The Path Towards Ultra-reliable Low-latency Communications via HARQ

João Pedro Battistella Nadas

Submitted in fulfilment of the requirements for the
Degree of Doctor of Philosophy

School of Engineering
College of Science and Engineering
University of Glasgow

April 2021

Abstract

Ultra-reliable Low-latency Communications (URLLC) is potentially one of the most disruptive communication paradigms offered by the next generation of wireless networks, 5G. This is easily demonstrated by the diverse set of applications it enables, such as autonomous driving; remote surgery; wireless networked control systems; mission-critical machine type communication; and many more. Basically, URLLC consists of the almost 100% guarantee of message delivery within a very short time interval. Furthermore, the pressure from climate change coupled with the massive growth of cellular networks expected to occur in the near future means that URLLC must also be energy efficient. On its own, achieving low-latency with high reliability is already a stringent requirement, but when that is coupled with the need for resource efficiency, it becomes even more challenging. That is the motivation behind this thesis: to study URLLC in the context of resource efficiency. Thus, a study of the counterintuitive use of retransmissions, more specifically Hybrid Automatic Repeat Request (HARQ), in the scenario of URLLC is proposed and carried out. HARQ is very attractive in terms of resource efficiency, and that is the motivation behind using it even when stringent time constraints are imposed. Four contributions are made by the present work. **Firstly**, a mathematical problem is presented and solved for optimizing the number of allowed retransmission rounds considering HARQ in URLLC, considering both energy efficiency as well as electromagnetic irradiation. This representation relies on a few assumptions in order to be realizable in practical scenarios. Namely, these assumptions are regarding the possibility of early error detection for sending the feedback signals and on not having to consider medium access control introduced delays. **Secondly**, we consider one important aspect of wireless systems, which is that they can be greatly optimized if they are designed with a specific application in mind. Based on this, a study of the use of HARQ specifically tuned for Networked Control Systems is presented, taking into account the particular characteristics of these applications. Results here show that fine-tuning for the specific characteristics of these applications yields better results when compared to using the results from the previous contribution, which are more application-agnostic. These improved results are possible thanks to the exploitation of application-specific characteristics, more specifically the use of a packetized predictive control strategy jointly designed with the communication protocol. Next, the concept of

HARQ for URLLC is extended to a larger scale in an effort to relax the aforementioned assumptions. This is studied within the framework of self-organizing networks and leverages machine learning algorithms in order to overcome those strict assumptions from the first contribution. This is demonstrated by developing a digital twin simulation of the city of Glasgow and generating a large dataset of users in the cellular network, which is a **third** contribution of this thesis. Then, machine learning (more specifically long short-term convolutional neural networks) is applied for predicting message failures. Lastly, a protocol to exploit such predictions in combination with HARQ to deliver downlink URLLC is applied, resulting in a **fourth** contribution. In summary, this thesis presents a latency aware HARQ technique which is shown to be very efficient. We show that it uses up as much as 18 times less energy than a frequency diversity strategy and that it can emit more than 10 times less energy electromagnetic field radiation when compared to the same strategy. We also propose joint design techniques, where communication and control parameters are tweaked at the same time, enabling wireless control systems with a three-fold reduction in required bandwidth to achieve URLLC requirements. Lastly, we present a digital twin of the city of Glasgow which enables us to create a prediction algorithm for predicting channel quality with very high accuracy—root mean square error on the order of 10^{-2}. This ties into the rest of the contributions as it can be used to enable early feedback detection, which in turn can be used to make sure the latency aware protocol can be employed.

List of Publications

Journal Papers

- **J. P. B. Nadas**, O. Onireti, R. D. Souza, H. Alves, G. Brante, and M. A. Imran, “Performance Analysis of Hybrid ARQ for Ultra-Reliable Low Latency Communications”, IEEE Sensors Journal, vol. 19, no. 9, pp. 3521–3531, May 2019.

- M. Ozturk, A. I. Abubakar, **J. P. B. Nadas**, R. N. B. Rais, S. Hussain, M. A. Imran, “Energy Optimization in Ultra-Dense Radio Access Networks via Traffic-Aware Cell Switching’ ’. Accepted for publication IEEE Transactions on Green Communications and Networking, 2021.

- F. A. Asuhaimi, S. Bu, **J. P. B. Nadas**, and M. A. Imran, “Delay-Aware Energy-Efficient Joint Power Control and Mode Selection in Device-to-Device Communications for FREEDM Systems in Smart Grids,” IEEE Access, vol. 7, pp. 87369–87381, 2019.

- P. V. Klaine, **J. P. B. Nadas**, R. D. Souza, and M. A. Imran, “Distributed Drone Base Station Positioning for Emergency Cellular Networks Using Reinforcement Learning,” Cognitive Computation, vol. 10, no. 5, pp. 790–804, May 2018.

Book Chapters

- **J. P. B. Nadas**, R. D. Souza, G. Brante, O. Onireti, H. Alves, and M. A. Imran, “Reducing EMF Emissions in Ultra-Reliable Low Latency Communications with HARQ’ ’. Low Electromagnetic Emission Wireless Network Technologies: 5G and Beyond, edited by M. A. Imran, IET, pp. 231-250, 2019@.

- **J. P. B. Nadas**, G. Zhao, R. D. Souza, and M. A. Imran, “Ultra Reliable Low Latency Communications as an Enabler For Industry Automation’ ’, Wireless Automation as an Enabler for the Next Industrial Revolution, edited by Q. H. Abbasi, Wiley, pp. 89-107, Dec. 2019@.

- **J. P. B. Nadas**, P. Valente Klaine, R. de Paula Parisotto, and R. D. Souza, “Intelligent Positioning of UAVs for Future Cellular Networks,” Enabling 5G Communication Systems to Support Vertical Industries, pp. 217–232, Jun. 2019.

- M. A. Imran, A. Turkmen, M. Ozturk, **J. P. B. Nadas**, and Q. H. Abbasi. “Seamless Indoor/Outdoor Coverage in 5G.’ ’, Wiley 5G Ref: The Essential 5G Reference, edited by R. Tafazolli, Wiley, pp. 1-23, 2019@.

Conference Papers

- **J. P. B. Nadas**, M. Jaber, S. v. d. Berghe, and M. A. Imran, “Towards Continuous Subject Identification Using Wearable Devices and Deep CNNs’ ’, 2019 IEEE International Conference on Communications (ICC), Jun. 2020@.

- M. Ozturk, **J. P. B. Nadas**, P. H. V. Klaine, S. Hussain, and M. A. Imran. “Clustering Based UAV Base Station Positioning for Enhanced Network Capacity.’ ’ 2019 International Conference on Advances in the Emerging Computing Technologies (AECT), pp. 1-6. IEEE, Feb. 2020.

- R. P. Parisotto, P. V. Klaine, **J. P. B. Nadas**, R. D. Souza, G. Brante, and M. A. Imran, “Drone Base Station Positioning and Power Allocation using Reinforcement Learning,” 2019 16th International Symposium on Wireless Communication Systems (ISWCS), Aug. 2019.

- **J. P. B. Nadas**, P. Klaine, L. Zhang, G. Zhao, M. Imran, and R. Souza, “Performance Analysis of Early-HARQ for Finite Block-Length Packet Transmission,” 2019 IEEE International Conference on Industrial Cyber Physical Systems (ICPS), May 2019.

- A. Rizwan, **J. P. B. Nadas**, M. A. Imran, and M. Jaber, “Performance Based Cells Classification in Cellular Network using CDR Data,” 2019 IEEE International Conference on Communications (ICC), May 2019.

- Y. A. Sambo, P. V. Klaine, **J. P. B. Nadas**, and M. A. Imran, “Energy Minimization UAV Trajectory Design for Delay-Tolerant Emergency Communication,” 2019 IEEE International Conference on Communications Workshops (ICC Workshops), May 2019.

- F. A. Asuhaimi, **J. P. B. Nadas**, and M. A. Imran, “Delay-optimal mode selection in device-to-device communications for smart grid,” 2017 IEEE International Conference on Smart Grid Communications (SmartGridComm), Oct. 2017.

- **J. P. B. Nadas**, M. A. Imran, G. Brante, and R. D. Souza, "Optimizing the energy efficiency of short term ultra reliable communications in vehicular networks," 2017 15th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt), May 2017.

Contents

Abstract **i**

List of Publications **iii**

List of Acronyms **xiii**

List of Symbols **xvi**

Acknowledgements **xxi**

Declaration **xxii**

1 Introduction **1**

1.1 Background 1

1.2 Related Works 3

1.2.1 HARQ for URLLC 3

1.2.2 Communication and Control JD 4

1.2.3 SON Enabled HARQ 5

1.3 Motivation 7

1.4 Contributions 8

1.5 Organization 11

2 Technical Background **12**

2.1 Large Scale Loss — Path Loss 12

2.2 Small Scale Loss – Multi-Path Fading 14

2.2.1 Wireless Communication Model 15

2.3 Error Probability at Finite Block Length 15

2.4 Diversity 16

2.5 Probability of Error for HARQ 19

2.6 Number of Transmission Attempts 20

2.7 Channel Estimation 21

2.7.1 Effect of Imperfect Channel Estimation 22

3 Latency Aware HARQ **24**
3.1 Introduction 24
3.1.1 Summary 26
3.2 System Model 26
3.2.1 Energy Consumption Model 27
3.2.2 Latency Constraint 30
3.3 Optimization 31
3.3.1 Optimizing Maximum Energy 32
3.3.2 Discussion Around Convexity 35
3.3.3 Derivation of (3.30) 37
3.3.4 Discussion 37
3.3.5 Optimizing EMF Radiation 38
3.4 Numerical Results 40
3.4.1 Energy Efficiency Optimization 40
3.4.2 EMF Radiation Optimization 44
3.5 Conclusion 50

4 Communication and Control Joint Design **52**
4.1 Introduction 52
4.2 System Model 54
4.3 Optimizing the Bandwidth 55
4.3.1 Joint Design with New Messages 56
4.3.2 Joint Design with MRC 58
4.4 Simulations 59
4.4.1 Baseline Scenarios 62
4.4.2 Comparison 62
4.5 Conclusion 67

5 SON Enabled HARQ in URLLC **69**
5.1 Introduction 69
5.2 Digital Twin 71
5.2.1 NS3 71
5.2.2 Mobile Network 72
5.2.3 Users 74
5.2.4 Channel Model 74
5.2.5 Collected Metrics 74
5.3 Processing the Data 75
5.3.1 Extracting Meaning from the Data 75
5.3.2 Matching Sampling Interval 76

5.4 LSTM 79
5.5 E-HARQ Protocol 80
5.5.1 Performance Evaluation Framework 82
5.6 Numerical Results 84
5.7 Conclusion 89

6 Conclusion and Future Research 91
6.1 Conclusion 91
6.2 Future Research 92

Bibliography 94

List of Tables

3.1 CC-HARQ Energy Efficiency Optimization Simulation Parameters 41
3.2 CC-HARQ EMF Radiation Minimization Simulation Parameters 46

4.1 WNCS Simulation Parameters 61

5.1 Information about the eNBs 73
5.2 Statistical Analysis of the predictions. 86

List of Figures

2.1 Effect of imperfect channel estimation on the average SNR for different values of ρ and $\bar{\gamma}_{\mathrm{d}}$. A peak power constraint is considered such that the maximum $\bar{\gamma}_{\mathrm{d}}$ is 40dB, which is the same value for $\bar{\gamma}_{\mathrm{p}}$. 23

3.1 Performance of using CC-HARQ, showing the diversity gains and the average number of transmissions. In this example $R = 1$ and $m = 1$ are used. 25

3.2 Illustration of the reason why a maximum latency constraint imposes a minimum rate. In this example, $z = 3$. 31

3.3 Average energy used ($\bar{E}_{\mathrm{b}}$) and maximum possible energy required ($\hat{E}_{\mathrm{b}}$) versus z for various strategies, considering $m = 1$ and $m = 3$. As a reminder, R_{min} is the rate which ensures the latency constraint, $\bar{\gamma}_{\mathrm{d}}$ is the SNR of the data forward transmition, $\bar{\gamma}_{\mathrm{max}}$ is the maximum possible SNR obtained from peak power limitations and the link budget, $\bar{\gamma}_{\mathrm{min}}$ is the average SNR to guarantee the target reliability, and $\bar{\gamma}^{\star}$ is the optimal SNR in terms of energy considering the link budget, obtained numerically. 43

3.4 Performance for different λ with $T_{\mathrm{out}} = 10^{-5}$ and comparison between proposed scheme and the benchmark. The plot for $m = 1$ starts at a higher value because the URLLC contraints could not be met with the link budget and the peak power limitation. 44

3.5 Performance for different T_{out} with $\lambda = 6.48$ ms and comparison between proposed scheme and the benchmark. The plot for $m = 1$ starts at a higher value because the URLLC contraints could not be met with the link budget and the peak power limitation. 45

3.6 EMF radiation energy for different values of z, considering various LOS conditions and two different lengths of L_{T}. 46

3.7 Value of $z^{\star}$ for different maximum latencies λ. Here the predicted value is the one obtained from (3.58) for $\hat{z}$. 47

3.8 Value of the optimal number of attempts, $z^{\star}$, for different distances d. . . . 48

3.9 Value of the optimal number of attempts, $z^{\star}$, for different LOS conditions, represented by different values of m. 49
3.10 Average $\bar{\tau}$ and maximum $\hat{z}^{\star}$ number of transmission attempts for various target latencies λ. 49
3.11 EMF radiation for both the frequency diversity scheme and the proposed CC-HARQ scheme. 50

4.1 System model for a PPC solution using a wireless link between the controller and the actuators. 54
4.2 Illustration of PPC. 57
4.3 Illustration of the MRC approach to PPC. 60
4.4 Bandwidth required for different distances considering both JD strategies, a control only optimization and a communication only approach (with CC-HARQ). λ is set at 1 ms and $\epsilon_{\mathrm{c}} = 10^{-9}$. Here, $m = 1$. 63
4.5 Bandwidth required for different values of latency considering both JD strategies, a control only optimization and a communication only approach (with CC-HARQ). d is set at 10 m and $\epsilon_{\mathrm{c}} = 10^{-9}$. Here, $m = 1$. 64
4.6 Bandwidth required for different values of reliability considering both JD strategies, a control only optimization and a communication only approach (with CC-HARQ). d is set at 10 m and $\lambda = 1$ ms. Here, $m = 1$. 64
4.7 Minimum required bandwidth for each JD approach and various pairs of z and K, considering non-LOS ($m = 1$). 65
4.8 Minimum required bandwidth for each JD approach and various pairs of z and K. Here we consider that there is some LOS, and thus $m = 3$. 66
4.9 Bandwidth required for different distances considering both JD strategies, a control only optimization and a communication only approach (with CC-HARQ). λ is set at 1 ms and $\epsilon_{\mathrm{c}} = 10^{-9}$. Here, $m = 3$. 67
4.10 Bandwidth required for different values of reliability considering both JD strategies, a control only optimization and a communication only approach (with CC-HARQ). d is set at 10 m and $\lambda = 1$ ms. Here, $m = 3$. 68

5.1 eNBs and cells around Glasgow on the position of the digital twin. The arrows indicate the orientation of the antennae, while their colour indicates a particular downlink frequency and are included to show highlight how complete the digital twin is with respect to reality. 72
5.2 Instantaneous SINR for 3 pairs of Cell/UE. 77
5.3 First 50 samples of the original signals. 78
5.4 Signals with normalized sampling interval to 1 ms. 78
5.5 Averaged signals using a 100 samples moving average filter. 79

5.6 Training of the LSTM recurrent neural network, using two thirds of the signal to train. The remaining one third is used to test the learned behavior. 80

5.7 Example of how using R obtained from (5.1) makes it possible to meet the target latency. Note that here $z = 3$ and $n(\lambda) = (\lambda - \delta)B/z$, such that in total δ seconds are being used to decode the messages. 82

5.8 SINR Predictions for the pair (28,3). 85

5.9 SINR Predictions for the pair (54,2). 85

5.10 SINR Predictions for the pair (88,5). 86

5.11 Block error rate versus SNR for CC-HARQ and E-HARQ for $z = 2$ and $z = 4$ and with $m = 1$. 87

5.12 Block error rate versus SNR for CC-HARQ and E-HARQ for $z = 2$ and $z = 4$ and with $m = 3$. 88

5.13 Performance comparison between CC-HARQ and E-HARQ for $z = 3$ transmission attempts and different values of λ. 88

5.14 Performance comparison between CC-HARQ and E-HARQ for $\bar{\gamma} = 0$ dB and different values of z. 89

5.15 Performance comparison between CC-HARQ and E-HARQ at high SNR ($\bar{\gamma} = 10$ dB) and different values of z. 90

List of Acronyms

3GPP	3^{rd} Generation Partnership Project
5G	Fifth Generation of Mobile Communications
6G	Sixth Generation of Mobile Communications
ACK	Acknowledgement
API	Application Programming Interface
AWGN	Additive White Gaussian Noise
CC-HARQ	Chase combining HARQ
CNN	Convolutional Neural Network
CSI	Channel State Information
DL	Downlink
EMF	Electromagnetic Field
eNB	Evolved Node B
EURO-COST	Co-operative for Scientific and Technical research
EVA	Extended Vehicular A
FBL	Finite Block-length
HARQ	Hybrid Automatic Repeat Request
ID	Identifier
IMSI	International Mobile Subscriber Identity
INR	Incremental Redundancy
JD	Joint Design
KKT	Karush-Kuhn-Tucker
LOS	Line-of-sight
LSTM	Long short-term Memory

LTE	Long-term Evolution
MAC	Medium Access Control
METIS	Mobile and Wireless Communications Enablers for the Twenty-twenty Information Society
MIMO	Multiple-input Multiple-output
ML	Machine Learning
MPC	Model Predictive Controller
MRC	Maximum Ratio Combining
NACK	Non-acknowledgement
NLOS	Non-line-of-sight
NOMA	Non-orthogonal Multiple Access
NS3	Network Simulator 3
PA	Power Amplifier
PDCP	Packet Data Convergence Protocol
PDF	Probability Density Function
PHY	Physical
PL	Path Loss
PPC	Packetized Predictive Control
RF	Radio-frequency
RLC	Radio Link Control
RSRP	Reference Signal Received Power
S-ARQ	Automatic Repeat Request
SC	Selection Combining
SINR	Signal to Interference plus Noise Ratio
SNR	Singal to Noise Ratio
SON	Self-organizing Network
TS	Technical Specification
UE	User Equipment
UL	Uplink
URLLC	Ultra-reliable Low-latency Communications

WNCS	Wireless Networked Control System

List of Symbols

α	Path loss exponent
$\bar{\gamma}$	Average SNR
$\bar{\gamma}_{\mathrm{d}}$	Average SNR used into data transmission
$\bar{\gamma}_{\mathrm{eff}}$	Effective average SNR
$\bar{\gamma}_{\mathrm{eff}}^{\mathrm{freq}}$	Effective average SNR for the frequency diversity scheme
$\bar{\gamma}_{\mathrm{fb}}$	Average SNR for the feedback message
$\bar{\gamma}_{\mathrm{max}}^{\mathrm{freq}}$	Maximum average SNR for the frequency diversity scheme
$\bar{\gamma}_{\mathrm{p}}$	Average SNR used for channel estimation
$\bar{\tau}$	Average transmission attempts
$\bar{\varepsilon}$	Average error probability
$\bar{E}$	Average communication energy
$\bar{E}_{\mathrm{b}}$	Average communication energy per successful bit
$\bar{E}_{\mathrm{f}}$	Average communication energy using frequency diversity
β_{cc}	Auxiliary function for PPC optimization
β_{s}	Auxiliary function for PPC optimization
Δ	Auxiliar function of theorem 1
δ_{fb}	Time to decode feedback message
δ_{fw}	Time to decode forward message
δ_{s}	Decoding time for the E-HARQ case
δ_{e}	Decoding time for the S-ARQ case
ϵ_{c}	Error rate tolerable by the control plant
η	PA efficiency
Γ	Complete Gamma function

γ	SNR
γ_0	SNR threshold
Γ_{inc}	Lower incomplete Gamma fucntion
Γ_{inc}^{-1}	Inverse incomplete Gamma function
$\hat{E}_{\text{b}}$	Maximum communication energy per successful bit
$\hat{z}^{\star}$	Optimal number of transmission attempts
λ	Maximum latency for message exchange
λ'	Maximum latency of the application
Ω	Auxiliar function of theorem 1
Φ	Auxiliar function for objective function
ϕ	Number of operations per bit required for decoding messages
ρ	Number of pilot symbols
$\tilde{h}$	Estimated channel
ε	Target error probability
Ξ	Auxiliar function for the optimization problem
ζ	Auxiliar function used in the channel estimation
A	Auxiliar function for P_{PA}
a	COST231 auxiliary function
A_0	Loss at reference distance
C	Channel capacity
c	Speed of light
C_{m}	COST231 additional loss parameter
d	Link distance
d_0	Reference distance
$E_{\text{p,tx}}$	Energy used by the transmitter for channel estimation
E_{PA}	Energy irradiated by the transmitter
E_{st}	Startup energy
E_{tx}	Energy used by the transmitter
f	Auxiliar function to prove convexity

f_{apu}	Clock frequency of the arithmetic logic unit
f_{c}	Carrier frequency
g	Auxiliar function to prove convexity
G_{t}	Total antenna gain
$h(t)$	Channel gain
h_{base}	Height of base station antenna
h_{mobile}	Height of mobile device antenna
K	Length of the prediction horizon
L_{b}	Path loss in dB
L_{D}	Data length
L_{fb}	Feedback message length
L_{fw}	Forward message bit length
L_{H}	Header length
L_{T}	Total bit length
m	Parameter for the Nakagami-m distribution
M_{c}	Coding margin
M_{l}	Link margin
n	Channel uses
N_0	Noise power spectral density
n_{d}	Channel uses used for data transmission
n_{e}	Number of channel uses to decode the message for E-HARQ
n_{fb}	Number of feedback channel uses
n_{fw}	Number of forward channel uses
n_{s}	Number of channel uses to decode the message for S-ARQ
P	Power per channel use
p_{γ}	PDF of γ
$P_{\mathrm{el,rx}}$	Passband circuit power consumption at the receiver
$P_{\mathrm{el,tx}}$	Baseband RF circuit power consumption at the transmitter
P_{el}	Circuit energy consumption

P_{naka}	Nakagami-m distribution PDF
P_{out}	Outage probability
$P_{\text{PA, d}}$	PA power consumption for the payload message
$P_{\text{PA, fb}}$	PA power consumption for the feedback message
$P_{\text{PA, p}}$	PA power consumption for the channel estimation
P_{PA}	PA power consumption
p_{ray}	Rayleigh distribution PDF envelope
$P_{\text{rf,max}}$	Max instantaneous radiated power
P_{rf}	Radiated power
P_{r}	Received power
P_{t}	Transmitted power
$P_{\text{out},i}$	Outage probability at the i^{th} attempt
$P_{\text{out},z}^{\text{CC}}$	Outage probability of CC-HARQ
$P_{\text{out},z}^{\text{INR}}$	Outage probability of INR-HARQ
$P_{\text{out},z}^{\text{S}}$	Outage probability of S-HARQ
Q^{-1}	Inverse Q function
R	Rate in bpcu
$r(t)$	Received message
$R^{\star}$	Optimal communication rate
R_{b}	Rate in bits per second
R_{max}	Rate that guarantees the target outage at the maximum possible SNR
R_{min}	Minimum communication rate
$s(t)$	Message at the transmitter
T_{out}	Target outage
V	Channel dispersion
v	Optimal ratio of power for channel estimation
W	Bandwidth
$w(t)$	AWGN
W_0	Upper branch of the main Lambert-W function

W_{-1}	Lower part of the main Lambert-W function
z	Maximum number of transmissions

Acknowledgements

I want to thank everyone who was in any way involved in the elaboration of this thesis. I am not going to be able to name everyone, but I will do my best.

My supervisor, Professor Imran, who was always a patient mentor and who provided me with countless opportunities during the past 4 years.

All the people I met who were colleagues at first but then became very close friends: Abed, Aman, Andrea, Bruno, Giancarlo, Metin, and Paulo.

Richard who pushed to succeed from day one, Kayode who was always there to offer sound advice, Lei who was always friendly and encouraging, Mona who gave me the opportunity to explore new areas which ultimately gave me skills to find a good job, Yusuf who is one of the kindest persons I have ever known, Shuja whose inputs and suggestions saved me towards the finish line, and Linus Torvalds who created Linux and git.

My parents—Beatriz and Christian—who always invested heavily in my education, but who have also always provided any type of support that I needed. My brother Rafael who refuses to use LaTeX, and my cat Neo, whose emotional support was priceless.

Lastly, but most importantly, my wife Jessiane, who always supported me with comprehension, patience, and love.

Declaration

With the exception of chapters 1 and 2, which contain introductory material, all work in this thesis was carried out by the author unless otherwise explicitly stated.

Moreover, throughout this thesis, I employ the use of the plural “we” as a recognition for the input and valuable discussion I had with collaborators when researching the topic. Nonetheless, the work presented is mine, unless otherwise indicated.

Chapter 1

Introduction

1.1 Background

The path towards ultra-reliable and low-latency communications (URLLC) is still being outlined. URLLC is perhaps the most disruptive novelties brought by the current generation of mobile networks (5G) and despite many countries already having active 5G deployments, it still is a hot research topic. This is due to the fact that URLLC is still being standardized and the applications which will leverage it are still being matured.

Very few would have foreseen the role smartphones would play in our daily lives, greatly in part due to being always connected, when the first mobile data connections were being deployed. Yet, merely 30 years after the first mobile networks capable of data connectivity were deployed, billions of people throughout the world have the web on their fingertips running applications that were unimaginable at that time. Analogously, we can argue that the most disruptive URLLC applications are still to be crafted, once the capabilities are provided by the network.

That said, the roster of applications that is envisioned today presents tremendous potential to change our lives for the better in a multitude of ways. Remote drone control [1], wireless factory automation [2], remote surgery [3], smart-grid automation [4], and self-driving cars [5] are some examples of these applications.

As the name suggests, remote drone control consists of drones being controlled from a remote location, wirelessly. This already exists today with current network capabilities, whereby an operator can send a trajectory to a drone, who will compute the necessary control signals literally on the fly based on on-board sensors and take the actions accordingly. What URLLC brings to the mix, is the possibility of performing those computations on a remote controller [1]. This remote controller also receives the sensory information wirelessly and sends only the control signals to the flying device. This has a very important benefit that it reduces the cost and, perhaps more significantly, the weight and battery usage, of the drone significantly. However, for it to be possible, URLLC quality of service (QoS) standards must be guaranteed. Furthermore, if we can efficiently deploy a large scale of URLLC enabled drones, we could have them cooperating to perform a task that is orchestrated at the central controller, essentially transforming the system into a decoupled set of inexpensive actuators [6].

This is an interesting bridge for the next set of applications we are going to discuss: wireless factory automation [2, 7]. In this use-case, a set of controllers interact with the factory automation elements through wireless channels [8]. The wireless link may be the sensor to controller link, on the controller to actuator link, or on both links at the same time [9]. Several control plants require URLLC levels of QoS in these wireless links. Moreover, once we remove the need for sensors and actuators to be connected via wires, the engineer designing the application can explore new degrees of freedom, and treat a collective of these elements as the system, in a multi-variable control setup, that is flexible and easy to deploy in practice. However, this is only possible if an efficient URLLC strategy can be employed, otherwise, resource requirements on the control nodes can be prohibitive.

Similarly, remote surgery [3] is part of a class of applications that require haptic communications [10], which consist of tactile feedback for an operator. The operator is typically dictating the movements and actions of a certain system. Interestingly, it can be viewed as a derivation of the factory automation use-case, where the controller is replaced by a person who is providing the inputs to the system. When speaking about remote surgery, a surgeon operates using robotic arms from a distance. In the context of wireless communications, this could be the case, for instance, in a search and rescue mission where it is not possible to set up a wired infrastructure and also where the surgeon cannot access the patient directly. It is paramount that the medical doctor performing the operation has the most up-to-date sensory information to react accordingly, and that their commands are applied by the robot operating with an ultra-reliable degree of certainty.

Smart-grid automation, by its turn, consists of the automation of handling control signals relevant to the operation of smart grid deployments [4]. For instance, the grid can be sub-divided into some local geographical areas and the surplus or demand of energy can be sent from one area to another, such that the grid can operate much more efficiently [11]. This must happen quickly and certainly, as any delays or missed signals can cause power elements to malfunction. More specifically, in the event of a power outage in a certain area, other areas must be warned about the reconfiguration required, otherwise, the strain put on the energy network might cause a cascade effect and large portions of the grid may suffer an outage and possibly long term costly damages. Preventing these events is what is known as smart-grid teleprotection [12], and it must have URLLC levels of service in order to be possible.

Regarding self-driving cars [5], and traffic automation in general, URLLC is a clear requirement [12]. The first very obvious case of using URLLC communication for traffic efficiency is in platooning, where a long queue of vehicles can drive at the same platooning speed, and tweak their individual acceleration according to a coordinated communication [13]. Another use-case is fully autonomous roads, wherein all the cars in a road or city are autonomous [14]. In this case, the velocity of the vehicles coupled with the need for the most up-to-date information about their surroundings and—more importantly—other vehicles in order to make safe decisions [14] imposes URLLC needs.

This is by no means a comprehensive list of potential applications which will be unlocked by URLLC in the coming years, however, it is enough to showcase the disruptive nature that it will have in our lives and moreover to realize that a large number of new radio powered devices are going to be produced and put in operation in the short future.

Despite all these benefits, URLLC still has strict challenges in order to be utilized in practice [12, 15]. For instance, achieving strict maximum latencies with very high reliability is not trivial, since those requirements are conflicting. Moreover, in the context of the network deployments of the future, communication needs to be resource-efficient. That may be in the context of energy requirements, spectrum usage, or electromagnetic field (EMF) radiation emission. This means that simply trying to enable URLLC requirements by using more resources is not a viable strategy, and thus clever protocols must be devised. Another important challenge is that URLLC applications must co-exist with what is already deployed. Meaning that MAC introduced delays are also challenging.

1.2 Related Works

This thesis has three technical chapters, one which deals with hybrid automatic repeat request (HARQ) in URLLC, one which incorporates communication and control joint design (JD) alongside retransmissions for control applications, and one which considers using self-organizing networks (SON) to enable HARQ in URLLC applications. Next, we present a literature review divided by these three categories.

1.2.1 HARQ for URLLC

In [16], modulation order, transmit power, the number of transmission attempts, and code rate are optimized using a realistic energy consumption model for the case of truncated simple and Chase combining (CC)-HARQ [17]., considering fast and block-fading scenarios in a Nakagami-m channel. They do not consider latency in their modelling.

On the other hand, adaptive HARQ is studied in [18] and compared to traditional 1-bit feedback HARQ. Modelling the system with a Markov decision process, the authors derive optimal policies for truncated and persistent adaptive HARQ. An analysis using cooperation and simple HARQ is considered in [19], accounting for average delay constraints from a coding and modulation point of view. They propose a solution for power allocation and communication strategy that minimizes the overall power consumption of the system and their results show that the solution can reduce the overall energy consumption. However, average delays are not suitable for a wide range of URLLC applications, which have strict maximum latency deadline for any message exchanges. Moreover, the authors in [20–22] allocate power in order to improve energy efficiency considering truncated simple automatic repeat request (S-ARQ) and CC-HARQ in a Rayleigh block-fading channel. Furthermore, URLLC and the impact of finite block-length (FBL) in channel capacity are considered, while they present a formal description of the optimization problem and solve it in closed-form using the Karush-Kuhn-Tucker (KKT) conditions; they also show that power allocation in HARQ is a good strategy to improve the system energy efficiency. However, they do not analyze the effect of retransmissions on latency.

Sun *et.al.* [23] analyzed improving the overall energy efficiency of a 5G URLLC network by considering a resource allocation policy from a queuing perspective. Also considering queue delays, in [24] the authors showed that the policy has a great impact on the achievable latency for providing system-wide URLLC. However, they do not consider the effect of fading.

More recently, in [25], the authors obtain an expression for the effective energy efficiency in the FBL regime, which considers the latency into the channel capacity. They optimize this metric by minimizing the non-empty buffer probability. Lopez *et.al.* [26], by their turn, dig deep into the effect of imperfect channel state information (CSI) in URLLC communications, showing different research directions that can be explored even in the next generation of mobile networks, 6G. To avoid the problem of medium access control (MAC) added delays (in particular for the retransmission attempts) onto the communication in URLLC, Liu *et.al.* [27] propose in their work a grant free URLLC HARQ scheme. Albeit an interesting approach, its practical employability is somewhat limited due to regulatory constraints.

1.2.2 Communication and Control JD

There are quite a few strategies that tackle the problem of factory automation without considering JD into the system model. In [8], a study of an indoor factory automation with wirelessly enabled active actuators is performed and distributed multiple-input multiple-output (MIMO) techniques are employed to improve the guarantee of the necessary reliability. On the other hand, the authors in [28] present a framework that shows the probability of latency and reliability constraints being obeyed in URLLC systems and use this framework to test frequency and time diversity techniques. They compare their solution to the case of a single transmission (no time diversity) and show that their approach greatly outperforms it.

Zaki-Hindi *et.al.* [29] present a multi-tenancy approach that considers using licensed and unlicensed spectrum with URLLC constraints taken into account, meaning that attempts of using unlicensed spectrum have to be used sensibly.

Cooperative communication in the context of JD has been studied in [30], where the authors propose that the relay node could have a faster data rate when compared with the direct transmission, thus, incurring a higher probability of failure. They tweak the energy consumption of the relay node in conjunction with the length of the prediction horizon, to boost the reliability of the system. However, they do not take URLLC constraints into account in their model.

On the other hand, since the URLLC requirements towards WNCS are difficult to attain, JD techniques, where communication and control are taken into account simultaneously to enhance the system performance [31], are an interesting solution to provide efficient URLLC suitable for the applications. In the literature, some authors investigate the use of communication and control JD [32–34]. In [32], the authors consider optimizing the energy efficiency of the sensor-controller link via finding optimal sampling frequency, latency, and reliability parameters given a control cost. Lu *et.al.* [33] propose a JD technique selecting sampling rates of multiple control plants and solving an optimization problem to minimize the total control cost whilst respecting maximum delay constraints of all the plants. They do not consider the use of packetized predictive control (PPC) [35] in their analysis. The authors in [34], in turn, presented an analytical method to determine the best length of the prediction horizon in PPC to reduce the wireless resource consumption taking into account the effect of finite block-length. They do not, however, consider the effect of the wireless channel.

The authors in [36] base their work on [34]. They consider optimizing the energy used by the transmitter by using a well-crafted incremental redundancy (INR) HARQ scheme, whereby the transmitter sends new parity bits at every failed attempt. However, they do not consider a JD approach in their scheme.

1.2.3 SON Enabled HARQ

Digital twins have been a popular strategy in several cyber-physical systems in the recent decades [37–43]. They consist of having a digital simulation of a physical system, such that one can test algorithms in software. Despite all of this attention, particularly in the face of 6G, all of the works mentioned above consider a digital twin of a specific digital twin application incorporated into a traditional system-level simulation for the communication network.

Other works [44–46], conversely, are able to implement digital twins considering the telecommunication system as the simulated object. In [44], a digital twin is proposed for recreating the real-world environment based on data provided by users in real-time for the purpose of decision making. On the other hand, [45] developed a digital twin architecture reference model for cyber-physical systems, to identify several key aspects of the system during design time. However, [44, 45] do not consider a whole cellular network as part of the digital twin system.

In [46], a digital twin of the edge network is proposed, to operate alongside the actual network deployment. Similarly in [47], another digital twin is designed with edge processing in mind. Their solution consists of a dynamic digital twin to capture the time-varying characteristics of the network in order to provision edge resources in a drone-powered network. Both approaches can be used alongside the network deployment but do not serve to simulate the whole system in order to derive policies offline.

A comprehensive survey of digital twin networks has been presented in [48], and several works have been listed. However, none of them considers the creation of a digital twin which considers the entire cellular network as the cyber-physical system in order to generate simulation data that is more closely resembling of reality. This is an important gap since the lack of data availability is a stringent challenge in the design of machine learning (ML) systems towards telecommunications. Moreover, traditional system-level simulations rely on several assumptions which can taint the results.

Next, related work in the area of SON is summarized. The authors in [49] propose using a long short-term memory (LSTM) convolutional neural network (CNN) to predict the number of user elements (UEs) in a network and thus allocate resources preemptively. These types of neural networks have also been used by [50] to predict the number of users for a base station, and then using that information to determine MIMO system parameters to optimize the energy efficiency of the network. Moreover, the work in [51] also uses LSTM networks, but they predict measurement reports from users in a cellular network from real-world data. On the other hand, [52] predicts user mobility in a vehicular communication setup. Noteworthy, the OCEAN channel state information predictor is presented in [53]. It uses several features obtained from the network and feeds that into a deep learning algorithm in order to determine the channel state information. This work differs from all of the above as it proposes to use LSTM networks to predict the average signal to interference plus noise ratio (SINR) of users in a network and to use those predicted values as decision-making inputs for HARQ retransmissions. In other words, we go one step further and present an algorithm that leverages the predictions obtained directly to enable HARQ in URLLC.

Using ML to predict failure in HARQ rounds is presented in [54], wherein the authors propose a supervised learning approach using the outcome of the decoder to train a model and hopefully predict when a message will fail without the need to attempt the decoding, similar to what we do in this thesis. The main difference is that we are doing our prediction indirectly, based on a predicted channel quality, and thus, we are able to do it at the receiver, avoiding the need for the feedback altogether.

As mentioned earlier Lopez *et.al.* [26] considers the effect of the imperfect CSI at the receiver into URLLC applications. By contrast, we get around this issue by predicting the channel quality using ML and reacting to poor channel conditions by preemptively sending new message attempts.

In this thesis, we circumvent the problem of the long processing times at the receiver to decode the message by predicting the channel quality and automatically triggering retransmission attempts. The authors in [55], on the other hand, explore other combination techniques at the receiver, to achieve the same goal. Moreover, the strategy proposed here can also be used to avoid added latency due to securing a transmission slot as a result of MAC policies. On the other hand, in [56], another approach regarding MAC using non-orthogonal multiple access (NOMA) is considered, which shows promising results, but for the uplink. Here, we tackle the issue for the downlink communication. Another example of using NOMA and HARQ to enable URLLC can be found in [57], where the authors study it in the context of the Internet of Things. They use a fixed number of allowed attempts as a strategy to save latency, however, they could combine what they did with the approach proposed here and still have the latency guarantees and at the same time use an optimal number of transmission attempts allowed.

Other attempts to predict poor channel conditions, not based on ML, also exist in the literature. For instance, in [58] explore adaptation to 5G new radio which enhances the channel quality reporting, particularly in HARQ rounds, in order to better assess when a deep fade state will occur, and thus enable URLLC with a more efficient HARQ.

1.3 Motivation

This thesis shows a study of resource-efficient HARQ for URLLC and how it can be applied in practical scenarios. We focus on enabling efficient URLLC applications, such that all the aforementioned applications, and many more, can be implemented sustainably. However, URLLC is not easy to achieve, given the contradictory nature of increasing reliability whilst at the same time reducing latency [59–61]. At the same time, HARQ has been a long-term staple of efficient wireless communication design, being able to increase diversity in the communication whilst saving resources [59]. The trade-off in question is that typically HARQ incurs a higher latency in order to provide its resource efficiency, while in URLLC higher latencies cannot be tolerated. This makes the study of HARQ in the context of URLLC non-trivial, and thus an interesting topic to study. Therefore, in this work, we explore ”The Path Towards URLLC via HARQ”, as the title says.

The research objectives of this work can be outlined as:

I Determine whether or not using HARQ in URLLC can yield better performance in terms of energy efficiency when compared to other diversity strategies.

II Similarly, determine if HARQ in URLLC can be used to reduce the amount of EMF radiation, with a similar comparison.

III Investigate how communication and control JD techniques used alongside retransmission strategies in order to improve the spectral efficiency of WNCS URLLC applications.

IV Tackle the challenge of scheduling added delays in order to enable HARQ to URLLC applications.

1.4 Contributions

First, we use a comprehensive energy consumption scheme, first proposed by [16], in order to formulate an optimization problem for maximizing the energy efficiency of a generic URLLC application. We extend the model by including the effect of the imperfect channel estimation and URLLC constraints. The latter is done by finding a communication rate that guarantees the target latency and reliability, from an information theory perspective[1].

Then, we show that the energy efficiency can be optimized by employing an optimal policy in terms of data rate and the maximum number of allowed attempts. We also derive a floating-point equation that can be used to determine this policy and which yields insights into the system behaviour. Also in this context, we use the same model to find the optimal policy to minimize the maximum EMF radiation in URLLC applications. Both these approaches are compared with another strategy that considers using frequency diversity strategies in order to achieve the QoS required by the target applications, and we show how our proposed approach outperforms this strategy by using an application as an example. These contributions have been published in [59, 62].

In other words, we investigate the energy efficiency of URLLC enabled CC-HARQ and a novel optimization strategy is proposed via optimizing the maximum number of allowed transmission attempts, for a block-fading Nakagami-m channel[2], whilst guaranteeing a maximum latency. We analyze the trade-off between rate and diversity, showing that significant energy savings can be obtained. Similarly, in [63] the number of allowed attempts is also optimized, but the focus is to reduce the required bandwidth for the URLLC application.

Unlike [16, 18, 20, 21], we guarantee a maximum latency with a target reliability. The work in [19], by its turn, considers energy consumption using HARQ but only accounts

[1] Considering achievable rates as opposed to considering coding and modulation techniques.

[2] The Nakagami-m distribution is interesting to model wireless communication systems, as its parameter can be tweaked to reflect various LOS conditions.

for average delays. Average delays have limited applicability in URLLC applications, as these typically consider a strict deadline on all messages. A power allocation scheme is not considered in this work because it requires longer feedback messages, which can be a problem in URLLC. Moreover, since we consider peak power constraints, the applicability of power allocation strategies is limited.

Next, we investigate how can we improve resource utilization even further if we consider application-specific characteristics for wireless networked control systems (WNCS). More specifically, we explore how a wireless system employed in conjunction with a PPC strategy can be improved by tweaking both control and communication parameters concomitantly, in what is known as JD [64]. We first extend the proposed strategy in [34] to consider the random effect of the wireless channel and optimize the system bandwidth by determining the optimal policy in terms of allowed transmission attempts—a wireless communication parameter—and length of control horizon—a control parameter. In this strategy, the plant has a buffer that keeps a sequence of the next control signals to use, and if a communication outage occurs, the received message is discarded and the buffer is used. Part of this contribution has been published by us in [9].

Next, we propose our own strategy, which relies on keeping the copies of the messages when an outage occurs and performing maximum ratio combining (MRC) in order to increase the chances of a successful decoding attempt. Note that this is essentially (CC)-HARQ. Both these approaches are then compared with two non-JD strategies, one that employs our earlier proposed policy, which is designed purely from a communication perspective, and another which considers only tweaking the length of the control horizon, but keeping the number of retransmission attempts constant. Through this comparison, we can show that the JD techniques greatly outperform the other approaches for this particular problem, in terms of required bandwidth to meet the URLLC constraints.

The approaches to CC-HARQ proposed earlier rely on a few assumptions which greatly influence latency, such as early message error detection and readily available MAC schedule for retransmission attempts. These assumptions are not trivially realized in real deployments. This may result in challenges when attempting to employ the proposed approaches in practical scenarios. Thus, in order to relax these assumptions, we propose using ML to enable CC-HARQ in a cellular network, in an implementation of SON. SON can be defined as mobile networks proactively learning how they should behave without their actions being explicitly programmed [65]. With this contribution, one important research gap that we aim to close is the development of a digital twin for a full-blown cellular network. This can be used in order to generate a large dataset for the creation of SON applications.

Our proposed implementation consists of using the temporal series of channel quality information for users in a network and employing an LSTM CNN which can predict when the channel is going to be in a poor state for a given link. Then, the transmitter does not have to wait for a non-acknowledgement (NACK) in order to send the next copy of the message, thus avoiding the problem of early error detection. On the other hand, the transmitter can also preemptively allocate MAC resources when it predicts that a certain link is going to be stringent, such that this assumption is also relaxed. We have first published some results on the potential gains of having a strategy that can enable rapid retransmissions in [66].

This SON implementation is made possible by the creation of a realistic digital twin of the city of Glasgow, which is used in a network simulation to generate a large amount of realistic user data. That is then used to train the LSTM model, which can then predict poor channel quality. To the best of the author's knowledge, this is the first work of this kind that presents a realistic model of the cellular deployment.

To re-iterate, the contribution of this thesis can be summarized as showing how can URLLC be made more efficient and used in practice by employing HARQ. This is achieved by:

1. Proposing and solving an optimization problem to minimize the energy efficiency of URLLC applications using CC-HARQ, which finds the optimal policy with respect to data rate and the maximum number of allowed transmission attempts.

2. Proposing and solving a similar problem, but that instead looks at minimizing the maximum EMF radiation output of URLLC applications, finding the optimal policy in the same regard as before.

3. Proposing a JD technique that uses MRC in PPC for WNCS.

4. Extending an existing JD technique to consider the effect of the wireless channel and comparing this approach with the MRC solution, as well as with non-JD techniques.

5. Create a digital twin of the city of Glasgow and perform a realistic simulation, making the resulting dataset available for researchers.

6. Using the obtained dataset, propose and implement an ML algorithm to predict the channel quality for the users based on LSTM CNNs.

7. Leveraging the predicted channel quality information, which enables efficient early feedback techniques. This in turn can be used to relax some of the assumptions made by the CC-HARQ optimization protocols proposed.

In summary, this thesis presents: a latency aware HARQ technique which for typical application parameters uses up to 18 times less energy when compared to using the same spectrum but with frequency diversity instead of time diversity; How using the same latency aware technique it is possible to reduce the (EMF) radiation, showing that it is possible to produce more than 10 times less EMF radiation when compared to the same baseline; Communication and control joint design techniques using wireless networked control systems specific characteristics in conjunction with clever retransmission techniques, yielding up to 3 times less bandwidth required for application typical parameters, when compared with non-joint design techniques; A digital twin for a cellular network, which mimics the city of Glasgow with real user trajectories, base station positions and the full LTE stack. This digital twin is used to generate a dataset that is made publicly available; Lastly, an implementation of long short-term memory convolution neural networks using the aforementioned dataset which is able to predict channel conditions with very high accuracy, achieving root mean square error on the order of 10^{-2}. The predicted channel quality is an enabler for early failure detection which enables the other solutions proposed in this work.

1.5 Organization

The remainder of this thesis is organized as follows: Chapter 2 outlines some important base concepts from the literature, which served as a basis for the work carried out; Chapter 3 shows the proposal for the latency aware HARQ and the optimization of the energy efficiency and maximum EMF radiation minimization, while the related work regarding these contributions is outlined in Subsection 1.2.1. The joint design study is outlined in Chapter 4 with its respective literature survey presented in Subsection 1.2.2. Meanwhile, The digital twin, the SON, and the early feedback work is shown in Chapter 5, with related work presented in Subsection 1.2.3. Lastly, Chapter 6 has the conclusion of this thesis and presents future research directions.

Chapter 2

Technical Background

In this chapter, we present relevant telecommunications engineering background information which served as a foundation to this thesis. We present well-established concepts from wireless communications theory, which enabled the work performed in this thesis. Namely, we cover the communication system model and the mathematical equations which underpin it.

The work of this thesis consists of, from an information theory perspective, providing resource optimization for URLLC by the use of retransmissions, namely HARQ. Thus, in this chapter, we cover the large and small scale losses and provide an overview of the possible diversity strategies which can be leveraged in order to enable URLLC, with a focus on HARQ. We also present what is the effect of imperfect channel estimation in the system. Moreover, we cover the achievable rates at FBL, and also the system performance for HARQ at a high signal-to-noise ratio (SNR).

Studying communication systems from the lens of information theory provides engineers with important guidelines in terms of performance as to what can be achieved by a system. Finding strategies for optimal resource utilization from this perspective is thus an important and current subject in wireless communications, as evidenced by the literature review presented in the previous chapter. The foundation laid on this chapter is then subsequently used by the following chapters, which include the various contributions of this thesis.

2.1 Large Scale Loss — Path Loss

Losses in radio wave propagation are typically divided into large-scale loss, or path loss (PL), and small-scale loss, or *fading.* The former is the characterization of the average signal strength that reaches a receiver and it increases with the distances between transmitter and receiver. Not only is the signal attenuated by the propagation, but also reflections, scattering, and diffraction play an important role in reducing the average received signal strength [67].

There are various models for characterizing the PL, which is always defined as the ratio between the received power P_{r} and the transmitted power P_{t}. Free space PL is the most basic model and it considers that there is a direct line of sight (LOS) between transmitter and receiver and that the signal propagates via a straight line [68]. The PL in the free space model is expressed as [68]

$$\frac{P_{\mathrm{r}}}{P_{\mathrm{t}}} = \left(\frac{\sqrt{G_{\mathrm{t}}}c}{4\pi d f_{\mathrm{c}}}\right)^2, \tag{2.1}$$

where d is the link distance, G_{t} is the total gain of the antennas, c is the speed of light and f_{c} is the carrier frequency.

When only one ground reflection and the direct LOS dominate the behaviour of the system, the two-ray model can be used to represent the PL [68], which is based on ray tracing the significant paths of the signal from transmitter to receiver. Building upon this idea, Amitay [69] proposed a model with ray tracing which considers 10 rays and is suitable for microcells in urban environments and relies on computational simulations. Several other general ray-tracing models have also been proposed based on computational simulations [68]. However, they are complex, rely heavily on having a good model of the environment, and are scenario specific.

Empirical models are very popular because they do not rely on having extensive knowledge of the environment and can often be generalized to similar contexts. These types of models are the most common in urban telecommunications and date as early as 1968 when Okumura did an extensive collection of data in Tokyo and derived an accurate PL model [70] for radio communication in urban scenarios. This model was then extended by Hata [71] in the '80s and, more recently, by the European cooperative for scientific and technical research (EURO-COST) [72] to be more complete.

The PL in the EURO-COST, or COST231, [73] model, is expressed as

$$\begin{aligned} L_{\mathrm{b}} = 46.3 + 33.9\log(10^{-6}f_{\mathrm{c}}) - 13.82\log(h_{\mathrm{base}}) \\ - a(h_{\mathrm{mobile}}) + (44.9 - 6.55\log(h_{\mathrm{base}}))\log(d/10^3) + C_{\mathrm{m}}, \end{aligned} \tag{2.2}$$

where L_{b} is the path loss (in dB), h_{base} and h_{mobile} are the base station and mobile device antennae heights in meters. C_{m} assumes 0 dB for medium sized cities and suburban environments and 3 dB for dense urban environments. Lastly, $a(h_{\mathrm{mobile}})$ is determined via

$$a(h_{\mathrm{mobile}}) = 1.1\log(10^{-6}f_{\mathrm{c}}) - 0.7)h_{\mathrm{mobile}} - (1.56\log(10^{-6}f_{\mathrm{c}}) - 0.8). \tag{2.3}$$

These empirical models are very precise for the scenarios for which they are designed. However, they can be complicated and rather intractable when combined in complex systems [68]. For that reason, simpler models, which are also able to predict average received powers with reasonable accuracy but which offer more friendly mathematical representation have been proposed. In this case, most commonly the PL is represented as [68]

$$\frac{P_\mathrm{t}}{P_\mathrm{r}} = A_0 \left(\frac{d}{d_0}\right)^\alpha, \tag{2.4}$$

where A_0 is the loss at a reference distance for the antenna far field d_0 and α is the path loss exponent. A_0, d_0 and α can be obtained either from measurements or analytically [68].

2.2 Small Scale Loss – Multi-Path Fading

In wireless communications, unlike its wired counterpart, the received signal is typically attenuated by random elements in the medium. This effect is called *fading* and it occurs mainly due to reflection, scattering, and diffraction of the electromagnetic wave used for communication. It arises from the fact that multiple copies of the transmitted signal with different phases and amplitudes reach the receiver at any particular time. This is an extremely complex interaction and therefore it is typically characterized as a random process [67].

When there is no LOS, the channel envelope can be viewed as normally distributed because of the completely random nature of the reflections. When this is the case, the phase and quadrature components of the signal have a normally distributed gain, and the resulting power will be a random variable that follows a Rayleigh distribution [67]. The probability density function (PDF) of the channel power for a Rayleigh distribution channel has a is expressed as [68]

$$p_\mathrm{ray}(x) = \frac{1}{P_\mathrm{r}} \mathrm{e}^{-\frac{x}{P_\mathrm{r}}}. \tag{2.5}$$

However, because of the non-LOS assumption, this is not general and it does not fit values obtained empirically [68]. Thus, a more general model was envisioned, considering that the phase and quadrature components experience a Gamma distributed channel power. This is known as a Nakagami-m channel and has its channel power with PDF [68]

$$P_\mathrm{naka}(x) = \frac{m^m x^{m-1}}{\Gamma(m) P_\mathrm{r}^m} \mathrm{e}^{\left(-\frac{mx}{P_\mathrm{r}}\right)}, \tag{2.6}$$

where $\Gamma(.)$ is the complete Gamma function[1]. This is a more general approach as now there is a parameter that can be used to represent different environments, m. When m is 1, the channel is Rayleigh and when m is larger than 1, practical measurements show that

[1] $\Gamma(y) = \int\limits_0^{+\infty} x^{y-1} \mathrm{e}^{-x} dx$ [74].

it represents a stronger LOS component between transmitter and receiver, whereas when m is less than 1 it represents a more severe fading condition than Rayleigh.

2.2.1 Wireless Communication Model

Regarding the wireless link, messages $s(t)$ are sent to the receiver such that $r(t)$, the received message, is expressed by

$$r(t) = h(t)s(t) + w(t), \tag{2.7}$$

where t indicates time, $h(t)$ is the channel gain and $w(t)$ is the additive white Gaussian noise (AWGN). Note that here we consider a normalized power.

Considering an infinite block-length, the error probability is determined according to the channel capacity $C(\gamma)$, which indicates the attainable rate R (in bits per channel use) for a given instantaneous SNR γ, as

$$C(\gamma) = \log_2(1 + \gamma). \tag{2.8}$$

If we a variable substitution here, replacing γ with γ_0, such that γ_0 represents a threshold of SNR at which error-free communication is possible, now C will represent the rate at which we can have error-free communication. Thus, (2.8) becomes

$$\gamma_0 = 2^R - 1, \tag{2.9}$$

after solving for γ_0.

The probability of failing can then be determined by calculating the probability of the instantaneous SNR γ is smaller than γ_0. By its turn, γ is obtained by using (2.6) into (2.7). Finally, the probability of failure is expressed as [75]

$$P_{\text{out}} = \frac{\Gamma_{\text{inc}}\left(m, m\frac{\gamma_0}{\bar{\gamma}}\right)}{\Gamma(m)}, \tag{2.10}$$

where $\Gamma_{\text{inc}}(.)$ is the lower incomplete Gamma function[2] and $\bar{\gamma}$ is the average SNR.

2.3 Error Probability at Finite Block Length

The analysis performed in the previous section considers that an infinite block can be used for coding. However, advances in information theory in the last decade showed that this

[2] $\Gamma_{\text{inc}}(y, b) = \int\limits_0^b x^{y-1}\mathrm{e}^{-x}dx$ [74].

assumption can be dropped and the length of the block can be taken into account on the formulations, in what is known as the error probability at FBL. The achievable rate under the FBL regime has been studied by Polyanskiy [76] *et.al.* and a closed-form approximation has been derived on the achievable rate for the AWGN channel. Considering normalized noise power and n channel uses, the achievable rate is tightly approximated by

$$R \approx C(P) - \sqrt{\frac{V(P)}{n}} Q^{-1}(\varepsilon), \tag{2.11}$$

where ε is the target error probability, P is the power per channel use, $C(P) = \log_2(1+P)$ is the Shannon capacity, $V(P)$ is the channel dispersion and Q^{-1} is the inverse Q function.

As discussed in [77], (2.11) can be utilized to express the achievable rates in the block fading regime by setting the instantaneous SNR γ to be equal to $P|h|^2$, such that the channel dispersion is given by

$$V(\gamma) = \frac{\gamma(\gamma+2)}{2(\gamma+1)^2}(\log_2 \mathrm{e})^2, \tag{2.12}$$

Isolating ε in (2.11), yields

$$\varepsilon = Q\left(\sqrt{\frac{n}{V(\gamma)}}(C(\gamma) - R)\right), \tag{2.13}$$

which is the error probability for one channel realization. The average error probability $\bar{\varepsilon}$ can be obtained as

$$\bar{\varepsilon} = \int_{\mathbb{R}+} Q\left(\sqrt{\frac{n}{V(\gamma)}}(C(\gamma) - R)\right) p_\gamma \mathrm{d}\gamma. \tag{2.14}$$

Note that, due to the random nature of the $h(t)$, γ will be a random variable. Here, p_γ indicates the PDF of γ.

Considering a Nakagami-m channel, (2.14) does not have a closed-form solution. However, it has been shown in [77] that it can be tightly approximated in block fading regime when considering high SNR—as in the case of URLLC—by the outage probability P_{out}.

2.4 Diversity

In wireless communications, diversity (or antenna diversity) consists of sending copies of the message as a way of enhancing the chances of correctly decoding it at the receiver. There are several strategies to achieve antenna diversity and mitigate the effect of small-scale fading. If the channels used to send the copies are uncorrelated, one order of diversity is gained per copy of the message. As presented in [68], examples of diversity strategies are:

- **Spatial Diversity** can be implemented using multiple antennae on the receiver, the transmitter, or on both. Proper spacing between the antennae ensures that the multiple channels are uncorrelated. Depending on the carrier frequency wavelength, space limitations present an important challenge for this type of strategy as the spacing between the antennae grows proportionally with the carrier frequency wavelength. Another way of achieving physical diversity is to use cooperation. Exploiting the broadcast nature of wireless communications, a relay node receives the message and forwards it to the destination. Although this avoids the issue of separating the antennae, as the relay and the source nodes are not in the same location, the channel between the two becomes another source of error.

- **Frequency Diversity** is achieved by sending the multiple copies of the message using uncorrelated frequency channels. It can also be implemented by sending the message over a wide spectrum and employing a forward error correction technique, thus exploiting non-flat regions of the frequency spectrum. As spectrum is an expensive commodity in telecommunications, frequency diversity can be expensive when used in a large scale.

- **Polarization diversity** is another strategy that uses multiple antennae, however, instead of relying on the distance between them to produce uncorrelated channels, uncorrelation is obtained by using different polarization for each antenna. This technique is less scalable than the others since channels become more correlated as the phase difference in the polarization of the antennae is reduced.

- **Time diversity** consists of sending multiple copies of the message at different time instances, exploiting the time variability of wireless channels. This is especially interesting in block-fading environments, where the channel varies independently from one block to the other and thus the copies of the message are sent via uncorrelated channels. The big incentive in using time diversity is that interesting savings can be attained when a message is successfully decoded in the early attempts since the resources for the following ones are not required. The trade-off, however, is that when they fail, the transmission takes longer.

- **Combined strategies** are also possible, employing multiple sources of antenna diversity at the same time and thus achieving higher diversity orders, while mitigating the downsides of individual strategies.

As mentioned, the above-mentioned diversity strategies are methods for conveying multiple copies of the messages to the receiver. On top of that, the strategy used for combining the copies of the message at the receiver can also influence the performance of the technique employed. There are several techniques for combining messages at the receiver side, we present below a summary of three popular strategies as they are presented in [68]:

- **Selection Combining (SC)** consists of only considering the best copy received and discarding all the others. In applications where continuous communication is maintained, SC systems may require each antenna to be monitored continuously to evaluate its SNR, and thus it might not benefit from keeping only one radio-frequency (RF) chain turned on [68]. Another downside of using an SC strategy is that the information contained in the other copies of the message are discarded and thus the performance is sub-optimal.

- **Threshold Combining**, also known as switch and stay combining, can be used as a simpler alternative to SC. The branches are monitored independently and sequentially. Once one channel is performing above a predefined threshold, this becomes the active signal at the receiver and it only changes if its SNR falls below the threshold. This means that, regardless of the type of system, only one RF chain is active during the communication. Even though the information of the other channels is not being used in both schemes, the implementation of threshold combining is more efficient than SC considering the aforementioned cases where SC systems have to continuously monitor each antenna.

- **Maximum Ratio Combining (MRC)** is the optimal way of combining the signals from the branches to leverage all the information that has been exchanged. It consists of weighting each branch according to its strength relative to the others. It achieves full diversity and does not waste any of the received information. However, knowing the instantaneous SNR at each branch is required to implement the scheme properly. Moreover, in order to implement MRC one RF chain must be present in each antenna in order to receive the various echoes of the message and properly combine them before trying to decode.

The above-mentioned techniques are not employed in time diversity schemes directly, as they require that all the copies of the message are available to the receiver at the same time. However, there are analogous implementations using HARQ which have mathematically the same performance as the aforementioned strategies, in terms of information theory. Moreover, to avoid wasting resources, in HARQ the receiver sends a feedback message to the transmitter, and a retransmission is only issued if the previous attempt has failed. This is what gives HARQ its strength, using the alternative branches only when they are required, thus optimally using the available resources.

The simplest implementation consists of trying to decode the message at the receiver and upon failure discarding the failed message and sending a NACK via the feedback channel. When the transmitter receives a NACK, it immediately sends a new copy of the message which the receiver tries to use for decoding. This is known as simple HARQ (S-HARQ) and in practice, the number of transmission attempts is halted, such that when a maximum number of rounds z has occurred an outage is declared, which in turn stops retransmissions. The performance of S-HARQ is the same as SC, with the advantage of achieving full diversity without having to spend the resources for all of the branches. The cost of doing so is the cost involved in sending the feedback messages, however, a single bit feedback is typically enough to declare a decoding failure making the feedback messages short and economic.

2.5 Probability of Error for HARQ

For S-HARQ, the outage probability is the same as that of selection combining [68]. Since each attempt is performed in an independent channel, the probability of outage considering a maximum of z transmissions in simple HARQ $P_{\text{out},z}^{\text{S}}$ is found by multiplying the probability of each individual failing, which yields

$$P_{\text{out},z}^{\text{S}} = \left(\frac{\Gamma_{\text{inc}}\left(m, m\frac{\gamma_0}{\bar{\gamma}}\right)}{\Gamma(m)} \right)^{z}. \tag{2.15}$$

Equation (2.15) can be tightly approximated at high SNR by [67]

$$P_{\text{out},z}^{\text{S}} \approx \left(\frac{\left(\frac{m\gamma_0}{\bar{\gamma}}\right)^{m}}{\Gamma(m+1)} \right)^{z}. \tag{2.16}$$

Considering CC-HARQ, the outage probability after z rounds $P_{\text{out},z}^{\text{CC}}$ then will be that of maximum ratio combining [75]

$$P_{\text{out,z}}^{\text{CC}} = \frac{\Gamma_{\text{inc}}\left(zm, m\frac{\gamma_0}{\bar{\gamma}}\right)}{\Gamma(zm)}. \tag{2.17}$$

Lastly, $P_{\text{out,z}}^{\text{CC}}$ can be tightly approximated at high SNR [78] by

$$P_{\text{out,z}}^{\text{CC}} \approx \frac{\left(\frac{m\gamma_0}{\bar{\gamma}}\right)^{mz}}{\Gamma(zm+1)}. \tag{2.18}$$

Another more elaborate type of HARQ is known as INR-HARQ. It consists of sending additional parity bits at every new attempt, as opposed to sending identical copies of the message. For INR-HARQ there are no closed-form solutions for the outage probability after z attempts $P_{\text{out,z}}^{\text{INR}}$, but it can be solved numerically by evaluating

$$P_{\text{out,z}}^{\text{INR}} = \Pr\left(\prod_{i=1}^{z}(1+\gamma_i) < 2^R\right). \tag{2.19}$$

Also note that INR-HARQ achieves a better performance than CC-HARQ in several cases, however encoding the messages in a way that new parity bits can be sent for each new attempt is a challenging task and adds to the latency overhead. For this reason, we do not use INR-HARQ in the contributions of this thesis, as its applicability in practical setups would be extremely limited, if not prohibited by latency constraints.

2.6 Number of Transmission Attempts

In practice, the maximum allowed number of transmission attempts z is limited by the application, in what is known as truncated HARQ. If that was not the case, it would be theoretically possible to have error-free communication, albeit with no bound in latency, as the transmitter could continue to attempt delivering messages until the message was decoded.

When considering truncated HARQ, if the receiver succeeds in decoding a message it sends back an acknowledgement (ACK), otherwise, it stores the signal and sends back a NACK. For simple and CC-HARQ, upon receiving a NACK, the transmitter resends the message. For simple HARQ, the receiver discards the previous message and tries to decode the new message whereas in CC-HARQ, the receiver combines all messages using MRC before trying to decode. In the case of INR HARQ, the receiver sends more parity bits instead of another copy of the message, such that the receiver may now succeed in decoding the message due to having a stronger code. This process is repeated until success or z attempts have been carried out. If after z rounds the message has not been correctly decoded, an error is declared. The probability of errors occurring determines the reliability of the communication.

Based on the probability of error of each attempt, it is possible to calculate the average number of transmissions $\bar{\tau}$ as

$$\bar{\tau} = 1 + \sum_{i=1}^{z-1} P_{\text{out},i}, \tag{2.20}$$

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

Note that time diversity is not possible unless the channel varies from one attempt to the other. If the channel has slow variability, slow frequency hopping can be employed between consecutive attempts to ensure that each round experiences a different channel realization, gaining diversity. In practice, it can be achieved by using a different channel of the communication standard for every attempt. This imposes that h must be estimated at the beginning of each attempt. Also, note that imperfect channel estimation causes a degraded performance. This will be discussed next.

2.7 Channel Estimation

In order to perform coherent detection, the receiver must estimate the channel. Furthermore, as discussed in the previous subsection, in our set-up this estimation has to occur before each transmission round. It can be done via in-band pilot training, where the first ρ symbols of each attempt are used to estimate the CSI.

Thus, the received signal for $1 \leq k \leq \rho$ is expressed as

$$r(k) = hp(k) + w(k) \tag{2.21}$$

where $p(k)$ is the sequence of pilot symbols. Using any established channel estimation technique, for example minimum mean-square error [79], the receiver obtains an estimate $\tilde{h}$ of the channel, which differs from the actual channel realization. Therefore, during the remaining $n_{\mathrm{d}} = n - \rho$ channel uses (which are used for data transmission), when $\rho < k \leq n$, we have

$$r(k) = \tilde{h}s(k) + (h - \tilde{h})s(k) + w(k). \tag{2.22}$$

Although it is dependent on the signal and not Gaussian, this effect of imperfect channel estimation ($(h - \tilde{h})s(k)$) can be well modelled as Gaussian and can be combined with $w(k)$ into an effective noise perceived by the system, as discussed in the literature [80,81]. This effective noise provides a worst case scenario [81] and is modeled as a lower effective average SNR for the purpose of system performance analysis, such that [80]

$$\bar{\gamma}_{\mathrm{eff}} = \frac{\rho \bar{\gamma}_{\mathrm{d}} \bar{\gamma}_{\mathrm{p}}}{1 + \bar{\gamma}_{\mathrm{d}} + \rho \bar{\gamma}_{\mathrm{p}}}, \tag{2.23}$$

where $\bar{\gamma}_{\mathrm{d}}$ is the average SNR used to transmit the message and $\bar{\gamma}_{\mathrm{p}}$ is the average SNR used for the pilots. To account for the channel estimation error, we use the effective average SNR $\bar{\gamma}_{\mathrm{eff}}$ obtained via (2.23) for the purpose of evaluating the outage probability, which as discussed, is a good approximation for the error probability considering the block fading channel at high SNR.

Next, as in Gursoy [81], for determining channel capacity there are two possible scenarios, without and with a peak power constraint. For the first scenario, it is shown in [81] that the best strategy is to use only one pilot and there is an analytical solution that yields the best power allocation strategy.

Denoting v as the fraction of power used for data transmission, such that

$$\bar{\gamma}_{\mathrm{d}} n_{\mathrm{d}} = v\bar{\gamma}n \text{ and } \bar{\gamma}_{\mathrm{p}}\rho = (1-v)\bar{\gamma}n, \tag{2.24}$$

we write the optimal value $v^{\star}$ as [80]

$$v^{\star} = \zeta - \sqrt{\zeta(\zeta-1)}, \tag{2.25}$$

where ζ is expressed by

$$\zeta = \frac{(n-1)(1+\bar{\gamma}n)}{\bar{\gamma}n(n-2)}. \tag{2.26}$$

However, if we consider a peak power constraint, as in most practical applications, allocating the power optimally to 1 pilot might violate it. In this case, more than one pilot must be used to obtain optimal performance. As in [81], the optimal value of ρ is then obtained numerically considering that each pilot uses the maximum allowed power, according to the peak limitation. The case with peak power limitations is more relevant in practical scenarios, as this is often the case.

2.7.1 Effect of Imperfect Channel Estimation

In Fig. 2.1 we have plotted the SNR loss due to imperfect channel estimation considering peak power constraints versus the number of pilots used and $\bar{\gamma}_{\mathrm{d}}$. In this example, the peak power limitations alongside the path loss yields $\bar{\gamma}_{\mathrm{d}} \leq 40$dB. As discussed before, the pilot transmit power is the maximum allowed such that $\bar{\gamma}_{\mathrm{p}} = 40$dB. We can observe that the effect of imperfect channel estimation is more pronounced when trying to obtain an estimate for the channel with $\bar{\gamma}_{\mathrm{d}}$ close to the pilot SNR $\bar{\gamma}_{\mathrm{p}}$ and fewer pilots are used. This relates to the fact that the power used in estimation is too small in comparison to the signal power. Therefore, we conclude that when choosing the number of pilots, it is important to consider how far from the peak power limitation is the data going to be transmitted. In general, it is beneficial to use more pilots when the average SNR of the transmission is close to the maximum allowed by the system. On the other hand, the SNR loss is relatively small and a well-designed link margin can be enough to account for it.

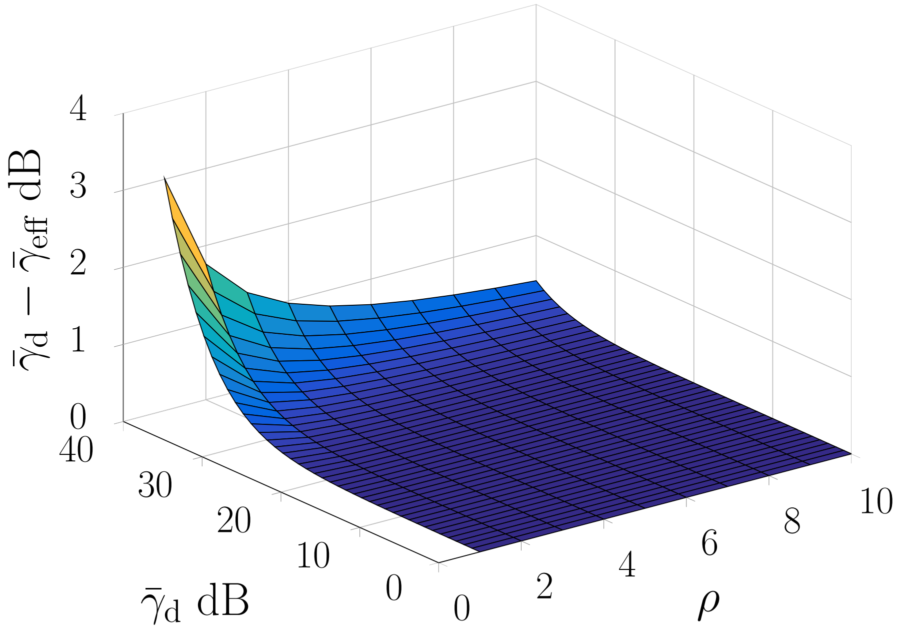


Figure 2.1: Effect of imperfect channel estimation on the average SNR for different values of ρ and $\bar{\gamma}_{\mathrm{d}}$. A peak power constraint is considered such that the maximum $\bar{\gamma}_{\mathrm{d}}$ is 40dB, which is the same value for $\bar{\gamma}_{\mathrm{p}}$.

Chapter 3

Latency Aware HARQ

3.1 Introduction

The question we want to answer in this chapter is the following: Can we leverage HARQ in a URLLC application in such a way that we can have a more efficient communication?

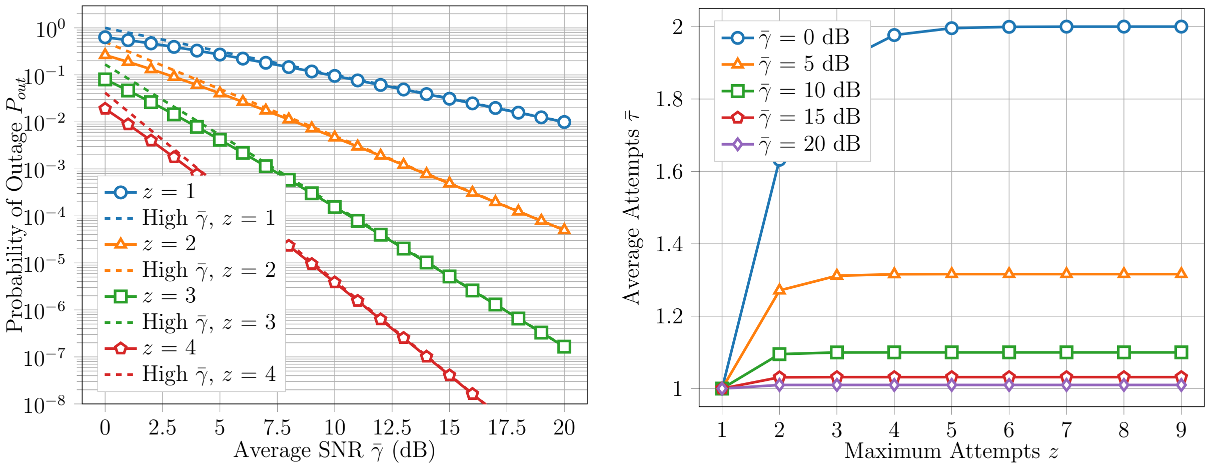
The short answer to this question is surprisingly yes. The long answer is detailed throughout the next pages of this work, where we offer mathematical proof to this counter-intuitive notion, alongside numerical simulations which illustrate the concept.

We have chosen to analyse the problem in two different aspects. The first is regarding energy efficiency. It is a very relevant metric in communications, which has a particularly important relevance when considering the degree of greenhouse emissions of cellular deployments as well as the energy constraints of battery-operated devices. The second metric has to do with human health, and it is regarding minimizing the EMF radiation output of the wireless communication.

There are several strategies to improve efficiency in wireless communications, many of which revolve around mitigating the effect of fast-fading through diversity. In systems without latency constraints, HARQ is well known to improve the energy efficiency [16, 18] by providing time diversity.

The whole controversy around the idea of using HARQ in a URLLC system boils down to the fact that the trade-off in HARQ schemes is providing antenna diversity at the cost of longer transmit times. On the one hand, we gain reliability by adding more copies of the channel, at the cost of imposing longer latencies into the system. In this work, however, we are able to assimilate this excess latency by imposing an increased data rate. This will, by its turn, also lower the reliability of the system, but critically not to the same amount as the increased reliability gained by the extra diversity, and this is the trade-off that we explore.

On top of that, the presence of the feedback channel ensures that we, on average, are not going to be using all of the resources to achieve full diversity. This is the real strength



(a) Outage probability.

(b) Average number of transmissions.

Figure 3.1: Performance of using CC-HARQ, showing the diversity gains and the average number of transmissions. In this example $R = 1$ and $m = 1$ are used.

of HARQ, and by the end of this chapter, the reader should have the same confidence that HARQ can be used for efficient communication design in the context of URLLC applications.

To illustrate the benefits of using HARQ, we have plotted in Fig. 3.1 the probability of outage in CC-HARQ for different levels of average SNR and the average number of transmissions for different z. As we can see in Fig. 3.1a, the diversity gains are substantial and result in greatly reducing the requirements in transmit power. At the same time, the average number of attempts is kept very close to one, even for large z, as we can see in Fig. 3.1b. This is because the probability of failing at the later stages is much smaller than that of failing at the early stages [1] and thus revealing the real strength of using HARQ: high diversity at very low cost.

But then another question arises, what is the best number of attempts to allow, in order to optimize the system? If the number of attempts is too high, the impose data rates will be so large that achieving the target reliability will be too costly. On the other hand, if it is too low, there will be a small diversity gain which means that the full potential of HARQ is not being leveraged. In this chapter, we answer this question mathematically by proposing and solving two optimization problems. The former optimizing the energy efficiency of the communication while the latter minimizes the EMF radiation of the system.

The motivation to optimize the average energy usage is two-fold. Firstly, the huge burst of novel applications enabled by URLLC is bound to include several use-cases where battery constraint systems are employed. The second motivation is environmental. The rationale goes as follows: Green-house emissions are one of the biggest concerns of our

[1] This can be intuitively inferred if one considers that to fail at the later stages, the earlier attempts must have already failed, such that the later transmissions actually occur.

time; Moreover, consider that the telecommunications industry uses a significant chunk of the global energy; Coupled with that, future networks are predicted to have a massive number of new deployments, particularly in the context of Ultra-dense networks; All these factors mean that future protocols must be designed with energy efficiency in mind.

In order to provide a sense of practical application, the proposed energy efficiency solution is further evaluated in a smart grid teleprotection scenario, as described in the mobile and wireless communications enablers for the twenty-twenty information society (METIS) test case number 5 [12]. It consists of reliably delivering messages within a tight latency constraint between substations for the purpose of triggering protection mechanisms when faults occur, preventing damage to the grid. The results show that using different channels for each HARQ round achieves relevant energy savings—up to 18 times more efficient—compared to using all channels in parallel to achieve frequency diversity. This is the case even when accounting for higher data rates required to meet latency constraints in CC-HARQ.

3.1.1 Summary

What we are presenting in this chapter can be summarized as follows. Considering a URLLC scenario, this Chapter assesses the trade-off in terms of energy consumption between achieving time diversity through retransmissions and having to communicate at a higher rate due to latency constraints. The analysis herein considers Nakagami-m block-fading channels with CC-HARQ. A fixed-point equation to determine the best number of allowed transmission attempts considering the maximum possible energy spent is derived, which yields insights into the system behaviour. Furthermore, by comparing the energy consumption of the proposed approach against direct transmission with frequency diversity, this chapter shows that even in the face of stringent latency and reliability constraints, time diversity can be an efficient strategy when compared with using the same spectrum on average to a frequency diversity technique. This is evidenced by the results, which are presented in two optics: energy efficiency and EMF radiation. The results show substantial benefits of using retransmissions in both scenarios when selecting the maximum number of transmission attempts according to the proposed approach.

3.2 System Model

Using the communication model presented in Chapter 2 as a starting point to model the system, we include the context of URLLC as proposed by us in [59], wherein a message has to be delivered with very high reliability within a maximum latency λ'.

3.2.1 Energy Consumption Model

In this Chapter, we model the energy consumed in the transmission of one message considering the radio startup energy, the pre-transmission processing energy, the energy involved in powering the passband receiver elements, and the electromagnetic radiation, similar to [16]. However, we consider the exchange of short messages, typical of URLLC. Therefore, the data rates are relatively low, which in turn causes the arithmetic processing unit clock speed to also be low [16]. This, in turn, causes the energy required for encoding and decoding messages to be small, especially when compared to other consumptions. Thus, we disregard the baseband coding/decoding energy from our model simplifying the presentation without harming the analysis, following [82]. On the other hand, since the energy consumed to transmit pilot symbols becomes relevant with respect to the total energy when the information packets are shorter, we explicit their contribution to the energy consumption in the following analysis.

Energy Used by the Transmitter

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

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

$$n_{\mathrm{fw}} = \frac{L_{\mathrm{H}} + L_{\mathrm{D}}}{R}. \tag{3.1}$$

As in [16, 59, 62], the consumption of baseband and RF circuits is assumed to be constant and equal to $P_{\mathrm{el,tx}}$. Also, the power used to energize passband receiver elements $P_{\mathrm{el,rx}}$ is assumed to be invariant. However, the electromagnetic radiation energy depends on the PA's consumption P_{PA}, which is a function of its average drain efficiency η and of the radiated power P_{rf} [16]

$$P_{\mathrm{PA}} = \frac{P_{\mathrm{rf}}}{\eta}. \tag{3.2}$$

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

$$P_{\mathrm{rf}} = N_0 W M_{\mathrm{l}} M_{\mathrm{c}} A_0 d^{\alpha} \bar{\gamma}, \tag{3.3}$$

where N_0 is the noise power spectral density, W is the bandwidth in Hz, the link margin is M_{l}—which includes the noise figure and other unforeseen losses—, and M_{c} is the coding margin (further explained in Section 3.2.2). Combining (3.2) and (3.3), we have

$$P_{\mathrm{PA}} = Ad^{\alpha}\bar{\gamma}/\eta, \tag{3.4}$$

where $A = N_0 W M_{\mathrm{l}} M_{\mathrm{c}} A_0$.

Lastly, to obtain the PA power consumption for the feedback $P_{\mathrm{PA,fb}}$, data transmission $P_{\mathrm{PA,d}}$ and channel estimation $P_{\mathrm{PA,p}}$ phases we use (3.4) with the respective average SNR for each phase, $\bar{\gamma}_{\mathrm{fb}}$ for the feedback, $\bar{\gamma}_{\mathrm{d}}$ for the data transmission and $\bar{\gamma}_{\mathrm{p}}$ for the channel estimation. Here, the transmit power used for estimation and feedback is always the maximum[2], such that $\bar{\gamma}_{\mathrm{fb}} = \bar{\gamma}_{\mathrm{p}}$ and $P_{\mathrm{PA,fb}} = P_{\mathrm{PA,p}}$.

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

$$n_{\mathrm{fb}} = \frac{L_{\mathrm{fb}}}{R}. \tag{3.5}$$

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

$$E_{\mathrm{tx}} = E_{\mathrm{st}} + \frac{\tau}{W}\left[n_{\mathrm{fw}}(P_{\mathrm{el,tx}} + P_{\mathrm{PA,d}}) + E_{\mathrm{p,tx}} + n_{\mathrm{fb}}P_{\mathrm{el,rx}}\right], \tag{3.6}$$

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

Maximum EMF Radiation

By its turn, the maximum EMF energy output—the one that considers all z allowed transmission attempts—of each exchange E_{PA}, can be obtained by multiplying the number of allowed transmission attempts, the broadcast time of each attempt and P_{PA} from 3.4, yielding

$$E_{\mathrm{PA}} = z\frac{L_{\mathrm{T}}}{RW}P_{\mathrm{PA}}. \tag{3.7}$$

Here, the broadcast time is obtained by considering the total number of bits being sent L_{T} at the rate R with bandwidth W.

[2] As shown earlier, using as much energy as possible for channel estimation yields better results in terms of reducing the effect of the imperfect channel estimation. The feedback is also considered to use the maximum possible power, to ensure that the feedback signal has the highest chance of being decoded.

Energy Used by the Receiver

Assuming receiver and transmitter use identical radios, the energy used by the former is similar to the one used by the latter. Following the same steps[3], the energy used by the receiver for τ attempts is

$$E_{\mathrm{rx}} = E_{\mathrm{st}} + \frac{\tau}{W}\left[n_{\mathrm{fb}}(P_{\mathrm{el,tx}} + P_{\mathrm{PA,p}}) + (n_{\mathrm{fw}} + \rho)P_{\mathrm{el,rx}}\right]. \tag{3.8}$$

Average Energy per Successful Bit

The average energy $\bar{E}$ is obtained by considering the average number of transmissions $\bar{\tau}(z)$ and adding (3.6) with (3.8), yielding

$$\bar{E} = 2E_{\mathrm{st}} + \frac{\bar{\tau}(z)}{W}\left[n_{\mathrm{fw}}(P_{\mathrm{el}} + P_{\mathrm{PA,d}}) + (\rho + n_{\mathrm{fb}})(P_{\mathrm{el}} + P_{\mathrm{PA,p}})\right] \tag{3.9}$$

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

In order to obtain $\bar{E}_{\mathrm{b}}$, the average energy per successful bit, we normalize the result in (3.9) by the payload length times the probability of success after z attempts, yielding

$$\bar{E}_{\mathrm{b}}(z) = \frac{\bar{E}}{L_{\mathrm{D}}(1 - P_{\mathrm{out},z})}. \tag{3.10}$$

Energy Consumption of Frequency Diversity

In order to compare the scheme which we are proposing to a system with frequency diversity, we also need a model for the energy consumption of such a scheme. Consider that we have $\lceil\tau\rceil$ channels of bandwidth W to send copies of the message, which are combined by the receiver using MRC. In this case, to achieve the target error probability, the effective average SNR of each message is given by [59]

$$\bar{\gamma}_{\mathrm{eff}}^{\mathrm{freq}} = \frac{m\left(2^R - 1\right)}{\left(T_{\mathrm{out}}\Gamma(\lceil\tau\rceil m + 1)\right)^{1/\lceil\tau\rceil m}}. \tag{3.11}$$

The maximum average SNR for each channel $\bar{\gamma}_{\mathrm{max}}^{\mathrm{freq}}$ is reduced accounting for the radiated power in all channels as

$$\bar{\gamma}_{\mathrm{max}}^{\mathrm{freq}} = \bar{\gamma}_{\mathrm{max}}/\lceil\tau\rceil, \tag{3.12}$$

to account for the peak power constraint.

Moreover, making

$$\bar{\gamma}_{\mathrm{p}} = \bar{\gamma}_{\mathrm{max}}^{\mathrm{freq}} \tag{3.13}$$

[3] Note that we consider the wake-up energy both at transmitter and receiver, thus we assume scheduled-rendezvous [83].

and calculating the effective average SNR using (3.11), we determine the average SNR of the data. Next, following similar steps as for the CC-HARQ case, the energy per successful bit for the case of frequency diversity is

$$\bar{E}_{\mathrm{f}}(\lceil\tau\rceil) = \frac{2E_{\mathrm{st}} + \frac{\lceil\tau\rceil}{W}\left[n_{\mathrm{fw}}(P_{\mathrm{el}} + P_{\mathrm{PA,d}}) + \rho(P_{\mathrm{el}} + P_{\mathrm{PA,p}})\right]}{L_{\mathrm{D}}(1 - P_{\mathrm{out},\lceil\tau\rceil})}. \tag{3.14}$$

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

3.2.2 Latency Constraint

The goal of this chapter is to investigate the impact of CC-HARQ on the energy consumption of a point-to-point URLLC system. Traditionally, it is well understood that HARQ improves energy efficiency at the cost of higher latencies [84]. On the other hand, when considering URLLC, a strict maximum latency is imposed [85]. Therefore it is not obvious that CC-HARQ improves the energy efficiency in this scenario.

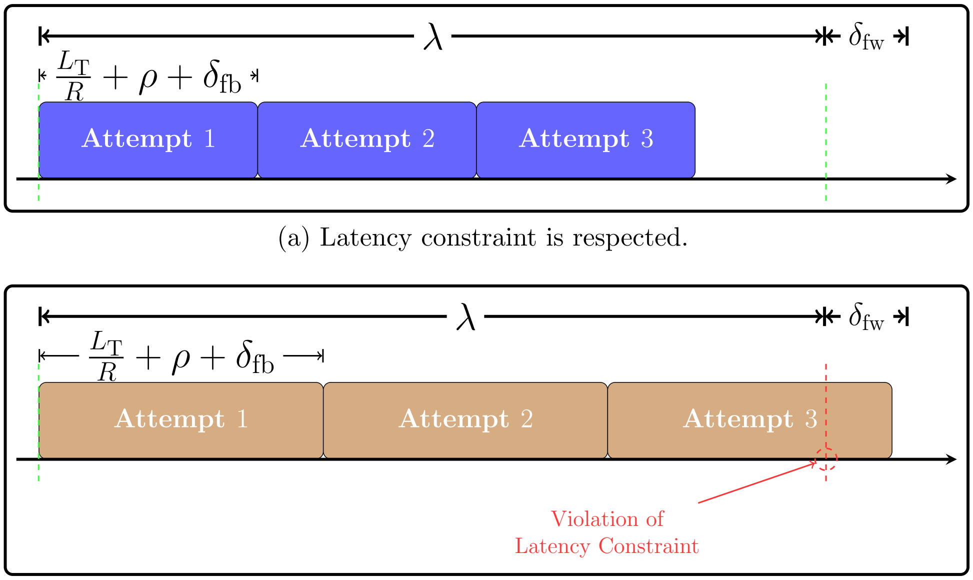
Here the transmitter must fit all z transmission attempts—and their associated acknowledgements, decoded—within a maximum latency λ' seconds using a bandwidth W. Note that, as mentioned before, the receiver estimates the channel at each attempt. If the accumulated SNR is below a threshold, it decodes the header and sends back a NACK immediately after all the symbols have been received and stored. Therefore, the entire message only has to be decoded once, saving latency. The time to decode it δ_{fw} is deducted λ', such that all transmission attempts have to fit within $\lambda = \lambda' - \delta_{\mathrm{fw}}$ seconds. This can be viewed as a constraint on the minimum communication rate R_{min} in bits per channel use. Fig. 3.2 illustrates this idea in a case where $z = 3$ and with normalized bandwidth. Note that in Fig. 3.2a, the communication rate is higher than the minimum and it is possible to transmit L_{T} bits z times in less than λ seconds. Conversely, in Fig. 3.2b, the rate is lower than R_{min} and attempting to communicate L_{T} bits using up to z transmission attempts violates the constraint.

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

$$W\lambda = z(n_{\mathrm{fw}} + n_{\mathrm{fb}} + \rho + W\delta_{\mathrm{fb}}), \tag{3.15}$$

where δ_{fb} is the time it takes for the transmitter to decode the feedback packet. Next, we substitute n_{fw} and n_{fb} defined in (3.1) and (3.5), respectively, in (3.15) while using $R = R_{\mathrm{min}}$. Then using $L_{\mathrm{T}} = L_{\mathrm{H}} + L_{\mathrm{D}} + L_{\mathrm{fb}}$, and solving for R_{min}, we arrive at

$$R_{\mathrm{min}} = z\frac{L_{\mathrm{T}}}{W(\lambda - z\delta_{\mathrm{fb}}) - z\rho}. \tag{3.16}$$



(a) Latency constraint is respected.

(b) Latency constraint is violated.

Figure 3.2: Illustration of the reason why a maximum latency constraint imposes a minimum rate. In this example, $z = 3$.

Note from (3.16) that using a larger z imposes a higher R_{min}, resulting in a trade-off between diversity and rate. Moreover, despite having excellent performance in terms of error rates, turbo codes, the most commonly used in long-term evolution (LTE), have a significant complexity [86] and thus may not be suitable for CC-HARQ considering URLLC applications. Instead, polar codes are good candidates for encoding feedback signals, as they are simple to implement and can be used to encode and decode short feedback messages within negligible time [86], such that $\delta_{\text{fb}} \ll \lambda$. The trade-off is a slight loss in terms of error rate [87], which we have added to the path loss model as M_{c}.

3.3 Optimization

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

$$\begin{aligned} \underset{z\in\mathbb{N}^*}{\text{minimize}} \quad & \bar{E}_{\text{b}}(z) && \text{(3.17a)}\\ \text{subject to} \quad & P_{\text{out},z} \le T_{\text{out}}, && \text{(3.17b)}\\ & P_{\text{rf}} \le P_{\text{rf,max}}, && \text{(3.17c)}\\ & R \ge R_{\text{min}}. && \text{(3.17d)} \end{aligned}$$

3.3.1 Optimizing Maximum Energy

Although it can be numerically verified that the objective function is convex with respect to z, and thus has one unique global optimal solution, to the best of our knowledge, it is not possible to prove it analytically due to the shape of $\bar{\tau}$ and where it appears in (3.17a).

Thus, we propose an alternative approach, where we optimize the maximum energy consumption[4] $\hat{E}_{\text{b}}$ in the same setup. The value of $\hat{E}_{\text{b}}$ is obtained by replacing $\bar{\tau}(z)$ by z in (3.10), such that

$$\hat{E}_{\text{b}}(z) = \frac{2E_{\text{st}} + \frac{z}{WR}\Phi}{L_{\text{D}}(1 - P_{\text{out},z})}, \tag{3.18}$$

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

This approach simplifies the problem as it removes the average number of attempts—obtained via (2.20)—, which is not trivially tractable, from the objective function. Moreover, this allows us to design a protocol with the worst-case scenario in mind, which is a sensible approach in URLLC, and we also show numerically that this result yields almost the same performance as a solution obtained numerically via the problem in (3.17). Moreover, we show that in this case, the optimal rate $R^\star$ and when R_{min} exceeds this optimal, the optimal number of transmission attempts $\hat{z}^\star$ can be obtained via a floating-point equation. Furthermore, using the obtained result, we are able to show that when the link budget is more stringent, as is the case of URLLC, $R^\star$ becomes smaller, and using the obtained $\hat{z}^\star$ becomes advantageous.

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

$$R^\star = \min\left(\frac{W_0\left(\frac{\Omega}{\mathrm{e}}\right) + 1}{\ln(2)}, R_{max}\right), \tag{3.19}$$

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

[4] By optimizing the maximum energy consumption, we mean minimize the maximum possible consumption, considering that all z attempts are used.

main Lambert-W function [88],

$$\Omega = \frac{\frac{L_{fb}}{L_{fw}}(P_{el} + Ad^{\alpha}\bar{\gamma}_p) - \Delta + P_{el}}{\Delta} \tag{3.20}$$

and

$$\Delta = \frac{Ad^{\alpha} m\Gamma(mz)}{\Gamma_{inc}^{-1}(mz, T_{out})}, \tag{3.21}$$

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

Proof. Considering an URLLC scenario,

$$P_{\text{out},z} \leq T_{\text{out}} \ll 1, \tag{3.22}$$

and thus $\hat{E}_{\text{b}}$ can be well approximated by simplifying the denominator of (3.18), as

$$\hat{E}_{\text{b}}(z) \approx \frac{2E_{\text{st}} + \frac{z}{WR}\left[L_{\text{fw}}(P_{\text{el}} + P_{\text{PA,d}}) + L_{\text{fb}}(P_{\text{el}} + P_{\text{PA,p}})\right]}{L_{\text{D}}}, \tag{3.23}$$

considering perfect channel state information at the receiver, for tractability. Using (3.4) and computing the derivative with respect to $\bar{\gamma}_{\text{d}}$,

$$\frac{\partial \hat{E}_{\text{b}}}{\partial \bar{\gamma}_{\text{d}}} = \frac{z}{WR} L_{\text{fw}} A d^{\alpha} > 0, \tag{3.24}$$

thus, increasing the transmit power to obtain a better (larger) SNR results in a larger $\hat{E}_{\text{b}}$. Therefore, we assume that the transmit power used is the one that guarantees the target outage, such that solving (2.17) for $\bar{\gamma}$ results in

$$\bar{\gamma}_{\text{d}} = \frac{m\left(2^R - 1\right)\Gamma(mz)}{\Gamma_{\text{inc}}^{-1}(mz, T_{\text{out}})}. \tag{3.25}$$

Next, we obtain $P_{\text{PA,d}}(z)$ using (3.4) and (3.25) and replace it into (3.23). Finally, we solve $\partial \hat{E}_{\text{b}}/\partial R = 0$ yielding

$$\Delta 2^{R^{\star}}(\ln(2)R^{\star} - 1) = \Omega, \tag{3.26}$$

after simple algebraic manipulations. Lastly, we use the upper part of the main branch of the Lambert-W function to solve (3.26) for $R^{\star}$, arriving at

$$R^{\star} = \frac{W_0\left(\frac{\Omega}{\text{e}}\right) + 1}{\ln(2)}. \tag{3.27}$$

However, when accounting for peak power limitations, the SNR in (3.25) is limited at

$\bar{\gamma}_{\text{max}}$, such that setting $P_{\text{rf}} = P_{\text{rf,max}}$ in (3.4) and solving for $\bar{\gamma}$ yields

$$\bar{\gamma}_{\text{max}} = \frac{P_{\text{rf,max}}}{Ad^{\alpha}}. \tag{3.28}$$

In other words, $R^{\star}$ is limited by R_{max}, which is obtained using $\bar{\gamma}_{\text{max}}$ in (2.17) and solving for R with $P_{\text{out},z} = T_{\text{out}}$, as

$$R_{\text{max}} = \log_2\left(1 + \bar{\gamma}_{\text{max}} \frac{\Gamma_{\text{inc}}^{-1}(mz, T_{\text{out}})}{m\Gamma(mz)}\right). \tag{3.29}$$

Lastly, combining (3.27) and (3.29) yields (3.19).

□

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

$$2m\hat{z}\left(\ln(2)\frac{L_T}{W\lambda}\hat{z} - 1\right) + \ln(m\hat{z}) = 1 - \ln(T_{out}^2 2\pi) \tag{3.30}$$

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

Proof. The optimization problem is defined as

$$\begin{aligned} \underset{z \in \mathbb{N}^*}{\text{minimize}} \quad & \hat{E}_{\text{b}}(z) && (3.31\text{a}) \\ \text{subject to} \quad & P_{\text{out},z} \leq T_{\text{out}}, && (3.31\text{b}) \\ & P_{\text{rf}} \leq P_{\text{rf,max}}, && (3.31\text{c}) \\ & R \geq R_{\text{min}}. && (3.31\text{d}) \end{aligned}$$

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

$$R = R_{\text{min}} \approx z\frac{L_{\text{T}}}{W\lambda}, \tag{3.32}$$

can be obtained from (3.16), by considering that the time of transmitting the pilots is negligible in comparison to the data transmission.

Then, obtaining $P_{\text{PA,d}}(z)$ as in the proof of Theorem 1 and replacing (3.32) into (3.23) we arrive at

$$\hat{E}_{\text{b}} \approx \frac{2E_{\text{st}} + \frac{\lambda}{L_{\text{T}}}\left[L_{\text{fw}}P_{\text{PA,d}}(z) + L_{\text{fb}}P_{\text{PA,p}} + L_{\text{T}}P_{\text{el}}\right]}{L_{\text{D}}}. \tag{3.33}$$

Then, considering that $z^{\star}$ can assume real values, we derive $\hat{E}_{\text{b}}$ with respect to $\hat{z}$ and equate to zero. This can only be done because we can prove the convexity of $\hat{E}_{\text{b}}$ with respect to $\hat{z}$, which we will show next. Assuming high spectral efficiency, $2^R \gg 1$, which is true in many practical applications even for small R, it is possible to rewrite $\partial \hat{E}_{\mathrm{b}}/\partial \hat{z} = 0$ as (3.30), as we will demonstrate shortly. □

3.3.2 Discussion Around Convexity

Convexity of $\hat{E}_{\mathrm{b}}$

In this section, we perform algebraic manipulations with $\hat{E}_{\mathrm{b}}$ in order to determine its convexity with respect to $\hat{z}$. The methodology of this proof is based on the literature, mainly on [89]. From (3.33), we note that the only function of $\hat{z}$ on the right hand side of the equation is $P_{\mathrm{PA,d}}$, which is directly proportional to $\bar{\gamma}_{\mathrm{d}}$. Therefore, it is explicit that proving the convexity of $\bar{\gamma}_{\mathrm{d}}$ with respect to $\hat{z}$ is sufficient proof that $\hat{E}_{\mathrm{b}}$ is convex with respect to $\hat{z}$.

Setting the target outage as the probability of outage[5], such that $P_{\mathrm{out}} = T_{\mathrm{out}}$ and using $\bar{\gamma} = \bar{\gamma}_{\mathrm{d}}$ into (2.18)(the high SNR approximation for the probability of outage) we can solve it for $\bar{\gamma}_{\mathrm{d}}$, yielding

$$\bar{\gamma}_{\mathrm{d}} = \frac{m(2^R - 1)}{(T_{\mathrm{out}}\Gamma(mz+1))^{\frac{1}{mz}}}. \tag{3.34}$$

Next, considering using the Stirling approximation[6] for factorials [90]

$$x! \approx \sqrt{2\pi x}\left(\frac{x}{\mathrm{e}}\right)^x, \tag{3.35}$$

where e is Euler's constant, coupled with the fact that $\Gamma(x+1) = x!$ [74], allows us to rewrite (3.34) as

$$\bar{\gamma}_{\mathrm{d}} \approx \frac{m2^{\hat{z}\frac{L_{\mathrm{T}}}{W\lambda}}}{\left(T_{\mathrm{out}}\sqrt{2\pi m\hat{z}}\right)^{\frac{1}{m\hat{z}}}\frac{m\hat{z}}{\mathrm{e}}}, \tag{3.36}$$

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

Equation (3.36) can be written in the form of $f(\hat{z})g(\hat{z})$, with

$$f(\hat{z}) = m2^{\hat{z}\frac{L_{\mathrm{T}}}{W\lambda}} \tag{3.37}$$

and

$$g(\hat{z}) = \frac{1}{\left(T_{\mathrm{out}}\sqrt{2\pi m\hat{z}}\right)^{\frac{1}{m\hat{z}}}\frac{m\hat{z}}{e}}. \tag{3.38}$$

[5] The reason this can be done is further explored when we discuss the convexity of the constraints.

[6] Note that we are not loosing accuracy here, as we want to find a proof of convexity by using this approximation. Moreover, the approximation is used for tractability.

Moving on, (3.38), by its turn, can be represented via

$$g(\hat{z}) = \frac{1}{g_1(\hat{z})g_2(\hat{z})}, \tag{3.39}$$

where

$$g_1(\hat{z}) = \left(T_{\text{out}}\sqrt{2\pi m\hat{z}}\right)^{\frac{1}{mz}} \tag{3.40}$$

and

$$g_2(\hat{z}) = m\hat{z}/\mathrm{e}. \tag{3.41}$$

First, to show that $g_1(\hat{z})$ is log-concave, we must have [89]

$$\frac{\partial^2 \ln(g_1(\hat{z}))}{\partial \hat{z}^2} \leq 0. \tag{3.42}$$

Since

$$\frac{\partial^2 \ln(g_1(\hat{z}))}{\partial \hat{z}^2} = \frac{4\ln\left(T_{\text{out}}\sqrt{2\pi m\hat{z}}\right) - 3}{2(m\hat{z})^3}, \tag{3.43}$$

and $m\hat{z} \neq 0$, (3.42) becomes

$$0 < \hat{z} \leq \frac{\mathrm{e}^{3/2}}{2\pi T_{\text{out}}^2}, \tag{3.44}$$

after some algebraic manipulations. Because T_{out} is always positive and very small and $\hat{z} \geq 1$, (3.44) holds in any practical scenario. Thus, $g_1(\hat{z})$ is log-concave.

Moreover, because $g_2(\hat{z})$ is both log-convex and log-concave [89], the product of $g_1(\hat{z})$ and $g_2(\hat{z})$ is also log-concave [89] and therefore its inverse, $g(\hat{z})$, is log-convex [89].

Finally, since $f(\hat{z})$ is an exponential function, it is log-convex [89]. The product of two log-convex functions is also log-convex [89], thus $\bar{\gamma} = f(\hat{z})g(\hat{z})$ is also log-convex and therefore convex, concluding the proof. □

Convexity of the constraints in (3.31)

I It is possible to observe in (2.18) that $P_{\text{out,z}}$ is monotonically decreasing with respect to $\bar{\gamma}$. Moreover, analyzing (3.31a), we can see that $\hat{E}_{\text{b}}$ is monotonically increasing with respect to $\bar{\gamma}$. Thus, $P_{\text{out},z}$ should be as high as possible in order to minimize (3.31a) and therefore to respect (3.31b), we must have

$$P_{\text{out},z} = T_{\text{out}}. \tag{3.45}$$

Therefore, the constraint can be incorporated into the objective function and its convexity does not bear any importance to the solution of the optimization problem.

II The inequality in (3.31c) can be rewritten as

$$\bar{\gamma}_{\mathrm{d}} \leq \frac{P_{\mathrm{rf,max}}}{Ad^{\alpha}}, \tag{3.46}$$

by using (3.3).

This means that the proof of convexity of $\bar{\gamma}_{\mathrm{d}}$, presented earlier, proves that (3.31c) is a convex constraint.

III In (3.32) we can clearly observe that R_{min} is linear with respect to z, which means that (3.31d) is convex with respect to z.

3.3.3 Derivation of (3.30)

In possession of the proof that $\hat{E}_{\mathrm{b}}$ is convex with respect to $\hat{z}$, we can find the solution to our optimization problem by solving

$$\frac{\partial \hat{E}_{\mathrm{b}}}{\partial \hat{z}} = 0, \tag{3.47}$$

which yields

$$\begin{gathered}\frac{\partial \frac{2E_{\mathrm{st}}+\frac{\lambda}{L_{\mathrm{T}}}\left[L_{\mathrm{fw}}Ad^{\alpha}\bar{\gamma}_{\mathrm{d}}+L_{\mathrm{fb}}P_{\mathrm{PA,p}}+L_{\mathrm{T}}P_{\mathrm{el}}\right]}{L_{\mathrm{D}}}}{\partial \hat{z}} = 0 \\ \partial\bar{\gamma}_{\mathrm{d}}/\partial\hat{z} = 0.\end{gathered} \tag{3.48}$$

Computing the derivative above results in

$$\Xi\left(2\ln(2)\frac{L_{\mathrm{T}}}{W\lambda}m\hat{z}^{2} - 2m\hat{z} + \ln(\hat{z}) + \ln\left(\frac{2\pi T_{\mathrm{out}}^{2}m}{\mathrm{e}}\right)\right), \tag{3.49}$$

where

$$\Xi = \frac{\mathrm{e}2^{\hat{z}\frac{L_{\mathrm{T}}}{W\lambda}-\frac{1}{2m\hat{z}}-1}}{m\left(T_{\mathrm{out}}\sqrt{\pi m\hat{z}}\right)^{\frac{1}{m\hat{z}}}}. \tag{3.50}$$

Since

$$\Xi \neq 0, \tag{3.51}$$

replacing (3.49) in (3.48) and diving both sides by Ξ results in (3.30).

3.3.4 Discussion

The solution presented in this section can be utilized to obtain insights into the behaviour of the optimization problem. For instance, because in URLLC the values of T_{out} are always positive and much smaller than one, the left-hand side of (3.30) always yields a positive value. Therefore, fixing the values of T_{out}, m, L_{T} and W in (3.30) and choosing a smaller value for λ—considering a more stringent latency—causes $\hat{z}^{\star}$ to be smaller. In other words, having a more severe constraint in terms of latency imposes that fewer maximum attempts are optimal, which can be explained by having less time for more attempts. With the same rationale, but now fixing λ and decreasing the number of bits to communicate (L_{T}), yields a larger $\hat{z}^{\star}$. This is because the duration of each attempt is shorter, due to fewer bits being conveyed. Similarly, decreasing m results in a larger $\hat{z}^{\star}$, which is due to the diversity gains being more relevant in a worse channel condition. Besides, considering a communication channel with smaller bandwidth has a similar effect as considering a stricter latency in the solution of (3.30). Lastly, keeping the parameters on the right-hand side of (3.30) fixed and increasing the value of T_{out}, (*i.e.* considering a less reliable communication) yields a smaller $\hat{z}^{\star}$, which is explained by the fact that the gains in diversity are less important for a more relaxed reliability.

3.3.5 Optimizing EMF Radiation

Depending on the application, optimizing the energy efficiency of the communication might not be the most interesting strategy. For instance, in health and safety applications, we might want to minimize the EMF exposure instead.

We can thus write the following optimization problem:

$$\begin{aligned}
\underset{z \in \mathbb{N}^*}{\text{minimize}} \quad & E_{\text{PA}} = z\frac{L_{\text{T}}}{RW}\frac{N_0 W M_{\text{l}} M_{\text{c}} A_0 d^{\alpha}\bar{\gamma}}{\eta} && \text{(3.52a)}\\
\text{subject to} \quad & P_{\text{out},z} \leq T_{\text{out}}, && \text{(3.52b)}\\
& P_{\text{rf}} \leq P_{\text{rf,max}}, && \text{(3.52c)}\\
& R \geq R_{\text{min}} && \text{(3.52d)}
\end{aligned}$$

To solve (3.52), since the reliability requirement is stringent, we assume that we are operating at the high SNR region, which is true in several URLLC applications [20,21,59,91], and thus can we use (2.18) to determine $P_{\text{out,z}}$. It is possible to observe in (2.18) that $P_{\text{out,z}}$ is monotonically decreasing with respect to $\bar{\gamma}$. Moreover, analyzing (3.52a), we can see that E_{PA} is monotonically increasing with respect to $\bar{\gamma}$. Thus, $P_{\text{out},z}$ should be as high as possible in order to minimize (3.52a) and therefore to respect (3.52b), we must have

$$P_{\text{out},z} = T_{\text{out}}. \tag{3.53}$$

Next, we solve for $\bar{\gamma}$ yielding

$$\bar{\gamma} = \frac{m(2^R - 1)}{(T_{\text{out}}\Gamma(mz+1))^{\frac{1}{mz}}}. \tag{3.54}$$

Then, replacing (3.54) into (3.52a) we prove that E_{PA} is monotonically increasing with respect to R and thus the smallest value possible of R yields the optimal solution to (3.17). This coupled with (3.17d) yields

$$R = R_{\min}. \tag{3.55}$$

Here, we disregard the pilot used for estimation—such that we can simplify z in the objective function[7]—and write the minimum rate as

$$R_{\min} = z\frac{L_{\mathrm{T}}}{W\lambda}. \tag{3.56}$$

The next step is to replace (3.56) into (3.55) and then into (3.52a), yielding

$$\underset{z\,\in\,\mathbb{N}^{*}}{\text{minimize}} \quad E_{\mathrm{PA}} = \frac{\lambda N_0 W M_{\mathrm{l}} M_{\mathrm{c}} A_0 d^{\alpha} \bar{\gamma}}{\eta}. \tag{3.57a}$$

The problem above can be solved identically to the optimization problem for energy efficiency. However, since we have the guarantees that the minimum operating rate is optimal, and moreover because we want to optimize the optimal energy in this type of application, the solution is now exact. Thus, solving

$$2m\hat{z}\left(\ln(2)\frac{L_{\mathrm{T}}}{W\lambda}\hat{z} - 1\right) + \ln(m\hat{z}) = 1 - \ln(T_{\mathrm{out}}^2 2\pi), \tag{3.58}$$

for $\hat{z}$ will yield an optimal value in terms of EMF radiation minimization. Note that, because we found the solution by relaxing z to assume real values, the actual policy implemented $z^{\star}$ is obtained by choosing the closest integer to $\hat{z}^{\star}$. Albeit this solution is in the form of a floating-point equation, we can still use it to make predictions regarding the optimal value of z with respect to the system parameters. For instance:

I A more stringent latency causes $z^{\star}$ to be smaller;

II Relaxing the target outage causes an increase in $z^{\star}$;

III Having access to less spectrum causes $z^{\star}$ to decrease;

IV When there are fewer bits to exchange, $z^{\star}$ grows;

V When the LOS conditions are worse, z increases;

VI The large scale path loss does not influence the value of $z^{\star}$.

[7] Note that this assumes that the energy of sending the pilots is much smaller than the energy of data transmission, which is typically the case when one symbol is used for channel estimation.

3.4 Numerical Results

In this Section, we evaluate several aspects of the proposed approach.

3.4.1 Energy Efficiency Optimization

In this Subsection, we are using parameters from the METIS test case #5 [12]: smart grid communications. The parameters considered are summarized in Table 3.1. This example fits within the block-fading model because there is almost no mobility and consecutive transmissions are assumed to be performed in uncorrelated frequency channels. In addition, the Nakagami-m channel model describes the scenario well since it correctly depicts the variability of situations encountered in smart grids, with varied LOS conditions [92]. Additionally, to perform slow frequency hopping, we assume that the transmitter uses sub-carriers with independent and identically distributed channel gains and bandwidth W for each transmission attempt, such that the average consumed bandwidth is $\bar{\tau}W$ Hz per message.

First, we show in Fig. 3.3a and Fig. 3.3b how $\bar{E}_\mathrm{b}$ and $\hat{E}_\mathrm{b}$ vary with respect to z for various strategies regarding SNR and using the rate according to[8]

$$R = \max\left(R^\star, R_\mathrm{min}\right). \tag{3.59}$$

Moreover, the SNR strategies are characterized by: 1) A benchmark obtained numerically, by optimizing the average SNR to determine its optimal value $\bar{\gamma}_\mathrm{d}^\star$, 2) using the average SNR that guarantees the target outage $\bar{\gamma}_\mathrm{d,min}$ and 3) using the maximum average SNR according to the maximum transmit power $\bar{\gamma}_\mathrm{d,max}$. Furthermore, we also show $\hat{E}_\mathrm{b}$ using $\bar{\gamma}_\mathrm{d}^\star$ and $R = R_\mathrm{min}$ to illustrate how using a higher rate impacts the value of $\hat{E}_\mathrm{b}$. Note that as the curve of $\bar{E}_\mathrm{b}$ is a benchmark, it has been generated using a numerically obtained optimal number of pilots for channel estimation, while the other curves represent real implementations and therefore use a fixed $\rho = 6$.

In Fig. 3.3a, we show the performance for the case with a Nakagami parameter of $m = 1$. We can notice that because there is no LOS, the diversity gains are not as impactful and the average SNR has little impact on $\bar{E}_\mathrm{b}$, such that the curve with $\bar{\gamma}_\mathrm{d,min}$ performs very close to the benchmark, more so for the smaller values of z. This means that using the results of Corollary 1, which assumes that using the lowest possible SNR will yield a good performance in such scenarios. Additionally, we observe that optimizing the rate has little to no effect on $\hat{E}_\mathrm{b}$, as due to the stringent link budget characteristics, $R^\star$ will be close to R_min. Moreover, the curve for $\hat{E}_\mathrm{b}$ with $\bar{\gamma}_\mathrm{d}^\star$ matches exactly with the one with $\bar{\gamma}_\mathrm{d,min}$ (both for Figs. 3.3a and 3.3b), as predicted analytically in (3.24).

[8] If $R_\mathrm{max} < R_\mathrm{min}$ the link cannot be closed for λ and T_out.

Table 3.1: CC-HARQ Energy Efficiency Optimization Simulation Parameters

Parameter	**Value**
Target Outage (T_{out})	10^{-5} [12]
Typical Link Latency (λ)	6.48 ms†
Time to decode feedback (δ_{fb})	0.0213 ms [87]
Maximum Transmit Power ($P_{\mathrm{rf,max}}$)	-16 dBW [92]
Bandwidth (W)	180 KHz [92]
Distance (d)	100 m
Spectral Noise Power Density (N_0)	-204 dBW/Hz
Link Margin (M_{l})	15 dB
Coding Margin (M_{c})	3 dB [87]
Path Loss Exponent (α)	2.5 [67]
Attenuation at reference (A_0)	38.5 dB [67]‡
Header Length (L_{H})	16 bits [16]
Feedback Length (L_{fb})	17 bits††
Payload Length (L_{D})	1216 bits [12]
Radio Startup Energy (E_{st})	0.125 nJ [82]
Average PA Efficiency (η)	50% [93]
Electronic Power (P_{el})	-15.69 dBW [16]

† $\lambda' = 8$ms [12] and δ_{fw} =1.5ms [87].

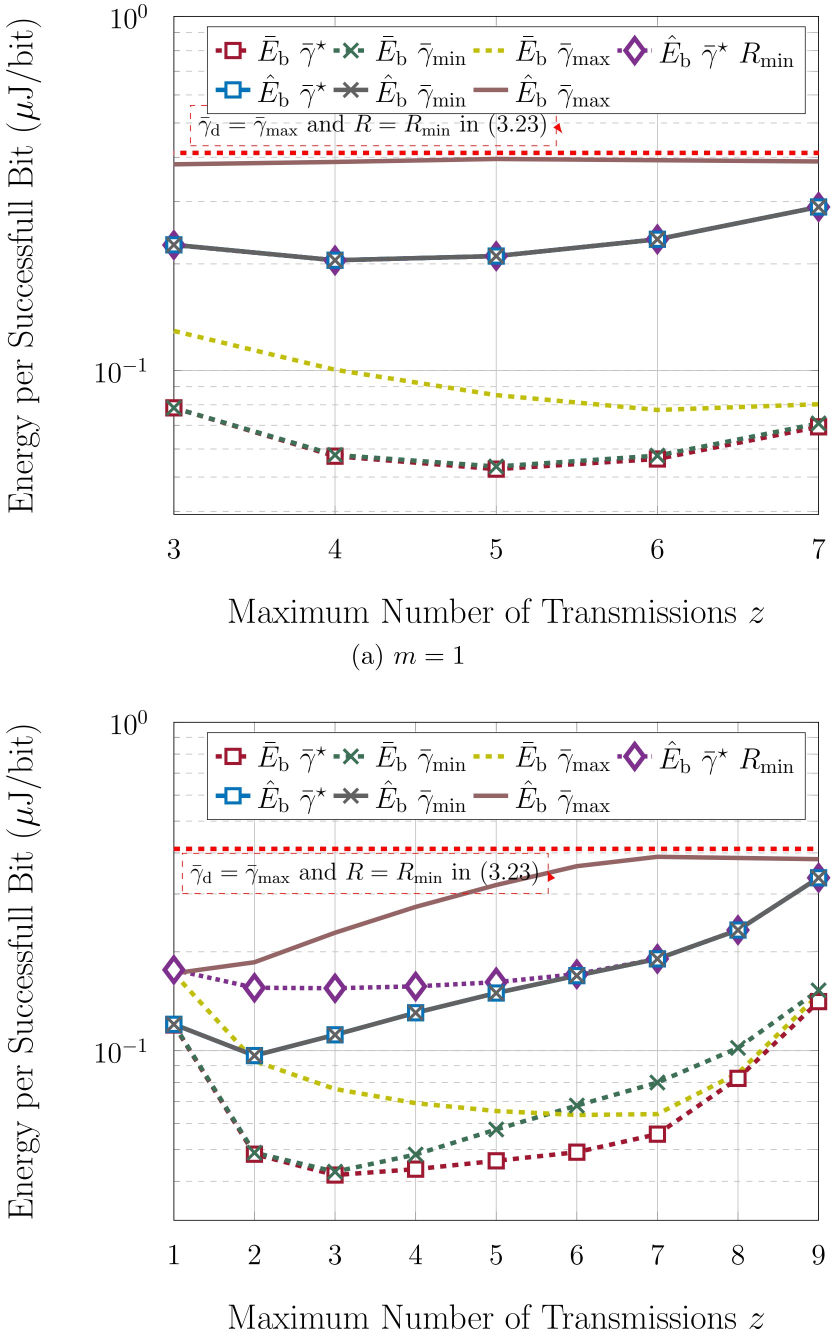
‡ We consider a reference distance of 1 m, unit gain on the antennas and a carrier frequency centered at 2 GHz.

†† We consider 1 bit feedback, such that $L_{\mathrm{fb}} = L_{\mathrm{H}} + 1$.

On the other hand, in Fig. 3.3b, $m = 3$ is shown, and thus the link is less stringent. Here, the diversity gains of using higher z benefit more from using $\bar{\gamma}_{\mathrm{d}}^{\star}$ in terms of $\hat{E}_{\mathrm{b}}$. However, when z is relatively small (≤ 4), using $\bar{\gamma}_{\mathrm{d,min}}$ still performs very close to the benchmark. Also note that, because of the more relaxed link budget, there is more freedom to increase the rate, and using the results from Theorem 1 results in performance gains, in particular where z is smaller. The red dotted line in both Fig. 3.3a and Fig. 3.3b is the result of using $\bar{\gamma}_{\mathrm{d}} = \bar{\gamma}_{\mathrm{max}}$ and $R = R_{\mathrm{min}}$ in (3.23). It represents a ceiling in the maximum consumption when $R^{\star} < R_{\mathrm{min}}$ and we operate at the maximum transmit power.

Further, in order to evaluate the performance of the proposed solution, we compare it with the case of a single transmission. To make the comparison fair, we allow the direct transmission to use $\bar{\tau}$ channels to send copies of the message and perform MRC at the receiver, exploiting frequency diversity. However, because it is not possible to use fractions of a sub-carrier and $\bar{\tau}$ is not always an integer, we use the next closest integer for consistency. All channels are estimated separately and this cost is taken into account when computing the energy used to send the message with frequency diversity $\bar{E}_{\mathrm{f}}$, presented in Section 3.2.1. Moreover, the optimal rate considering frequency diversity is similar to the one obtained in (3.19), and the value of R is determined according to (3.59).

Fig. 3.4 shows the ratio between the energy consumption considering frequency diversity and that of our proposed solution. The energy ratio considering the benchmark (obtained numerically) is determined by calculating $\bar{E}_{\mathrm{f}}(\lceil\bar{\tau}\rceil)/\bar{E}_{\mathrm{b}}(z^{\star})$, while the ratio which considers the results in Theorem 1 and Corollary 1 is calculated as $\bar{E}_{\mathrm{f}}(\lceil\bar{\tau}\rceil)/\bar{E}_{\mathrm{b}}(\hat{z}^{\star})$. As we can observe, for the target latency of 6.48 ms and $m = 3$, the proposed solution outperforms the frequency diversity by a factor of more than 2. Note that here, since $R^{\star} > R_{\mathrm{min}}$, considering a more stringent latency has little effect on the performance as the higher rate was already guaranteeing a more stringent latency. On the other hand, when $m = 1$, the link is more stringent and $R^{\star} < R_{\mathrm{min}}$, such that changes in λ incur in changes to the performance. Also regarding $m = 1$, we can see that for the target latency of 6.48 ms, the proposed scheme is about 8 times better, since the HARQ approach is able to achieve higher orders of diversity using the same spectrum on average. Because we use the next closest integer to $\bar{\tau}$ when choosing the number of channels to use for frequency diversity, our approach uses less bandwidth on average, in other words, our approach constitutes a better way of using the spectrum. Despite using less bandwidth on average, we can often achieve higher orders of diversity, resulting in energy savings. For example, for the case where $z^{\star} = 2$, the diversity orders are the same but because $\bar{\tau}$ is close to one, the frequency diversity approach uses almost twice the bandwidth. This result shows the strength of HARQ, achieving high diversity orders at low cost, which is possible because on average it requires few attempts and thus uses a small number of channels. Conversely, when considering frequency diversity, the cost of obtaining a large diversity is to use all channels in



(a) $m = 1$

(b) $m = 3$

Figure 3.3: Average energy used ($\bar{E}_\mathrm{b}$) and maximum possible energy required ($\hat{E}_\mathrm{b}$) versus z for various strategies, considering $m = 1$ and $m = 3$. As a reminder, R_min is the rate which ensures the latency constraint, $\bar{\gamma}_\mathrm{d}$ is the SNR of the data forward transmition, $\bar{\gamma}_\mathrm{max}$ is the maximum possible SNR obtained from peak power limitations and the link budget, $\bar{\gamma}_\mathrm{min}$ is the average SNR to guarantee the target reliability, and $\bar{\gamma}^\star$ is the optimal SNR in terms of energy considering the link budget, obtained numerically.

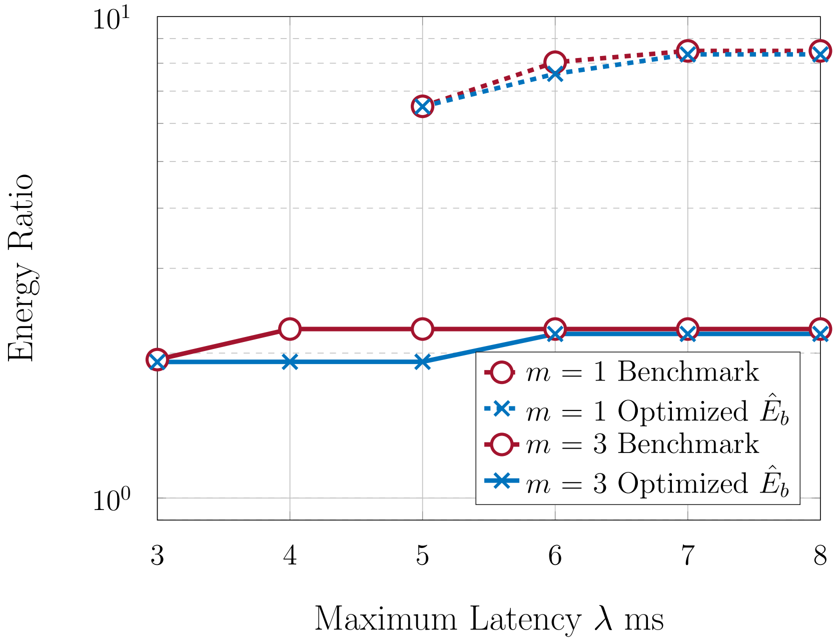


Figure 3.4: Performance for different λ with $T_{\text{out}} = 10^{-5}$ and comparison between proposed scheme and the benchmark. The plot for $m = 1$ starts at a higher value because the URLLC contraints could not be met with the link budget and the peak power limitation.

every transmission, which is less energy and spectral efficient.

In Fig. 3.5, we show the performance when considering different levels of reliability, considering the same ratios as before with $\lambda = 6.48$ ms. As we can observe, when we have good channel conditions ($m = 3$), increasing the reliability has less impact in the frequency diversity strategy compared to the case where $m = 1$, because the need for diversity is not as pressing and thus only a few channels are used each time. However, when there is no LOS ($m = 1$), increasing the reliability comes at a high cost for a direct transmission because of the need for diversity to reach the stringent reliability requirement. Since the CC-HARQ strategy does not need to use all of those resources in every attempt, it yields better performance when considering more reliable specifications. For instance, if we consider no LOS and $T_{\text{out}} = 10^{-6}$, our scheme outperforms the direct transmission with frequency diversity by more than 18 times. Furthermore, we can observe from this numerical example, both in Figs. 3.4 and 3.5, that optimizing $\hat{E}_{\text{b}}$ has very similar performance compared to the benchmark solution which uses numerically obtained optimal number of pilots, SNR, and z.

3.4.2 EMF Radiation Optimization

Here we show simulation results based on generic URLLC parameters. In these results, it is possible to confirm the predictions made above as well as show, with a numerical example, the benefits of using HARQ in terms of reducing the EMF radiation output. Unless

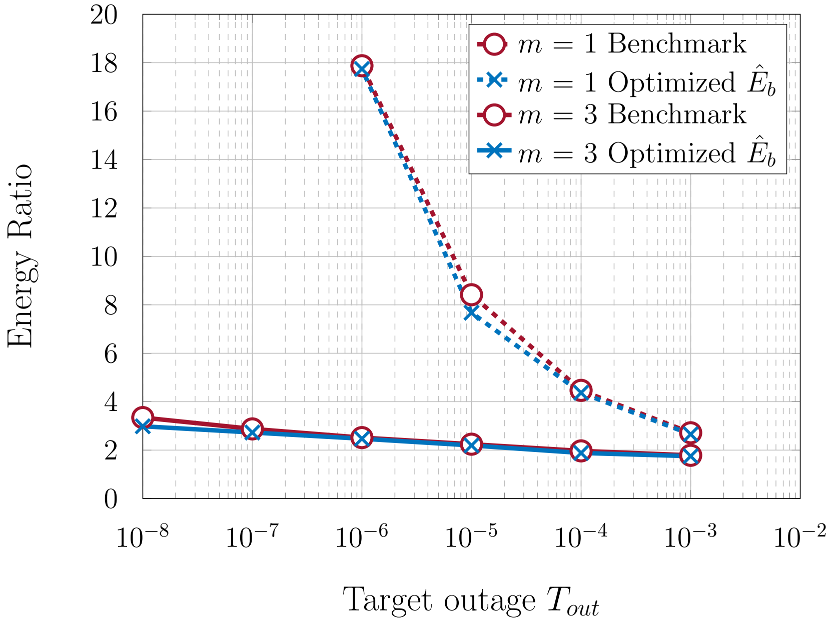


Figure 3.5: Performance for different T_{out} with $\lambda = 6.48$ ms and comparison between proposed scheme and the benchmark. The plot for $m = 1$ starts at a higher value because the URLLC contraints could not be met with the link budget and the peak power limitation.

otherwise stated, the parameters used to generate the plots are presented in Table 3.2.

The value of E_{PA} is shown in Fig. 3.6 considering two different lengths of bits and varied LOS conditions, encompassing sub-Rayleigh ($m = 0.5$) and Rayleigh ($m = 1$) as non-line-of-sight (NLOS) conditions as well as some LOS ($m = 3$). Fig. 3.6a contains the three conditions for $L_{\text{T}} = 512$ while Fig. 3.6b plots the same setup but with $L_{\text{T}} = 1024$. We can observe in both plots the trade-off involved in increasing z without violating the latency and reliability constraints. At first, the diversity gains make a large difference, in particular, because the target outage is so stringent, and the amount of energy required is severely higher in all cases when $z = 1$. As z increases, there is a point ($z^\star$) where the increased rates imposed to meet λ overcome the diversity gains and more energy is required for the communication.

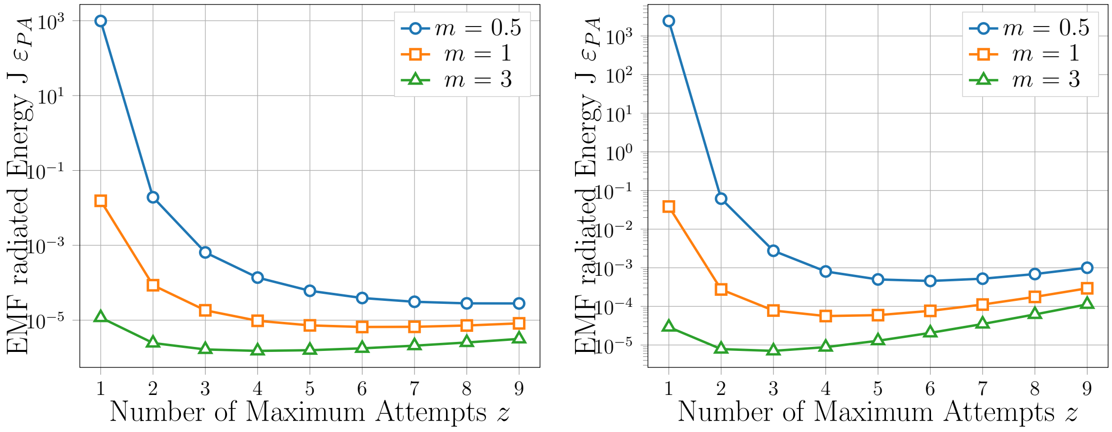
As predicted by analyzing (3.30), when m is larger (*i.e.* there is more LOS), $z^\star$ will be smaller. This can be explained by the fact that when the link is less stringent the diversity is not as important as when the link suffers more severely from the small scale fading. At the same time, we can also observe by comparing Fig. 3.6a and 3.6b that increasing the payload causes the tipping point to occur for smaller z. This is tied to the fact that the imposed rates are higher for larger L_{T} and the cost of increasing z is higher.

To validate the analytical results, we have plot in Fig. 3.7 the value of $z^\star$ found using (3.30) (before rounding to the closest natural number) alongside a value obtained using exhaustive search. As we can see, the values agree quite well and moreover, the prediction that a larger λ would incur in a larger $z^\star$ was also valid. The reason for this is that

Table 3.2: CC-HARQ EMF Radiation Minimization Simulation Parameters

Parameter	**Value**
Target Outage (T_{out})	10^{-5} [12]
Typical Link Latency (λ)	5 ms [12]
Bandwidth (W)	180 KHz [92]
Distance (d)	100 m [12]
Spectral Noise Power Density (N_0)	-204 dBW/Hz
Link Margin (M_{l})	15 dB
Coding Margin (M_{c})	3 dB [87]
Path Loss Exponent (α)	2.5 [67]
Attenuation at reference (A_0)	38.5 dB [67][‡]
Payload Length (L_{T})	1024 bits
Average PA Efficiency (η)	50% [93]

[‡] We consider a reference distance of 1 m, unit gain on the antennas and a carrier frequency centered at 2 GHz.



(a) $L_{\text{T}} = 512$

(b) $L_{\text{T}} = 1024$

Figure 3.6: EMF radiation energy for different values of z, considering various LOS conditions and two different lengths of L_{T}.

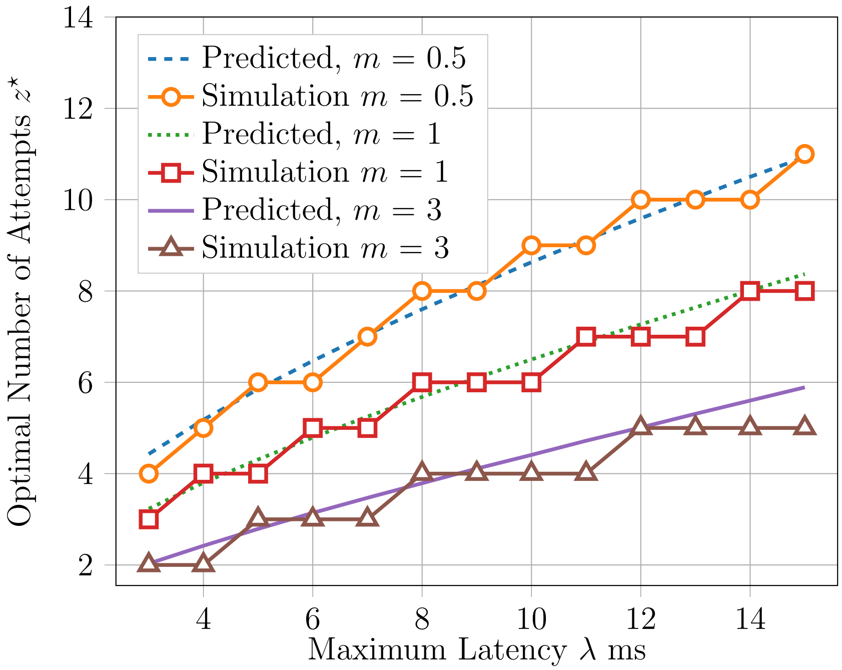


Figure 3.7: Value of $z^{\star}$ for different maximum latencies λ. Here the predicted value is the one obtained from (3.58) for $\hat{z}$.

when the latency is more relaxed, the cost of fitting more attempts within λ is somewhat mitigated.

Perhaps one of the more surprising conclusions of (3.30) is that large-scale path loss conditions do not influence the optimal policy to be used with regards to the number of allowed attempts. This is evidenced by the resulting derivation, which does not contain any elements of the large-scale path loss. It is important to note, however, that the actual amount of maximum EMF radiation will be higher for more severe large-scale path loss scenarios. What is not influenced is the policy with regards to the number of allowed attempts to minimize the EMF radiation. To illustrate this idea we have run the simulations considering one of the path loss parameters, the distance, and present the results in Fig. 3.8. Note that regardless of the distance the value of $z^{\star}$ remains the same. However, it is important to keep in mind that even though $z^{\star}$ does not change, E_{PA} grows with the distance, as to keep operating at the target reliability the transmit power has to compensate for the large scale loss. Moreover, increasing the link distance might cause changes in the LOS conditions, which would then change the value of m and that could have an impact on $z^{\star}$.

Here we can clearly observe one of the consequences of the real relaxation of z in order to obtain the predicted value of $z^{\star}$, expressed as $\hat{z}^{\star}$. As a result, this may introduce small errors in the optimal predicted value. Another consequence is that sometimes $\hat{z}^{\star}$ is higher and sometimes it is smaller than $z^{\star}$, as we can see in Fig. 3.8. These are artefacts introduced by the fact that $\hat{z}$ is a Real relaxation of z.

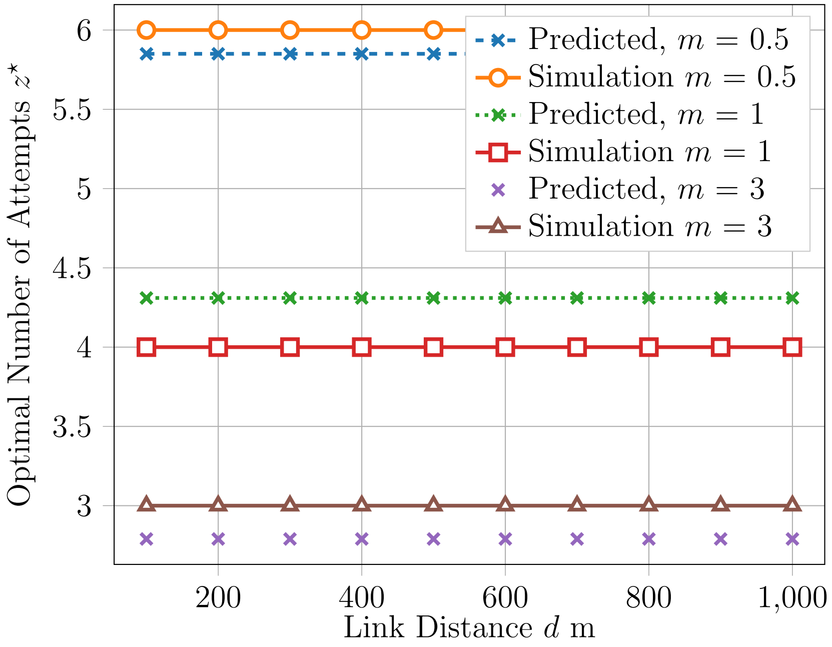


Figure 3.8: Value of the optimal number of attempts, $z^\star$, for different distances d.

To explore this further, in Fig. 3.9 we show $\hat{z}$ and z for different LOS conditions, expressed as different m values. Note how since the integer values of z changes in steps while the Real solution to (3.58) is continuous. Since the approximation curve sometimes crosses the discrete steps of $z^\star$, the predicted value is not a bound to the actual optimal policy. Also note that for a few values of m—for instance 3.8—the closest integer to $\hat{z}^\star$ does not coincide with $z^\star$. This is the aforementioned error.

As discussed before, the strength of HARQ lies in the fact that it uses on average a small number of resources. We can see that from Fig. 3.10, where we plot the optimal number of attempts and the average number of attempts that arises from using that optimal. Note how, even when allowing a large number of transmission attempts, the average number of transmissions is kept low. Also note that because we are increasing the rate for a higher z, no latency is added as a result of increasing z.

Motivated by this efficient use of resources, in Fig. 3.11, we compare the energy involved in using a frequency diversity scheme obtained via (3.14) with average SNR obtained from (3.12) with the energy for the proposed CC-HARQ communication. This allows us to showcase the benefits of using retransmissions even when low latencies are considered. In particular, when the link is more stringent (NLOS scenario, $m = 1$), the CC-HARQ can be used much less energy than the frequency diversity, despite having to increase its data rate to compensate for the added time diversity. This does not scale indefinitely, so depending on the parameters using the frequency diversity strategy might be beneficial, in particular for the most stringent latency situations. Note that, however, for those cases, the average number of transmissions for the CC-HARQ scheme was almost one, so the

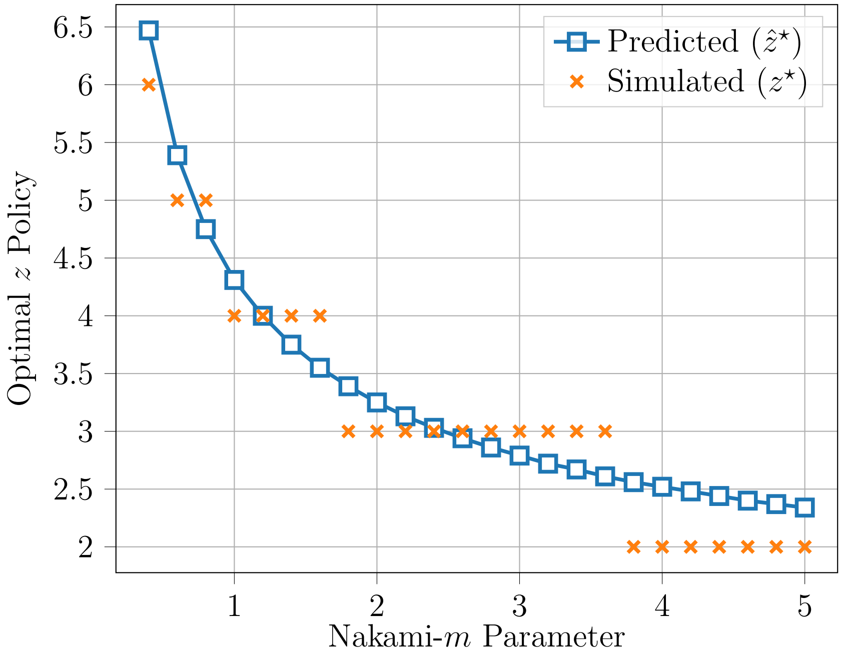


Figure 3.9: Value of the optimal number of attempts, $z^\star$, for different LOS conditions, represented by different values of m.

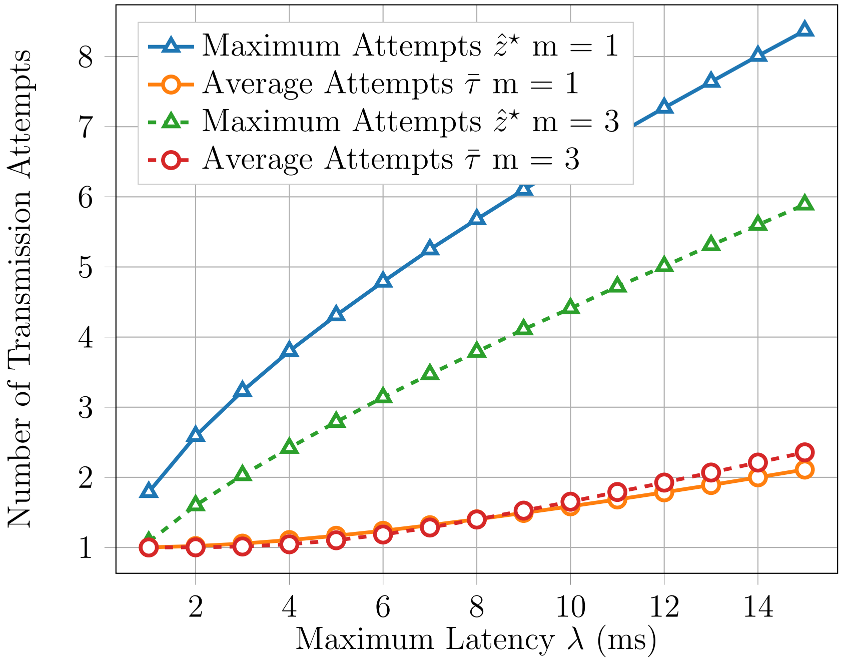


Figure 3.10: Average $\bar{\tau}$ and maximum $\hat{z}^\star$ number of transmission attempts for various target latencies λ.

spectrum usage would be much smaller than when using the frequency diversity with two replicas of the message (because we are rounding up) and thus there is still an argument for using CC-HARQ. It is also interesting to observe the behaviour of the curves around 13ms. The frequency diversity strategy has a discrete step at that point. This happens

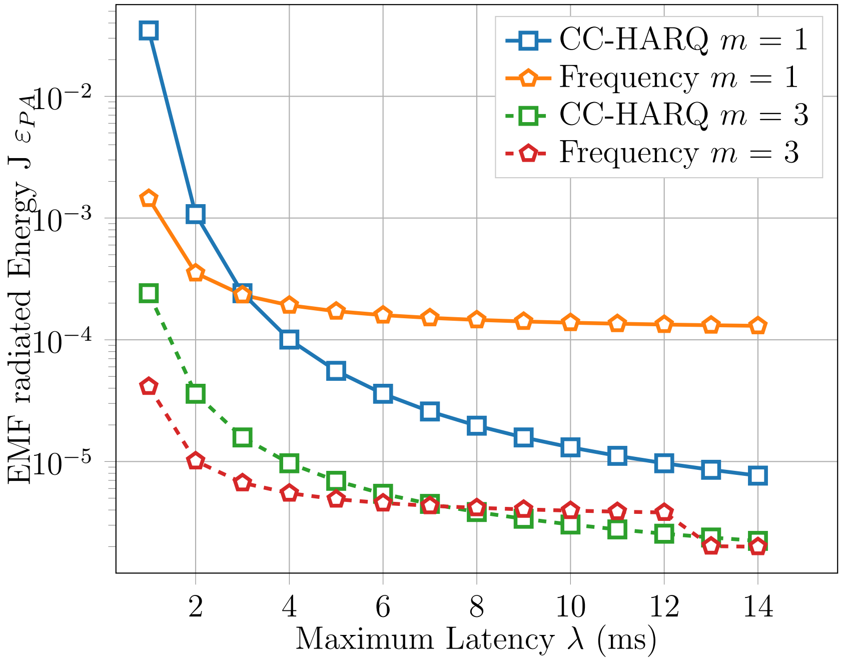


Figure 3.11: EMF radiation for both the frequency diversity scheme and the proposed CC-HARQ scheme.

because the number of allowed frequency channels is determined according to the average number of transmissions, to have a fair comparison. Moreover, it can only assume integer values, whereas the average number of transmissions is a real value. Hence, the former has discrete variations as observed in Fig. 3.11.

Lastly, the results showed here are similar to the ones in [59] in terms of $z^{\star}$, with the difference that here the result obtained analytically is optimal while in [59] it represents an approximation which can be used to obtain a good performance. This comes from the fact that in [59] the circuit consumption is taken into account for the complete problem (optimizing the energy efficiency) and here we only consider reducing the maximum EMF output.

3.5 Conclusion

In this chapter, we have analyzed the use of truncated CC-HARQ in order to meet the stringent reliability and latency requirements of URLLC whilst increasing energy efficiency in a block-fading Nakagami-m channel. We demonstrated that the energy consumption depends on the choice of allowed transmission attempts and can be minimized by optimally tuning this parameter. Moreover, we have shown how to minimize the EMF radiation of an application by doing a similar process. Further, we proposed solving an optimization problem that allows us to arrive at a fixed point equation to determine the optimum number of maximum transmission attempts, which can be used to obtain several insights in either scenario.

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

It is also interesting to note that the analysis herein relies on a couple of assumptions, namely that the decoding time is not prohibitively large, the feedback channel is perfect, and there is always a slot available for retransmission immediately, with respect to MAC. In Chapter 5 we present a solution that relies on SON in order to get around these assumptions in a practical deployment.

Moreover, the design proposed in this Chapter is application agnostic. However, depending on application-specific characteristics it is possible to have an even better operation, as we will see next, in Chapter 4.

Chapter 4

Communication and Control Joint Design

4.1 Introduction

URLLC comprises a wide range of applications, and therefore it is interesting to investigate whether or not we can benefit from application-specific characteristics when designing a communication protocol to overcome its stringent constraints.

More specifically, in the context of this Thesis, we want to study the design of HARQ protocols in order to support WNCS to the utmost effectiveness. In other words, this chapter aims to answer the question: Can we improve the performance of the HARQ communication protocol by taking into consideration WNCS application-specific parameters?

In the previous chapter, we managed to determine the optimal number of transmission attempts to allow for a generic URLLC application. Here, we want to take that a step further and explore what happens when we consider devising our communication strategy taking into account the benefits of using PPC.

Bringing wireless connectivity into industry automation processes brings many benefits as it is faster to deploy and easier to maintain when compared to solutions relying on wires to connect sensors, actuators, and controllers. However, several WNCS applications also have stringent requirements with regards to latency and reliability, typically on the order of sub-millisecond and 10^{-9}, respectively [94].

This means that we must be able to provide a URLLC in the wireless links to leverage the benefits of incorporating wireless technologies into control plants.

To achieve robust stability in WNCS, PPC can be used [95], such that the plant stores in a buffer a vector of future control signals determined via model predictive control whenever a message is successfully relayed to it. This way, when packets fail to be delivered, a predicted control signal can be used and the stability of the plant can be maintained over lossy networks.

In PPC, a model predictive controller is used to determine a vector of control signals that will be sent to the actuator. The vector, consisting of control signals, is also referred to as the prediction horizon. Those vectors are buffered, such that if a message is lost, there are still control signals available at the actuator, and the plant can be controlled more precisely.

To illustrate this, consider a remotely controlled car. Once the trajectory is determined, a controller can numerically determine a sequence of inputs (steering wheel angles and accelerations) that will make the car follow that trajectory. This sequence of inputs is then sent to the car, which keeps sending back to the controller updated information about its state. However, if a message with a sequence of inputs fails to be delivered, the car still maintains all the other control signals in the vector. Despite them being suboptimal (in the sense that they were computed with outdated sensory information about the car's state), they will still yield a trajectory that will go towards the intended goal.

In this context of PPC, we study two strategies. The first was proposed by [34] in the context of an AWGN channel and is extended by us to consider the effect of the wireless channel. It consists of optimizing the length of the prediction horizon together with the maximum allowed attempts, given that the controller discards the failed messages.

On the other hand, the second solution—proposed by us—works similarly but attempts to increase its chances of decoding failed messages by keeping them at the actuator and combining those using MRC for new attempts. The motivation behind this scheme is to try and leverage the amount of mutual information already sent to the buffer which was not decoded with success, such that the chances of decoding of the subsequent attempt can be increased.

Furthermore, both approaches are also compared with communication and control solutions, without joint design. The performance is measured in terms of consumed bandwidth for each controller/actuator link as a function of the distance, the maximum link latency, and the target reliability.

Our results show that for typical networked control systems parameters the joint design techniques can achieve far superior usage of spectrum, with even better performance for more stringent characteristics. Moreover, the first solution yields better results due to a higher number of attempts being allowed for the same length of the prediction horizon, even though it is somewhat simpler than the second one. This is an interesting conclusion of this chapter, showing how the age of information is not only relevant to the application itself, but also to the wireless communication protocol.

The contributions of this chapter are detailed as follows: we derive an expression for determining the minimum bandwidth required for operation considering the protocol proposed in [34] with the random effect of the wireless channel and propose our own protocol. The approach proposed by us combines messages at the buffer using MRC. In [34], the authors only considered an AWGN channel, which cannot be applied in practice. Here we have extended that to include the effect of the wireless channel by incorporating fast fading into the model. Furthermore, we compare both approaches, showing that using the more complex solution does not yield better results due to the limitations in the number of attempts. Moreover, we also illustrate how JD techniques can be superior to single design techniques when the channel conditions are poor. On the other hand, our simulation results show that when there is a stronger LOS component, non-JD techniques can outperform the JD ones, and moreover, our proposed protocol outperforms that of [34].

The analytical challenges introduced here, in comparison to the previous chapter, are regarding taking into account the application-specific characteristics, which affect the communication. To elaborate a bit further, we show how increasing the payload length in order to gain more error tolerance provides improvements in terms of required communication bandwidth. Our numerical simulations show that for typical WNCS parameters these improvements can result in up to 40% less bandwidth being required to reach the application QoS requirements when comparing the JD technique with using the solution proposed in the earlier chapter. This is due to the fact that the former does not leverage the nature of the control system application.

4.2 System Model

We consider a WNCS with a wireless link between the controller and the plant, such as the one presented in Fig. 4.1. Moreover, PPC is used such that a buffer stores a vector of K control signals, and if a message fails to be delivered a signal from the buffer is used. In PPC, the stability of the plant depends on the length of the prediction horizon, and a larger K means a more stable plant [95]. However, solving the finite horizon optimization becomes more complex as K increases, and furthermore in terms of communication sending longer prediction horizons is more demanding as the length of payloads increase.

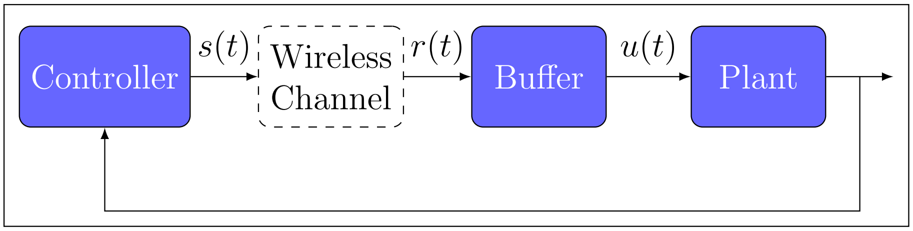


Figure 4.1: System model for a PPC solution using a wireless link between the controller and the actuators.

Regarding the wireless link, the model is the one presented in Chapter 2, where the message which arrives at the buffer is represented by (2.7). Both the controller and the plant are stationary, such that $h(t)$ remains constant over all symbols. In order to avoid deep fading, the transmitter switches to a different frequency channel at every new message, justifying the use of the block-fading model.

Furthermore, WNCS applications are strictly bound by a maximum latency λ, such that the plant can operate accordingly. If the next control signal is not available after λ seconds of the previous one, an error is declared. The plant can only tolerate errors with a rate less than ϵ_{c}, typically on the order of 10^{-9} [94].

Considering the log distance large scale path loss model presented in Chapter 2, the average SNR $\bar{\gamma}$ can be obtained by solving 3.3 for $\bar{\gamma}$, such that

$$\bar{\gamma} = \frac{P_{\mathrm{t}}}{W N_0 M_{\mathrm{l}} M_{\mathrm{c}} A_0 d^{\alpha}}. \tag{4.1}$$

4.3 Optimizing the Bandwidth

The goal of this chapter is to study joint design techniques to optimize resource utilization in a WNCS setting. Given that the control model for the proposed work considers a wireless link between the controller and the actuators, it is reasonable to shift the focus from energy efficiency to bandwidth optimization, as typically actuators are not energy-constrained. On the other hand, if we have a very optimized solution in terms of bandwidth, that directly translates into being able to utilize more actuators and controllers on the same plant, without the risk of collision in the messages being used by each of them.

In this context, we propose two approaches to joint design in the optics of PPC. The first considers optimizing the length of the control prediction horizon in conjunction with the number of allowed attempts before an outage is declared. Since each new forward message contains one extra control signal, the messages cannot be combined at the receiver. Thus, it makes more sense to send a more up-to-date control signal vector. This means that the error performance will be the same as the one for S-ARQ.

The second approach proposes a more elaborate protocol, which considers storing the old messages and performing MRC between attempts, in order to have a higher chance of decoding. This results in an error performance equal to that of CC-HARQ since we also make use of the feedback channel. In a generic application, such as what we studied in 3, the trade-off of using this approach would be a slightly more complex protocol to obtain a better error performance. However, in the case of WNCS with PPC, introducing this protocol means that the value of the allowed number of attempts z is bound by the length of the control signal vector. This is a direct effect of having to send less up-to-date messages, in order to allow MRC to be performed at the actuator.

4.3.1 Joint Design with New Messages

The notion of how PPC works is very well illustrated in Fig. 4.2, where the i^{th} plant state is represented by x_i, from which the control signal vector k_i is computed using a model predictive controller (MPC) and sent wirelessly to the buffer on message s_i, which then tries to decode the received message r_i. If it fails, the buffer can still use a previously successfully decoded control signal vector in order to send the correct instruction to the plant. For instance, in this example, this can be seen in the failure to decode r_2 and r_3. Since the buffer had enough control signals from decoding r_1 properly, it was still able to send the information from k_1 to the plant until it correctly decoded r_4. In this example, this is possible if we consider a value of $K > 2$. Conversely, if K was equal to 2, in the same example, an outage would have occurred when the receiver failed to decode r_3.

In other words, a constraint on the number of attempts z is imposed by the length of the prediction horizon. Because we must decode the message before we run out of control signals at the buffer, we must have

$$z < K. \tag{4.2}$$

As discussed earlier, the approach proposed here consists of applying PPC and optimizing the length of the prediction horizon together with the number of allowed failed attempts that still meet target latency and reliability requirements of the application, with the least bandwidth possible.

The URLLC requirements are incorporated by ensuring the plant has access to control signals at every sampling interval. This translates to ensuring messages are decoded before the buffer is empty.

Next, we write the optimization problem as

$$\underset{K \in \mathbb{N}^*, z \in \mathbb{N}^*}{\text{minimize}} \quad W(z) \tag{4.3a}$$

$$\text{subject to} \quad z < K, \tag{4.3b}$$

The problem in (4.3) can be solved numerically, as long as we have an expression for the minimum required bandwidth, which we derive next. Since we are considering a Nakagami-m channel, $P_{\text{out,z}}$, is well approximated at high SNR[1] by (2.16). In order to meet the application requirements, we must have

$$P_{\text{out,z}} \leq \epsilon_{\text{c}}. \tag{4.4}$$

[1] The approximation is done here so we can solve for W in closed-form.

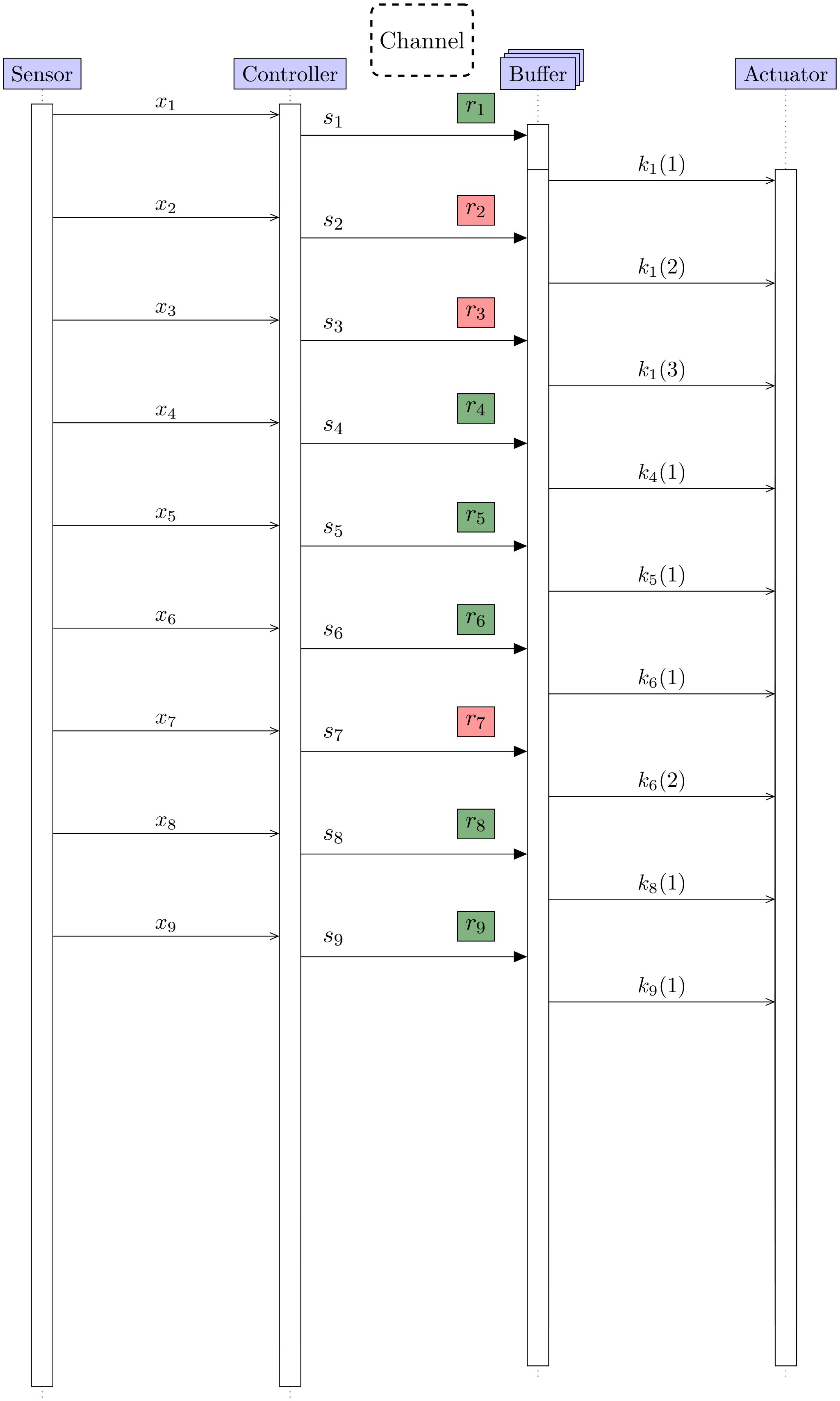


Figure 4.2: Illustration of PPC.

Thus, combining (4.1) and (2.16) in (4.4) we arrive at

$$W\left(2^{\frac{R_\mathrm{b}}{W}}-1\right) \leq \frac{\left(\epsilon_\mathrm{c}^{1/z} m!\right)^{1/m} P_\mathrm{t}}{m N_0 M_\mathrm{l} A_0 d^\alpha} = \beta_\mathrm{s}(z), \tag{4.5}$$

where $\beta_\mathrm{s}(z)$ is an auxiliary function.

To simplify the algebraic manipulations in this chapter, we use the rate in bits per second R_b instead of R. One can be easily obtained from the other by normalizing by the bandwidth.

Considering high spectral efficiency—which is a reasonible assumption considering that we typically have bandwidths on the orders of KHz—, such that $2^{R_\mathrm{b}/W} >> 1$, (4.5) can be solved for W as

$$W(z) \geq \frac{-R_\mathrm{b}\ln(2)}{W_{-1}\left(\frac{-R_\mathrm{b}\ln(2)}{\beta_\mathrm{s}(z)}\right)}, \tag{4.6}$$

where W_{-1} is the lower part of the main Lambert-W function [88].

To guarantee the target latency λ, R_b is obtained as

$$R_\mathrm{b} = \frac{L_\mathrm{H} + L_\mathrm{fb} + K L_\mathrm{u}}{\lambda}, \tag{4.7}$$

where L_H, L_fb and L_u are the lengths, in bits, of the header, the feedback signal and of one control signal, respectively. Note that, for simplicity, we consider that the bit rate for the feedback messages is the same as that of the forward messages[2]. Also note that, unlike in the derivations from Chapter 3, here each message must be exchanged in the space of λ seconds. This is possible because of the nature of the application, where the buffer at the plant stores the future signals, which can be used for reliability.

4.3.2 Joint Design with MRC

This approach is similar to the former, however, to increase the chances of decoding messages, we keep a copy of failed attempts at the receiver and combine those with the following messages using MRC at the receiver, as in CC-HARQ [17].

The approach is illustrated in Fig. 4.3, where x_i represents the i state of the plant, while s_i^j and r_i^j represent the sent and received j^th attempt for the i^th state, and $k_i(q)$ represents the q^th control signal from the vector generated for the i^th state. In the figure, a red background corresponds to a failed message and a green one to a successful one. Note how when the receiver fails to decode r_2^1, the next transmitter resends s_2, such that at the next sampling interval the receiver can perform MRC between r_2^1 and r_2^2. However,

[2] As is normally the case with commercial radios, which offer a pre-defined set of coding, rate, and modulation parameters.

despite having failed to decode r_2^1, the buffer still has k_1 stored in memory, such that it can send $k_1(2)$ to the actuator. After yet another failure, $k_1(3)$ is used, and when r_2^3 is combined with the two previous attempts, k_2 is now successfully decoded at the buffer, which means that now $k_2(3)$ can be sent to the plant. Upon receiving the ACK for s_2^3, the controller is free to send a new vector of signals, s_5^1, and the cycle continues.

Therefore, a new restriction on z is imposed, since now whenever we successfully decode a message after τ failed attempts we no longer obtain K useful control signals, but rather $K - \tau$. This imposes that $z \leq \lceil K/2 \rceil$. To prove this, let us assume that $z > \lceil K/2 \rceil$. After having failed $\tau > \lceil K/2 \rceil$ times, only $K - \tau$ control signals are not outdated and therefore the next packet has to be successfully decoded in $K - \tau$ attempts. However, by definition $K - \tau < \lceil K/2 \rceil$, thus it is not possible to guarantee the target error probability and maximum latency for $z > \lceil K/2 \rceil$.

Therefore, the optimization problem written as

$$\underset{K \in \mathbb{N}^*, z \in \mathbb{N}^*}{\text{minimize}} \quad W(z) \tag{4.8a}$$

$$\text{subject to} \quad z \leq \lceil K/2 \rceil. \tag{4.8b}$$

Limiting the number of attempts in this manner can be justified by an attempt in increasing the chances of decoding the message and indeed the probability of failing changes to (2.18), at high SNR.

Following the same steps as before, we arrive at

$$W(z) \geq \frac{-R_\text{b} \ln(2)}{W_{-1}\left(\frac{-R_\text{b} \ln(2)}{\beta_\text{cc}(z)}\right)}, \tag{4.9}$$

given that

$$\beta_\text{cc} = \frac{(\epsilon_\text{c} \Gamma(mz+1))^{1/mz} P_\text{t}}{m N_0 M_\text{l} A_0 d^\alpha}, \tag{4.10}$$

where β_cc is another auxiliary function and Γ is the complete gamma function.

4.4 Simulations

To evaluate the performance of the proposed methods, we have performed numerical simulations using parameters from typical WNCS applications, which are summarized in Table 4.1.

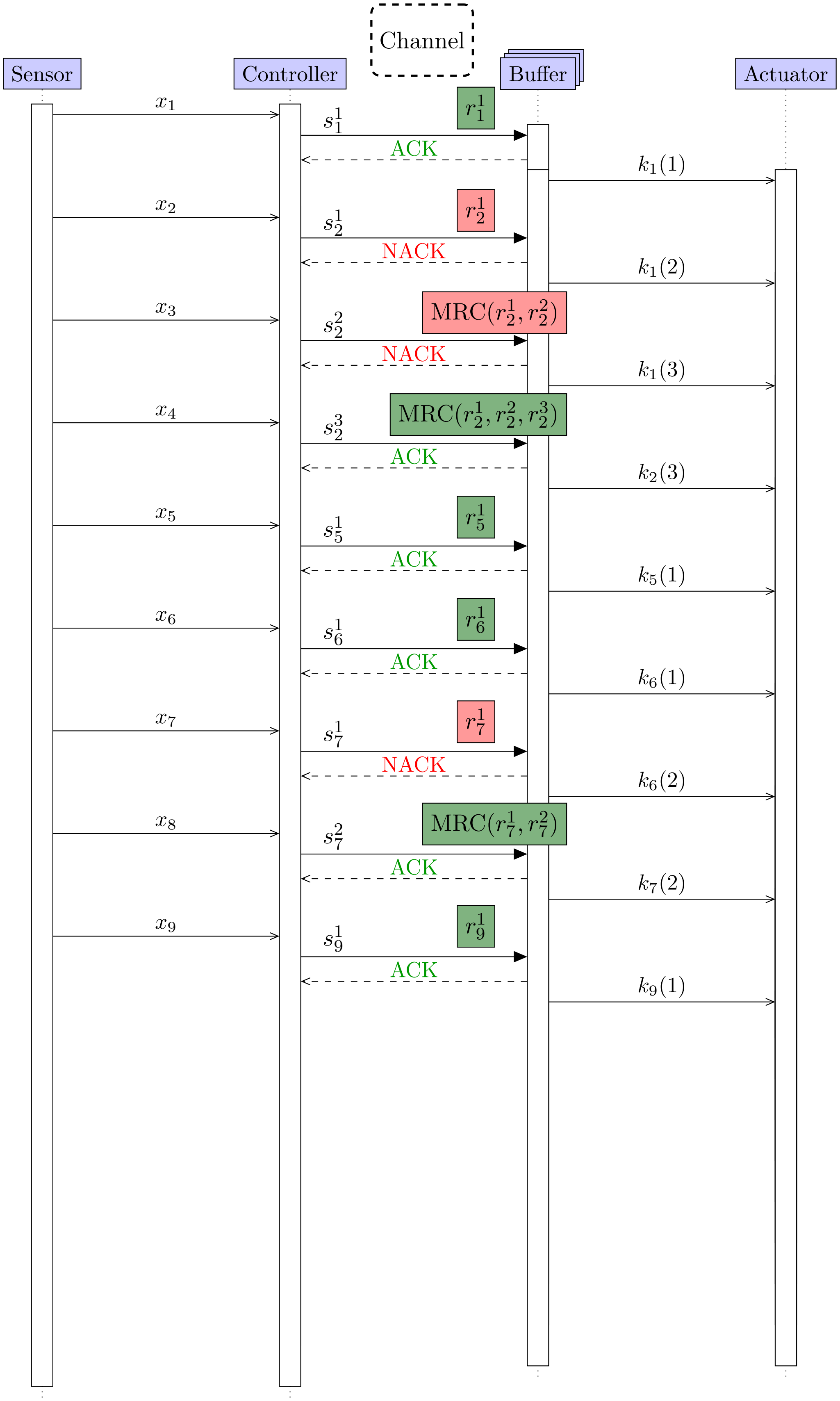


Figure 4.3: Illustration of the MRC approach to PPC.

Table 4.1: WNCS Simulation Parameters

Parameter	**Value**
Target Outage (ϵ_{c})	10^{-9} [94]
Link Margin (M_{l})	22 dB [†]
Coding Margin (M_{c})	3 dB [87]
Path Loss Exponent (α)	2 [96][‡]
Attenuation at Reference (A_0)	46.76 dB [96][††]
Noise Spectral Power Density (N_0)	-204.0 dBW/Hz
Transmit Power (P_{t})	-6.0 dBW [97]
Nakagami Parameter (m)	1 [96]
Maximum Latency (λ)	1.0 ms [94]
Control Signal Length (L_{u})	16 bits
Header Length (L_{h})	16 bits [98]
Feedback Length (L_{fb})	17 bits [‡‡]
Distance (d)	10 m [96]

[†] Noise figure at the device is 9 dB [96]. The extra 13 dB accounts for unforseen losses.

[‡] Experimental measurements show that free space path loss can be used for $d < 15$m [99].

[††] Value for a carrier frequency of 5.2GHz [96].

[‡‡] header plus 1 bit feedback.

4.4.1 Baseline Scenarios

The JD solutions presented in this chapter are further compared with 2 other techniques, one involving only optimizing control parameters and another which only considers the communication aspect.

Control Only Optimization

In this approach, z is set fixed at 2, and CC-HARQ is performed. The control optimization is performed by tunning K numerically in order to find the value which yields the smallest bandwidth which meets the application constraints in terms of latency and reliability.

Communication Only Optimization

Here, the approach proposed in Chapter 3 is used. That is, the rate is adjusted to make the CC-HARQ latency aware and meet the targets of the application. Since it consists of a communication only approach, no MPC is considered and therefore all messages must be delivered within the target latency and error probability. This causes the rate to be increased proportionally to z, such that more attempts can be made within λ seconds, as in Chapter 3. It also means that $K = 1$ is considered since no PPC is performed.

4.4.2 Comparison

Rayleigh Fading ($m = 1$)

One typical scenario of WNCSs is when there is no LOS between a wireless controller and the plant. In those cases, fading can be modelled as Rayleigh, which in other words translates to $m = 1$. This situation is what we study first, to compare all four approaches.

In Fig. 4.4, we plot the required bandwidth for various distances between the controller and the actuator. The proposed JD technique outperforms all other strategies regardless of the distance, followed closely by the JD technique using CC-HARQ. For a typical distance of 10 m, our approach requires around 60% of the bandwidth from the non-JD techniques and around 77% of the bandwidth required by the JD technique with CC-HARQ. Moreover, as the distances increase (resulting in larger path losses), the gap between both JD techniques and the control-only optimization increases. This is because the control-only solution does not have varying diversity, which suffers more in less favourable links. For instance, at 25 m the JD technique requires less than half the bandwidth when compared to a control-only strategy. On the other hand, the CC-HARQ only solution maintains its gap fairly constant as distances increase, as it can offer more orders of diversity in face of more stringent links.

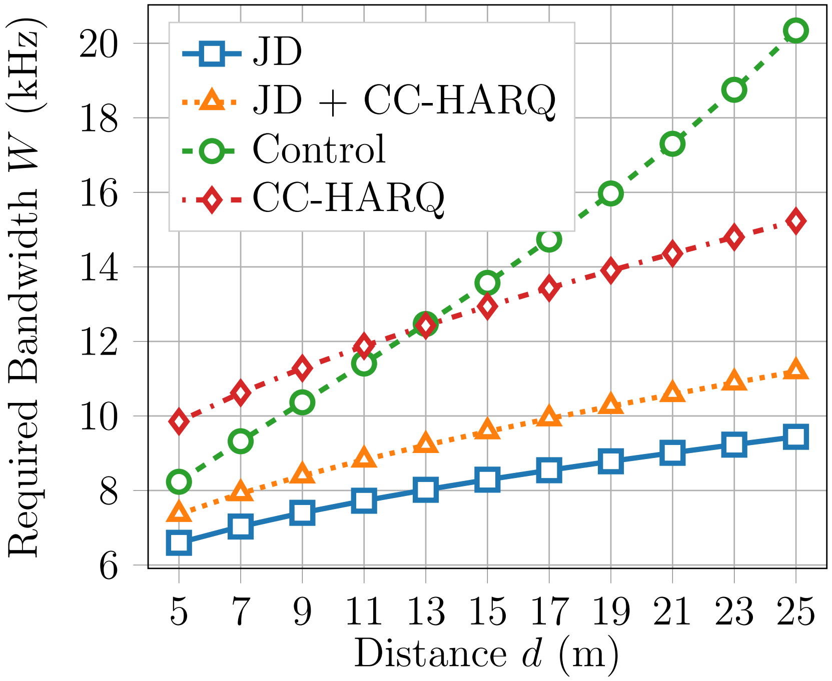


Figure 4.4: Bandwidth required for different distances considering both JD strategies, a control only optimization and a communication only approach (with CC-HARQ). λ is set at 1 ms and $\epsilon_c = 10^{-9}$. Here, $m = 1$.

Next, we analyze the effect of latency in the required bandwidth for the four strategies. As we can observe in Fig. 4.5, when λ is relaxed, all strategies perform similarly, requiring between 1.5 kHz and 2 kHz of bandwidth regardless of the protocol used. However, typical WNCS applications have requirements of latency that are much more stringent, often less than 1 ms. In those cases, we observe a similar trend as with the curve for the distance. The JD solutions outperform the non-JD ones and when λ is more stringent the gap increases considerably, *e.g.* at 0.5 ms. This is the case because in the JD case, there is more freedom to use different values of z and K, since the protocol is jointly designed. Furthermore, the solution without CC-HARQ is also always superior to the one with it.

Another relevant parameter to consider is the target outage probability. Figure 4.6 shows the effect of considering other values for ϵ_c. Making it more stringent has a similar effect as increasing the distance, the control only strategy becomes much worse than all others and the gap among the other three is maintained, with the JD approach being the best for all the range. On the other hand, relaxing ϵ_c to 10^{-7} causes the control only optimization to match the performance of the CC-HARQ one and the gap between both JD strategies becomes narrow. The communication-only solution remains consuming around 50% more bandwidth than the other approaches.

We have demonstrated how d, λ, and ϵ_c affect the performance of the system in terms of the required bandwidth to meet the requirements in each case. Regardless of the parameters, the JD solution with CC-HARQ yielded a poorer performance than the other

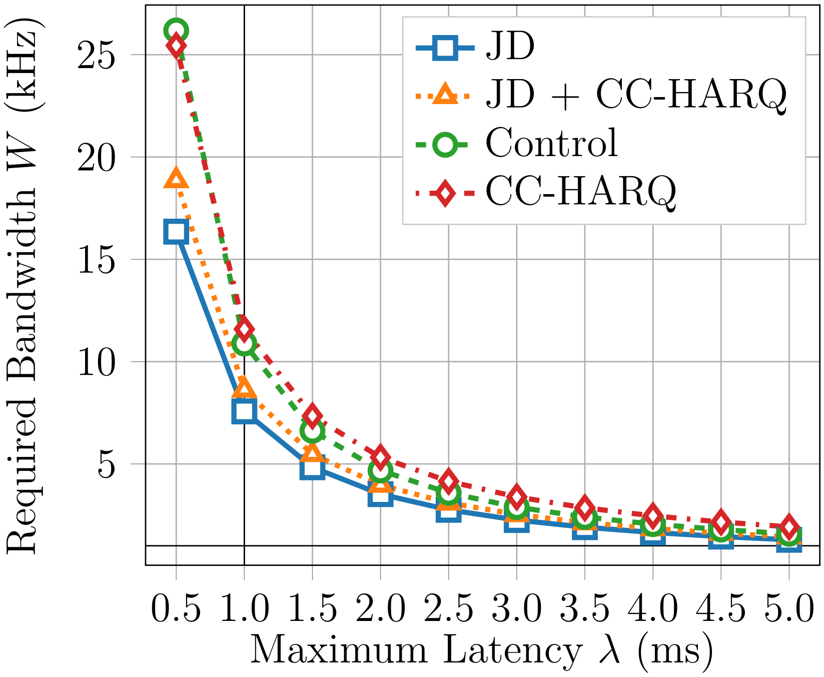


Figure 4.5: Bandwidth required for different values of latency considering both JD strategies, a control only optimization and a communication only approach (with CC-HARQ). d is set at 10 m and $\epsilon_c = 10^{-9}$. Here, $m = 1$.

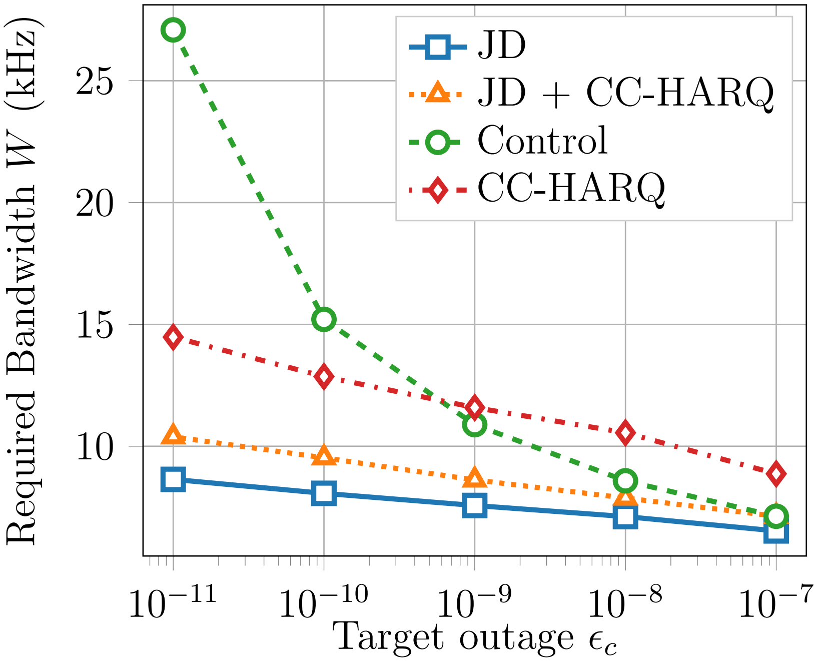
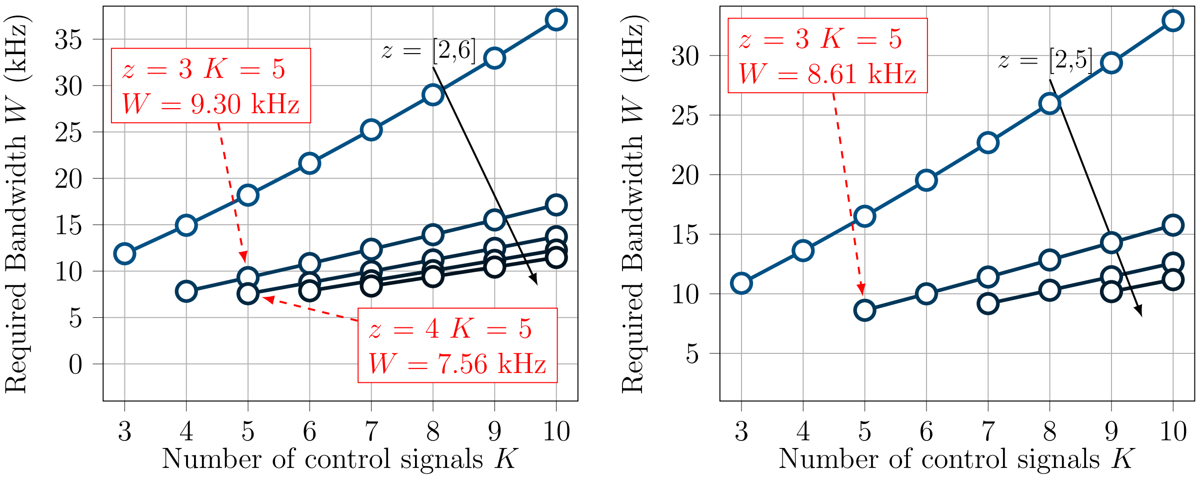


Figure 4.6: Bandwidth required for different values of reliability considering both JD strategies, a control only optimization and a communication only approach (with CC-HARQ). d is set at 10 m and $\lambda = 1$ ms. Here, $m = 1$.



(a) Joint Design.

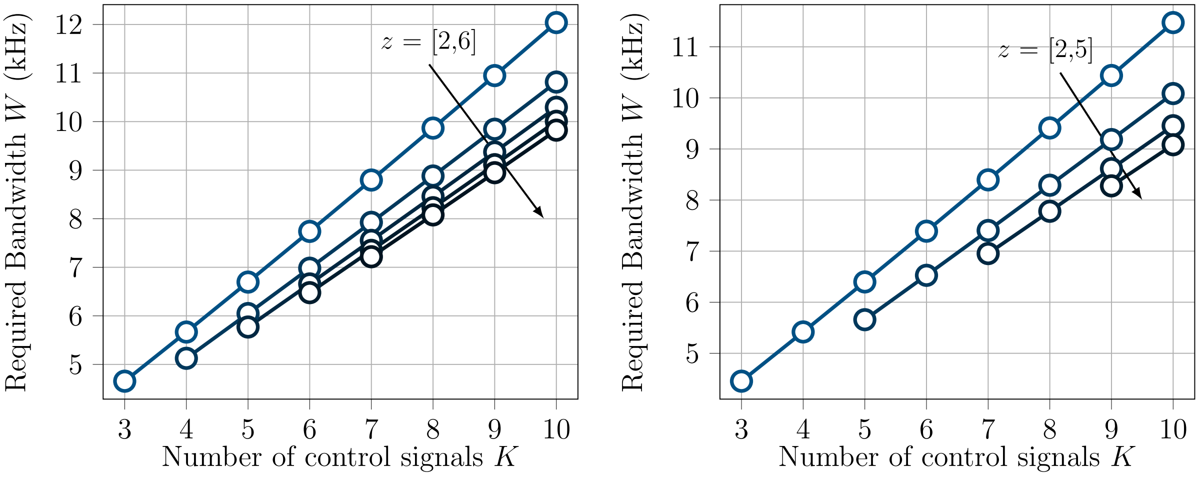
(b) Joint Design with CC-HARQ.

Figure 4.7: Minimum required bandwidth for each JD approach and various pairs of z and K, considering non-LOS ($m = 1$).

JD solution, which seems counterintuitive, as the CC-HARQ is more complex. However, this can be explained due to the allowed z in each case. The price to pay to perform MRC at the receiver is that when we fail and subsequently decode the message successfully in the following attempt, we decode K-1 useful control signals. If we fail 2 times and decode the message on the third attempt, we obtain K - 2 useful control signals, and so on. On the other hand, in the JD strategy, because newer signals are being sent, we always decode K useful control signals, which allows us to use a higher z for the same K and meet the latency and reliability requirements, yielding a more efficient use of the spectrum.

In order to dive deeper and showcase why it is the case that the more robust protocol does not yield the best performance, we have plotted in Fig. 4.7 the minimum required bandwidth for both schemes for a wide range of allowed (z, K pairs). Each line corresponds to a different value of z, growing in the direction indicated by the black arrow.

We can clearly see that for the same (z, K) pair, the CC-HARQ approach has better performance, as evidenced by the example of $z = 3$ and $K = 5$, whence the JD approach requires $B = 9.3$ kHz, while the CC-HARQ solution can guarantee the URLLC QoS requirements with only 8.61 kHz of bandwidth. However, when we consider $K = 5$, the JD strategy can allow up to 4 retransmissions, and with that, it can operate on 7.56 kHz. Meanwhile, the CC-HARQ approach is capped at $z = 3$ when $K = 5$, and thus it cannot outperform the former. So when we compare the absolute best that each protocol can achieve, the JD approach performs better because of its more relaxed constraint on z.



(a) Joint Design.

(b) Joint Design with CC-HARQ.

Figure 4.8: Minimum required bandwidth for each JD approach and various pairs of z and K. Here we consider that there is some LOS, and thus $m = 3$.

Introduction of a LOS component ($m = 3$)

Despite the Rayleigh scenario being a very representative case of WNCSs, there are also several situations where some LOS might be present in factory automation scenarios. Moreover, this might be an interesting considering when doing antenna placement, to facilitate the communication. Therefore, we dedicate this last sub-section to studying the case where there is some LOS between the controller and the plant. This is expressed mathematically in our model by considering $m = 3$.

Let's start by analyzing a similar figure to Fig. 4.7, which compares both the JD and the JD with CC-HARQ approaches. The comparison is presented in Fig. 4.8, where we show the required bandwidth for different pairs of z and K. Note how now that the channel conditions are better with some LOS, increasing K and z is no longer beneficial since the reliability and latency constraints will be met with a lower bandwidth even for a small K and z. For this particular example, this results in the optimal policy being that of $z = 2$ and $K = 3$, which is a pair allowed in both strategies. The result is that now the JD with CC-HARQ outperforms the simpler approach.

Next, we compare the JD techniques with the non-JD ones in the same LOS conditions. In Fig. 4.9 we can see that having better channel conditions, causes the diversity gains for a single shot transmission to be more efficient than using either of the JD techniques. Note how the CC-HARQ curve (without JD) outperforms all the other ones in terms of required bandwidth up until $d = 33$m, when the more challenging link budget makes it not the best alternative anymore. It is also interesting to note how the JD only technique is outperformed by the JD technique which uses CC-HARQ as well. This is what we predicted when analyzing Fig. 4.8. Lastly, note how the control only approach has identical

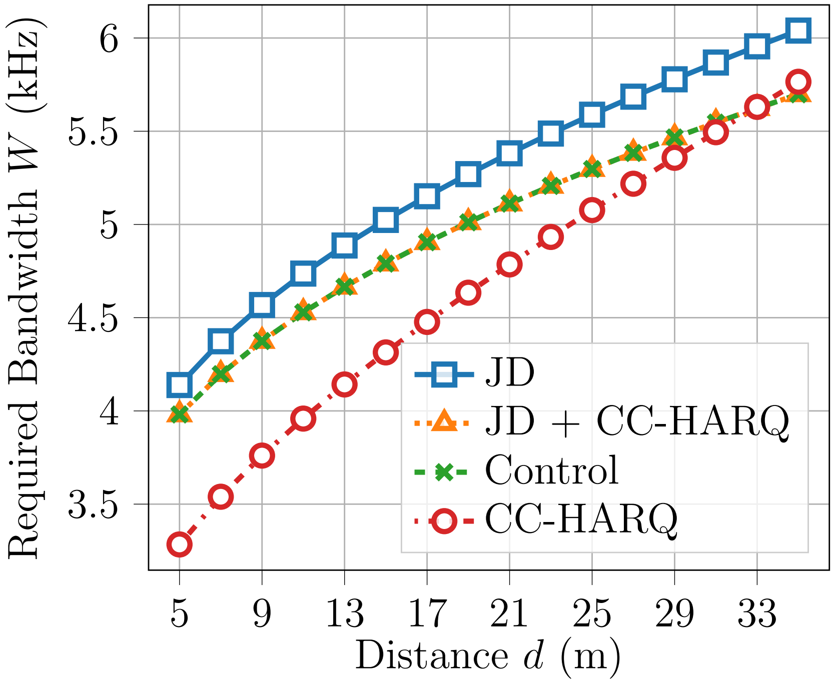


Figure 4.9: Bandwidth required for different distances considering both JD strategies, a control only optimization and a communication only approach (with CC-HARQ). λ is set at 1 ms and $\epsilon_c = 10^{-9}$. Here, $m = 3$.

performance to the one with JD with MRC. This is the case because, as explained before, for the control only approach we consider MRC with $z = 2$, and that happens to be the optimal approach.

In a similar analysis, we compare in Fig. 4.10 the required bandwidth for all the techniques considering different levels of reliability. Again, the same observations can be made. This time, however, the fact that the communication only approach is only beneficial when the constraints are easier to attain is more pronounced. This is evidenced by the sharp slope going down on the communication only strategy, which requires more bandwidth for T_{out} up to 10^{-11} and after that starts to outperform the other strategies.

These analysis considering better LOS conditions are very interesting to highlight that when evaluating the strategy to be used, the environment needs to be taken into account. If less challenging conditions are present, going for the approach presented in Chapter 3 is the better solution.

4.5 Conclusion

In the Introduction of this chapter, we asked the question: Can we improve the performance of the HARQ communication protocol by taking into consideration WNCS application-specific parameters?

The answer is yes, particularly when channel conditions are stringent. We were able to

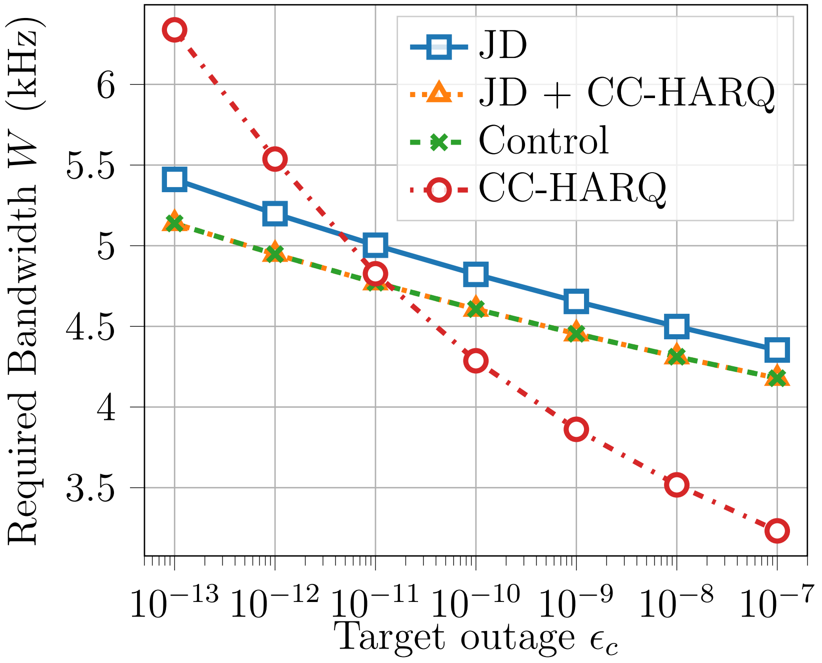


Figure 4.10: Bandwidth required for different values of reliability considering both JD strategies, a control only optimization and a communication only approach (with CC-HARQ). d is set at 10 m and $\lambda = 1$ ms. Here, $m = 3$.

show that taking into consideration WNCS application-specific aspects when designing the protocol can bring benefits in terms of resource utilization, more specifically bandwidth.

Another somewhat surprising result is that bringing CC-HARQ into the solution, which is a more complex protocol but that normally out-performs its simpler counterpart, is not beneficial in the case of PPC if both are designed with an awareness of the buffer at the receiver.

Moreover, it is clear from the results demonstrated in this chapter that for WNCS, joint design strategies greatly outperform naive approaches when stringent communication conditions are at play, which only consider communication or control parameters separately.

On the other hand, by evaluating the example numerically for less stringent channel conditions, we were able to show that when the constraints are more relaxed, going for a communication only approach can outperform the JD approaches in terms of required bandwidth. This last point highlights the strength of the approach proposed in Chapter 3.

Chapter 5

SON Enabled HARQ in URLLC

5.1 Introduction

Enabling URLLC is a complex task, as evidenced by the time it took for it to be included in mobile network specifications. Moreover, this is particularly challenging when considering that 5G must be more efficient than previous iterations of mobile networks in terms of resource utilization (be that bandwidth, monetary, or energy consumption) in order for our society to move forward sustainably, particularly in the face of the scale and ubiquity with which mobile networks have been deployed up until now, and more so are projected to be.

As we showed in Chapter 3, exploring time diversity—in the form of HARQ—can be an interesting approach to fulfil those latency and reliability requirements at the same time, while being resource-efficient.

On the other hand, for those results in [59] to be possible in practical deployments, we still must satisfy some assumptions. Notably, a short message decoding time is assumed, scheduling problems with the repeated message is not taking into account and the effect of an imperfect feedback channel is dismissed.

In this chapter, we will show how can those assumptions be relaxed in practice by leveraging the use of SON. This is achieved by using ML algorithms that can predict the quality of the channel beforehand and determine when retransmissions will be needed. The approach proposed here tackles all those issues. Firstly, it eliminates the need for the transmitter to wait for the feedback message before deciding to send a retransmission, which takes care of the imperfect feedback channel and the time for the receiver to decode the messages and request a feedback. Secondly, it provides the network with ahead of time predictions of when retransmissions will be needed, such that it can provision the required resources accordingly.

In the context of this chapter, the network is learning how to predict the quality of the channel. However, in order for such an algorithm to be designed, data must be available, and this is a fundamental issue in SON: obtaining real-world telecommunications data.

In order to circumvent this issue, a few strategies might be employed. For instance, researchers may team up with telecommunications companies and form partnerships which provide the access to datasets. Alternatively, researchers might assemble testbeds and generate a synthetic data-set based on experiments. Both approaches suffer from scalability problems, as they require the research groups to be established and with access to larger amounts of funding. A third, and attractive, option is to create a digital twin [100] of the desired scenario. It is much more than a simple simulation, it consists of a model which closely mimics the scenario of interest [100] and allows us to obtain huge amounts of data without the issues of the other mentioned approaches. The reason this goes further than a typical system-level simulation is because it incorporates real-world information which is not easily mimicked in the latter. For instance, instead of a king-walk mobility pattern and a Manhattan grid city model, a digital twin typically incorporates the actual topology of the city in question and real-world mobility paths.

This last alternative is what we chose to do, and we present our approach in this chapter. We implement a mobile network digital twin of the city of Glasgow is implemented and a comprehensive data-set is generated. Alongside the cellular network, users driving vehicles are deployed in the digital twin scenario and they follow real-world paths along the simulated area while maintaining wireless communication with the base stations. A large amount of communication and localization data is recorded while the digital twin is being run. We create this framework by combining network simulator 3 (NS3) [101], localization application programming interfaces (APIs), and the publicly available telecommunications data from cell-mapper[1]. The dataset is very close to real-world scenarios, as NS3 deploys the entire LTE stack, and the scenario itself contains real-world user trajectories and actual BS parameters (placement, DL frequency, antenna orientation, etc).

In order to showcase its usefulness, we then process the data-set and use it to tackle the aforementioned problem of predicting the channel quality. That, in turn, is used to enable the strategy presented in Chapter 3: HARQ enabling efficient URLLC. This shows how HARQ can be used in practical scenarios and still provide important resource savings while guaranteeing URLLC performance.

Therefore, the contributions of this chapter are three-fold and can be summarized as follows:

- A digital twin implementation of the city of Glasgow running a mobile network with real base station parameters and accurate routes for vehicular users connected to the network and running an application.
- The dataset obtained from running this digital twin, which can be freely accessed

[1] Obtained from www.cellmapper.net.

and used to derive novel results.

- The algorithm which uses the generated dataset to enable URLLC via HARQ and yield important resource savings while guaranteeing stringent latency and reliability constraints.

In summary, this chapter outlines a strategy to enable URLLC with a good resource utilization which can be deployed in realistic scenarios, effectively demonstrating how to realize the gains presented in Chapter 3. This is achieved by leveraging the power of SON, a powerful new paradigm that is being deployed within 5G. Moreover, a digital twin of an actual mobile deployment for the city of Glasgow, in Scotland, is presented to showcase how can SON algorithms be devised even in the absence of real-world data. Next, the generated dataset is used to train an LSTM CNN which is used to predict signal quality at the receiver. Those predictions are in turn used to enable HARQ, which delivers excellent resource utilization, while still meeting the stringent URLLC constraints, thanks to an algorithm that uses the predicted signal quality to determine the need for retransmission before a feedback signal is received.

5.2 Digital Twin

As mentioned, in this chapter we propose a digital twin to simulate a mobile network in the city of Glasgow with users moving and accessing network services. In this section, we describe how this was conceived to be as closely related to reality as possible.

5.2.1 NS3

The system is created using NS3, an evolution of the popular network simulator 2 designed to be as close to reality as possible [101], which makes the software an ideal candidate for a digital twin. More specifically, NS3 is a fully-fledged network simulator programmed in C++ which provides the ability to simulate several layers of the communication stack with a very high level of customization. Moreover, when it comes to the users, it has a wide array of mobility patterns that can be easily extended. Furthermore, it already provides several communication standards out of the box. For instance, its LENA module implements all the LTE stack and provides 3$^{\text{rd}}$ generation partnership project (3GPP) channel models for the simulation, while its LENA-5G module offers 5G capabilities. These, of course, can be programmatically extended to produce any desired algorithm. In the digital twin designed in this work, LENA was used, since 5G deployments are still somewhat in their infancy at the time of the creation of the simulations, so the coverage would not be ideal for a large-scale simulation. This, however, does not implicate in loss of generality on the

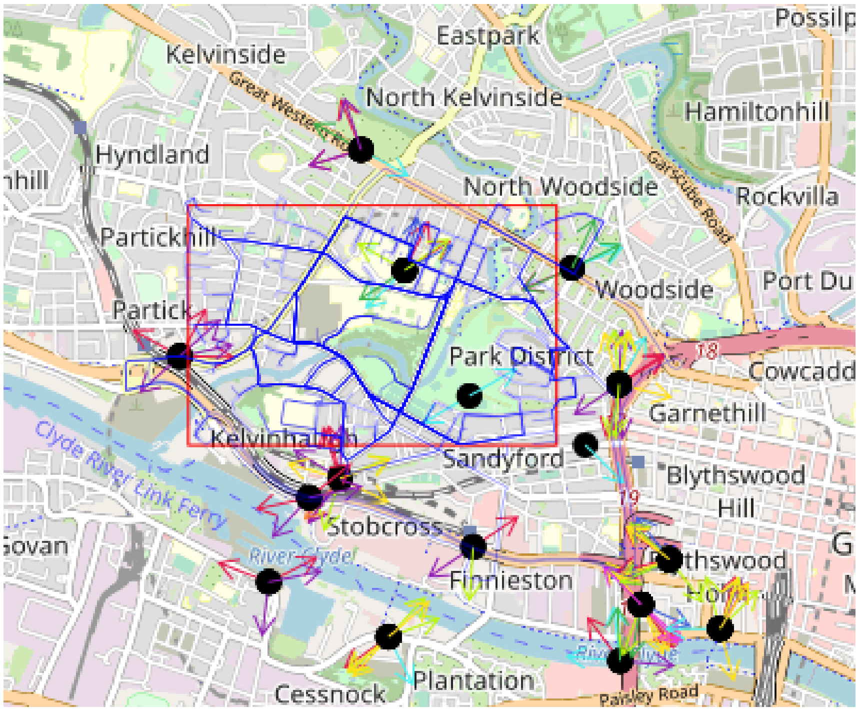


Figure 5.1: eNBs and cells around Glasgow on the position of the digital twin. The arrows indicate the orientation of the antennae, while their colour indicates a particular downlink frequency and are included to show highlight how complete the digital twin is with respect to reality.

results here obtained, since they consist of using the obtained measured data after the fact, effectively extending the LTE capabilities of the network in a SON manner.

5.2.2 Mobile Network

In order to make the digital twin proposed in this work as close to reality as possible, Evolved Node B (eNB) parameters from a real-world deployment in Glasgow were used. The data input into the model consists of the parameters from 17 different eNBs [2] (in a total of 115 cells) for one provider in a square area in the city of Glasgow. This is illustrated in Figure 5.1, which contains a representation of the eNB and the cells used in the simulation, overlaid on the map of the part of Glasgow that is being simulated. Each eNB has some cells and the arrows indicate the orientation of each cell, while the colour of the arrows indicates unique downlink frequencies. The data reflect real-world measurements obtained from www.cellmapper.net, and the cells used in this work are all of the cells from one provider.

For the reader's benefit, we have included in Table 5.1 with the information from the eNBs. It shows, by eNB location, the cells which are a part of the eNB in a list of frequencies (in MHz) and orientation (degrees with respect to North).

[2] All the eNBs from one provider which was in the area of interest.

Table 5.1: Information about the eNBs

Location	**Cell Info (frequency MHz, Orientation)**
(55.869403, -4.306505)	(1846.7, 316°), (1846.7, 164°), (1846.7, 254°), (1861.1, 303°), (1861.1, 93°), (1861.1, 273°)
(55.878740, -4.291900)	(1851.7, 344°), (1846.7, 118°), (1846.7, 203°), (1866.1, 164°), (1861.1, 118°), (1861.1, 207°)
(55.873287, -4.288382)	(1846.7, 282°), (1851.7, 91°), (1846.7, 251°), (1863.8, 287°), (1866.1, 102°), (1863.8, 239°), (2680, 165°), (2680, 280°)
(55.867609, -4.283159)	(1851.7, 105°), (1851.7, 285°)
(55.873389, -4.274889)	(1851.7, 293°), (1846.7, 112°), (1851.7, 289°), (1866.1, 289°), (1866.1, 111°), (1866.1, 250°)
(55.863864, -4.293605)	(1846.7, 81°), (1846.7, 213°), (1846.7, 207°), (1861.1, 80°), (1861.1, 167°), (1861.1, 227°), (798.5, 86°), (2680, 62°), (2680, 292°), (2662.9, 341°), (2662.9, 151°), (2662.9, 294°)
(55.863005, -4.296087)	(1846.7, 123°), (1846.7, 276°), (1861.1, 88°), (1861.1, 303°)
(55.865389, -4.273814)	(1851.7, 0°)
(55.868134, -4.271110)	(1846.7, 59°), (1846.7, 283°), (1851.7, W (254°), (1861.1, 41°), (1861.1, 228°), (2680, 55°), (2680 , SW (234°), (2680, 220°), (2662.9, 329°), (2662.9, 236°), (2662.9, 209°)
(55.868134, -4.271110)	(1846.7, 281°), (1861.1, 101°), (2680, 49°)
(55.859192, -4.299337)	(1846.7, 156°), (1846.7, 288°), (1861.1, 52°), (1861.1, 303°)
(55.860782, -4.282858)	(1846.7, 281°), (1861.1, 101°), (2680, 49°)
(55.856712, -4.289681)	(1851.7, 18°), (1846.7, 100°), (2680, 295°), (2680, 93°), (2680, 268°), (2662.9, 87°), (2662.9, 95°), (2662.9, 274°)
(55.855628, -4.271013)	(1851.7, 140°), (1846.7, 188°), (1851.7, 218°), (1861.1, 320°), (1866.1, 222°)
(55.858173, -4.269335)	(1846.7, 5°), (1846.7, 2°), (1846.7, 8°), (1866.1, 55°), (1863.8, 339°), (2680, 3°), (2680, 65°), (2680, 184°), (2662.9, 21°), (2662.9, 14°), (2662.9, 186°), (2162.2, 2°), (2162.2, 356°), (2162.2, 181°)
(55.860214, -4.267042)	(1846.7, 170°), (1846.7, 122°), (1846.7 , S (178°), (1863.8, 187°), (1866.1, 109°), (1863.8, 179°), (2680, 7°), (2680, 110°), (2680, 178°), (2662.9, 172°), (2662.9, 116°), (2662.9, 173°)
(55.857083, -4.262965)	(1851.7, 135°), (1846.7, 241°), (1846.7, 256°), (1861.1, 248°), (2680, 210°), (2680, 246°), (2680, 270°), (2662.9, 30°), (2662.9, 244°), (2662.9, 257°)

On these eNBs, the entire LTE stack has been installed. This means that users moving through the scenario imitate with a high degree of similarity users moving in that part of the city of Glasgow.

5.2.3 Users

The strength of the digital twin proposed in this work is the ability it has to mimic a real-world scenario with a high resemblance. With that in mind, 100 vehicles are considered as the users. For the trajectory of the users to be as realistic as possible, the Google Maps API was used to generate trajectories. The trajectories of the vehicles can also be observed in Fig. 5.1, illustrated by the blue lines which follow the streets on the map.

Moreover, the 100 users were given each a start and an endpoint inside the depicted red square. Then trajectories are then converted into waypoints inside NS3 using Route Mobility Model [102]. These waypoints are then used in the simulation using the traffic speeds obtained from the API in order to determine the velocity of each user moving from one waypoint to the next. This way the vehicles will move at realistic speeds for the city.

On top of that, for testing purposes, each user is running an application that is attempting to download one packet per second.

5.2.4 Channel Model

The path loss model considered in the digital in this Chapter is the well-established and empirically validated COST231 [73], which has been presented in Chapter 2 and is characterized by (2.2).

On top of this, small-scale fading is also taken into account. For that, excess tap delays and relative power of taps are considered according to 3GPP technical specification (TS) 36.104 Annex B.2. An average user speed of 60 km/h is used, and several channel realizations are pre-computed using the extended vehicular A (EVA) model. Those values are stored and used by the digital twin to determine the instantaneous SINR for each channel realization of the simulation, in addition to the path loss obtained by using (2.2).

5.2.5 Collected Metrics

The simulation was run for the 100 users for 100 seconds and several traces were obtained. All the results can be fully accessed at [103]. The traces are the ones generated by LENA from within NS3, and they are presented in their documentation [104]. For the reader's benefit, the available data is outlined next.

The following traces are available, both for downlink (DL) as well as for uplink (UL). Radio link control (RLC), packet data convergence protocol (PDCP), MAC, and physical

(PHY) layer information. Particularly regarding the PHY traces, they are divided into the following:

- DL reference signal received power (RSRP) and SINR;
- UL SINR for the user equipment (UE);
- Interference values per resource block;
- UL and DL transmission data;
- UL and DL Reception data;

All the available fields for each trace can be viewed at [103]. Notably, the SINR traces are generated per cell/UE pair and per millisecond, while MAC, RLC, and PDCP traces are—naturally—only available for cell/UE pairs which constitute an allocation pair (that is, a UE which has been allocated to that cell), and are available with a different frequency. These factors are expected due to the nature of the information being collected. Depending on the type of application, adjustments must be made to make the data useful for the desired purpose, as is the case with this work. These steps are outlined next.

5.3 Processing the Data

Up to this point, this chapter's contribution has been regarding the generation and availability of the dataset. Starting from this section, said contributions are regarding utilizing the traces obtained from the digital twin towards URLLC, thus outlining the other contributions of this work.

The first task required to utilizing the traces towards any meaningful application is to select which data is of relevance for the desired outcome and to process this data such that it can be used to achieve its goal.

We propose to enable efficient URLLC by allowing HARQ to be performed. Moreover, we use the data to predict channel conditions at a given time, such that the probability of failure can be determined a priori, and retransmissions can be triggered without the need for the cell to receive a NACK from the UE.

To that end, first, the data must be sorted, then processed, and lastly fed into the proposed ML algorithm which yields useful information that can be used to operate the proposed protocol.

5.3.1 Extracting Meaning from the Data

Since the objective is to predict the probability of failure from the channel quality, the best possible metric is to predict the instantaneous SINR. If that could be achieved, it would be possible to determine with an extreme degree of certainty whether or not a packet would fail.

However, this is an extremely daunting task, due to the nature of the wireless channel. This is particularly difficult considering that there is a high level of mobility in this type of application, making the channel variations too random to be predicted with meaningful accuracy. Thus, the next best thing is to predict the average SINR for a given window and to use well-established PHY layer knowledge in order to determine probabilities of failure.

Moreover, knowing the average SINR with a good degree of accuracy allows for a very efficient use of the strategies proposed in [59], which showed very interesting gains in terms of energy used when comparing a direct transmission with frequency diversity versus CC-HARQ for URLLC.

As with any supervised learning model, predictions are made on top of existing datasets which contain the desired quantities to be predicted [65]. Since we wish to predict the average SINR for UE/cell allocation pairs, the MAC traces in conjunction with the SINR traces are leveraged to obtain values of instantaneous SINR per UE/cell pair only for the time instants[3] where there is a link between the two.

Fig. 5.2 illustrates the result of the process described above for three pairs of cell/UE. Each line represents the instantaneous DL SINR for each instant of the simulation for each pair of cell/UE. The first and second numbers in the legend indicate the international mobile subscriber identities (IMSIs) and cell identifiers (IDs), respectively, for each signal taken from [103]. It is clear that each vehicle is moving on a different trajectory around the area, which results in a different behaviour with respect to its DL SINR. However, despite that being possible to perceive by looking at the trends of the lines, there is still a substantial amount of randomness in the values, which is to be expected since they show the instantaneous SINR.

5.3.2 Matching Sampling Interval

Since the signals are obtained whenever there is communication between the vehicles, they need to be processed before they can be fed into any prediction algorithm. Notably, most algorithms do not work with an explicit notion of time, instead, the time information is encoded into the sampling interval, and samples are compared to one another. This is standard practice not just in ML algorithms, but also in any digital signal processing strategy.

On the other hand, the sampled data for each cell/user pair does not have a constant

[3] Due to how the traces are saved from NS3, the SINR time instants always have six digits, while the MAC traces are truncated at three decimal points (it records milliseconds). Thus, the SINR time instants are truncated at milliseconds in order to be able to match the two logs.

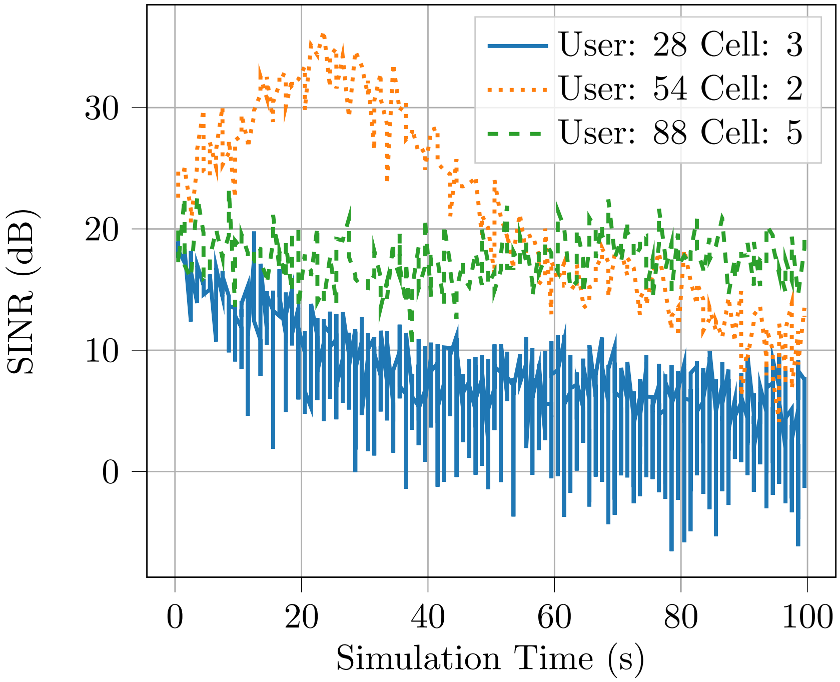


Figure 5.2: Instantaneous SINR for 3 pairs of Cell/UE.

sampling interval. This is mainly due to instantaneous characteristics of the network (such as MAC strategies, that is, if a resource is busy, the user will have to wait for a free resource block in order to receive its data packet). To illustrate this in the data, Fig. 5.3 shows the first 50 samples of the signals portrayed in Fig. 5.2. Note how for the pair (28,3) the 50 samples span over less than 10 seconds, while for the other two pairs it takes 25 seconds for them to be acquired.

Thus, to be able to apply a prediction algorithm on the input signals, the sampling instants must be matched. Since the smallest sampling period on the data is 1 ms (which comes from LTE's numerology), the signals on this work have their sampling intervals converted to 1 ms. Fig. 5.4 shows what the same signals look like after processing such that their sampling intervals are matched at 1 ms. The strategy to match the sampling intervals consists of performing a linear interpolation between the samples, thus ensuring the reconstructed signal matches reality as best as possible.

After the sampling interval is matched, the only operation that must be performed on the data is to compute the average SINR. This is done by computing a moving average filter, considering 100 samples, over the regularized signals. The result for the 3 signals in question is displayed in Fig. 5.5. Note how the sharp edges of the signal for the pair (28, 3) get smoother on the averaged signal. Also, note that the other two signals look very similar after the averaging. This is due to the fact that, for this particular interval, they are varying quite slowly compared to the length of the filter.

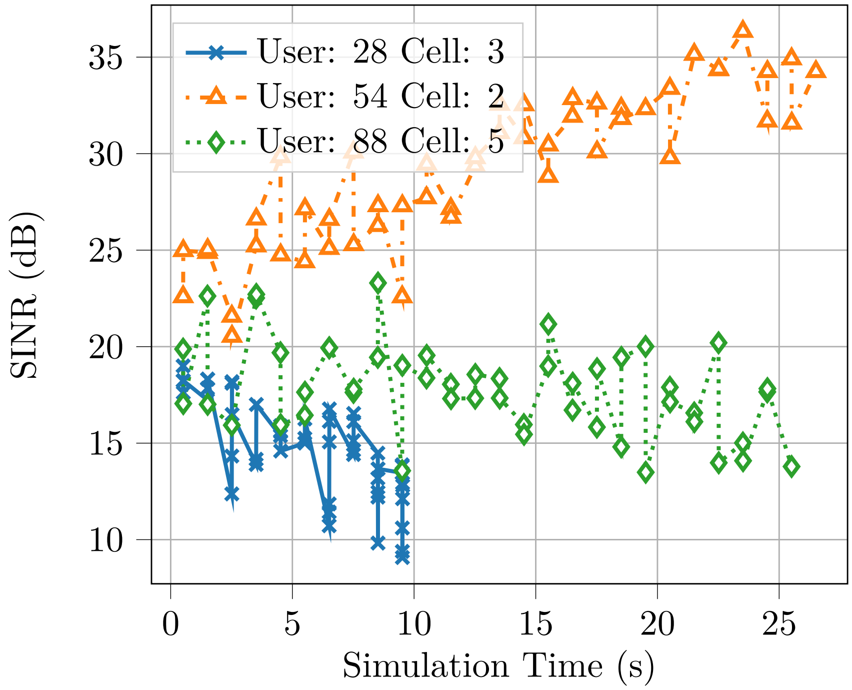


Figure 5.3: First 50 samples of the original signals.

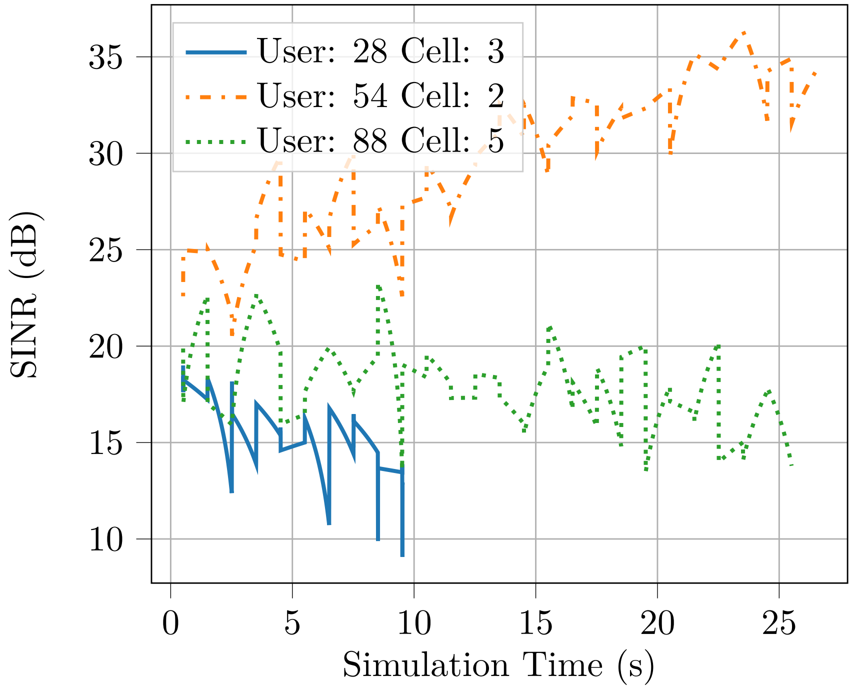


Figure 5.4: Signals with normalized sampling interval to 1 ms.

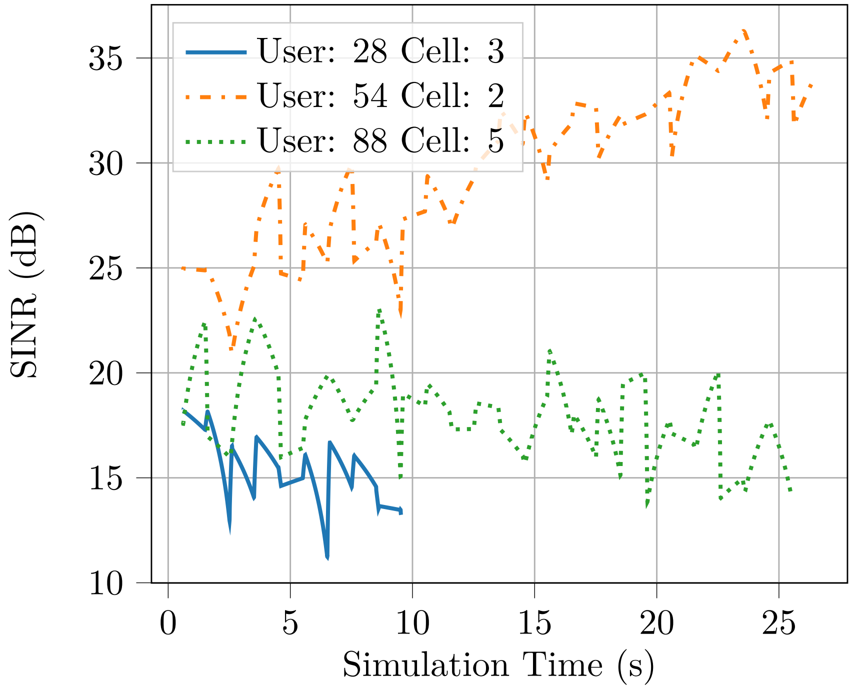


Figure 5.5: Averaged signals using a 100 samples moving average filter.

5.4 LSTM

One of the most suitable ML tools for working with time series and trying to predict future values based on past inputs is an LSTM recurrent neural network. That is because it operates on the signal in question, without having information about future values of the series, and finds trends in the data which help predict with good accuracy what is the next value.

It is capable of achieving that, as it finds hidden patterns encoded in the time series'. This is realized by training the network on known inputs, and then using the set of learned weights to predict values of other series. In other words, if two series have enough behavioral similarities, it is possible to use an LSTM to predict the behaviour of a completely unseen signal by exploiting the environmental and systemic familiarities.

Furthermore, users of a cellular network—particularly of a similar application—inherently possess a similar pattern in the way their signals behave. That is because they are all inserted in the same medium, and have quite similar mobility characteristics. This will be the case, even more, when considering users with similar radios and antenna placement. Therefore, it is possible to exploit these similarities and use an LSTM to predict information about the user SINR.

Critically, it is possible to train the network on one user and exploit the learned traits to predict the SINR level of several different users. In this work, such a model is proposed. The LSTM network proposed here has 4 units and uses the past 2 samples as look-back, in other words, it looks at the previous two values of SINR in order to predict the next. We

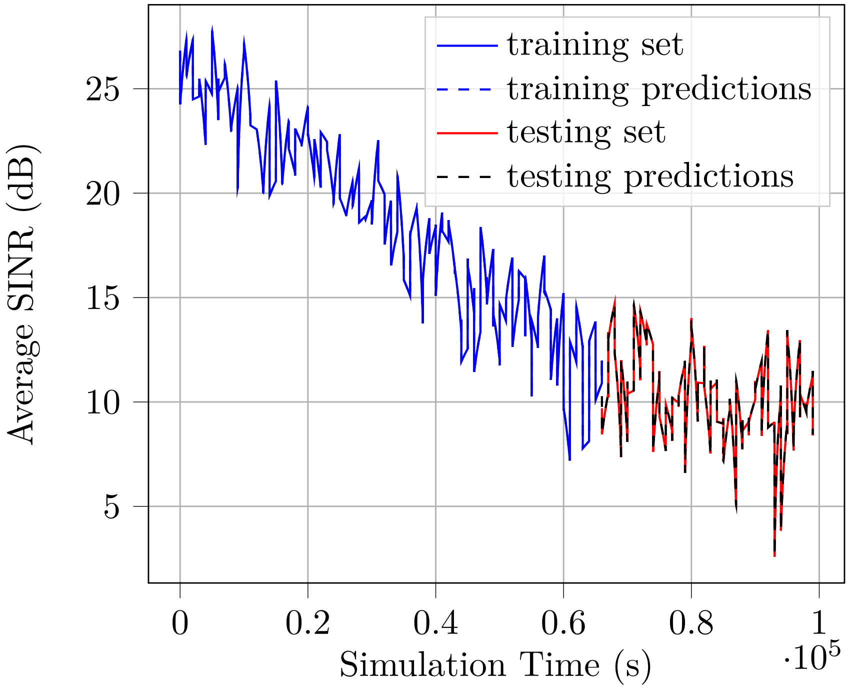


Figure 5.6: Training of the LSTM recurrent neural network, using two thirds of the signal to train. The remaining one third is used to test the learned behavior.

chose 4 units as it is a good trade-off in terms of the number of possible knobs to tweak in order to represent complex functions while at the same time not incurring too much training overhead. Moreover, we selected 2 samples as look-back such that the model can work on short term trends, which is typically how SINR behaves. The error function is the mean squared error and the optimizer is the Adam optimizer[4]. Furthermore, to enhance the predictions, the signals are scaled between -1 and 1 to compute the predictions. This is an inexpensive mathematical operation that can be easily undone to use the obtained value.

Fig. 5.6 shows the training of the LSTM. It learns for two-thirds of the time how the signal behaves, and then applies the learned values to the remainder of the signal. It is evident that the learned weights are able to predict what is the SINR behaviour for that given user. Further, on 5.6, the predictions of completely unseen data, of other users, are presented.

5.5 E-HARQ Protocol

One of the most efficient techniques in terms of resource utilization is HARQ [59]. It is a well-established technique that has been studied extensively in traditional network

[4] The Adam optimizer has been empirically shown to produce better results when compared to other adaptive gradient descent methods [105].

architectures, and it is present in traditional LTE networks. However, when considering a URLLC application, it is not intuitive that the gains in terms of resource efficiency can still be realized. This is particularly the case because several of the assumptions made for traditional types of communication do not hold in stringent low latency and high-reliability scenarios.

A few aspects of the novel mode of communication are especially troublesome. For instance, the time to decode the message—particularly when using the well-established turbo codes—is non-negligible, and may cause the link budget to be unachievable for a given application (for instance, if the time to decode the message is on the order of a few milliseconds, that can eat up all the latency available, as this needs to be performed at every attempt). Furthermore, MAC constraints must be taken into account. That is, when a message fails to be decoded, the transmitter must request another resource block, which might not be readily available. This, in turn, results in one of two scenarios. Either the network must always keep some resource blocks available for when messages fail on one of its links, or the latency requirement will not be satisfied alongside the reliability requirement.

Both of these issues can be mitigated by the strategy proposed in this work, which aims to enable HARQ within the context of URLLC applications. We established in Chapter 3 that HARQ can be extremely attractive to save network resources, provided one can get around those aforementioned concerns. The strategy proposed here is to leverage the predicted values of average SINR and use error rate curves in order to determine when a given message will fail with high probability (defined according to the application). When that is the case, the transmitter does not wait for the feedback message, and instead sends the next attempt immediately. This is one implementation of E-HARQ, first hinted by us in [66].

Notably, it will remain doing that for every attempt. This is made possible by the intelligence which is encoded into the LSTM network, which is tuned by previous users of the network. At this point, one might ask, what is the difference of this strategy when compared to traditional power allocation or diversity techniques? That is, one could instead of communicating through HARQ leverage the obtained values of predicted average SINR and increase the transmit power in order to guarantee a successful delivery. Alternatively, the desired diversity could be achieved by allocating several frequency channels in parallel and using those for the transmission, when the predicted average SINR is low.

The answer to this is that, even in this case, there are advantages in using HARQ, as proposed here. This rests on the fact that, despite the transmitter not waiting for the NACK in order to send the next attempt, the receiver might still decode the first attempt successfully. When that is the case, it will simply discard any further attempts and send the ACK back. The transmitter, by its turn, will receive this feedback signal and stop

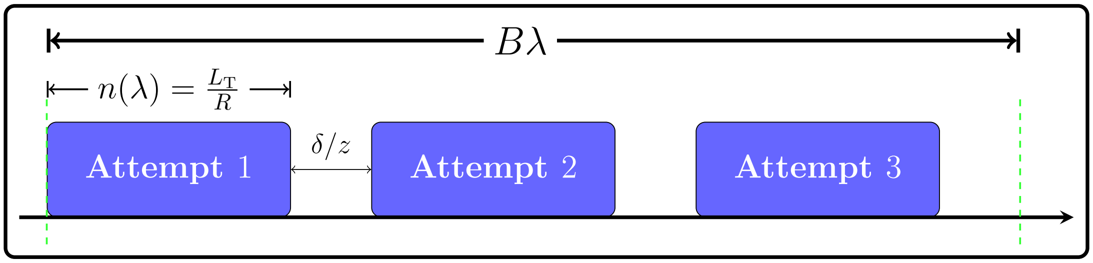


Figure 5.7: Example of how using R obtained from (5.1) makes it possible to meet the target latency. Note that here $z = 3$ and $n(\lambda) = (\lambda - \delta)B/z$, such that in total δ seconds are being used to decode the messages.

transmitting new attempts. Indeed, there are still some wasted resources in the useless message which was transmitted, however, it is still possible to achieve full diversity at a lower cost.

Still, on this point, consider the following example. Assume that the transmission time of a forward message is 2ms and that the receiver takes 0.5ms to decode the message. Consider a 10.5ms latency budget. If the transmitter is not waiting for the receiver to send the feedback message, it is possible to achieve diversity 5 using HARQ, since each attempt will take one-fifth of the latency budget. Now, assume that the receiver can correctly decode the message at the third attempt, thus the entire message takes 6.5ms to be transmitted. However, it is key to understand that the diversity order is still maintained at 5. Conversely, if a frequency diversity strategy is used, the transmitter will have no choice but to send all 5 copies of the message at once, on uncorrelated frequency channels, in order to obtain the same level of diversity, thus using more resources. This effect, when considered for the entire network, yields a net gain in resource utilization while guaranteeing the desired latency at the target outage probability.

5.5.1 Performance Evaluation Framework

In order to further demonstrate how the proposed E-HARQ technique yields performance gains, we demonstrate this evaluation framework, which is a mathematical reasoning of the aforementioned example.

Considering L_T as the total number of bits to be transmitted, both for forward and feedback messages, we determine the minimum code rate (in bits per channel use) required to meet the target latency (λ) as

$$R = \frac{L_T}{n(\lambda)}, \tag{5.1}$$

which is a function of n, the number of available channel uses for each attempt. Fig. 5.7 illustrates the idea for the case with $z = 3$ and considering that δ seconds are being used in total to decode messages.

The proposed framework consists of determining the number of channel uses available for each attempt according to the desired scheme and then determining the minimum communication rate according to (5.1). Next, the values of n and R are used for the purpose of measuring system performance under the desired metric. Below we show how to determine n for S-ARQ and E-HARQ in order to evaluate their performance.

In S-ARQ, when the receiver successfully decodes the message it sends back an ACK, whereas if it fails it sends a NACK instead. Upon receiving a NACK, the transmitter sends a new copy of the message and the receiver discards the signal from the first message and repeats the process until z attempts have elapsed or it successfully decodes the message. In S-ARQ the receiver has to decode all the bits sent by the transmitter at every attempt before deciding to send an ACK or NACK, such that it has to use up to

$$\delta_{\mathrm{s}} = z\frac{L_{\mathrm{T}}\phi}{f_{\mathrm{apu}}} \tag{5.2}$$

seconds out of the latency budget to decode messages. Here, ϕ is the number of operations per bit required for decoding the messages, and f_{apu} is the arithmetic logic unit clock frequency. Therefore, $n_{\mathrm{s}}(\lambda)$, the number of available channel uses for each S-ARQ attempt, becomes

$$n_{\mathrm{s}}(\lambda) = (\lambda - \delta_{\mathrm{s}})\,\frac{W}{z}. \tag{5.3}$$

On the other hand, when considering E-HARQ, the transmitter uses our proposed strategy to predict whether or not an error will occur and sends the next attempt without waiting for the NACK, thus saving important latency resources. We propose to predict the average SNR and if $\bar{\gamma}$ is below a certain threshold, the transmitter sends the next attempt, without the need to wait for a NACK. Since the average SNR is low, there is a high probability of failure, and wasting time with complex decoding algorithms might not be the best approach, in particular when latency is so critical. This way, the payload only has to be decoded once. Therefore, δ_{e}, the amount of time required for decoding messages in the E-HARQ case, is determined by adding the time to decode the payload once with the amount of time to decode the remaining bits (headers, overhead and feedback signals) z times, yielding

$$\delta_{\mathrm{e}} = \frac{(z(L_{\mathrm{T}} - L_{\mathrm{P}}) + L_{\mathrm{P}})\phi}{f_{\mathrm{apu}}}, \tag{5.4}$$

where L_{P} is the payload length. Thus, the number of channel uses available in each E-HARQ attempt, $n_{\mathrm{e}}(\lambda)$, is

$$n_{\mathrm{e}}(\lambda) = (\lambda - \delta_{\mathrm{e}})\,\frac{W}{z}. \tag{5.5}$$

We can use the equations derived here to make certain predictions on the performance of each scheme. For instance, analyzing (5.3) and (5.5) asymptotically, when $\lambda \to \inf$, $n_{\mathrm{s}} \approx n_{\mathrm{e}}$ and the performance of both schemes will be very similar. This explains why the protocols designed without latency in mind are sub-optimal when considering URLLC.

Moreover, taking the partial derivative of n_s with respect to z yields

$$\frac{\partial n_\mathrm{s}}{\partial z} = -\frac{W\lambda}{z^2}, \tag{5.6}$$

while for n_e we have

$$\frac{\partial n_\mathrm{e}}{\partial z} = -\frac{W\left(\lambda - \frac{L_\mathrm{P}\phi}{f_\mathrm{apu}}\right)}{z^2}. \tag{5.7}$$

Considering that in any system $L_\mathrm{P} > 0$, it is possible to see that (5.6) decreases faster than (5.7), thus proving that S-ARQ will always have access to fewer channel uses when compared to E-HARQ. Therefore, the former requires a larger coding rate to deliver the same latency performance. This provides mathematical guarantees that E-HARQ outperforms S-ARQ for any z.

Although we are comparing E-HARQ with S-ARQ in this example, the proposed framework could be used for more complex ARQ mechanisms such as CC-HARQ or INR-HARQ. In those cases, n and R would be determined in a similar fashion, the difference would be in the function used to determine the error.

5.6 Numerical Results

Applying the prediction algorithm to the processed signals for the three signals that have been analyzed throughout this work yields the following results. Figs. 5.8, 5.9, and 5.10 display the prediction results for the pairs (28,3), (54,2), and (88,5), respectively. Note how, in all 3 figures the proposed LSTM architecture is able to predict the values of average SINR closely, thus rendering it a great candidate to be used for poor channel conditions.

Looking more closely at Fig. 5.8, it is possible to see that this UE started close to the eNB, but moved away from it as time went on. It's also possible to observe that the edges which are quite sharp on the original signal, in Fig. 5.2 are much smoother after all the post-processing, due to the averaging effect. On the other hand, the user on 5.9 started somewhat close to the eNB, moved even closer, and then moved away. Lastly, the user for the signal depicted in Fig. 5.10 stayed more or less in the same region, with respect to the eNB in question. In the plot, that can be verified by a smaller variation on the SINR throughout the simulation. However, it still varies around 7 dB and that is predicted by the proposed algorithm.

The algorithm was able to predict the average SINR throughout the simulation with very good accuracy. The results from Figs. 5.8, 5.9, and 5.10 have also been analysed statistically, to provide a better understanding of the significance of the results. The metric chosen was the root mean square error, and the results are depicted in Table 5.2. As we can see by this metric, the accuracy of the predictions is very high.

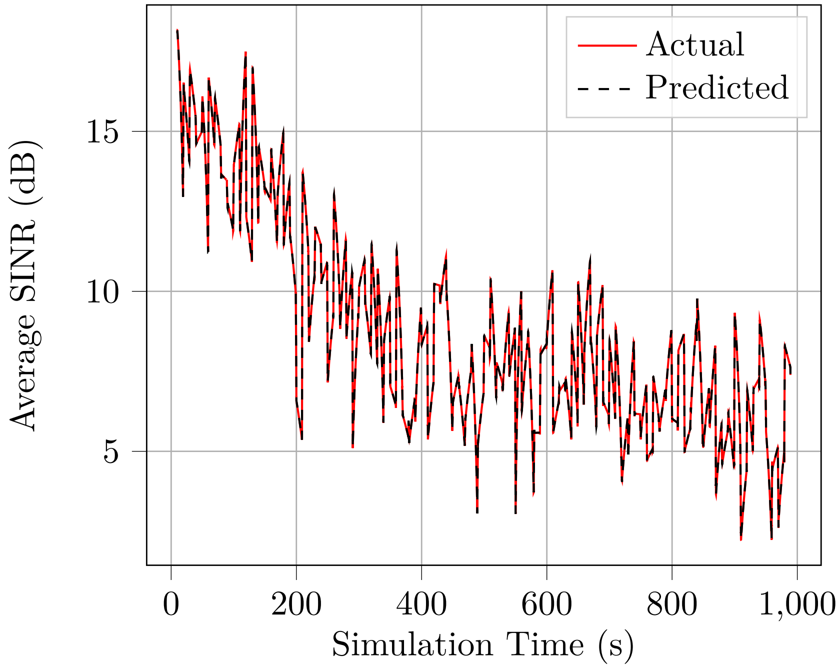


Figure 5.8: SINR Predictions for the pair (28,3).

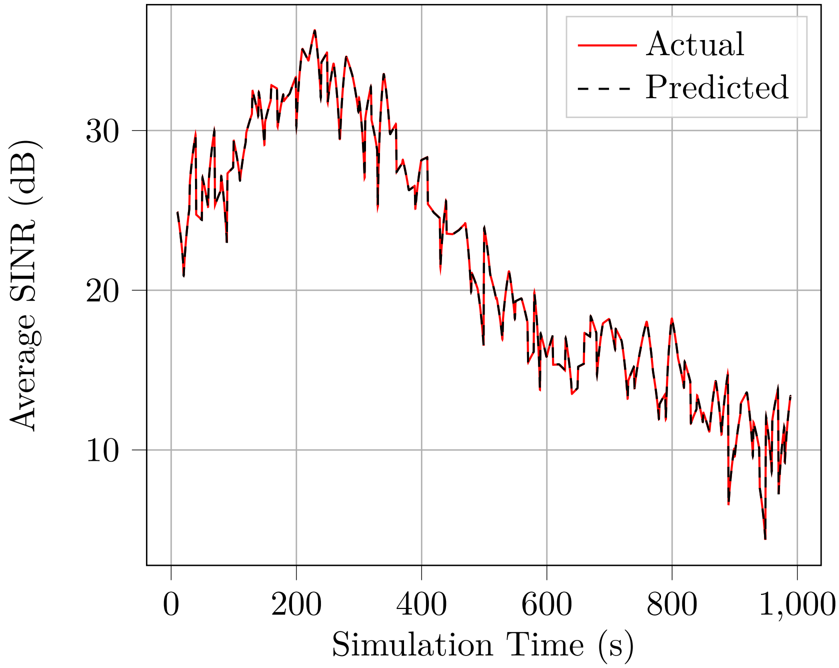


Figure 5.9: SINR Predictions for the pair (54,2).

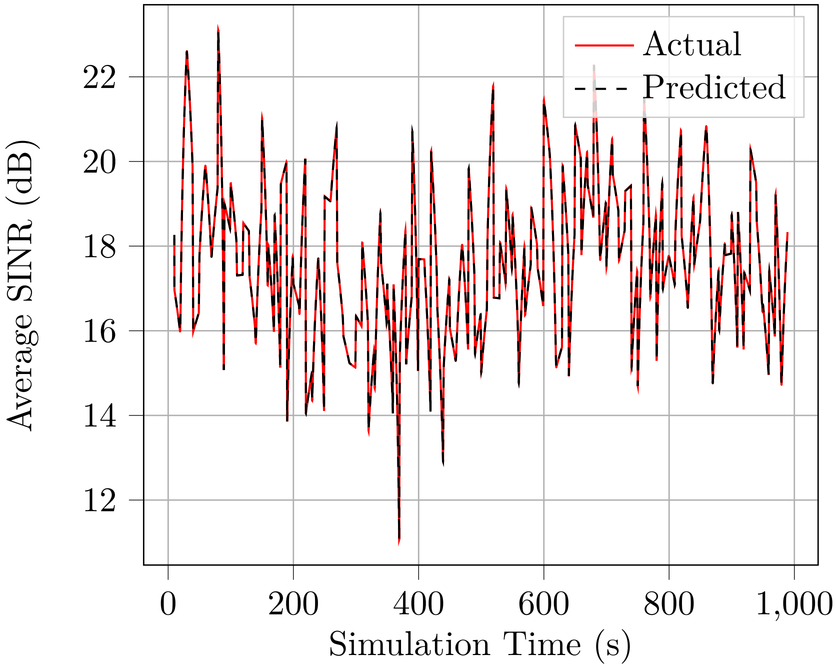


Figure 5.10: SINR Predictions for the pair (88,5).

Table 5.2: Statistical Analysis of the predictions.

UE/Cell Pair	**Root Mean Square Error**
(28, 3)	0.014066
(54, 2)	0.016603
(88, 5)	0.011529

Next, the potential savings of using the proposed prediction results in conjunction with the CC-HARQ protocol are showcased, such that the gains of using SON-enabled HARQ are highlighted. Fig. 5.11 shows the average error probability as a function of the average SNR for CC-ARQ and E-HARQ, for the NLOS ($m = 1$) case, while Fig. 5.12 has the same information when there is some LOS ($m = 3$). In both figures, the rate is determined for each line by using (5.1) and the appropriate value of n depending on the scheme. As it can be observed, the benefits of E-HARQ are more pronounced for larger values of z, since the difference in coding rate is larger. It is particularly interesting to observe the benefits of the proposed scheme by observing how increasing the diversity via allowing more transmission attempts actually incurs in worse performance for traditional CC-HARQ. This is due to the fact that the imposed rates become too high in order to meet the latency requirement. Particularly when z grows, the required rates grow considerably due to the time to decode each attempt on the traditional approach. Conversely, the proposed method of this work avoids wasting that precious time by predicting the channel quality, and thus results in a much better performance. Moreover, when m grows, meaning that there is more LOS, the

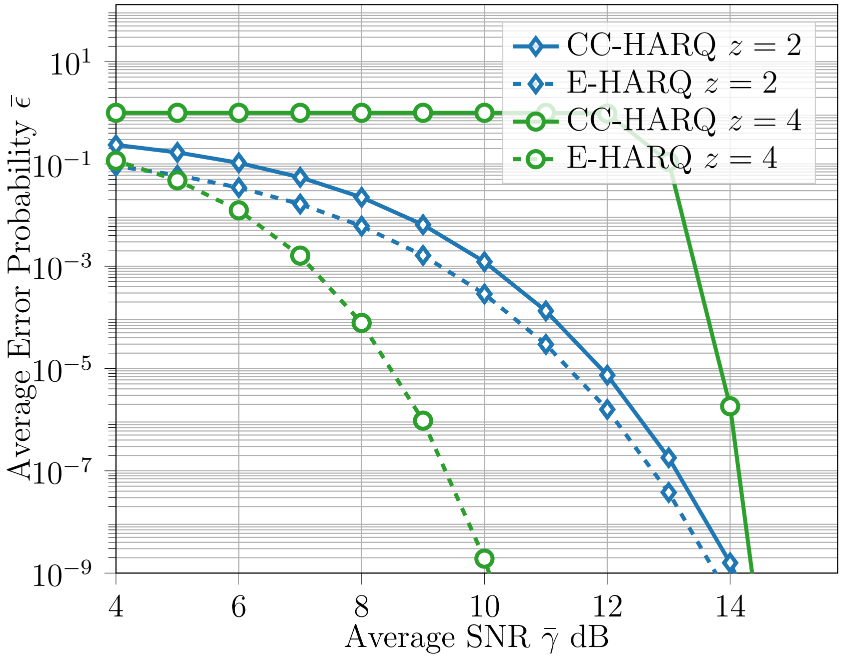


Figure 5.11: Block error rate versus SNR for CC-HARQ and E-HARQ for $z = 2$ and $z = 4$ and with $m = 1$.

performance gap also increases. This is due to the fact that when the channel is better, those diversity gains become even more pronounced, and adding diversity is much more beneficial in E-HARQ than it is in traditional CC-HARQ, for the aforementioned reasons.

Next, in Fig. 5.13, the error rate performance of both schemes when the target latency varies is compared. This is done for two levels of average SNR (0 and 10 dB) and for $z = 3$. As it can be observed, for a more strict latency the performance difference is greater since the latency budget is more stringent. When the target latency increases, both schemes tend to the same performance, as predicted by the asymptotic analysis presented earlier. Moreover, the gap is almost equivalent regardless of the average SNR considered, showing that the gains of using the proposed scheme can be used in different scenarios, such as applications with stringent power limitations (*e.g.* cognitive radio) or applications with access to more energy (*e.g.* cyber-physical systems).

To show the impact of increasing the number of allowed attempts z, the average error probabilities for both schemes when $m = 1$ and considering $\bar{\gamma} = 0$ dB and -8 dB, respectively, are presented in Figs. 5.14 and 5.15 In Fig. 5.15 it is possible to clearly observe that increasing the number of allowed attempts does not scale indefinitely, as at some point the required coding rate will overcome the added gains from increased diversity. The more stringent the link budget (*e.g.* smaller average SNR) the earlier this tipping point will occur. In Fig. 5.14, on the other hand, since it depicts a higher SNR scenario ($\bar{\gamma} = 0$ dB), the tipping point is only plotted for CC-HARQ, as it occurs for larger z in the case of E-HARQ. This further explains the results from Figs. 5.11 and 5.12, where having a

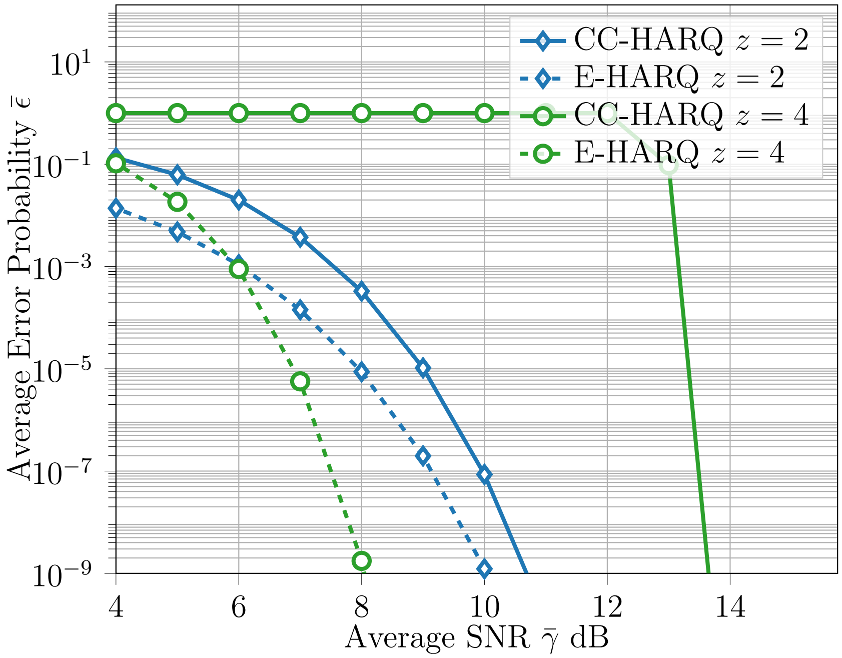


Figure 5.12: Block error rate versus SNR for CC-HARQ and E-HARQ for $z = 2$ and $z = 4$ and with $m = 3$.

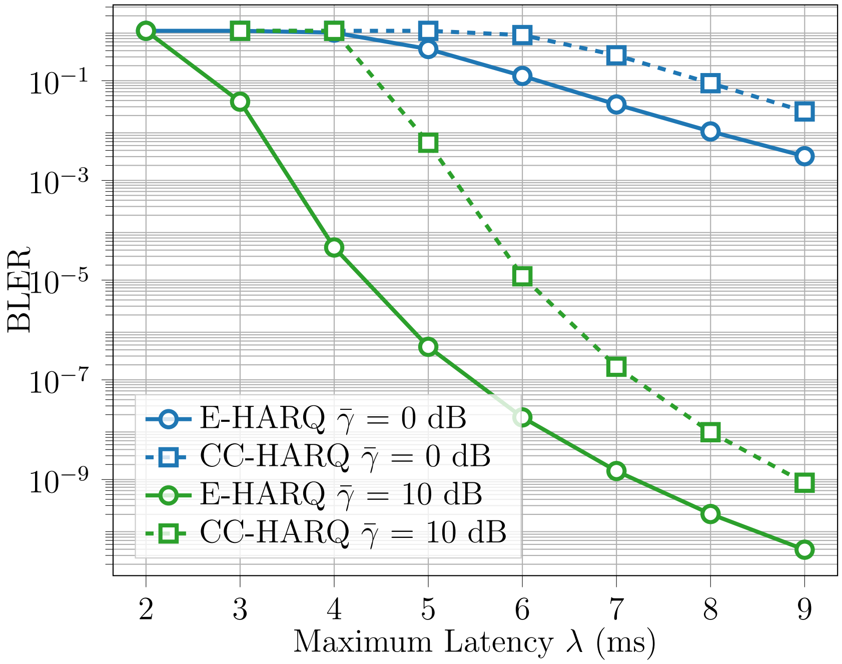


Figure 5.13: Performance comparison between CC-HARQ and E-HARQ for $z = 3$ transmission attempts and different values of λ.

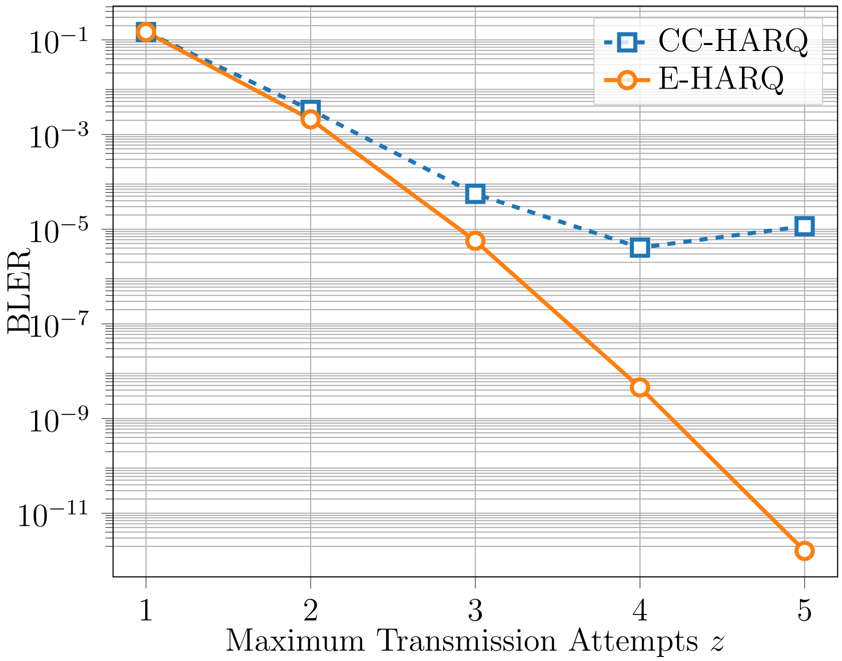


Figure 5.14: Performance comparison between CC-HARQ and E-HARQ for $\bar{\gamma} = 0$ dB and different values of z.

higher z yielded a worse error performance.

5.7 Conclusion

In this chapter, a digital twin was proposed and implemented, showing how SON applications can be devised even when there is data scarcity, which enables solutions to be tested and developed more quickly and efficiently, without the need for real models in the design phase. Furthermore, leveraging the generated data-set, an LSTM was designed to predict the instantaneous SINR for simulated users inside the digital twin, resulting in high levels of precision in the prediction of average SINR. Moreover, the clever use of pre-processing techniques to the available data coupled with domain-specific knowledge was crucial to extract meaning from the data. The refined information was then fed into a proposed algorithm to enable a resource-efficient operation of URLLC applications, thus showing how can SON be a facilitator for important applications which rely on the latency and reliability promises of 5G.

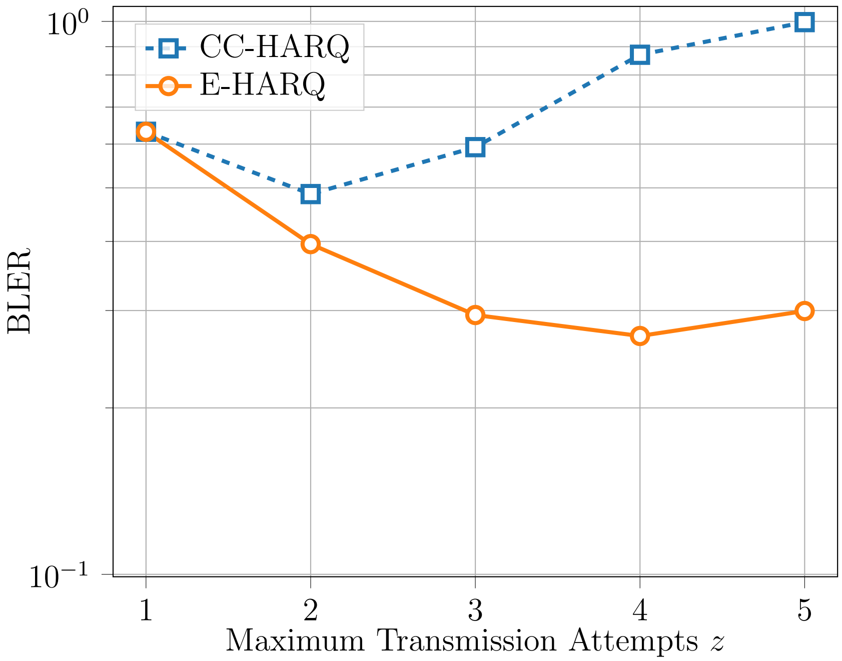


Figure 5.15: Performance comparison between CC-HARQ and E-HARQ at high SNR ($\bar{\gamma} = 10$ dB) and different values of z.

Chapter 6

Conclusion and Future Research

6.1 Conclusion

URLLC applications pose a few of the most exciting possibilities of the next generations of mobile communications and the challenges of devising resource-efficient communication protocols tailored to those are rooted in the fact that latency and reliability are typically antagonistic to one another.

On the other hand, time diversity is a proven strategy to provide efficient resource utilization in traditional applications, but its use in URLLC scenarios is counter-intuitive due to the fact that it could incur an increased link latency.

In this thesis, we have shown how this increased latency can be compensated by the gains in diversity and result in efficient utilization.

First, we explored this in the context of a generic URLLC application, showing important energy efficiency improvements when comparing with a similar approach using a different diversity approach to enable the stringent requirements. We derived a detailed energy consumption model and showed how to effectively determine the communication rate and the number of allowed attempts that would provide optimal energy efficiency. Furthermore, we also showed how a similar strategy could be used to reduce the EMF radiation, an important metric usually tied to increasingly relevant human health concerns. This approach relied on a series of assumptions, which if overcome could lead to very interesting practical implementations.

Next, we explored how taking into account application-specific requirements when designing the wireless communication protocol could mean that a more efficient solution could be attained. More specifically, we showed that for WNCS, using designing a PPC solution and tweaking its parameters in conjunction with the communication settings led to savings in the required operational bandwidth for the challenging QoS requirements of industry automation applications—which are amongst the most stringent of URLLC applications. We compared two approaches, one which does not take into account the mutual information accumulated at the receiver on each attempt, and one which does. We were able to show that, unlike with traditional applications, the simpler approach had a better performance, due to the less restrictive imposition on the number of allowed attempts.

The final investigation performed in this thesis was regarding a way to relax the assumptions made earlier in an attempt to make the proposed HARQ approach implementable in practical scenarios. This was done in the context of SON, whereby ML is applied to mobile communications in order to tweak its operation. We were able to relax assumptions made earlier by predicting when a message would fail and not waiting for feedback from the receiver before performing a retransmission. To design such protocol, we first created a digital twin of the city of Glasgow and gathered realistic simulation data. The data was then processed and fed into the ML algorithm which was, in turn, used to predict the failure of messages by predicting the channel quality using an LSTM convolutional neural network.

We conclude this work by presenting just a few of the large array of possible extensions of future research that could spin off from this thesis.

6.2 Future Research

Several directions could be taken to continue this study. For instance, the investigation from Chapter 4 could be extended to consider real control deployments, and instead of performing joint design with a single optimization function, the control cost could be introduced as part of the optimization problem, in a multi-objective optimization problem.

Secondly, a data collection campaign could be done in order to collect real-world data and the SON implementation proposed in Chapter 5 could be performed using data from an actual deployment. Furthermore, yet another interesting approach would be to use the learned weights from the digital twin and run the LSTM without training it a priori, to show that the parameters can be optimized even without having long-term access to the environment. Still, regarding Chapter 5, the dataset generated in this work can be used for a myriad of applications, by multiple researchers in several communication domains. In the context of URLLC, for example, other ML algorithms could be explored to have even better performance.

The third line of research which could spur from this thesis consists of taking the study from Chapter 3 further by studying how the proposed protocol behaves under real applications. It can be done by considering a coding and modulation approach, as opposed to the information theory line used in this work. This would show how the designed approaches fair against commercially available solutions and highlight the attainable gains of using the protocols we designed.

Moreover, the study of non-orthogonal multiple access techniques in order to reduce the scheduling problems with HARQ feedback scheduling could be considered.

Lastly, one interesting analytical extensions of this work involves considering other fading scenarios to evaluate the proposed algorithms in Chapters 3 and 4.

Bibliography

[1] H. Ren, C. Pan, K. Wang, Y. Deng, M. Elkashlan, and A. Nallanathan, “Achievable data rate for URLLC-enabled UAV systems with 3-d channel model,” *IEEE Wireless Communications Letters*, vol. 8, no. 6, pp. 1587–1590, Dec. 2019.

[2] H. Ren, C. Pan, Y. Deng, M. Elkashlan, and A. Nallanathan, “Joint power and blocklength optimization for URLLC in a factory automation scenario,” *IEEE Transactions on Wireless Communications*, vol. 19, no. 3, pp. 1786–1801, Mar. 2020.

[3] A. M. Lacy, R. Bravo, A. M. Otero-Piñeiro, R. Pena, F. B. D. Lacy, R. Menchaca, and J. M. Balibrea, “5g-assisted telementored surgery,” *BJS*, vol. 106, no. 12, pp. 1576–1579, Sep. 2019.

[4] R. Ricart-Sanchez, A. C. Aleixo, Q. Wang, and J. M. A. Calero, “Hardware-based network slicing for supporting smart grids self-healing over 5g networks,” in *2020 IEEE International Conference on Communications Workshops (ICC Workshops)*. IEEE, Jun. 2020.

[5] S. Ansari, J. Ahmad, S. A. Shah, A. K. Bashir, T. Boutaleb, and S. Sinanovic, “Chaos-based privacy preserving vehicle safety protocol for 5g connected autonomous vehicle networks,” *Transactions on Emerging Telecommunications Technologies*, vol. 31, no. 5, Apr. 2020.

[6] S. R. Pokhrel, J. Ding, J. Park, O.-S. Park, and J. Choi, “Towards enabling critical mMTC: A review of URLLC within mMTC,” *IEEE Access*, vol. 8, pp. 131 796–131 813, 2020.

[7] B. Holfeld, D. Wieruch, T. Wirth, L. Thiele, S. A. Ashraf, J. Huschke, I. Aktas, and J. Ansari, “Wireless communication for factory automation: An opportunity for LTE and 5G systems,” *IEEE Commun. Mag.*, vol. 54, no. 6, pp. 36–43, 2016.

[8] M. Alonzo, P. Baracca, S. R. Khosravirad, and S. Buzzi, “Urllc for factory automation: an extensive throughput-reliability analysis of d-mimo,” in *WSA 2020; 24th International ITG Workshop on Smart Antennas*, 2020, pp. 1–6.

[9] J. P. B. Nadas, G. Zhao, R. D. Souza, and M. A. Imran, “Ultra reliable low latency communications as an enabler for industry automation,” pp. 89–107, dec 2019.

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

[11] F. A. Asuhaimi, S. Bu, J. P. B. Nadas, and M. A. Imran, “Delay-aware energy-efficient joint power control and mode selection in device-to-device communications for FREEDM systems in smart grids,” *IEEE Access*, vol. 7, pp. 87 369–87 381, 2019.

[12] P. Popovski, V. Braun, H. Mayer, P. Fertl, Z. Ren, D. Gonzales-Serrano, E. Ström, T. Svensson, H. Taoka, P. Agyapong *et al.*, “Scenarios requirements and KPIs for 5G mobile and wireless system,” *ICT-317669-METIS/D1. 1, ICT-317669 METIS project*, 2013.

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

[14] V. Jain, S. Lapoehn, T. Frankiewicz, T. Hesse, M. Gharba, H. Cao, S. Gangakhedkar, J. Eichinger, A. Ramadan Ali, K. Ganesan *et al.*, “5G enabled cooperative collision avoidance: System design and field test,” in *IEEE Int. Symp. World of Wireless, Mobile and Multimedia Networks (WoWMoM)*, Jun. 2017, pp. 1–6.

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

[16] F. Rosas, R. D. Souza, M. E. Pellenz, C. Oberli, G. Brante, M. Verhelst, and S. Pollin, “Optimizing the code rate of energy-constrained wireless communications with HARQ,” *IEEE Trans. Wireless Commun.*, vol. 15, no. 1, pp. 191–205, 2016.

[17] D. Chase, “Code combining - a maximum-likelihood decoding approach for combining an arbitrary number of noisy packets,” *IEEE Trans. Commun.*, vol. 33, no. 5, pp. 385–393, May 1985.

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

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

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

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

[22] E. Dosti, M. Shehab, H. Alves, and M. Latva-Aho, "Ultra reliable communication via optimum power allocation for HARQ retransmission schemes," *IEEE Access*, vol. 8, pp. 89 768–89 781, 2020.

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

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

[25] M. Shehab, H. Alves, E. A. Jorswieck, E. Dosti, and M. Latva-aho, "Effective energy efficiency of ultra-reliable low latency communication," *IEEE Internet of Things Journal*, pp. 1–1, 2021.

[26] O. L. A. Lopez, N. H. Mahmood, H. Alves, C. M. Lima, and M. Latva-aho, "Ultra-low latency, low energy, and massiveness in the 6g era via efficient CSIT-limited scheme," *IEEE Communications Magazine*, vol. 58, no. 11, pp. 56–61, nov 2020.

[27] Y. Liu, Y. Deng, M. Elkashlan, A. Nallanathan, and G. K. Karagiannidis, "Analyzing grant-free access for urllc service," 2020.

[28] P. Nouri, H. Alves, M. A. Uusitalo, O. A. López, and M. Latva-aho, "Machine-type wireless communications enablers for beyond 5g: Enabling URLLC via diversity under hard deadlines," *Computer Networks*, vol. 174, p. 107227, jun 2020.

[29] A. Zaki-Hindi, S.-E. Elayoubi, and T. Chahed, "Multi-tenancy and URLLC on unlicensed spectrum: Performance and design," *Computer Networks*, vol. 177, p. 107311, aug 2020.

[30] S. Xie, B. Chang, G. Zhao, Z. Chen, and Y. Huang, "Optimal power allocation for relay-assisted wireless packetized predictive control," in *2019 24th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA)*. IEEE, sep 2019.

[31] P. Park, S. C. Ergen, C. Fischione, C. Lu, and K. H. Johansson, “Wireless network design for control systems: A survey,” *IEEE Commun. Surveys Tuts.*, vol. PP, no. 99, p. 1, 2017.

[32] P. Park, J. Araújo, and K. H. Johansson, “Wireless networked control system co-design,” in *Networking, Sensing and Control (ICNSC), 2011 IEEE International Conference on.* IEEE, 2011, pp. 486–491.

[33] C. Lu, A. Saifullah, B. Li, M. Sha, H. Gonzalez, D. Gunatilaka, C. Wu, L. Nie, and Y. Chen, “Real-time wireless sensor-actuator networks for industrial cyber-physical systems,” *Proc. IEEE*, vol. 104, no. 5, pp. 1013–1024, 2016.

[34] X. Tong, G. Zhao, M. A. Imran, Z. Pang, and Z. Chen, “Minimizing wireless resource consumption for packetized predictive control in real-time cyber physical systems,” in *2018 IEEE International Conference on Communications Workshops (ICC Workshops).* IEEE, 2018.

[35] D. E. Quevedo, J. Ostergaard, and D. Nesic, “Packetized predictive control of stochastic systems over bit-rate limited channels with packet loss,” *IEEE Trans. Autom. Control*, vol. 56, no. 12, pp. 2854–2868, Dec. 2011.

[36] F. E. da Silva, A. L. V. Iaremczuk, R. D. Souza, G. Brante, G. L. Moritz, and S. Hussain, “Hybrid ARQ in wireless packetized predictive control,” *IEEE Sensors Letters*, vol. 4, no. 5, pp. 1–4, may 2020.

[37] E. Glaessgen and D. Stargel, “The digital twin paradigm for future NASA and u.s. air force vehicles,” in *53rd AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics and Materials Conference.* American Institute of Aeronautics and Astronautics, apr 2012.

[38] J. Lee, E. Lapira, B. Bagheri, and H. an Kao, “Recent advances and trends in predictive manufacturing systems in big data environment,” *Manufacturing Letters*, vol. 1, no. 1, pp. 38–41, oct 2013.

[39] J. Lee, B. Bagheri, and H.-A. Kao, “A cyber-physical systems architecture for industry 4.0-based manufacturing systems,” *Manufacturing Letters*, vol. 3, pp. 18–23, jan 2015.

[40] S. H. Khajavi, N. H. Motlagh, A. Jaribion, L. C. Werner, and J. Holmstrom, “Digital twin: Vision, benefits, boundaries, and creation for buildings,” *IEEE Access*, vol. 7, pp. 147 406–147 419, 2019.

[41] F. Tao, H. Zhang, A. Liu, and A. Y. C. Nee, “Digital twin in industry: State-of-the-art,” *IEEE Transactions on Industrial Informatics*, vol. 15, no. 4, pp. 2405–2415, apr 2019.

[42] O. E. Marai, T. Taleb, and J. Song, “Roads infrastructure digital twin: A step toward smarter cities realization,” *IEEE Network*, vol. 35, no. 2, pp. 136–143, mar 2021.

[43] Y. Lu, X. Huang, K. Zhang, S. Maharjan, and Y. Zhang, “Low-latency federated learning and blockchain for edge association in digital twin empowered 6g networks,” *IEEE Transactions on Industrial Informatics*, vol. 17, no. 7, pp. 5098–5107, jul 2021.

[44] Z. Wang, X. Liao, X. Zhao, K. Han, P. Tiwari, M. J. Barth, and G. Wu, “A digital twin paradigm: Vehicle-to-cloud based advanced driver assistance systems,” in *2020 IEEE 91st Vehicular Technology Conference (VTC2020-Spring)*. IEEE, may 2020.

[45] K. M. Alam and A. E. Saddik, “C2ps: A digital twin architecture reference model for the cloud-based cyber-physical systems,” *IEEE Access*, vol. 5, pp. 2050–2062, 2017.

[46] K. Zhang, J. Cao, S. Maharjan, and Y. Zhang, “Digital twin empowered content caching in social-aware vehicular edge networks,” *IEEE Transactions on Computational Social Systems*, pp. 1–13, 2021.

[47] W. Sun, P. Wang, N. Xu, G. Wang, and Y. Zhang, “Dynamic digital twin and distributed incentives for resource allocation in aerial-assisted internet of vehicles,” *IEEE Internet of Things Journal*, pp. 1–1, 2021.

[48] Y. Wu, K. Zhang, and Y. Zhang, “Digital twin networks: a survey,” *IEEE Internet of Things Journal*, pp. 1–1, 2021.

[49] G. Ding, J. Yuan, J. Bao, and G. Yu, “LSTM-based active user number estimation and prediction for cellular systems,” *IEEE Wireless Communications Letters*, vol. 9, no. 8, pp. 1258–1262, aug 2020.

[50] P. Ge and T. Lv, “Energy-efficient optimized dynamic massive MIMO based on predicted user quantity by LSTM algorithm,” in *2018 IEEE/CIC International Conference on Communications in China (ICCC)*. IEEE, aug 2018.

[51] H. Zhou and S. Zhao, “Prediction of a new kind of MR data,” *Procedia Computer Science*, vol. 131, pp. 904–910, 2018.

[52] Q. Liu, G. Chuai, J. Wang, and J. Pan, "Proactive mobility management with trajectory prediction based on virtual cells in ultra-dense networks," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 8, pp. 8832–8842, aug 2020.

[53] C. Luo, J. Ji, Q. Wang, X. Chen, and P. Li, "Channel state information prediction for 5g wireless communications: A deep learning approach," *IEEE Transactions on Network Science and Engineering*, vol. 7, no. 1, pp. 227–236, jan 2020.

[54] S. AlMarshed, D. Triantafyllopoulou, and K. Moessner, "Supervised learning for enhanced early HARQ feedback prediction in URLLC," in *2020 IEEE International Conference on Communication, Networks and Satellite (Comnetsat)*. IEEE, dec 2020.

[55] O. L. A. Lopez, N. H. Mahmood, and H. Alves, "Enabling URLLC for low-cost IoT devices via diversity combining schemes," in *2020 IEEE International Conference on Communications Workshops (ICC Workshops)*. IEEE, jun 2020.

[56] R. Kotaba, C. N. Manchon, T. Balercia, and P. Popovski, "How URLLC can benefit from NOMA-based retransmissions," *IEEE Transactions on Wireless Communications*, vol. 20, no. 3, pp. 1684–1699, mar 2021.

[57] F. Ghanami, G. A. Hodtani, B. Vucetic, and M. Shirvanimoghaddam, "Performance analysis and optimization of NOMA with HARQ for short packet communications in massive IoT," *IEEE Internet of Things Journal*, vol. 8, no. 6, pp. 4736–4748, mar 2021.

[58] G. Pocovi, A. A. Esswie, and K. I. Pedersen, "Channel quality feedback enhancements for accurate URLLC link adaptation in 5g systems," in *2020 IEEE 91st Vehicular Technology Conference (VTC2020-Spring)*. IEEE, may 2020.

[59] J. P. B. Nadas, O. Onireti, R. D. Souza, H. Alves, G. Brante, and M. A. Imran, "Performance analysis of hybrid ARQ for ultra-reliable low latency communications," *IEEE Sensors J.*, p. 1, 2019.

[60] A. Azari, M. Ozger, and C. Cavdar, "Risk-aware resource allocation for URLLC: Challenges and strategies with machine learning," *IEEE Communications Magazine*, vol. 57, no. 3, pp. 42–48, Mar. 2019.

[61] Y. K. Tun, D. H. Kim, M. Alsenwi, N. H. Tran, Z. Han, and C. S. Hong, "Energy efficient communication and computation resource slicing for eMBB and URLLC coexistence in 5g and beyond," *IEEE Access*, vol. 8, pp. 136 024–136 035, 2020.

[62] J. P. B. Nadas, O. Onireti, R. D. Souza, H. Alves, G. Brante, and M. A. Imran, "Reducing EMF emissions in ultra-reliable low-latency communications with HARQ," in *Low Electromagnetic Emission Wireless Network Technologies: 5G and beyond.* Institution of Engineering and Technology, nov 2019, pp. 231–250.

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

[64] G. Zhao, M. A. Imran, Z. Pang, Z. Chen, and L. Li, "Toward real-time control in future wireless networks: Communication-control co-design," *IEEE Communications Magazine*, vol. 57, no. 2, pp. 138–144, feb 2019.

[65] P. V. Klaine, M. A. Imran, O. Onireti, and R. D. Souza, "A survey of machine learning techniques applied to self-organizing cellular networks," *IEEE Communications Surveys & Tutorials*, vol. 19, no. 4, pp. 2392–2431, 2017.

[66] J. Nadas, P. Klaine, L. Zhang, G. Zhao, M. Imran, and R. Souza, "Performance analysis of early-HARQ for finite block-length packet transmission," in *2019 IEEE International Conference on Industrial Cyber Physical Systems (ICPS).* IEEE, may 2019.

[67] T. S. Rappaport, *Wireless communications: principles and practice*, 2nd ed. prentice hall PTR New Jersey, 2002.

[68] A. Goldsmith, *Wireless communications.* Cambridge university press, 2005.

[69] N. Amitay, "Modeling and computer simulation of wave propagation in lineal line-of-sight microcells," *IEEE Trans. Veh. Technol.*, vol. 41, no. 4, pp. 337–342, 1992.

[70] Y. Okumura, "Field strength and its variability in vhf and uhf land-mobile service," *Rev. Elec. Comm. Lab.*, vol. 16, no. 9, pp. 825–873, 1968.

[71] M. Hata, "Empirical formula for propagation loss in land mobile radio services," *IEEE Trans. Veh. Technol.*, vol. 29, no. 3, pp. 317–325, 1980.

[72] P. E. Mogensen, P. Eggers, C. Jensen, and J. B. Andersen, "Urban area radio propagation measurements at 955 and 1845 mhz for small and micro cells," in *Global Telecommunications Conference, 1991. GLOBECOM'91.'Countdown to the New Millennium. Featuring a Mini-Theme on: Personal Communications Services.* IEEE, 1991, pp. 1297–1302.

[73] E. COST231, "Urban transmission loss models for mobile radio in the 900-and 1,800 mhz bands (revision 2), ser," *European Cooperation in the Field of Scientific and Technical Research. Netherlands: The Hague*, 1991.

[74] R. Beals and R. Wong, *Special Functions: A Graduate Text.* Cambridge university press, 2010.

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

[76] Y. Polyanskiy, H. V. Poor, and S. Verdú, "Channel coding rate in the finite blocklength regime," *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2307–2359, 2010.

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

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

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

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

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

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

[83] A. Keshavarzian, H. Lee, and L. Venkatraman, "Wakeup scheduling in wireless sensor networks," in *ACM Int. Symp. on Mobile Ad Hoc Networking and Computing.* ACM, 2006, pp. 322–333.

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

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

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

[87] M. Shirvanimoghaddam, M. S. Mohamadi, R. Abbas, A. Minja, B. Matuz, G. Han, Z. Lin, Y. Li, S. Johnson, and B. Vucetic, “Short block-length codes for ultra-reliable low-latency communications,” *arXiv preprint arXiv:1802.09166*, 2018.

[88] R. M. Corless, G. H. Gonnet, D. E. G. Hare, D. J. Jeffrey, and D. E. Knuth, “On the LambertW function,” *Advances in Computational Mathematics*, vol. 5, no. 1, pp. 329–359, dec 1996.

[89] S. Boyd and L. Vandenberghe, *Convex optimization.* Cambridge university press, 2004.

[90] G. Boole, *A treatise on the calculus of finite differences.* Macmillan, Cambridge, 1860.

[91] J. P. Nadas, M. A. Imran, G. Brante, and R. D. Souza, “Optimizing the energy efficiency of short term ultra reliable communications in vehicular networks,” in *Int. Symp. on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt).* IEEE, 2017, pp. 1–6.

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

[93] A. He, S. Srikanteswara, K. K. Bae, T. R. Newman, J. H. Reed, W. H. Tranter, M. Sajadieh, and M. Verhelst, “System power consumption minimization for multichannel communications using cognitive radio,” in *IEEE Int. Conf. Microwaves, Communications, Antennas and Electronics Systems*, Nov. 2009, pp. 1–5.

[94] S. Ashraf, Y. P. E. Wang, S. Eldessoki, B. Holfeld, D. Parruca, M. Serror, and J. Gross, “From radio design to system evaluations for ultra-reliable and low-latency communication,” in *Proc. European Wireless 2017; 23th European Wireless Conf*, May 2017, pp. 1–8.

[95] D. E. Quevedo and D. Nešić, “Robust stability of packetized predictive control of nonlinear systems with disturbances and markovian packet losses,” *Automatica*, vol. 48, no. 8, pp. 1803–1811, 2012.

[96] O. N. C. Yilmaz, Y. P. E. Wang, N. A. Johansson, N. Brahmi, S. A. Ashraf, and J. Sachs, “Analysis of ultra-reliable and low-latency 5g communication for a factory

automation use case," in *Proc. IEEE Int. Conf. Communication Workshop (ICCW)*, Jun. 2015, pp. 1190–1195.

[97] T. ETSI, "102 889-2 v1. 1.1 (2011-08) electromagnetic compatibility and radio spectrum matters (erm)," *System Reference Document*, 2011.

[98] *IEEE Standard for Low-Rate Wireless Networks*, IEEE Std. 802.15.4, 2015.

[99] B. Holfeld, D. Wieruch, L. Raschkowski, T. Wirth, C. Pallasch, W. Herfs, and C. Brecher, "Radio channel characterization at 5.85 GHz for wireless M2M communication of industrial robots," in *Proc. IEEE Wireless Communications and Networking Conf*, Apr. 2016, pp. 1–7.

[100] M. J. Kaur, V. P. Mishra, and P. Maheshwari, "The convergence of digital twin, IoT, and machine learning: Transforming data into action," in *Internet of Things*. Springer International Publishing, Jul. 2019, pp. 3–17.

[101] G. F. Riley and T. R. Henderson, "The ns-3 network simulator," in *Modeling and Tools for Network Simulation*. Springer Berlin Heidelberg, 2010, pp. 15–34.

[102] T. Cerqueira and M. Albano, "RoutesMobilityModel," in *Proceedings of the 2015 Workshop on ns-3 - WNS3 '15*. ACM Press, 2015.

[103] J. P. B. Nadas, S. Ansari, and M. A. Imran, "Glasgow cellular network digital twin with vehicular users," 2021. [Online]. Available: http://dx.doi.org/10.5525/gla.researchdata.1087

[104] NS3, "LENA NS3 user documentation." [Online]. Available: https://www.nsnam.org/docs/models/html/lte-user.html

[105] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic gradient descent," in *ICLR: International Conference on Learning Representations*, 2015, pp. 1–15.