receiver.

Bibliography

[1] Mark Weiser. The computer for the 21st century. *Mobile Computing and Communications Review*, 3(3):3–11, 1999.

[2] David Soldani and Antonio Manzalini. Horizon 2020 and beyond: on the 5g operating system for a true digital society. *IEEE Vehicular Technology Magazine*, 10(1):32–42, 2015.

[3] J. C. Guey, P. K. Liao, Y. S. Chen, A. Hsu, C. H. Hwang, and G. Lin. On 5G Radio Access Architecture and Technology [Industry Perspectives]. *IEEE Wireless Communications*, 22(5):2–5, October 2015.

[4] Petar Popovski. Ultra-reliable communication in 5g wireless systems. In *5G for Ubiquitous Connectivity (5GU), 2014 1st International Conference on*, pages 146–151. IEEE, 2014.

[5] Yury Polyanskiy, H Vincent Poor, and Sergio Verdú. Channel coding rate in the finite blocklength regime. *IEEE Transactions on Information Theory*, 56(5):2307–2359, 2010.

[6] Shahram Zarei. Channel coding and link adaptation. 2009.

[7] Mohammad T Kawser, Nafiz Imtiaz Bin Hamid, Md Nayeemul Hasan, M Shah Alam, and M Musfiqur Rahman. Downlink snr to cqi mapping

for different multipleantenna techniques in lte. *International Journal of Information and Electronics Engineering*, 2(5):757, 2012.

[8] Guillermo Pocovi, Beatriz Soret, Klaus I Pedersen, and Preben Mogensen. Mac layer enhancements for ultra-reliable low-latency communications in cellular networks. In *Communications Workshops (ICC Workshops), 2017 IEEE International Conference on*, pages 1005–1010. IEEE, 2017.

[9] R1-1611223. Performance evaluation of ul urllc schemes. Technical report, 3GPP TSG RAN WG1 Meeting #87, 2016.

[10] Claude E Shannon. A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, 5(1):3–55, 2001.

[11] Giuseppe Durisi, Tobias Koch, and Petar Popovski. Towards massive, ultra-reliable, and low-latency wireless communication with short packets. *arXiv preprint arXiv:1504.06526*, 2015.

[12] Claude E Shannon. A mathematical theory of communication, part i, part ii. *Bell Syst. Tech. J.*, 27:623–656, 1948.

[13] Lawrence H Ozarow, Shlomo Shamai, and Aaron D Wyner. Information theoretic considerations for cellular mobile radio. *IEEE transactions on Vehicular Technology*, 43(2):359–378, 1994.

[14] David Tse and Pramod Viswanath. *Fundamentals of wireless communication*. Cambridge university press, 2005.

[15] Theodore S Rappaport et al. *Wireless communications: principles and practice*, volume 2. prentice hall PTR New Jersey, 1996.

[16] Frank Frederiksen and Troels Emil Kolding. Performance and modeling of wcdma/hsdpa transmission/h-arq schemes. 1:472–476, 2002.

[17] 5G PPP Use Cases and Performance Evaluation Models. Technical Report v1, 5G PPP Europe, April 2016.

[18] Hamidreza Shariatmadari, Zexian Li, Sassan Iraji, Mikko A Uusitalo, and Riku Jäntti. Control channel enhancements for ultra-reliable low-latency communications. In *Communications Workshops (ICC Workshops), 2017 IEEE International Conference on*, pages 504–509. IEEE, 2017.

[19] R1-1609664. Comparison of slot and mini-slot based approaches for urllc. Technical report, GPP TSG-RAN WG1 #86bis,, 2016.

[20] Umberto Spagnolini. *Statistical Signal Processing in Engineering*. Draft to be published (Ask Polimi contacts for the pdf), 2015.

[21] Roshni Srinivasan. Ieee 802.16 m evaluation methodology document. *IEEE 802.16 m-08/004r5*, 2009.