Rahim Karim Shamsudin

Licenciado em Ciências da Engenharia Electrotécnica e de Computadores

Protocol for Extreme Low Latency M2M Communication Networks

Dissertação para obtenção do Grau de Mestre em

Engenharia Electrotécnica e de Computadores

Orientador: Luís Filipe Lourenço Bernardo, Professor Associado com Agregação, FCT-UNL

Júri

Presidente: Prof.Doutor Fernando José Almeida Vieira do Coito, FCT-UNL
Arguente: Prof.Doutor Rui Miguel Henriques Dias Morgado Dinis, FCT-UNL
Vogal: Prof.Doutor Luís Filipe Lourenço Bernardo, FCT-UNL

FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE NOVA DE LISBOA

Setembro, 2018

Protocol for Extreme Low Latency M2M Communication Networks

Este documento foi gerado utilizando o processador (pdf)LaTeX, com base no template "novathesis" [1] desenvolvido no Dep. Informática da FCT-NOVA [2].
[1] https://github.com/joaomlourenco/novathesis [2] http://www.di.fct.unl.pt

In memory of my Grandfather, Shamsudin Ahmad

Acknowledgements

I want to start by thanking Prof. Dr. Luís Bernardo, firstly for all his classes and lectures throughout the telecommunications area where he inspired and kindled my passion for the telecommunications, that motivated me to make my thesis in this area. He is the reason this dissertation was made possible, and I thank him for all his time and patience with me, always providing documentation, thoughtful insights and gems of knowledge, in times of great hardship his guidance made things seem simple when they didn't look like it. I was extremely fortunate to have him as my adviser.

This work was supported by *Instituto de Telecomunições* under the project VELOCE-MTC - UID/EEA/50008/2013. I thank them for the scholarship BIL/Nº140-30-10-2017-VELOCE-MTC-UID/EEA/50008/2013-2.

I'm thankful to *Departamento de Engenharia Electrotécnica e de Computadores* and *Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa* for creating the conditions that allowed me to complete my education.

I'd like to thank all of my colleagues and friends that accompanied me throughout this course, for all the moments and memories, making this journey memorable. An honorable mention goes to my colleague Tiago Miguel de Góis Raposo, that gave me a great help in the dissertation, from teaching how to properly use Latex to listening to me talk over and over about the dissertation, always keen in helping me find a solution for my problems and most of all for motivating me all the way until the end.

Last but not least, I'd like to dedicate this paragraph to my family. A heartfelt thank you to my parents and my brother, for all the support and patience shown, for always believing in me, standing by me in the bad times and good times. They are the reason I am who I am. They taught me to work hard for my goals, to be resilient when facing adversity, and to never stop learning and improving myself.

"I have not failed. I've just found 10,000 ways that won't work."

- Nikola Tesla

Abstract

As technology evolves, more Machine to Machine (M2M) deployments and mission critical services are expected to grow massively, generating new and diverse forms of data traffic, posing unprecedented challenges in requirements such as delay, reliability, energy consumption and scalability. This new paradigm vindicates a new set of stringent requirements that the current mobile networks do not support. A new generation of mobile networks is needed to attend to this innovative services and requirements - the The fifth generation of mobile networks (5G) networks. Specifically, achieving ultra-reliable low latency communication for machine to machine networks represents a major challenge, that requires a new approach to the design of the Physical (PHY) and Medium Access Control (MAC) layer to provide these novel services and handle the new heterogeneous environment in 5G. The current LTE Advanced (LTE-A) radio access network orthogonality and synchronization requirements are obstacles for this new 5G architecture, since devices in M2M generate bursty and sporadic traffic, and therefore should not be obliged to follow the synchronization of the LTE-A PHY layer. A non-orthogonal access scheme is required, that enables asynchronous access and that does not degrade the spectrum.

This dissertation addresses the requirements of URLLC M2M traffic at the MAC layer. It proposes an extension of the M2M H-NDMA protocol for a multi base station scenario and a power control scheme to adapt the protocol to the requirements of URLLC. The system and power control schemes performance and the introduction of more base stations are analyzed in a system level simulator developed in MATLAB, which implements the MAC protocol and applies the power control algorithm.

Results showed that with the increase in the number of base stations, delay can be significantly reduced and the protocol supports more devices without compromising delay or reliability bounds for Ultra-Reliable and Low Latency Communication (URLLC), while also increasing the throughput. The extension of the protocol will enable the study of different power control algorithms for more complex scenarios and access schemes that combine asynchronous and synchronous access.

Keywords: Machine to Machine; 5G;ultra-reliable low latency communications; MAC protocol; Power control; LTE-A.

Resumo

À medida que a tecnologia evolui, é expectável o crescimento massivo das redes *Machine to Machine* (M2M) e serviços para missão crítica (e.g. cirurgia remota), que irão gerar novas e diversas formas de tráfego de dados, criando novos desafios sem precedentes no que toca a requisitos como atraso, fiabilidade, consumo de energia e escalabilidade. Este novo paradigma reivindica um novo conjunto de requisitos, que a rede móvel atual não consegue dar suporte. Uma nova geração de redes móveis é necessária para suportar estes novos serviços e requisitos - As redes de 5G. Oferecer comunicação ultrafiável e de baixa latência (URLLC) para serviços M2M representa um enorme desafio que requere uma nova abordagem no desenho da camada física e do camada do protocolo de acesso ao meio, de modo a poder providenciar estes serviços e lidar com ambiente de rede heterogéneo que é introduzido no 5G. As atuais restrições de sincronismo e ortogonalidade da rede móvel LTE-A são obstáculos para a arquitectura da rede 5G, visto que os dispositivos em redes M2M geram tráfego esporádico e repentino, e por isso não devem ser obrigados a cumprir com a sincronização imposta pela camada física do LTE-A. É necessário um esquema de acesso não ortogonal, que permita um acesso assincrono e não degrade o espectro.

Esta dissertação aborda as necessidades do tráfego das comunicações URLLC M2M na camada MAC. Ela propõe uma extensão do protocolo M2M H-NDMA para o cenário de multi-antena e um esquema de controlo de potência para adaptar o protocolo aos requisitos do URLLC. O desempenho do sistema e esquema de controlo de potência são analisados usando um simulador de sistema desenvolvido em ambiente MATLAB, que implementa o protocolo MAC e aplica o algoritmo de controlo de potência.

Os resultados mostram que com o aumento do número de estações base o atraso pode ser reduzido significativamente, o protocolo suporta mais dispositivos sem comprometer requisitos de fiabilidade e latência das comunicações URLLC, e ao mesmo tempo, aumenta o débito. A extensão realizada ao protocolo irá permitir o estudo de diferentes esquemas de controlo de potência para cenários mais complexos, bem como o estudo de esquemas que combinem o acesso em modo síncrono e assíncrono.

Palavras-Chave: Máquina a máquina (M2M); 5G; comunicação ultrafiável e de baixa latência (URLLC); protocolo MAC; controlo de potência, LTE-A.

Contents

List of Figures

List of Tables

Acronyms

5G	The fifth generation of mobile networks.
AID	Association identifier.
AP	Access Point.
APC	Active Power Control.
ATL-S-MACA	Adaptive Traffic Load slotted MACA.
BER	Bit Error Rate.
B-NDMA	Blind Network Diversity Multiple Access.
BS	Base Station.
CDF	Cummulative Distribution Function.
CP	Cyclic Prefix.
C-RAN	Cloud Radio Network.
CDMA	Code Division Multiple Access.
CERA	Code Expanded Random Access.
CoMP	Coordinated multipoint.
CSI	Channel State Information.
CTS	Clear-To-Send.
DAS	Distributed Antenna System.
DC	Diversity Combining.
E2E	End-to-End.
ELL	Extreme Low Latency.
eMBB	enhanced Mobile Broadband.
EPUP	Energy Per Useful Packet.
FASA	Fast-Adaptive Slotted ALOHA.

f-OFDM	filtered OFDM.
FBMC	Filter Bank Multicarrier.
GFDM	Generalized Frequency Division Multiplexing.
H-ARQ	Hybrid Automatic Repeat Request.
H2M	Human to Machine.
H-NDMA	Hybrid automatic repeat request Network Diversity Multiple Access.
IB-DFE	Iterative Block Decision Feedback Equalization.
ICI	Inter-Cell Interference.
IFFT	Inverse Fast Fourier Transform.
IoT	Internet of Things.
ISI	Inter Symbol Interference.
ITU	International Telecommunication Unit.
LDS	Low Density Spreading.
LoRa	Long range.
LPWA	Low Power Wide Area.
LSAS	Large Scale Antenna Systems.
LTE	Long Term Evolution.
LTE-A	LTE Advanced.
M2M	Machine to Machine.
MAC	Medium Access Control.
MACA	Multiple Access with Collision Avoidance.
MBB	Mobile Broadband.
MIMO	Multiple-input-multiple-output.
MIMO-NOMA	MIMO-NOMA transmission technique.
mMTC	massive Machine-Type Communications.
MPA	Message Passing Algorithm.
MPD	Multi-Packet Detection.
MPR	Multipacket reception.
MT	Machine Terminal.
NDMA	Network Diversity Multiple Access.
NGMN	Next Generation of Mobile Networks.

NOMA	Non-Orthogonal Multiple Access.
OFDM	Orthogonal Frequency Division Multiplexing.
OFDMA	Orthogonal Frequency Division Multiple Access.
OOB	Out of Band.
PAPR	Peak-to-Average Power Ratio.
PDMA	Pattern Division Multiple Access.
PER	Packet Error Rate.
PHY	Physical.
QoE	Quality of Experience.
QoS	Quality of Service.
QPSK	Quadrature Phase Shift Keying.
RACH	Random Access Channel.
RAP	Random Pilot Sequences.
RAW	Restricted Access Window.
RFID	Radio Frequency Identification.
RTS	Request-To-Send.
S-MAC	Sensor MAC.
SC-FDE	Single Carrier with Frequency Domain Equalization.
SC-FDMA	Single Carrier Frequency Division Multiple Access.
SC FDM	Single Carrier Frequency Division Multiplexing.
SCMA	Sparse Code Multiple Access.
SE	Spectral Efficiency.
SIC	Successive Interference Cancellation.
SISO	Single-Input Single-Output.
SNR	Signal to Noise Ratio.
SPR	Single Packet Reception.
TDMA	Time Division Multiple Access.
TTTL	Time To Transmit Latency.
TUOS	Terminal Unique Orthogonal spreading-Sequence.
UEs	User Equipments.

UFMC	Universal Filtered Multicarrier.
URLLC	Ultra-Reliable and Low Latency Communication.
WSN	Wireless Sensor Network.
WUR	Wake-Up Radio.
WURx	Wake-Up Receiver.

Chapter 1

Introduction

In a largely connected world, the number of devices that access the Internet is increasing year by year. During the last decade, the increase in mobile traffic has mainly been caused by the global adoption of smartphones and the corresponding applications, which have caused the cellular networks to move from voice-centered to data-centered services. These applications require high data rate, global access to the Internet, and seamless mobility, which have been the main driver of cellular standards in the past [1].

As technology evolves and new technologies appear, e.g. Internet of Things (IoT), paving the way towards a paradigm of ubiquitous connectivity, the realization of smart cities in which homes, vehicles, and mundane objects are endowed with sensing and communication capabilities will accelerate towards 2020 [2]. Thanks to the advances in communications technology, machines can be connected and reached cost effectively and will soon become an integral part of the global information network [3].

M2M deployments and services are expected to grow exponentially in diverse areas such as transportation, utilities, health, and environment. M2M deployments will generate many new and diverse forms of data traffic with varying requirements in terms of delay, reliability, energy consumption and security. This new scenario poses a new set of requirements not currently supported or optimized by Long Term Evolution (LTE) and LTE-A cellular systems, which have the primary focus of Mobile Broadband (MBB) communications [1].

Therefore, a new generation of mobile networks is needed to attend these novel requirements and services - the 5^{th} generation 5G. 5G is expected to enable a fully mobile and connected society and to empower socio-economic transformations [4]. The Next Generation of Mobile Networks (NGMN) formulated the following vision of the fifth generation [4]:"***5G is an end-to-end ecosystem*** *to enable a* ***fully mobile and connected society****. It empowers value creation towards customers and partners, through existing and emerging* ***use***

***cases**, delivered with **consistent experience**, and enabled by **sustainable business models**."*

Hence in 5G, there is a need to push the envelope of performance to provide, where needed, for example, much greater throughput, much lower latency, ultra-high reliability, much higher connectivity density and higher mobility range. 5G will also operate in a highly heterogeneous environment characterized by the existence of multiple types of access technologies, multi-layer networks, multiple types of devices and multiple types of user interactions. In such an environment, there is a fundamental need for 5G to achieve seamless and consistent user experience across time and space [4].

In addition of handling a heterogeneous environment and supporting a wide range of devices, 5G must support a high variety of use cases like broad-band access in dense areas, massive IoT or massive M2M, extreme real-time communications, lifeline communications and ultra-reliable communications [4]. The latter is drawing quite some attention, as it is envisioned to be a use case not only with significant growth in areas as automotive, health and assisted living applications, but a new world in which industries from manufacturing to agriculture rely on reliable M2M. Another interesting application of ultra-reliable communications involve remote operation and control, that require extreme low latency as well [4]. This subsection is called the URLLC, a type of M2M communication with low latency and high reliability strict requirements.

URLLC is targeted at emerging applications in which data messages are time-sensitive and must be securely delivered end-to-end. However, balancing out low latency and high reliability presents itself as an immense Quality of Service (QoS) challenge, introducing several challenges in the wireless system design. Moreover, attaining the parameters established for this type of communication[1], and supporting these standards for an overwhelming number of devices is hard to accomplish [5].

The work developed in this dissertation intends to address the unique environment introduced in 5G, with the goal of extending an existing M2M Hybrid automatic repeat request Network Diversity Multiple Access (H-NDMA) MAC protocol to make it suitable for URLLC. The MAC protocol is adapted to work with multiple base stations and a power control algorithm is proposed.

1.1 Research goals and contributions

This dissertation explores the capacity enhancement through the spatial densification of the M2M H-NDMA protocol, aiming to provide ultra-reliable low latency communication services for M2M networks. A new version of the M2M H-NDMA protocol is proposed in this thesis, whose goals are:

- Make the M2M H-NDMA MAC protocol suitable for a multi base station scenario;

[1] 3GPP requirements for URLLC includes the hard latency of 1ms over the air interface and the system reliability in terms of successful packet rate delivery of $1 - 10^{-5}$

- Introduce power control schemes into the protocol and optimizing power configurations for the multi base station scenario.

Both of these goals were accomplished, contributing to:

- Analysis of the performance of the M2M H-NDMA protocol in a multi base station scenario and the effect of power spreading in such environment regarding ultra reliable low latency requirements.
- An evolved version of the M2M H-NDMA protocol was designed and a system level simulator was implemented using MATLAB.

1.2 Dissertation's outline

The dissertation structure is as follows: Chapter 2 overviews an literature review, describing the related work to the dissertation scope. It starts by describing existing access schemes that are already in use, and presents other access schemes and MPR techniques that are considered enablers for 5G URLLC. A small description of the state of the art on M2M MAC protocols is given, with a focus on mission critical M2M MAC protocols. Chapter 3 starts by describing the new scenario considered - a distributed antenna system. It then presents the previous M2M H-NDMA protocol and introduces the changes needed to be made for making the protocol suitable to work in the new environment. A detailed power study is conducted, to analyze the performance of various transmissions non orthogonal multiple access (NOMA) power levels and evaluate the best power gap between them, in order to guarantee reliability and reduce delay. Ensuing, it is presented a power control scheme, using different power levels distribution strategies, and an analysis is done of the performance of this algorithm in accordance to the URLLC requirements. Finally, it presents the evolved version of the M2M H-NDMA algorithms, and a user guide to assist future users of the simulator. Chapter 4 summarizes all conclusions made throughout this dissertation and presents suggestions for future work that can be done to improve the protocol.

Chapter 2

Related Work

2.1 Introduction

Until now, from the previous mobile generations up to the current 4G, or as it is called LTE-A, all had an application that propelled the mobile communication market growth, always searching for improvement of the previous generation to keep up with these new apps.

As stated in [6], in the past few years the development of a 5G vision led to the consensus that the latest generation of cellular systems will be driven by various recently rising use cases, related to major drivers such as IoT, Gigabit wireless connectivity, tactile internet and ultra-reliable communications [7], whereas the previous generations of cellular systems design mainly focused the increase of spectral efficiency to indulge bandwidth-hungry applications for human users. As we can tell from these drivers, there will be a new paradigm shift on communication types that 5G will need to address, associated with the encompassing of Human to Machine (H2M) as well as M2M communications.

With the upcoming of new use cases come new requirements. Based on the current trends, 5G mobile networks will have to address defying challenges that are not adequately addressed by state-of-the-art deployed networks [8]: higher capacity, higher data rate, lower End-to-End (E2E) latency, massive device connectivity, reduced capital and operations cost and consistent Quality of Experience (QoE) provisioning.

Specifically, for M2M communications, energy efficiency is a mandatory requirement, since a significant fraction of the devices involved in this type of communication is expected to be battery operated [9], and some of these devices will be deployed in remote locations, so having a long battery life is especially important. Most of the current solutions for low power are designed for short range connectivity, that sometimes might not meet the requirements of M2M applications. Therefore, Low Power Wide Area (LPWA)

technology has been specifically designed with the objectives of low energy consumption and wide coverage [10]. Currently, the Long range (LoRa) is one of the prevailing LPWA technology for building IoT networks, but all fail to provide the URLLC service level.

A novel approach to PHY layer design, with groundbreaking changes is required to provide the services portrayed above. With the ever-increasing growth of M2M communications, the new generation of mobile communications must be able to deal with different types of services with different types of requirements, i.e. providing connection to a user for downloading a video and at the same time allowing M2M that has completely different requirements, and handle this without compromising any of the services. 5G will need to support the diversity of devices and service requirements in a scalable and efficient manner [8].

It is pointed out in [11] that since there will be a need of co-existence of human centric and machine-type services, or perhaps even hybrid of these two, the International Telecommunication Unit (ITU) has established three representative service classes: enhanced Mobile Broadband (eMBB), massive Machine-Type Communications (mMTC), and URLLC. The latter attracted a lot of attention recently because it can provide a real-time interaction among machines and humans, so that services like tactile internet, automated driving and augmented reality can be realized in a near future [12].

Of these three services, the PHY layer design of the URLLC service is by far the most defying and problematic, mainly because it must fulfill two clashing requirements: low latency and ultra-high reliability. If we try to minimize latency, it will imply the use of a short packet, resulting in a serious degradation of the channel coding gain. Nonetheless, if we try to enhance reliability, more resources will be needed, resulting ultimately in latency increase [11].

One can conclude from what is depicted above, that the current mobile generation deployed will not suffice to the arise of these new challenges and requirements. For instance, although various efforts have been made to have the current deployed networks support the IoT needs, many of the envisaged new applications impose requirements of ultra-low latency and ultra-high reliability, which are not easily supported by current networks [8]. Also, mentioned in [11], 3GPP has decided that the physical layer latency of URLLC should not surpass 0.5 ms, and in order to ensure this, the Time To Transmit Latency (TTTL) should be in the order of hundreds of microsecond. Since the TTTL of the LTE-A networks is fixed to 1 msec, it simply cannot satisfy the PHY layer constraint imposed.

The Orthogonality and synchronization requirements of LTE-A PHY layer, based on Orthogonal Frequency Division Multiplexing (OFDM) and Single Carrier Frequency Division Multiplexing (SC FDM) are also obstacles for its use in the new 5G architecture [7]. Synchronism means that the senders operate with a common clock for their processing. Orthogonality means that no crosstalk occurs in the receiver's signal detection process. However, for cases like the mMTC scenario, sensor nodes usually transmit different types

of data asynchronously and traditional OFDM calls for different users to be highly synchronized. Otherwise it leads to substantial interference among neighboring bands [11]. This scenario is an example of a case where orthogonality is destroyed, and as it is referred in [7], this makes the noise pile up without bounds in the OFDM.

Devices and machines involved in M2M generate bursty and sporadic traffic, meaning that they are not always transmitting, and therefore should not be complied to follow the synchronization of the LTE-A PHY layer. Instead, they should only access the network when they have information to transmit [7]. An effective way to do this would be to have the sporadic traffic carried by non-orthogonal waveforms for asynchronous signaling in the uplink.

The LTE-A waveform, with the plenteous guard band to other legacy networks, severely deteriorates spectral efficiency and can prevent band usage [7], resulting from the strict synchronism and orthogonality constraints within the PHY layer. In this new paradigm of uncoordinated interference, new waveforms and access and reception schemes are required, that implement sharp frequency notches, tight spectral masks to not hinder with legacy systems and to deal with uncoordinated interference as well as asynchronous signaling.

This chapter describes the related work on this area. It starts by describing some current multiple access schemes, following by the introduction of Multipacket reception (MPR) and some techniques and concepts related to it, and finally giving a brief overview of the M2M communication requirements, with emphasis on URLLC.

2.2 Orthogonal multiple access schemes

2.2.1 Orthogonal Frequency Division Multiple Access

The expanding of wireless digital communication resulted in a demand for wireless systems that are reliable and have a high spectral efficiency. This demand for higher data rates led to a wider use of the bandwidth, resulting in severe frequency selectivity of the channel, consequently making the Inter Symbol Interference (ISI) more serious [13].

In a single carrier communication system, time equalization is done to get rid of ISI. However, in the wide band channel scenario, the length of the time domain filter to achieve equalization is prohibitively large since it linearly increases with the channel response length [13].

A way to lessen the frequency selective fading in a wide band channel is to use multicarrier techniques. A multicarrier technique [13] subdivides the whole bandwidth of the channel into smaller sub bands, around each subcarrier. With such techniques come many advantages that involve their robust communication and stable interference management, as well as facilitating dynamic frequency reuse techniques and exploring multiuser diversity at smaller granularities than those achievable in CDMA-based networks [14].

The OFDM is a multicarrier modulation technique which uses orthogonal subcarriers to transmit information. As explained in [13], in the frequency domain the bandwidth of a subcarrier is intended to be smaller than the coherence bandwidth of the channel, so each sub-channel is seen as flat fading channel, simplifying the equalization process. In the time domain, by splitting a high rate data stream into multiple lower rate data streams that are parallelly transmitted, OFDM effectively solves the ISI problem in wide band communications.

The LTE and LTE-A employ Orthogonal Frequency Division Multiple Access (OFDMA) as the multiple access scheme for the downlink [14]. Multiple Access is done by allocating different groups of subcarriers simultaneously to different users [13], allowing simultaneous low data rate transmission from numerous users. This technique improves OFDM sturdiness to fading and interference.

2.2.2 Single Carrier Frequency Division Multiple Access

In LTE and LTE-A OFDMA is used for the downlink, as mentioned above. For the uplink, Single Carrier Frequency Division Multiple Access (SC-FDMA) is also used and it is based on Single Carrier with Frequency Domain Equalization (SC-FDE) [13]. The SC-FDE is a single carrier modulation combined with frequency domain equalization, and it is a different solution to the ISI problem.

The main advantage of SC-FDMA versus OFDMA is this type of transmission has lower Peak-to-Average Power Ratio (PAPR), since SC-FDMA spreads the energy of one subcarrier over the range of all subcarriers, prior to the computing of the Inverse Fast Fourier Transform (IFFT). Consequently, the spectral nulls in the channel are reduced with averaging, and PAPR is reduced [13]. Having low PAPR in the transmit waveform results in less power consumption in the mobile station comparing to an OFDMA transmission, thus enhancing the system uplink throughput.

Another reason for using SC-FDMA is also given in [13]: The architecture of the receiver in the case of SC-FDMA is more complex than the architecture of the OFDMA receiver. However, the design of the power amplifier used in the transmitter is simpler in the case of SC-FDMA due to the lower value of PAPR in this case. This observation reinforces the use of SC-FDMA as an uplink scheme, since the power efficiency and complexity are more important for mobile stations than for the base station.

2.2.3 Code Division Multiple Access

Code Division Multiple Access (CDMA) is a technique that has been widely used in wireless communication since it has great bandwidth efficiency.

CDMA is a spread spectrum technique that assigns each node with a unique code to spread each bit of the data sequence. In CDMA networks, the channel can be partitioned

by different codes. A sub-channel can be represented by a spreading code to allow simultaneous transmissions to occur on the same frequency bandwidth without interference [15].

Another way to explain it is given in [16]: CDMA applies a set of orthogonal codes to encode the data from different users before transmitting in a shared communication media. Hence, it allows multiple users to access the communication media at the same time by separating data from different users in the code domain.

2.2.4 Other Modulations for OMA

In this section, we will briefly give an overview of some new types of modulation techniques envisioned for 5G networks. The modulations mentioned below will fall into two categories: Modulations based on pulse shaping and based on sub-band filtering.

Pulse shaping, which is also considered as a subcarrier-based filtering, can successfully reduce Out of Band (OOB) leakage [17]. According to the Heisenberg-Gabor uncertainty principle, the time and frequency widths of the pulses cannot be decreased at the same time. Hence, waveforms based on pulse shaping are typically non-orthogonal in both time and frequency domains to uphold high spectral efficiency [17]. Fundamentally, modulations based on pulse shaping try to confine transmit signals within a narrow bandwidth and in this manner alleviate the OOB leakage so that they can work in an asynchronous scenario with a narrow guard band [17]. Two types of pulse shaping modulation are Filter Bank Multicarrier (FBMC) and Generalized Frequency Division Multiplexing (GFDM).

Sub-band filtering is another technique to reduce OOB leakage. In general, modulations based on sub-band filtering achieve better performance in comparison to traditional OFDM [17]. Two types of sub-band filtering are Universal Filtered Multicarrier (UFMC) and filtered OFDM (f-OFDM).

All these access schemes are different and have their strengths and weaknesses, but they have one thing in common: none of them considers collisions or if collision happens, they simply discard the packets. In the following section, the collision problem will be tackled and shown how to make use of it.

2.3 Multipacket Reception

In traditional communication systems, receivers can only receive a packet from each source at a time. These types of systems are classified as Single Packet Reception (SPR). Under the collision model, the capacity of a wireless network is restricted for the most part by the concurrent packet transmissions. This means if there were to be concurrent multiaccess and simultaneous transmissions, it would result in fruitless collisions and a meaningful degradation of the network throughput, and retransmissions usually

worsen the situation, as the number of packets that collide escalate, so does the number of retransmissions, contributing to further degradation of the throughput [18].

However as stated in [19], there is no fundamental reason that collided transmissions cannot be recouped by other means, for example, coding and signal processing. The advent of multiaccess techniques such as CDMA and multiuser detection led to a new examination of random access under a multiuser PHY layer [19].

In the 90's, a new MPR model was proposed, where the main premise that differentiates from the collision model, is given simultaneous transmissions, the reception can be described by conditional probabilities instead of deterministic failure [19]. MPR systems are capable of simultaneous decoding of multiple packets from more than one source concurrently, even if a collision arises, it is still possible to decode the packets that were transmitted [18]. Figure 2.1 represents the classification of MPR techniques suggested in [18]. These techniques are divided in three main classes, corresponding to where the MPR should be enabled. This classification is given based on three perspectives:

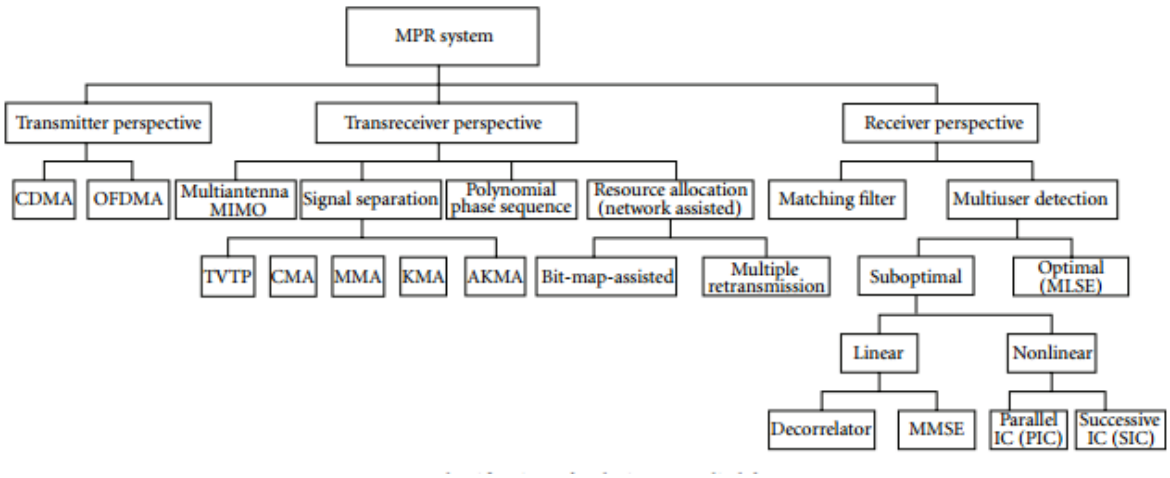


Figure 2.1: Categorization of techniques applied for MPR (adapted from [18])

- **Transmitter perspective**: this type of techniques demands a significant effort by the transmitter. The idea behind this technique is to separate different signals into orthogonal signaling dimensions, allowing multiple users to share the same channel. CDMA and OFDMA are examples of this class [18].

- **Transceiver perspective**: In this type of techniques, transmitters and receivers ought to collaborate on some operations to enable MPR. An example of this class is multi antenna MIMO.

- **Receiver perspective**: This type of techniques involves solely the receiver for decoding several packets simultaneously. Comparing to the previous classes, this one comes closer to the ideal of MPR, which is to shift the responsibility from transmitters to receivers [18].

A fleeting description of some the techniques that fall on the classes presented above is given in the following subsections. It will describe some techniques such as MIMO, Network Diversity Multiple Access (NDMA) and H-NDMA, that is a hybrid solution, which means that uses more than one of the techniques mentioned above. It will also be presented a brief description on Non-Orthogonal Multiple Access (NOMA) and some use cases.

2.3.1 MIMO

Multiple-input-multiple-output (MIMO) is a technique for sending and receiving various data signals simultaneously over the same radio channel, exploiting multipath propagation. A multi-antenna MIMO system can accomplish MPR by taking advantage of the spatial diversity of the transmissions. In such a system each antenna corresponds to a different channel characteristic *h(t)*, implying that a packet sent from one antenna can be distinguished from the one sent with another antenna [18].

MIMO has been an essential element for current mobile generations and wireless standards, and recently it has been considered a key enabler for 5G, but the concept is taken to a different level – massive MIMO. In [20], it is defended that massive MIMO or also known as Large Scale Antenna Systems (LSAS) along with the implementation of Random Pilot Sequences (RAP), will be an important enabler for 5G: By using many antennas at the Base Station (BS)s that concurrently serve many devices through spatial multiplexing, channel hardening can be achieved, eliminating small-scale fading as well as facilitating spatial multiplexing to many devices simultaneously. Each device effectively has an exclusive focused data beam, that does not suffer from small-scale fading and interference.

This is great for a mMTC scenario, but for the URLLC case, the authors in [20] suggest assigning specific pilot sequences to URLLC, since the number of URLLC devices would be smaller than the number of pilot sequences available. When critical low-latency data arrives, it can be sent with very high reliability, in a grant free kind of scheme. This means that LSAS can be an enabler for URLLC, due to the fact that it can provide a large diversity order as well as large spectral efficiency simultaneously [12], while meeting the latency and reliability requirements. In conclusion, LSAS can be an important enabler for URLLC in the 5th generation mobile networks.

2.3.2 NOMA

In the sections above, some OMA schemes were presented that are in use in today's current mobile generations. However, when we talk about massive simultaneous random access and low latency, two very important requirements for the 5G mobile generation, it is clear that current OMA schemes will not suffice.

The main reason for this is that in OMA techniques, users in each cell are allocated the resources exclusively [21]. This means that there is no user-interference, but it also

means that resources are not shared, meaning a user with poor channel conditions has a resource block for its exclusive use, and does not take advantage at its fullest, leading to poor Spectral Efficiency (SE). This combined with the scenario of having tens of thousands User Equipments (UEs), all trying to perform random access procedure for uplink access, leads to network congestion, unforeseen delay and high power consumption [22].

NOMA has been recently recognized as a promising technique for improving spectral efficiency. The main premise of this technique is the fact that it uses the power domain for multiple access, thus allowing users to use resources simultaneously. Taking the example of OFDMA, where subcarriers are allocated to specific users, if those have poor channel conditions it will affect the SE negatively. With NOMA, each user can access all subcarrier channels, therefore bandwidth resources assigned to users with poor channel conditions can still be accessed by the ones with stronger channel conditions, significantly improving the SE [23].

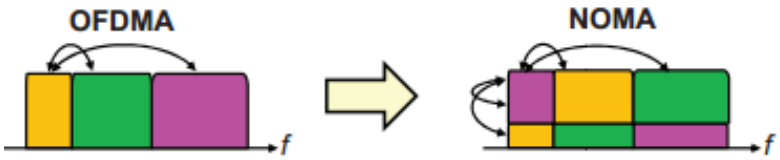


Figure 2.2: frequency/power domain user multiplexing using NOMA (adapted from [24])

This means NOMA can potentially provide the chance to meet the demanding 5G requirements, as we can tell by some of its key features [17]:

1. **Improved SE**: NOMA exhibits high SE, since each resource block can be exploited by multiple users.

2. **Ultra-high connectivity**: supporting multiple users within one resource block means NOMA can conceivably bolster massive connectivity for an immense number of smart devices, especially for IoT scenarios and M2M.

3. **Relaxed channel feedback**: In NOMA, perfect Channel State Information (CSI) is not required at the BS, needing only the received signal strength in the channel feedback.

4. **Low transmission latency**: for uplink NOMA, there is no need to schedule requests from users to the BS (in the time domain), like in OMA schemes. Therefore, a grant-free uplink transmission can be done in NOMA, lowering drastically the transmission latency.

Most of current NOMA schemes can be put into three categories: Power domain NOMA, code domain NOMA and multiplexing in multiples domains, meaning it uses 2 or more domains to multiplex users (i.e. power, code, spatial). Brief introductions to some use cases within these categories are given below.

2.3.3 Basic power domain NOMA

Power domain NOMA is about supporting multiple users within the same resource block, and distinguishing them with different power levels, resulting in more connectivity and higher throughput given limited resources [17].

Let us take figure 2.3 as an example. The BS will send a superimposed mixture containing two messages for two users. Unlike conventional schemes, NOMA users with poor channel conditions get more transmission power.

The message to the user with the weaker channel condition is allocated more transmission power, which guarantees that this user can distinguish its message directly by considering other user's information as noise. Still, the user with stronger channel condition needs to first detect the message for its partner, then subtract this message from its observation and finally decode its own information. This procedure is called Successive Interference Cancellation (SIC) [23].

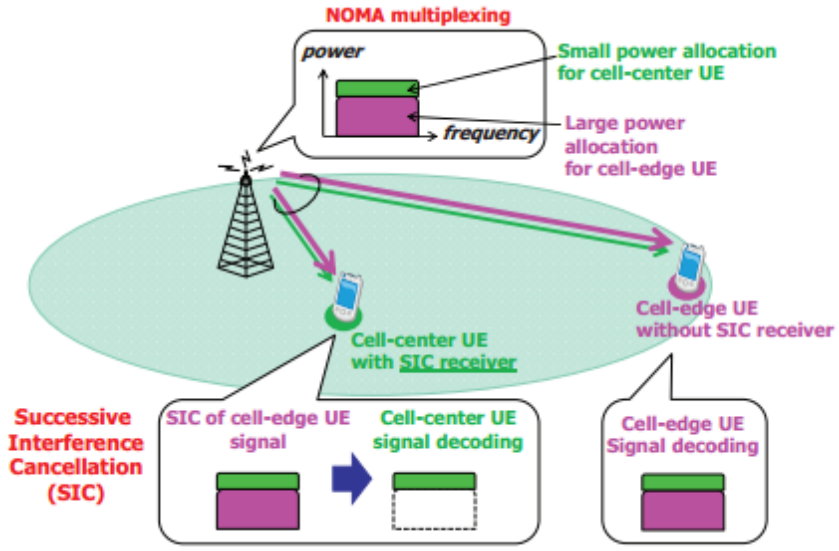


Figure 2.3: Illustration of downlink NOMA with SIC (adapted from [24])

Using NOMA with SIC has been proven to be beneficial. The results in [22] show that applying a protocol based on these two techniques to deal with random access, they achieved 30% more throughput compared with an orthogonal random access scheme, for a large number of user equipments.

Another use case of these two techniques is cooperative NOMA, where the users with stronger channel conditions act as relay to aid users with frailer channel conditions [17], [23]. After the stronger user carries out SIC for the weaker user, the stronger user acts as a relay to forward decoded information to this weaker user. So, it receives two copies of the message from different channels, thus improving the reliability of this user with the weaker channel.

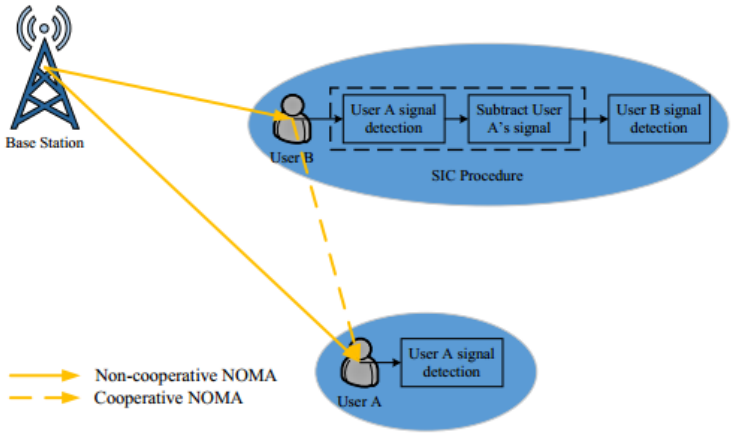


Figure 2.4: Illustration of cooperative NOMA (adapted from [23])

2.3.4 Code domain NOMA

Code domain NOMA is about supporting various transmissions within the same time-frequency resource block by doling out different codes to different users. This kind of NOMA technique has a certain spreading and shaping gain, at the expense of additional signal bandwidth as compared to power domain NOMA [17].

Some use cases of this type of technique are Low Density Spreading (LDS) and Sparse Code Multiple Access (SCMA). The basis of these techniques are essentially the same idea that one user's information is spread over numerous sub-carriers, but the number of subcarriers doled out to each user is lower than the aggregate number of subcarriers, and this low spreading (sparse) guarantees the number of users using the same subcarrier is not too big, so that the system complexity remains manageable [25].

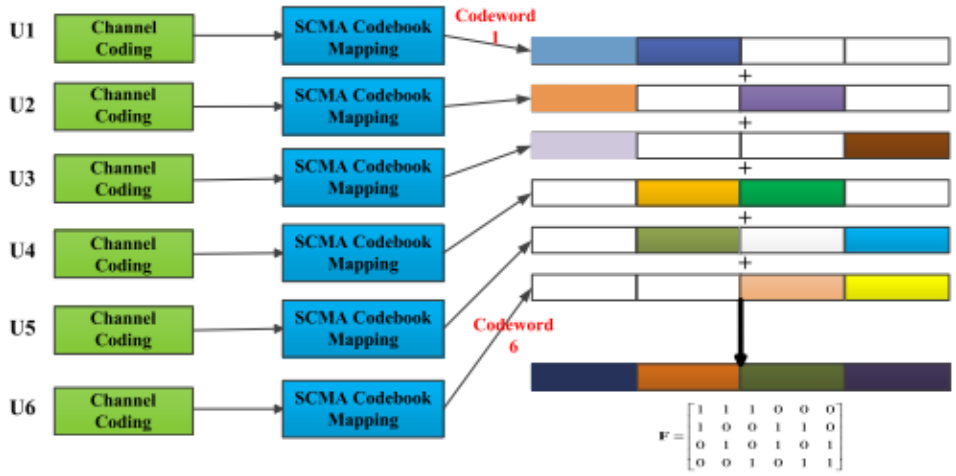


Figure 2.5: An example of an SCMA system with six users and four subcarriers (adapted from [25])

In figure 2.5, we can see an example of an SCMA system with six-users and four

subcarriers. The sparse feature is visible due to the fact each user only employs two subcarriers (the ones in each column of the matrix). It is important to note that since each user's message at different subcarriers are encoded together, SCMA requires joint decoding at the receiver, where a Message Passing Algorithm (MPA) is used to guarantee low complexity. This is an important feature of SCMA, setting it apart from power domain NOMA, as joint decoding rather than SIC is employed [25]. However, even with MPA, the complexity might still be high for user devices. In this case SCMA considers clustering based on the CSI and allocating different powers to different clusters. Thus when transmit powers vary among different clusters, SIC can be applied, which resembles power-domain NOMA, and within each cluster, MPA can be applied to distinguish users. Combining these two techniques can diminish the complexity of the receiver greatly [17].

Code domain NOMA techniques are able to support large levels of connectivity, but they also introduce redundancy, and for a massive device scenario with low latency requirements, it might inevitably degrade the SE at a certain point [21].

2.3.5 Multiplexing in multiple domains

NOMA techniques are not limited to use only one domain to multiplex users. Some solutions for NOMA have been proposed to multiplex in multiple domains, like power, code and spatial with the end goal to provide massive connectivity for 5G networks [17]. In this section, we will briefly introduce schemes that utilize more than one domain to multiplex users: Pattern Division Multiple Access (PDMA) and MIMO-NOMA transmission technique (MIMO-NOMA) transmission.

In PDMA, non-orthogonal patterns are allocated to different users to carry out multiplexing, which are designed in the multiple domains of power, code and space. PDMA resembles SCMA, being the main difference the fact that the amount of resource blocks used by each user can differ, i.e. the number of subcarriers occupied by one user is not necessarily much smaller than the total number of subcarriers [25]. This means one user could be able to transmit and receive on all subcarriers, unlike SCMA that strictly imposes the sparse feature.

MIMO-NOMA transmission basic idea is to extend the use of NOMA to the case in which both BS and users are equipped with multiple antennas, which results in this MIMO-NOMA combination [23]. The application of MIMO to NOMA is essential, since the spatial degrees of freedom empowered by MIMO are pivotal for meeting the performances requirements of 5G networks [25].

However, there some issues with this type of multiplexing. The main difficulty is the complexity of the receivers. The extension of NOMA with spatial multiplexing to more than two users with multiple carriers requires user clustering and resource allocation in multi-dimensional space, making that an analytical and computational challenge [23]. User ordering in MIMO-NOMA scenarios is also a challenging task, since in Single-Input Single-Output (SISO) case the user's channels are scalar, but when nodes are equipped

with multiple antennas, the user's channel are vectors or matrices, making user ordering in consonance with their channel conditions in a SISO fashion hard [25].

In conclusion, MIMO-NOMA techniques are great for providing more degrees of freedom and improving system throughput, but at the cost of more complexity at the user side, that in some scenarios might not be feasible (i.e. having multiple antennas in a small sensor making the receiver very complex).

2.3.6 Network Diversity Multiple Access

NDMA is a protocol that handles collisions by using time diversity multipacket reception.

NDMA obliges all terminals engaged in a collision of P packets to retransmit their packets P – 1 times. The P transmissions are required to adequately separate P colliding packets [26]. Therefore in NDMA, packets with collision are not discarded as in the conventional protocols, being instead stored in memory for further processing [27]. There are two classes of NDMA protocols, regarding the way they identify the number of terminals engaged in a collision: Classic NDMA applies terminal-specific orthogonal ID sequences in the packet headers. Blind Network Diversity Multiple Access (B-NDMA) protocols apply a collision multiplicity rank detection algorithm [26].

However, the use of these orthogonal IDs in classic NDMA may lead to bandwidth inefficiency as the user population increases, like in a mMTC scenario. Also, the NDMA MAC protocol is not able to deal with low Signal to Noise Ratio (SNR) scenarios. For these reasons, a H-NDMA (or Hybrid Automatic Repeat Request (H-ARQ) NDMA) protocol is proposed in [26] where terminals are presumed to be low resource battery operated devices and the BS is a high resource device that runs MPR algorithm with H-ARQ error control in real time. The H-NDMA techniques reuse packets with errors from past transmissions to improve packet reception on following transmissions. Figure 2.6 depicts the H-NDMA proposed scheme.

Results from simulations made in [26] show the proposed H-NDMA protocol improves the network capacity and also decreases packet delay and energy expenditure, compared to basic NDMA protocol, despite some performance degradation due to misdetections and false alarm errors. H-NDMA presents itself as a viable option for the 5G mobile networks. Still, since H-NDMA requires a somewhat strict synchronization in time for terminals to identify the epoch beginning, in the URLLC case this characteristic might be a disadvantage, causing some delay that might not be tolerable for mission critical M2M communications (i.e. remote surgery). However, some breakthrough was made in [28] where the authors adapted the H-NDMA protocol to provide Extreme Low Latency (ELL) for M2M services and advanced power saving mechanism with good results as long as the total load is controlled.

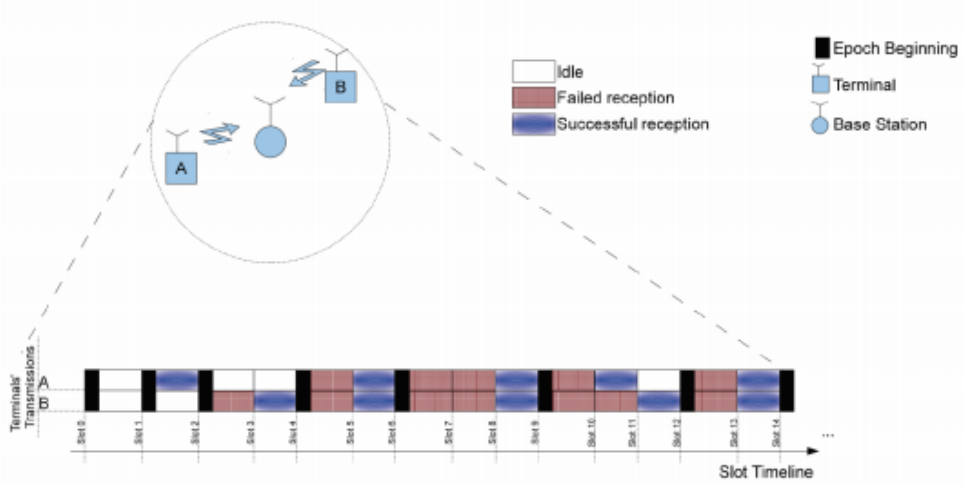


Figure 2.6: H-NDMA MPR scheme (adapted from [26])

2.4 MAC protocols for M2M communications

In the sections above, we have discussed some PHY layer-based techniques that could fulfill the requirements of M2M in 5G. However, to fully take advantage of the applications given by M2M communications, all layers must provide a certain level of support to attain the service requirements.

In this section, MAC layer requisites regarding M2M communications will be described, and classification of MAC protocols will be given, providing some examples for each class. We will address with some depth the mission critical MACs, for Wireless Sensor Network (WSN)s scenario. WSNs comprise of small self-sufficient devices called nodes or motes that gather information such as pressure, temperature or vibration from their physical surroundings [29]. Finally, an introduction to an energy saving concept – the Wake-Up Radio (WUR), will be given.

2.4.1 The M2M MAC requirements

The MAC layer is primarily in charge of channel access for nodes within a network that use a shared medium. The critical MAC layer challenge for M2M communications lies in facilitating channel access to an enormous amount of devices while supporting the diverse service requirements and unique traffic characteristics of devices in M2M networks [9].

Therefore, for supporting M2M the MAC protocols must be designed with a rich set of requirements to address the needs of the overlaying applications and scenarios. Those requirements are the ensuing [9]:

- *Data throughput* - MAC protocols for M2M communications must be highly efficient and have an adequate throughput, according to the application needs. Since channel/spectrum resources are limited and having many devices accessing the

channel, MAC protocols should attempt to minimize the time wasted due to collisions and control messages exchange. Equivalently, the throughput must be high to accommodate the very large number of devices.

- *Scalability* – Scenarios with M2M communications are anticipated to have several nodes. The node density is foreseen to increase as the deployment of application scenarios with M2M communications becomes more prevalent. Besides, the network conditions might always be in constant changing, having nodes leaving and entering. It is then critical that the MAC protocol be effortlessly adaptable and balanced smoothly to changing node densities with slight or no control information exchange and maintained fairness even after addition of new devices.

- *Energy efficiency* – This is one of the most important design thoughts for three reasons, which are: 1) the reality that many of the devices in M2M networks are predicted to be battery operated; 2) the economic impact of the power usage by the communication infrastructure; and 3) the environmental impact of the power expended. Collisions during channel access are a major cause of power consumption that ought to be lessened to the best degree conceivable. Conventional methods to reduce MAC energy consumption comprise of lessening collisions, sleep scheduling, power control and reducing idle listening.

- *Latency* - For many of the applications that depend on M2M communications, network latency is a crucial factor that rules the effectiveness and utility of the services, especially when we talk about applications or scenarios that rely on tight latency and reliability requirements to be successful, like intelligent transportation systems or e-health applications. In addition, the MAC protocol should guarantee long-term as well as short-term fairness, so that all devices can have an equal chance (or proportional to their priority) to send their message.

- *Cost effectiveness* – To make M2M communications-based systems a reality, the devices must be cost effective so that it is affordable to deploy them. A MAC protocol that has many desired properties but that relies on complex and pricey hardware is not convenient. Hence the MAC protocol should be designed to function effectively on simple hardware, especially considering the budget saving potential it can have when it comes to large scale deployments.

- *Coexistence* - A significant fraction of the access networks for M2M communications are expected to operate in the unlicensed bands. This means it is expected for many M2M access networks to be deployed near and independently in the same unlicensed band. Besides they also will have to coexist with other networks that traditionally operate in the unlicensed band (e.g. Wi-Fi and Bluetooth). Therefore, issues such as collisions due to the hidden terminals from neighboring networks must be addressed at the MAC layer.

Some of these requirements and issues that come with them, can already be seen in Radio Frequency Identification (RFID) technologies, where it is critical that the tags must be energy efficient, and the tags can be correctly identified without missing a single one [30], which is important to maintain the network scalability and stability. Also, in presence of unreliable channel conditions they still must guarantee throughput and latency thresholds, since the tags (if they are passive, being the case considered for comparisons) are powered for brief moments to exchange information and must do it with a high degree of success in that time. These issues, most commonly known as counting problems or missing-tag detection as well as energy-time tradeoffs are explained and contributions to them are made in [30], [31].

2.4.2 Classification of M2M MAC protocols

The MAC protocols can be classified on the basis of the underlying mechanism of collision avoidance and the organizational approach to sensor nodes [32]. One can classify MAC protocols based on architecture (distributed or centralized), based on the mode of operation (random access, slotted access, frame slotted), and many other types of classifications can be made [32].

For the sake of simplicity, most of the literature settles that current MAC protocols can be generally classified as contention-free, contention-based, and hybrid protocols that combine aspects of contention-based and contention-free, trying to leverage their strengths and mitigate their weaknesses [9].

2.4.2.1 Contention-based MAC protocols

Contention-based MAC protocols are among the simplest protocols in terms of implementation and setup [9]. In these protocols, the nodes compete for the channel in several ways to acquire the channel and transmit data. However, the biggest downside of this approach is its severe energy inefficiency due to collisions, idle listening, control packet overhead and overhearing[32].

This is especially perceived when the number of devices contending is very high, i.e. tens of thousands, causing serious latency due the increasing number of packets collisions, consequently snowballing energy consumption and affecting the scalability. Nevertheless, in [32] it is pointed out that being adaptive is the key to efficient energy usage, whether these adaptations are made at traffic management or topology level.

An example of a contention-based protocol is the Fast-Adaptive Slotted ALOHA (FASA). Originally proposed in [33], in this protocol, the network condition regarding the number of backlogged devices N_t is estimated by using drift analysis on the access results of the past slots [9]. Using this information, each node in the network is then given $1/N_t$ as its transmission probability in each slot. The BS is in charge of estimating the number of backlogged nodes and reporting the transmission probability to backlogged devices [9].

2.4.2.2 Contention-free MAC protocols

Contention-free or scheduled-based MAC protocols, erase the collision problem by allocating in advance transmission resources to the nodes in the network [9]. Some of the known contention-free protocols include CDMA and Time Division Multiple Access (TDMA).

The main advantage of contention-free MAC scheme is energy efficiency due its features such as the built-in duty cycle, being collision-free and the absence of control-packet overhead [32]. However, in the context of M2M communications, these protocols lack flexibility and scalability, having a hard time adapting when the number of nodes in a network varies or the load is bursty [9], [32]. Since a contention-free MAC scheme needs to rigidly follow the assigned communication time slots, it requires frame synchronization which involves the complicated tasks of slot allocation and schedule maintenance, leading to additional latency that decreases the throughput, and increases the nonproductive consumption of energy [32].

An example of a contention-free MAC protocol is cognitive polling, proposed in [34], a protocol that uses cognitive radio techniques and considers M2M devices secondary users [9]. The protocol is assumed to function in a scenario where network access is frame-based and the BS broadcasts the information regarding the resource allocation to the primary users in each frame [9]. In this protocol, M2M devices listen to these broadcasts and use the vacant resources for communicating among themselves. Still, the protocol does not grant any assurances on the throughput and delays the M2M devices undergo, and is inefficient for M2M devices with bursty traffic [9].

2.4.2.3 Hybrid MAC protocols

As it was discussed above, contention-based protocols adapt with ease to changing network scenarios and are more appropriate for networks with low loads. On the other hand, contention-free protocols get rid of collisions and can achieve a greater channel utilization at higher loads [9]. To capitalize on the advantages of both categories, hybrid protocols have been proposed that incorporate aspects of the contention-free and contention-based protocols.

These protocols work by switching between scheduled access at high loads and random access-based operation at low loads to avoid the degraded throughput and collisions of random access protocols at high loads and low channel utilization of schedule access at low loads [9]. Hybrid protocols seem a promising approach for designing MAC protocols for M2M communications. The main drawbacks of this type of protocols is their scalability, since many M2M scenarios are envisioned to have node densities that might be several orders of magnitude higher than the currently deployed ones. At such high densities, the frequency of collisions during the random access-based reservation phase of hybrid protocols turn into the bottleneck that keeps the network from accomplishing

a high utilization [9]. In addition, the advantages hybrid protocols provide come at the cost of protocol complexity, and such complexity might not be feasible [32].

An example of a hybrid MAC protocol is the IEEE 802.11ah. Its access mechanism operates as follows [35]: every station connected to the Access Point (AP), receives a unique identifier – Association identifier (AID). The AID can be used to determine which stations can access the medium. In particular, IEEE 802.11ah presents the concept of Restricted Access Window (RAW) [35], during which just certain stations are permitted to contend based on their AIDs. This restriction improves the access efficiency, corroborated by the number of stations whose transmissions do not experience collisions. RAW is split into slots, and the number of them as well as their duration and assignment are indicated in the beacon frame that is periodically sent by the AP to the stations. Every slot can be assigned to a single or multiple stations, and only active stations contend in the assigned slots [35]. An example of a RAW slot is presented in figure 2.7.

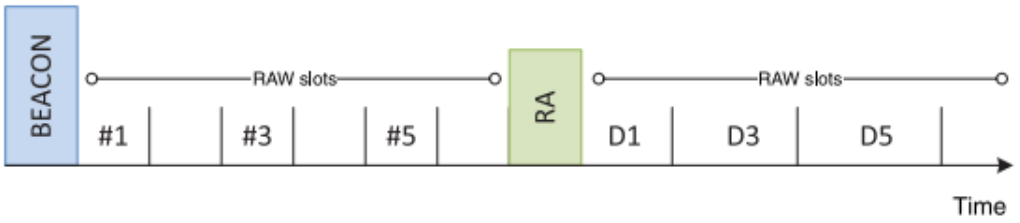


Figure 2.7: Beacon frame announcing a RAW with six slots followed by a RA frame that allocates slots for stations 1,3 and 5 (adapted from[35])

After viewing the three types of MAC protocols, one can conclude that all of them have their strengths and weaknesses, but when it comes to fulfilling M2M communication requirements, hybrid protocols clearly have an edge over the other types, although they still need to be refined and enhanced. A taxonomy of more M2M specific protocols and a table comparing some of them according to various parameters are presented below respectively in figures 2.8 and 2.9.

One can conclude from figure 2.9, that the M2M protocols presented meet and miss some of the parameters. The Code Expanded Random Access (CERA) protocol proposed in [36], is based on an alteration of the LTE dynamic Random Access Channel (RACH), and intends to provide support for a greater number of devices as compared to LTE, without increasing the resource requirements [9]. Although it achieves high throughput and low latency, and can handle bursts, it is somewhat hard to scale, and it is not very energy efficient. The IEEE 802.11ah, already presented above, does have a better scalability but has a higher latency than CERA. FASA is very simple to implement and has a low cost, but it performs worse in the rest of the parameters compared to other protocols in the table.

Cognitive polling is able to achieve high energy efficiency due to M2M devices listening to the broadcasts mentioned above and using unoccupied resources to communicate

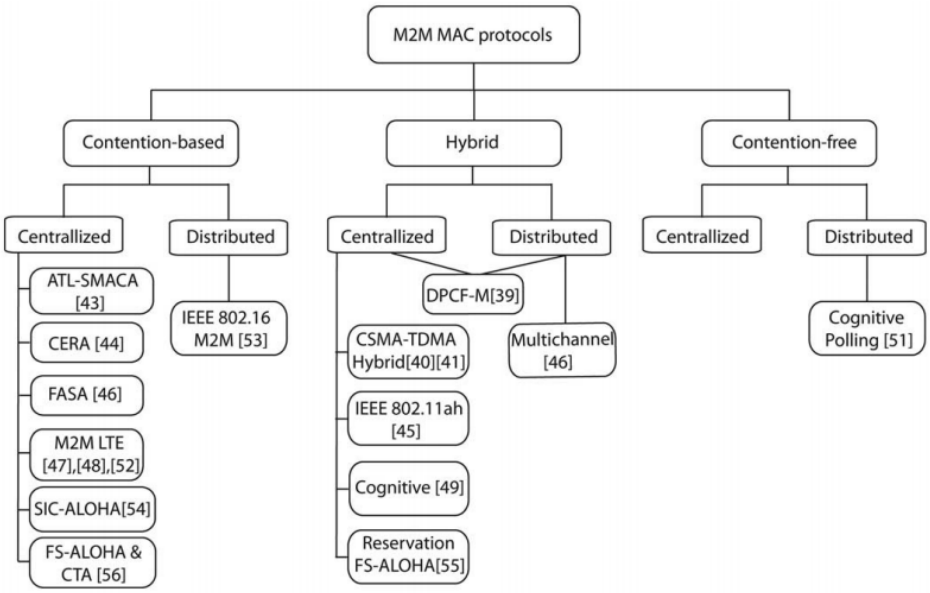


Figure 2.8: Taxonomy of M2M MAC protocols (adapted from [9])

Protocol	Throughput/ utilization	Scalability	Energy efficiency	Latency	Cost	Burst Handling
CERA [4]	High	Moderate	Moderate	Low	High	Yes
IEEE 802.11ah [4]	High	High	Moderate	Moderate	Low	No
FASA [4]	Low	Low	Low	High	Low	No
Cognitive polling [4]	Moderate	Low	High	High	High	No
M2M LTE [4]	Moderate	Low	Moderate	Moderate	High	No
ATL-SMACA [4]	Low	Low	Low	High	Low	No

Figure 2.9: Comparison table of M2M MAC protocols (adapted from [9])

among themselves, having no need to do spectrum scanning to find unused resources [9]. However, this protocol has a very high latency. M2M LTE protocol proposed in [37], suggests simplifications that improve the MAC layer in LTE-A efficiency and avoids unnecessary control overhead. However, contention resolution and collision possibility are not discussed, thus having low scalability. Adaptive Traffic Load slotted MACA (ATL-S-MACA) protocol proposed in [38], is a modification of the Multiple Access with Collision Avoidance (MACA). Nonetheless, when comparing to the rest of the protocols, one can see it performs just as worse as the FASA protocol.

One interesting observation is that if it is desired a high throughput, low latency can be provided but it is impossible to have high energy efficiency. If it is desired high energy efficiency, throughput will go down and latency will go up, as it is shown in figure 2.9 by the red markings. This means, that if it is desired high throughput, low latency and on top of that high energy efficiency, the three big requirements for URLLC, the M2M protocols shown cannot provide that. In the following section, the mission critical case

will be described, and MAC protocols designed for WSNs that fulfill the requirements will be presented.

2.4.3 URLLC: Mission critical MAC for WSNs

As sensors have limited energy resources and WSNs are deployed in places where it is not expected any human intervention to happen for a long time, for this application domain (e.g. environmental monitoring), the main design goal is to have energy efficiency [29]. Therefore, current network protocols excel in the energy efficiency but only providing simple best-effort data delivery.

With WSNs extending their support to a number of other application domains like patient monitoring, industrial process monitoring or military target tracking, the aforementioned best-effort data delivery is no longer suitable for these application areas [29]. Alternatively, improved data delivery is needed since the applications can only operate appropriately if data arrive in a timely and reliable fashion, e.g. a WSN for a battlefield surveillance loses its usefulness if information regarding an incoming enemy does not arrive in time [29].

These new types of WSN applications are so-called mission critical applications. In [29], mission-critical WSN applications are defined *as applications demanding data delivery bounds in the time and reliability domains*. To support these applications the network must provide both timely and reliable data delivery. Delay, reliability and throughput must be considered during network design.

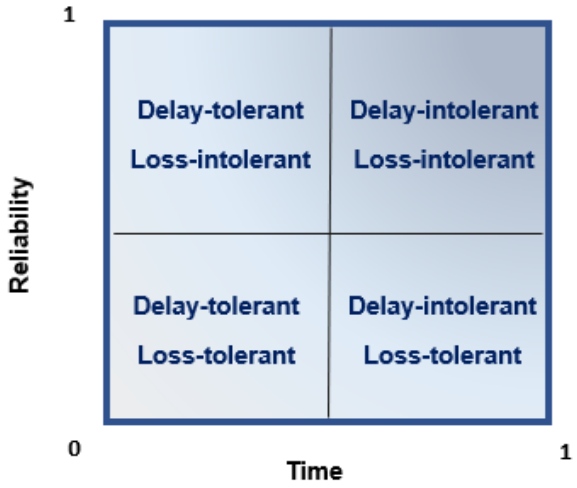


Figure 2.10: Different existing application classes (adapted from [29])

Figure 2.10 illustrates the different existing application classes. Attending to the definition of mission-critical applications provided above, one can easily infer that mission-critical applications are in the delay-intolerant, loss-intolerant class. This class demands

strict performance in both time and reliability domains [29], e.g. monitoring transmission pipelines in an oil refinery. Should a sensor detects an hazardous overpressure in a pipe, the alarm message must be sent in a reliable and timely manner to an actuator that handles a shutter valve.

Therefore, these particular WSN applications have two characteristics in common [29]:

1. Energy efficiency cannot be the only design concern;

2. Best-effort data delivery is not sufficient.

In [29], a myriad of MAC protocols are reviewed and placed in the classes, represented in the figure 2.10, regarding if they are a delay-aware protocol, designed to decrease or guarantee message transfer delay [29], a reliability-aware protocol, designed to increase or guarantee message transfer reliability, or if they are both - a delay and reliability-aware protocol, hence suited for mission critical scenarios. From all the reviewed protocols in [29], two of them stand-out : the Burst MAC protocol and the GinMAC protocol. These protocols have been studied by comparing them to the Sensor MAC (S-MAC) protocol. For the sake of completion, a brief explanation of S-MAC is given before discussing the two mentioned protocols.

S-MAC had energy efficiency as its main design goal [29], introducing the periodic duty-cycle principle to decrease idle listening. A node organizes its fixed active/sleep period with neighbors using SYNC packets. During an active period, nodes comply to the IEEE 802.11 standard to convey messages. The protocol uses and exchange of Request-To-Send (RTS) and Clear-To-Send (CTS) packets as a contention mechanism in the active period [29]. The main drawback is that packets carried by S-MAC might experience high delays in a multi-hop network as they probably will need to be queued at a node until the next active period [29]. Figure 2.11 below illustrates the functioning of the S-MAC protocol.

The Burst protocol, named this way in [29], is a static scheduling algorithm that accomplishes both reliable as well as timely data delivery. This algorithm computes and guarantees a delay upper bound of the end-to-end delay per periodic stream (streams are assumed to be periodic), by taking into account the maximum burst length *Bmax* (metric for capturing the stationarity of the link quality) and link interference [29]. More so, it allocates sufficient transmission slots per each link to successfully surpass link burstiness and interference problems, providing supplementary slots for possible retransmission. It also picks a least-burst route that diminishes the sum of *Bmax* over all links in the route. Thus, Burst is a viable option for mission-critical applications, as it accomplishes end-to-end assurances of data delivery in the delay as well as in the reliability domains [29]. However, for Burst protocol to be effective, a wise network planning is necessary before an actual deployment [29], meaning that if the expected load or its density suddenly

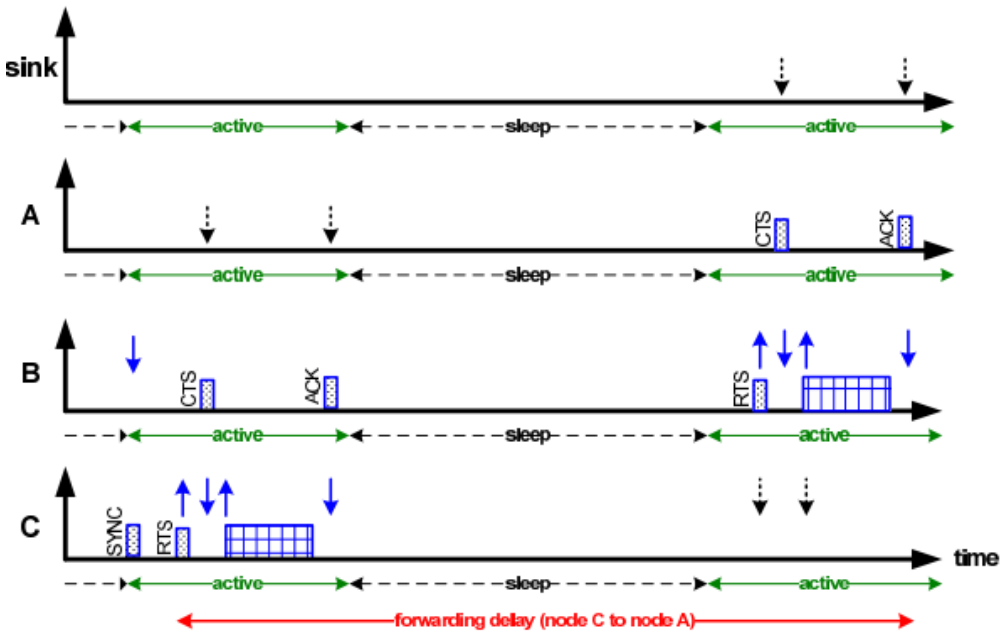


Figure 2.11: S-MAC protocol (adapted from [29])

increases with some orders of magnitude, or if it is deployed in a random network access scenario Burst will not be able to deliver.

GinMAC aims to support a control loop in an industrial process automation system [29]. In this particular setting, sensor data needs to be forwarded to the sink in a manner that does not exceed a defined time bound, and likewise a command from the sink must be carried to an actuator by a deadline [29]. To satisfy these requirements, GinMAC encompasses three features: off-line network dimensioning, an exclusive TDMA schedule and delay conform reliability control.

First of all, in the off-line dimensioning process, channel characteristics, a tree topology, and application traffic are defined. Pre-deployment calculations are also conducted to evaluate the worst-case burst length *Bmax* of all transmission links, being this metric the same as in the Burst algorithm [29].

Ensuing, the result of the dimensioning process, the TDMA schedule, contains unique transmission slots per node, as we can see in figure 2.12, and has a fixed epoch length E. Within this length, each node can forward one message to each actuator [29].

Lastly, based upon the observed channel properties, redundant transmission slots are added in the frame for reliability control without trespassing the computed delay bound of E. These redundant slots are applied to improve reliability through two methods. The first one is to generate temporal transmission diversity through packet retransmission if there is loss. The other one is to obtain spatial and temporal transmission diversity by transmitting duplicates of a packet to another m disjoint tree topologies [29]. An example of a GinMac protocol transmission is represented in figure 2.12.

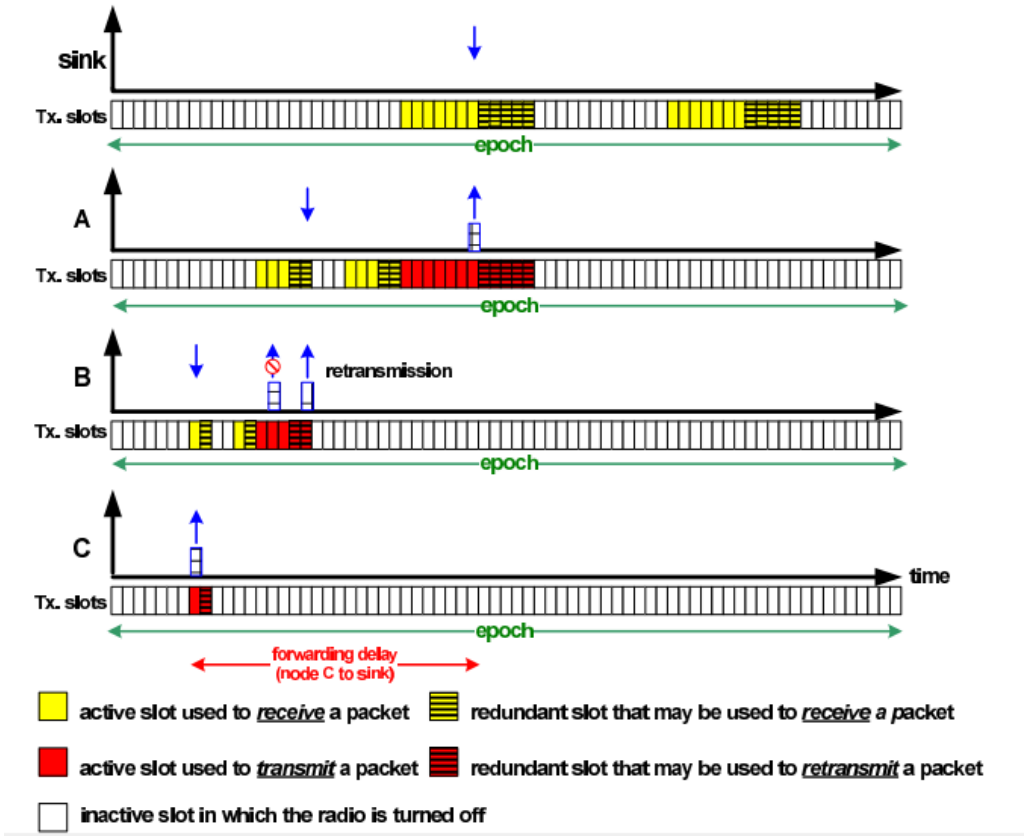


Figure 2.12: GinMAC protocol (adapted from [29])

GinMAC ensures end-to-end guarantees of reliability as well as delay while simultaneously accomplishing energy efficiency [29]. Despite the fact that GinMAC is able to support mission-critical data delivery it presents some limitations, since it is custom-made to a control loop setting, where data must be forwarded to the sink, meaning it follows a specific traffic pattern. Other mission-critical applications, for instance as battlefield tracking have a probability to create distinctive traffic patterns and for that reason cannot be supported by GinMAC [29].

Still, alike the Burst algorithm, GinMAC was designed for a specific scenario, so for a big and dense network, the slot allocation for all nodes might be hard to achieve if there are no slots to give out to all nodes, and with the increase in the number of nodes the time to transmit would be smaller for each node, possibly resulting in numerous collisions. In addition, the delay that would be introduced to make the slots reservation (since it is a TDMA-based protocol), would cause the low latency requisite to fail. Nonetheless, GinMAC can overpass these issues by employing MPR techniques.

2.4.4 The wake-up radio concept

Since sensor nodes that comprise WSNs operate with batteries, replacing or recharging constantly is costly and infeasible [39]. Hence, reducing the power consumption of those

sensors is of utmost importance, enabling the extension of the batteries lifetime and reducing replacement costs. When looking at the energy consumption of the sensor's components, the radio transceiver is one of the highest power consumers, since it is required for it to be in its listening state for messages to be received. This leads to a lot a of idle listening, when the transceiver is listening to the channel to check for an incoming message, consequently causing a big energy waste [40].

Traditionally, this problem of lower power consumption has been addressed through duty cycling methods. Duty cycling is a technique to decrease energy usage in idle mode, by alternating between listening mode to sleep mode, lessening the transceiver's power expenditure [40]. Nonetheless, this method comes with a cost: when radios are switched off, they cannot receive messages. This severely limits the network reactivity [39].

Having the above in mind, we can pinpoint the problem as it being the following: To enable communication among two wireless nodes, the receiver node must be awake when the sender initiates the communication. This is referred as a rendezvous. There are three types of rendezvous schemes [40]:

- **Pure synchronous**: Node's clocks are pre-synchronized so that the wake-up time of each node is known in advance. This obliges to frequent time synchronization, consuming substantial energy, and sensors might wake up even if there is nothing to receive or transmit, leading to idle listening.

- **Pseudo-asynchronous**: Source nodes awake and transmit a preamble signal that indicating the intent of data transmission. Time synchronization for this strategy is not required, however sensors adopt a duty cycle and consume a significant amount of energy with preamble signaling.

- **Pure asynchronous**: The sensor nodes are in deep sleep mode and can be woken up by their neighbors when required with resort to very low power wake up receivers. At any given time a node intends to send a packet, it starts to wake up the destination node using a wake message and only then sending the packet. Hence, wake-up receivers are an answer to the redundant energy expenditure resulting from rendezvous.

This last scheme depicts the WUR concept. This concept uses a separate Wake-Up Receiver (WURx) to monitor the communication channel continuously, while the main radio is kept in sleep mode all the time it is not needed [40]. When a node wants to communicate, it sends a wake-up signal that is detected by WURx, that allows the CPU to wake the main radio to start the communication. An example is provided in figure 2.13.

The first works on these concept that established concrete goals for WUR functioning came from [41], where they show that a wake-up radio scheme outperforms other schemes in terms of a wireless network lifespan.

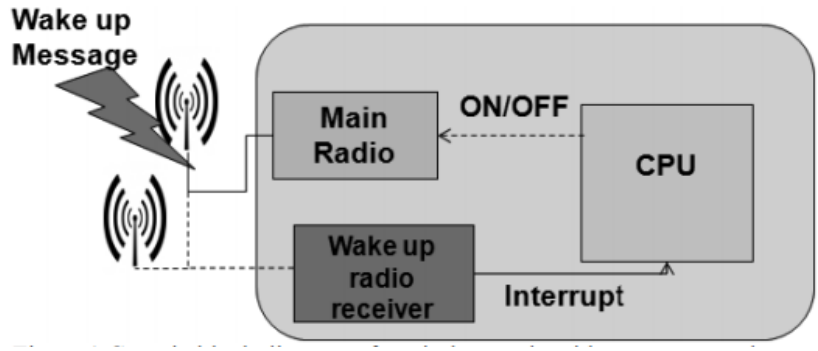


Figure 2.13: Generic node block diagram with a separate wake up radio receiver (adapted from [40])

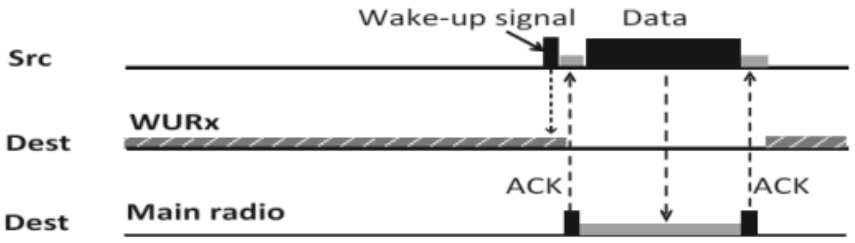


Figure 2.14: Asynchronous scheme using wake up radio (adapted from [40])

This work also considered the fact that one might not want to wake up a node when it is within range of the wake-up signals, even though the node is not actively involved in current events of interest. This phenomenon is called a false positive, and recent works from [39], [40] solved this problem by implementing addressing capability to the sensors. This means that the sender can specify what nodes it desires to wake up when sending the WUR signal. Plus, an addressing mechanism can significantly reduce the power consumption during a network formation phase. If a wake-up receiver can receive some command, The MAC data communication can be reduced for low power consumption [40].

WUR technology is extremely important for M2M URLLC communications, due to the energy saving potential it brings to WSNs, therefore an important enabler for 5G mobile networks. Currently in the IEEE 802.11 working group, the 802.11ba task group is working towards a standardized specification for WUR, solution that also includes important MAC features, making it more than just a power efficient receiver.

CHAPTER 3

System Description

3.1 Introduction

This chapter starts by describing the envisioned cellular network, where the BSs run a Coordinated multipoint (CoMP) MPR receiver algorithm in a DAS. The section starts with a description of the M2M H-NDMA protocol and the MPR receiver, which extend [28] to a multiple BS scenario. The use of multiple BSs connected to support CoMP in the system introduce new challenges and potential. Power control and power optimization approaches are studied for the M2M H-NDMA protocol. Finally the implementation of the simulator is described.

3.2 System rundown

This dissertation extends the architecture proposed in [28], by using multiple BSs to implement a DAS. The additional spatial diversity level introduced leads to increased scalability without degrading the reliability or the latency.

The system is composed by an amount of Machine Terminal (MT)s that send data to the base stations. Signal reception of all those terminals is accomplished by using a Cloud Radio Network (C-RAN) - a software defined radio architecture with the objective of taking away the signal processing functions from the BSs to put them in the cloud. CoMP is used - a form of cooperation in which multiple BSs communicate with each other to cancel out the interference and improve the overall system performance by jointly transmitting/receiving the user's data concurrently [42]. This coordination is done between the MTs associated/main BS - the BS the MT is associated to send data, and the secondary BSs - the BSs in the MT range that will also receive the data. This work focused mainly in the MAC design and the implementation of the system level simulator.

The implementation of the reception algorithms for signal reception handling is modeled using an existing H-NDMA receiver MATLAB model. For illustration purpose, figure 3.1 demonstrates an example of a C-RAN architecture.

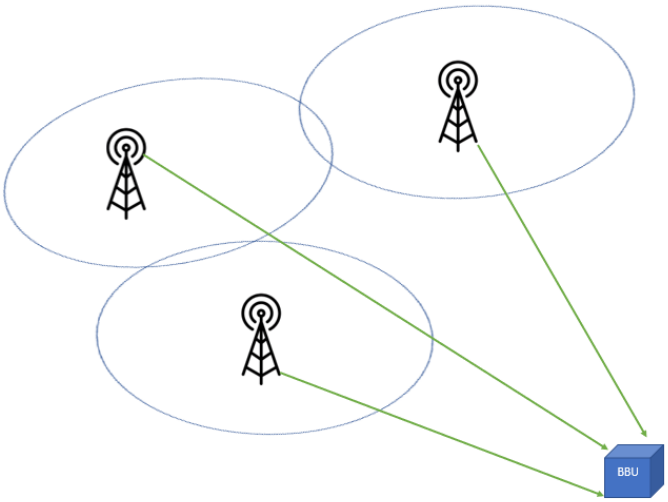


Figure 3.1: Cloud RAN

The MTs are low resource battery operated M2M devices, whereas the BSs are equipped with one or more antennas, and are connected to a cloud through the network fronthaul. Reception algorithms are run in the cloud, that can use Diversity Combining (DC) to deal with packet errors as a result of poor propagation conditions, Multi-Packet Detection (MPD) to deal with collisions and CoMP. In this document, the BSs designate the C-RAN assembly, with the antennas and the clouds.

Terminals access to the medium is implemented by an improved version of M2M-HNDMA protocol [28], which allows asynchronous access of isolated terminals, but coordinates access when multiple terminals intend to access the medium, thus maximizing the success conditions for packet reception.

BSs also have the capability of employing hybrid techniques that incorporate DC and MPD. MTs transmit data packets on time slots determined by the BSs, that also control the transmission power. Perfect channel estimation and synchronization is assumed. Colliding packets on each slot are also assumed to arrive simultaneously, meaning that time advance mechanisms exist to compensate different propagation times, making the offsets in all BSs below the Cyclic Prefix (CP) duration time.

Although the introduction of multiple BSs leads to Inter-Cell Interference (ICI), resulting in performance degradation on cell border MTs, the system handles it by turning ICI into a useful signal, by employing CoMP reception. The DAS uplink transmission is illustrated in figure 3.2, and an illustration presenting the difference between the system considered in this dissertation and the system in [28] is presented in figure 3.3. The interaction between the MT and the network is managed by multiple BSs. In the solution proposed, one BS acts as the main Associated BS, while the others are coordinated by it

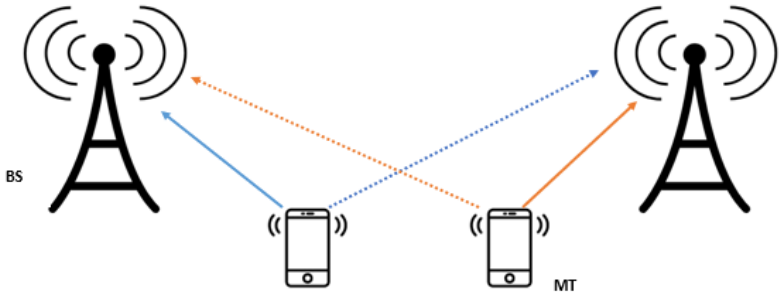
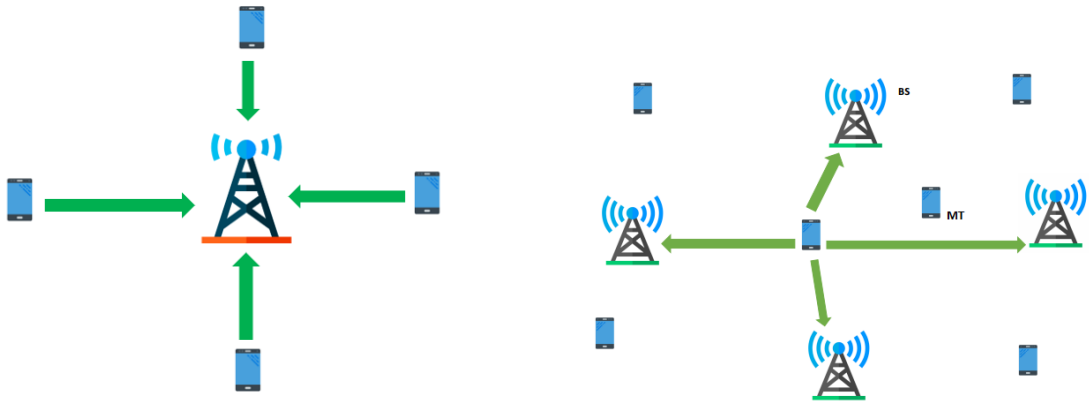


Figure 3.2: Example of uplink transmission in a DAS. The continuous line represents the transmission of the terminal to its associated BS, and the dotted line the transmission to the secondary BS.

and help in the communication.



(a) Previous system scenario (b) Current system scenario

Figure 3.3: Previous system scenario vs current system scenario

3.3 Impact of Spatial Diversity and Power Control

When introducing multiple BSs in the system, a new degree of freedom is unlocked, provided by the spatial diversity. This spatial diversity creates multiple spatial independent paths. For protocol effects, introducing this extra degree of freedom in the M2M H-NDMA protocol, directly correlates with creating more channels that do not require additional time slots. Thus a lower number of retransmissions is required, improving cell edge users data rate.

Combining this with NOMA, for η NOMA levels and κ BSs in a given cell, the number of terminals that could be solved per slot by the receiver would theoretically be $\eta \times \kappa$.

Therefore, the scalability of the system is increased, and overall key URLLC metrics, such as delay and reliability are improved.

However, the introduction of this degree of freedom results in scarcer control of the power received from the terminals. The different paths from each terminal to the BSs produce a spreading on the received power levels. In result, pure NOMA that was applied in a single cell scenario in [28] might no longer be feasible[43]. Let's take figure 3.4 for example, a two cell scenario of a cellular system where $U3$ and $U4$ are served by BS 1 and $U2$ and $U1$ are served by BS 2, assuming a two-user NOMA scheme so that $U3$ is paired with $U4$ and $U1$ is paired with $U2$.

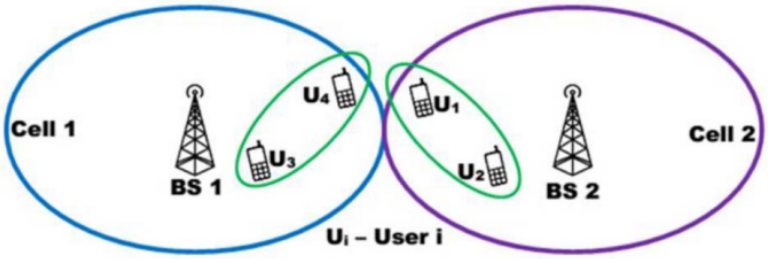


Figure 3.4: Example of network NOMA (adapted from [43])

The main problem is that $U4$ and $U1$ are both regulated only for their associated BSs. However, since they are nearly the same distance their non-main BSs, they will also receive their transmission, although with power not being regulated in this case. The power seen by those BSs would be

$$P_{MT} = E_{N_{MT_i}} - PL_{MT_{i,j}} + PL_{MT_{i,k}}. \tag{3.1}$$

where $PL_{MT_{i,j}}$ represents the pathloss of the MT_i to its associated base station j and $PL_{MT_{i,k}}$ the pathloss of the MT_i to the secondary BS k.

Hence, ICI occurs in downlink transmissions and mutual interference between $U4$ and $U1$ occurs in the uplink transmissions. Straightforward single-cell NOMA solutions are not well-suited to deal with an multi-cell scenario: network NOMA is required [43].

Taking into account the points made above and that in this dissertation the configuration is of a DAS, a power control policy implementation is required. The following sections describe the studies done and measures taken to implement an efficient power policy.

3.3.1 M2M H-NDMA protocol

The protocol used in this dissertation is an evolved version of the M2M H-NDMA protocol in [28], with the objective to provide URLLC for M2M. The protocol is still truthful to the

goals in [28], but it adapts it towards ultra reliable and low latency requirements and to a DAS environment. Hence, the revised objectives can be summarized in three main goals:

- Allow M2M devices to operate using the minimum power possible, without compromising the latency and reliability goals;
- Provide bounded delay and reliability to comply with URLLC requirements;
- High Scalability.

The following sections will describe the M2M H-NDMA protocol used and how the protocol was modified to be suited for these goals and new environment.

3.3.1.1 Protocol Characterization

M2M H-NDMA protocol is a random access protocol optimized to provide efficient MPR interactions and power saving. It inherits the epoch concept from H-NDMA - when a set of MTs begin transmitting, the rest of the MTs are prohibited from commencing transmissions before the end of the reception of the first set. The epoch is characterized by the contiguous set of data slots where the MTs keep on transmitting packets till either reaching success or the maximum retransmissions allowed [28]. Contrary to H-NDMA, no partial acknowledgments are sent. Alternatively, and like in NDMA, a tone transmission by the BSs signals that new retransmissions are in need in the course of an epoch. At any point, the BSs may suspend an epoch by simply turning the tone off. At the end of an epoch, the BSs generate a synchronization (SYNC) control frame, which acknowledges the packets received in the previous epoch, and determines the particular parameters to access the next one.

An epoch may start asynchronously, when a MT finds the medium idle (no tone) or it may start synchronously, after the BSs signal the end of an epoch by turning the tone off, just before sending the SYNC control frame. Thus, a wake-up radio method can be implemented to enhance the MT's power saving. The device can use a low power secondary radio to receive the tone, and turn off the main radio while they are waiting for the SYNC frame. Figure 3.5 depicts an example of this procedure, where the main radio stays off until a SYNC packet is captured by the secondary radio.

The MAC protocol algorithm procedure is depicted in figure 3.6, demonstrating a sucession of two epochs for a scheme with three MTs and one BS. The first epoch starts asynchronously having solely one MT transmitting, while the following epoch begins synchronized by the BS turning off the tone and runs MPR. Differently from NDMA, the receiver's algorithm is ran by the BS for every additional data slot [28]. The epoch is terminated when every single packet was properly received, which in this example was after three data slots.

One of the main limitations of this protocol derives from the requisite to have channels estimation. Before the transmission of the sequence of data frames in data slots, each MT

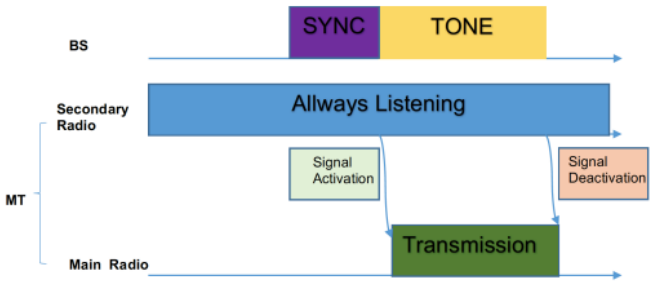


Figure 3.5: MTs wake up radio example (adapted from [28])

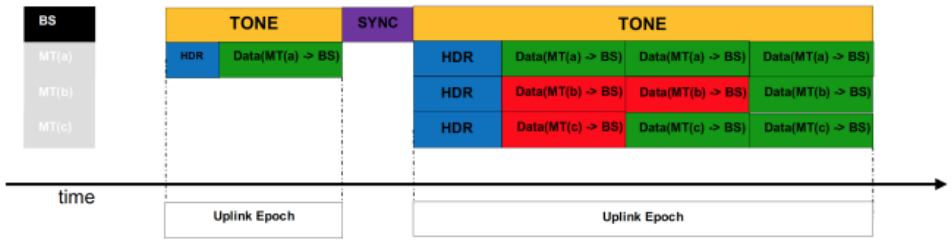


Figure 3.6: M2M H-NDMA protocol example (adapted from [28])

transmits a physical layer header (HDR) containing a pilot signal, that will be employed to estimate the channel. To receive all pilots transmitted concurrently, HDR are spread using unique orthogonal sequences [28], attributed by the BSs to each MT during a primary association stage. Hence, the HDR length increases linearly with the amount of MTs enrolled in the channel. The SYNC frame also includes an acknowledgment bit mask, which also grows linearly with the number of MTs. However, with the shift to a DAS, it is possible to mitigate slightly the delay that incurs from the overhead induced by the MTs HDR (more copies of the same packet transmitted without additional time slots), enabling to scale the number of MTs in the system without degrading a certain delay bound. On the other hand, with a DAS, new problems arise with tone control in multi-hop scenarios. This dissertation addresses only a simple DAS scenario where all BSs collaborate in the CoMP joint reception. More complex multi-hop scenarios were left for future study.

The solution proposed is for all neighbor BSs within the DAS to also send a tone in the same frequency. In a multi-hop scenario where the BSs are far away from each other (e.g. tens of kilometers away or more), a cell based approach would be needed to cover it. Frequency reuse would be defined based on a interference criteria: a maximum received power threshold would be required to reuse the frequency.

3.3.1.2 Power Control

In the previous version of the M2M H-NDMA, throughout the association stage the MTs receive the Terminal Unique Orthogonal spreading-Sequence (TUOS) and a list of potential SNR needed per MT, and measure the path loss to the BS. When transmitting

a packet, the MTs select a random power level from the list received during the initial association.

However, in this new environment with multiple base stations and increased spatial diversity, power control may become a problem since power regulation for each MT is only done by its associated base station (the nearest BS to the MT). Thus, the power fraction received in the associated BSs, is not controllable - it depends on the geographical distribution, being this power a fraction of the one of the controlling BS due to the great distance between the controlling and associated BSs. Hence, a pure NOMA approach based on strict power levels separation at the receiver is not manageable, due to the power spreading received from the various MTs.

On account of this effect, power is less controlled, and if there is a big imbalance in an epoch when it comes to MTs transmitting power levels (a lot of MTs transmitting in one power level and few in the other), a great deal of interference is generated. The protocol uses H-ARQ, that does not discard packets with errors like ARQ does, but stores them instead and asks for retransmissions of the packets until it has enough copies that combined together give the complete information needed to decode. So, can lead to numerous retransmissions to receive from for all terminals, increasing the delay and the spent average energy necessary to successfully receive a packet. In this dissertation, this energy is measured using Energy Per Useful Packet (EPUP), defined by the total amount of energy spent during all the epochs the MT participated over the number of successful transmissions.

Active Power Control (APC) policies that are epoch adaptive should be implemented and initial MT power distribution settings may be as well explored. Transmission power of the terminals must also be controlled - a maximum of 33 dBm (3dB) was considered for the UEs, which is equal to the LTE maximum. This restriction/verification was introduced in the model and in the simulator developed in this dissertation.

These policies should aim to reduce the EPUP and to improve the delay. Although having the MTs choosing randomly the power level on which they will transmit is not a bad approach, at the same time it might not be optimal for a DAS, since if the terminals that are far away from their associated BS happen to select the highest power to transmit and not changing it, the energy expenditure will be high. In case of imbalance, it may have to retransmit many times until success. For a terminal closer to a BS, the cost to change its transmission power is lower due to its proximity. In a given epoch, if there is an imbalance on the transmitting power levels, prioritizing the terminals that are closer to their associated BSs to change their transmission powers would impact minimally on their EPUP and eventually speed up the resolution of all the terminals involved in the epoch, possibly reducing the epoch length and the delay.

In this dissertation, an epoch APC is implemented and various MTs initial power distributions are assessed. The initial power distributions studied were the following:

- **Random**: The distribution is the same as in [28].

- **Low Power**: All MTs start with the lowest power level for transmission, which is updated in every epoch. The goal of this strategy is to lower the spent EPUP.
- **Evenly Distribution per Zone**: This strategy aims to guarantee a even power level distribution per CoMP zone.

The CoMP zones were defined as the level of isolation the MTs present to each BS. This algorithm was inspired from [44], and it is computed the isolation level for each terminal for all BSs by calculating their pathloss ratio,

$$\Psi_{\phi_k} = \frac{\sum\limits_{\phi \in \phi_k} PL_{\phi,k}}{\sum\limits_{1 \leq \phi \leq NBs} PL_{\phi,k}}. \tag{3.2}$$

After obtaining the pathloss ratios, for each terminal it was compared the ratios with the power gap, to determine the isolation level of the MTs to each base station. Considering a vector of Power Levels $Q = [Q_1, Q_2]$ the **power gap** is the value in dB of $Q_1 - Q_2$. So the power gap can be described as the separation between the NOMA power levels, and the higher the power gap, the higher the probability of successful receiver signal separation. If the MTs had pathloss ratios to multiple BSs that were within the power gap threshold, then they would be classified as being in the CoMP zone of those BSs. Therefore, the classification of MTs per zone is as following:

- Comp 1 : The terminal can only transmit to 1 base station, being that BS its associated one.
- Comp 2-N :The terminal can transmit to 2 or more base stations.

The APC is implemented with the goal to correct the epochs power levels ratio imbalance, and balance as close as possible to a 50/50 ratio, prioritizing the terminals that will need to increase their power level transmission in accordance to its proximity to their associated BS. This criteria was chosen for the reason that changing the MTs closest to their BSs requires a lower energy cost, since the further away the MT is, the higher the transmit power must be to be received at the BS. Hence, by using this criteria makes a minimal impact in the EPUP. So, the terminals to change first would be those only in the range of their associated BSs, and if the number of terminals in this condition present in a given epoch are not enough to balance the ratios, then the terminals in the CoMP zones, also prioritizing the choice regarding the proximity to their main associated BSs. The APC only gets into action if in the first transmission the terminals involved are not solved. The algorithm is applied in the second transmission. Due to the random access, information about which MTs are transmitting in the epoch can only be gained after the reception of all the preambles transmitted. The representation of this algorithm is depicted in figure 3.35.

Besides implementing the strategies and the APC policy described above, there is the need to study what would be the best transmitting power levels for the MTs to use, since in this new scenario with multiple BSs, power spreading and consequent interference is generated. Therefore, the **NOMA power levels** used by the terminals for transmission must be evaluated for this case, as well as the power gap used. All power values presented are relative to a **normalized power** used in the simulator. The transmission power T_x can be calculated from the relative one (MT_{E_N}) using,

$$T_x = Normalized_{Power} + MT_{E_N}. \tag{3.3}$$

Studies regarding this attributes are presented in section 3.5.

3.3.1.3 Packet Structure

This section describes the packets, their structure and the particular task they fulfill in the system.

The SYNC is a packet broadcasted by the BSs to signal the end and/or beginning of an epoch. It also confirms the reception of the packets sent by the MTs during the previous epoch, and defines access specific parameters for the next one. However in this DAS environment there is a problem regarding about which base station sends the SYNC. Since the packets should be transmitted concurrently, to avoid SYNC packets collisions, it is assumed the use of CoMP transmission for these packets. The main SYNC fields are:

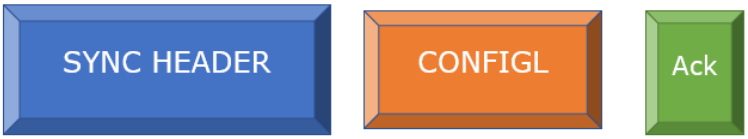


Figure 3.7: Main fields of the SYNC packet

- **SYNC Header**: BSs ID, pilot channel information;
- **Configuration List (CONFIGL)**: Includes the number of transmission powers and rules to be used (e.g. power control policies);
- **Acknowledgment (ACK)**: A bit field that acknowledges the packets received in the previous epoch. When the bit is 1, the acknowledge means that the packet was successfully received in the previous epoch. Since the system configuration considered is of a pure DAS, the ACK acknowledges all the concurrent transmissions in the cell.

When certain MTs transmission power level needs to be changed, the information of which MTs need to change and to what power level is sent embedded into the tone,

without the need to add another control packet, thus not aggravating the already existing overhead.

DATA packets contain the information the M2M devices want to transmit. Finally a header that contains the pilot signal, used to estimate the channel transmitted using the TUOS, is added on the first transmission within an epoch [28]. With the extension to a DAS, the management of the terminal IDs should be done at a DAS scale. New solutions are needed to handle a full distributed scenario, to avoid the bottleneck introduced by a solution based in a global network wide IDs, regardless of their location. This solution is impractical due to the huge spreading that would be applied to the preamble.

3.4 Multi-packet detection receiver performance

SC-FDMA is considered for the system uplink, based on the uncoded Iterative Block Decision Feedback Equalization (IB-DFE) MPR receiver from [45] for a Quadrature Phase Shift Keying (QPSK) constellation. Using an IB-DFE technique that performs SIC for each iteration, it allows the reception of more than one packet per slot in average [46]. The analytical expression for the Packet Error Rate (PER) in [45] is presented in this section.

3.4.1 Multi-packet detection receiver

A block of N symbols transmitted by a terminal p, can be expressed in the time domain as $\{s_{n,p}; n = 0...., N-1\}$ and on the frequency domain as $\{S_{k,p}; k = 0...., N-1\}$. At the DAS, the received signal from P terminals for a given transmission l is expressed in the frequency domain as $Y_k^{(l)} = \sum_{p=1}^{P} S_{k,p} H_{k,p}^{(l)} + N_k^{(l)}$, where $H_{k,p}^{(l)}$ is the channel response for the pth terminal at the lth transmission and $N_k^{(l)}$ is the background noise for the lth, modeled by a null average Gaussian random variable with variance $\sigma^2_{N^{(l)}}$.

For P terminals that transmit through L channels (e.g. L transmissions on a BS with a single antenna, or using m transmissions received by $\frac{L}{m}$ uncorrelated receiver's antennas,where CoMP processing is employed), the received L transmissions are characterized as $\boldsymbol{Y}_k = [Y_k^{(1)}, ..., Y_k^{(l)}]^T$, where $\boldsymbol{H}_k^T = [\boldsymbol{H}_{k,1}, ..., \boldsymbol{H}_{k,P}]$ and $\boldsymbol{H}_{k,p}^T = [H_{k,p}^{(1)}, ..., H_{k,p}^{(L)}]$ for $p = [1, ..., P]$, $\boldsymbol{S}_k = [S_{k,1}, ..., S_{k,P}]^T$ and $\boldsymbol{N}_k^T = [N_k^{(1)}, ..., N_k^{(L)}]$, where T represents the transpose of the matrix. So

$$\boldsymbol{Y}_k = \boldsymbol{H}_k^T \boldsymbol{S}_k + \boldsymbol{N}_k, \tag{3.4}$$

The expanded version of $\boldsymbol{Y}_k$ is

$$\begin{bmatrix} Y_k^{(1)} \\ \vdots \\ Y_k^{(L)} \end{bmatrix} = \begin{bmatrix} |\xi_{1,1}| H_{k,1}^{(1)} & \dots & |\xi_{1,P}| H_{k,P}^{(1)} \\ \vdots & \ddots & \vdots \\ |\xi_{l,1}| H_{k,1}^{(L)} & \dots & |\xi_{l,P}| H_{k,P}^{(L)} \end{bmatrix} \begin{bmatrix} S_{k,1} \\ \vdots \\ S_{k,P} \end{bmatrix} + \begin{bmatrix} N_k^{(1)} \\ \vdots \\ N_k^{(L)} \end{bmatrix}, \tag{3.5}$$

where $\left|\xi_{L,P}\right|$ denotes the path loss attenuation of a terminal p for a given transmission l. In the event where the pth terminal does not transmit for a given slot l, then $H_{k,P}^{(l)} = 0$.

3.4.2 IB-DFE Model

In IB-DFE, both the feedforward and the feedback filters are implemented in the frequency domain [46]. The IB-DFE receiver [45] runs N_{iter} iterations using the L channels, from the strongest to the weakest received signal power, to detect each of the P terminals. The estimated data symbol, $\tilde{S}_{k,p}^{(i)}$, for a given iteration i and terminal p is given by

$$\tilde{S}_{k,p}^{(i)} = \boldsymbol{F}_{k,p}^{(i)^T} \boldsymbol{Y}_k - \boldsymbol{B}_{k,p}^{(i)^T} \bar{\boldsymbol{S}}_k^{(i-1)}, \tag{3.6}$$

where $\bar{\boldsymbol{S}}_k^{(i-1)} = [\bar{S}_{k,1}^{(i-1)}, ..., \bar{S}_{k,P}^{(i-1)}]^T$ denotes the soft decision estimates from the previous iteration for all terminals. $\boldsymbol{F}_{k,p}^{(i)^T} = [F_{k,p}^{(i,1)}, ..., F_{k,P}^{(i,L)}]$ are the feedforward coefficients,

$$F_{k,p}^{(i,l)} = \frac{H_{k,p}^{(l)^*}}{\frac{\sigma_N^2}{\sigma_S^2} + \sum_{l=1}^{L} \left|H_{k,p}^{(l)}\right|^2}, \tag{3.7}$$

and $B_{k,p}^{(i)^T} = [B_{k,p}^{i,1}, ..., B_{k,p}^{i,P}]$ are the feedback coefficients,

$$B_{k,p}^{(i,1)} = \sum_{l=1}^{L} F_{k,p}^{(i,l)} H_{k,p}^{(l)} - 1, \tag{3.8}$$

calculated in [45] to minimize the mean square error at the receiver. The mean square error of terminal p at the ith iteration [45] is

$$\sigma_p^{2(i)} = \frac{1}{N^2} \sum_{k=0}^{N-1} \mathbb{E}\left[\left| \tilde{S}_{k,p}^{(i)} - S_{k,p} \right|^2 \right], \tag{3.9}$$

where $\mathbb{E}\left[\left| \tilde{S}_{k,p}^{(i)} - S_{k,p} \right|^2 \right]$ can be calculated using [45]. The Bit Error Rate (BER) of a terminal p at the ith iteration for a QPSK constellation is given by

$$BER_p^{(i)} \simeq Q\left(\frac{1}{\sigma_p^{(i)}} \right), \tag{3.10}$$

where $Q(x)$ denotes the Gaussian error function. For an uncoded system with independent errors, the PER for a fixed packet size of M bits is

$$PER_p^{(i)} \simeq 1 - (1 - BER_p^{(i)})^M. \tag{3.11}$$

Equation 3.11 provides a generic function that can be used to calculate the PER of any system given the channel response $\boldsymbol{H}_k$ and the bit energy to noise ratio $\frac{E_b}{N_0}$ for the received signal from each terminal. The energy received from terminal p during transmission l to the base stations is modeled by the $H_{k,p}^{(l)}$ coefficients, which account the attenuation

gains and different transmission powers. When a terminal does not transmit, the channel coefficient value is set to zero.

Figure 3.8 depicts the PER performance in function of the $\frac{E_b}{N_0}$, showing that the $\frac{E_b}{N_0}$ required to receive the MTs decreases when more transmissions are combined.

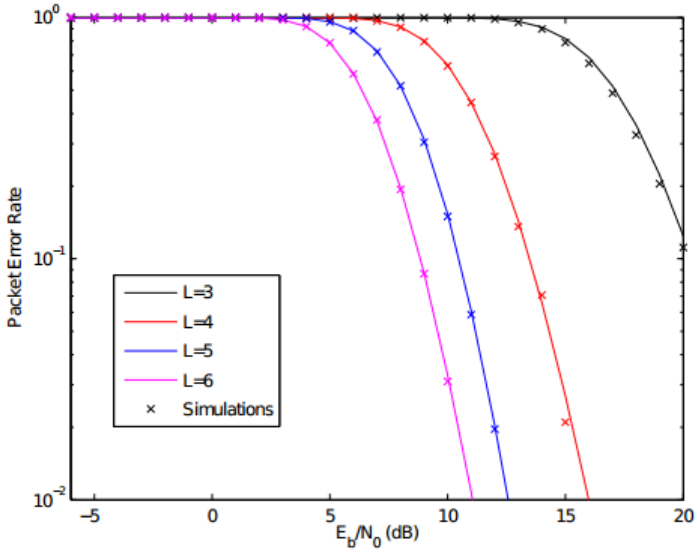


Figure 3.8: PER performance (adapted from [45])

3.5 Power separation and power levels analysis

In [28] was shown that using two power levels with a separation by 12dB and a minimum power of 3dB, and with terminals distributed evenly between the two power levels, was enough to assure an average PER below 10^{-3} and enable MPR, allowing the serial resolution from the signal with the highest power to the one with lowest power (assuming that SIC can be used successfully). This section analyses how the power spreading that results from DAS influences these values.

An initial set of tests was conducted to verify if the power separation used in [28] still satisfied the new system conditions introduced in this dissertation. The tests used a power separation of $\alpha = 12$dB, using two different vectors of power levels, $Q_1 = \{3, 15\}$ and $Q_2 = \{6, 18\}$ dB, which satisfied [28] conditions. the term power levels is used to designated the $\frac{E_b}{N_0}$ for the levels at the receiver. The number of base stations in the system varied between 1,2 or 4, the terminals aggregated load tested was $\Omega = [200\%, 400\%]$ and the number of terminals was $P = 20$. The values of Ω are related to the $\eta \times \kappa$ relation, e.g. $\eta \times \kappa = 3$ means that the system has 3 levels of freedom, so the amount of traffic load the system can handle is up to 3 times more than a system with only 1 level of freedom, even though due to the overhead of the preamble and other control fields, the effective capacity is inferior to 1 since only the data traffic contributes to the throughput. Anyhow, Ω reflects the overload possible in % attending to this relation.

In the simulations, terminals were distributed according to a 2D Poisson distribution, which models the position of each node as an independent event and is characterized by the average density value λ. The number of users in a area of the network, represented by a random variable X, is given by a Poisson distribution, defined by

$$P(X = c) = \frac{\left(\rho_{ON}\beta A_E\right)^c}{c!} e^{-\rho_{ON}\beta A_E}, c = 0, 1, ..., N, \tag{3.12}$$

where ρ_{ON} is the probability of the terminal being transmitting, A_E is the rectangle area, and β is the density of the terminals per m^2. In the simulations the values used were $\rho_{ON} = 1$, $A_E = 50$ m^2 and $\beta = 1$.

The Poisson distribution of the terminals was applied per quadrant: for each quadrant, a value of λ is drawn, and then divided by the number of quadrants (in this case there were 4), so the average number of terminals per quadrant q_i is

$$\lambda_{q_i} = \frac{\lambda}{4}, \tag{3.13}$$

and the total number of terminals transmitting in the whole area considered is

$$\lambda = \lambda_{q_1} + \lambda_{q_2} + \lambda_{q_3} + \lambda_{q_4}. \tag{3.14}$$

After obtaining the number of terminals transmitting per quadrant, their coordinates are generated using a uniformly distributed pseudo-random number generator to guarantee a random spacing between them.

All terminals generated must respect the maximum power transmit constraint given by,

$$P_{MT} = |PL_{MT}| + \sigma^2_{N_0} + G_0 + 10\log_{10} B - \Gamma \leq 3dB, \tag{3.15}$$

where,

$$\Gamma = E_{N_{BSassoc}} - E_{N_{MT}}, \tag{3.16}$$

being,

$$\sigma^2_{N_0} = -174 + 10\log_{10} H \tag{3.17}$$

the thermal noise for a bandwidth H, which in these simulations is 64 MHz (though this is a very high value of bandwidth for M2M, allows a much higher rate and throughput). B denotes the number of bits in a packet (B=2 in the case), and G_0 is the antenna gain.

Other parameters for this simulation were the block size $N = 64$ data symbols, normalized power = 21 dB, and no APC was used. All simulations were conducted in MATLAB.

Two metrics are proposed to evaluate the influence of the power gap and the power levels in the system performance:

- **Epochs with irregular PER (%)**: When the minimum power is set too low or the power gap is too narrow, H-NDMA is able to compensate partially by forcing more packet retransmissions during an epoch, increasing retransmissions. An epoch is classified with "irregular PER" when misdetections are found or if in the last transmission any terminals were found unresolved. This metric is presented as a percentage.

- **Epochs exceeding expected** ϵ: Considering the receiver can theoretically solve $\eta \times \kappa$ MTs per slot, for P terminals transmitting in an epoch, the value of ϵ should be $\lceil \frac{P}{\eta \times \kappa} \rceil$, where $\lceil \rceil$ denotes the ceil operation. An epoch is classified "exceeding expected ϵ", when the number of transmission slots in an epoch exceeds the value of ϵ. This metric complements the one above and is also presented as a percentage.

The results for this tests, in figure 3.9, show that for the lowest value of aggregated load, for 1 and 2 base stations, there were no errors. However, when the load starts to increase, it is clear that more epochs with irregular PER start to appear. Moreover, looking at the results with 4 base stations, it becomes evident that this power gap is insufficient, showing a higher percentage of epochs with irregular PER than the results with 1 or 2 base station.

Due the DAS configuration, if the terminals are in a CoMP Zone, the power regulation for those MTs is only done to their corresponding associated BS. The power reaching the other BSs that comprise that CoMP zone is not regulated directly, creating several intermediate power levels, "polluting" the NOMA power levels, and ultimately rendering SIC inefficient (the receiver cannot separate the signals properly) thus contributing to more errors.

Nonetheless, it was also verified that the epochs with irregular PER tend to decrease when the power transmission levels increase. It is also possible that by increasing the transmission power T_x value, and thus slightly increasing SNR, could prove to be helpful and enhance the odds of successful receiver separation. Therefore, a study was made where many transmission powers were tested with $normalized_{power} = [21, 24, 27, 30]$, $\alpha = [14, 16, 18, 20]dB$, and 2 power levels vectors Q_1 and Q_2 for each α, but testing only for 400% aggregated load, since it was verified that with more load, the errors increased significantly. The results are shown in figure 3.10, and it can be confirmed that for a normalized power of 21 dB, with the increasing number of base stations, it does not present a good performance. For normalized powers = 24 dB and 27 dB (i.e. respectively increasing 3 and 6 dB), there seems to be an increase in performance, and for normalized power = 30 dB, it is shown to have the best performance, with only one vector of power levels presenting an epoch with irregular PER.

However, from figure 3.10 it can only be concluded that one combination of power levels vector was not great, leaving the doubt about which of all other power levels combinations is the one with the probability to deliver best results. For that reason, a study regarding the metric *nº epochs exceeding expected* ϵ was conducted, to rule out from the

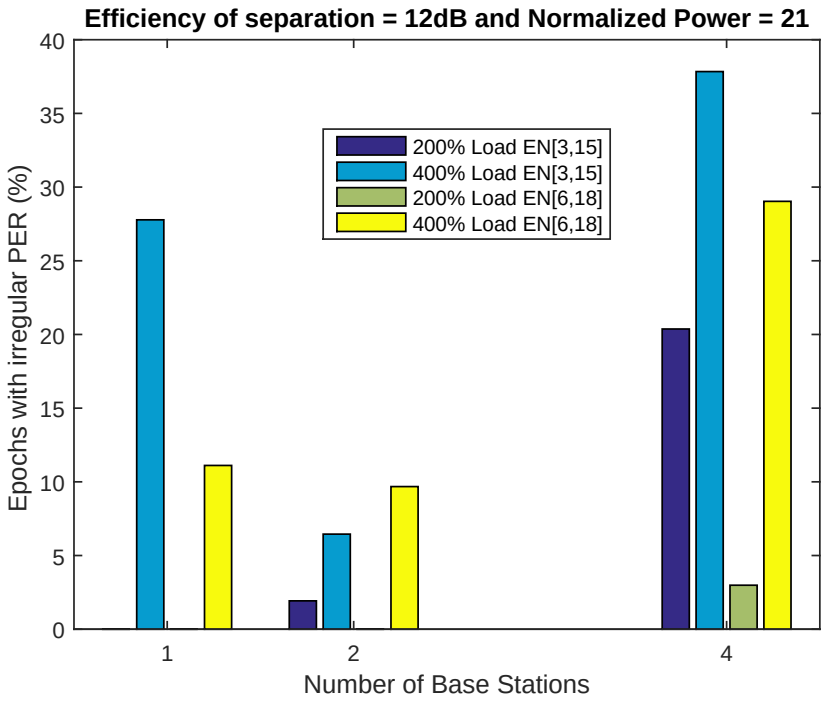


Figure 3.9: Influence on the number of base stations on PER with power gap = 12 dB.

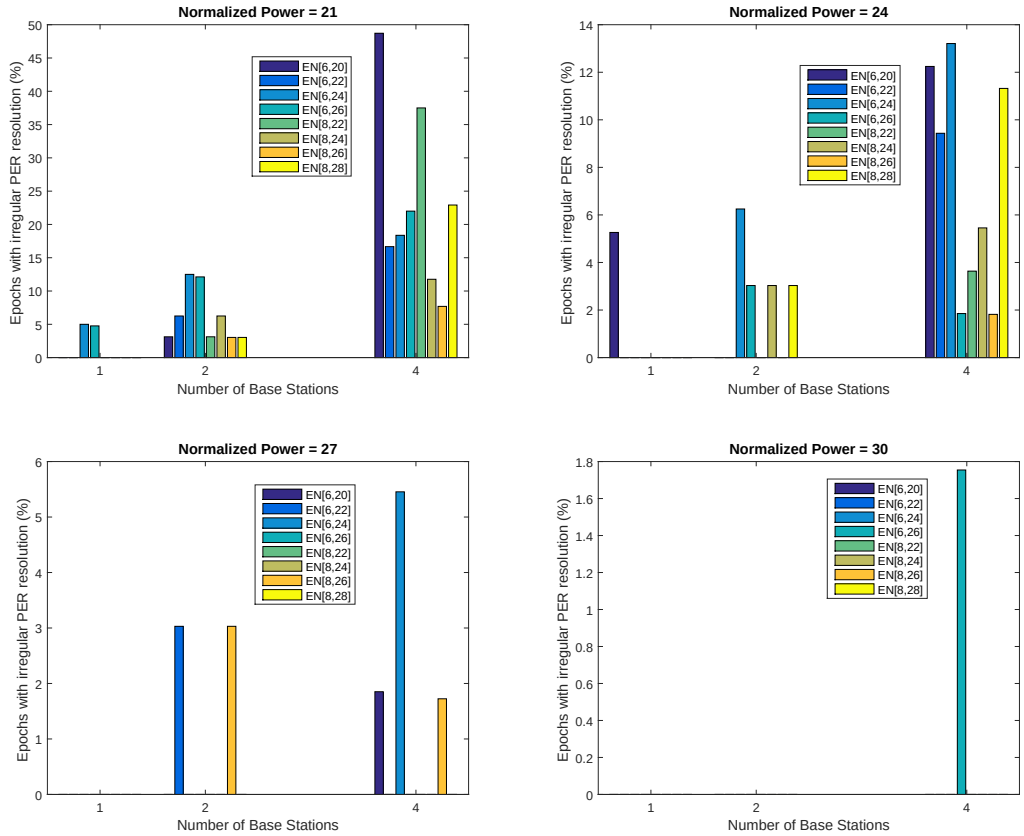


Figure 3.10: Relation between % of Irregular PER epochs and different power levels

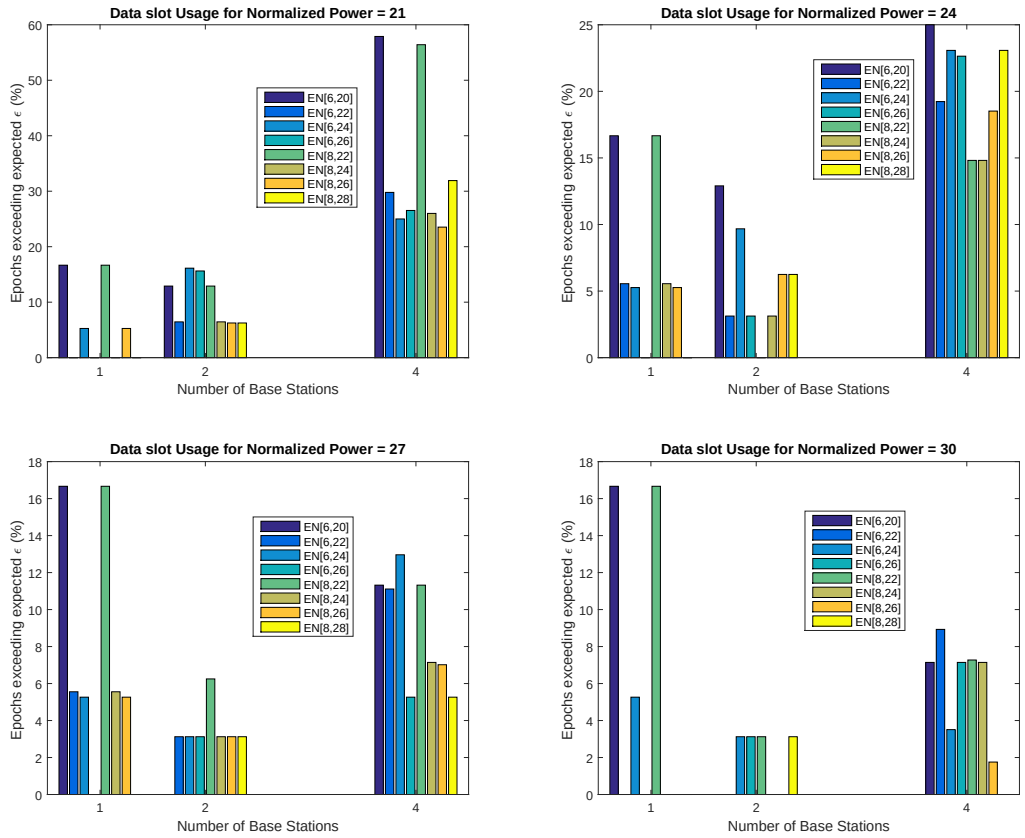


Figure 3.11: Relation between % of epochs exceeding ϵ and different power levels

best transmission power levels performance (the test with the best results was with normalized Power = 30 dB), the power gaps and power levels vectors that lead to epochs that exceeded the $\lceil \frac{P}{\eta \times \kappa} \rceil$ constraint. For the sake of completion, the same test was run for all the other data sets and conditions of the *Epochs with irregular PER (%)*, for comparison and verifying the most efficient power levels vector.

Results from the *nº epochs exceeding expected* ϵ study are presented in figure 3.11. On this figure, it is clear that as the transmission power is increased, the number of epochs not complying to the metric decreases. Looking closely, there is only one power level vector combination that showed no errors according to the criteria of this test: the power level vector $Q = \{8, 28\}$ dB.

In conclusion, taking into account the studies above, considering the initial conditions and parameterizations done as well as the scenario where these tests were conducted, in order to achieve the best reliability possible, the $\frac{E_b}{N_0}$ of the transmission power levels for the terminals should be $Q = \{28, 58\}$ dB.

Having the power gap and optimum power levels (considering 2 NOMA power levels), the following section will show the effects of a DAS configuration, on average delay, throughput and energy efficiency.

3.6 DAS effect on URLLC metrics

In this section, the impact of a DAS configuration on URLLC metrics and scalability is evaluated. This effect is analyzed, firstly, by variating the aggregated load Ω and secondly, by increasing the number of terminals J.

3.6.1 DAS effect on aggregated load

To test the effect of CoMP and the introduction of multiple base stations, a study was made with $P = 20$ and aggregated load $\Omega = \{100\%; 200\%; 300\%; 400\%; 500\%; 600\%\}$ for 1, 2 and 4 base stations. $\rho_{ON} = 1$ and $A_E = 50$ m^2 were considered for the simulations. The transmission power levels vector used, presented in the form of $\frac{E_b}{N_0}$ is $Q = \{28, 58\}$ dB.

Figure 3.12 depicts the impact of a DAS on the average delay scaling with the increment of the aggregated load. Results show that there is a significant gain in terms of delay tolerance. In another words, the average delay escalates more rapidly with less base stations and with the increase of the aggregated load. The difference is notorious, in particular when comparing the curves of 1 base station versus the 4 base stations curve. It is noticed easily that the average delay grows slower with more base stations, and for the same data points presents an overall gain, with much less average delay. This is likely due to the fact that with more base stations, more copies of the same packet are received from every transmission, so in fact, one transmission in a 4 base station DAS setup is almost equivalent to 4 transmissions in a 1 base station setup. It can therefore be concluded that with the increase of the number of BSs, the average delay tends to be lower.

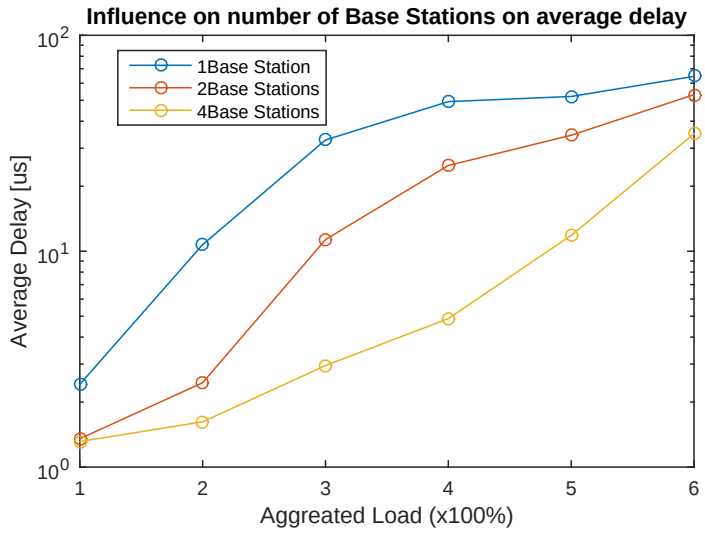


Figure 3.12: effect of Ω on average delay per number of base stations.

Figure 3.13 presents the effect of DAS configuration on the relation of average throughput and number of base stations. At first sight, it is noticeable that with the increasing number of base stations, and the use of CoMP, a lower number of retransmissions is needed, therefore enhancing the throughput. Having more base stations implies that the

value of $\eta \times \kappa$ is larger, enhancing the receiver capacity to resolve more terminals per data slot. In terms of aggregated load, this translates into being able to have a gain in terms of saturation. As it can be observed in figure 3.13, for 2 base stations the throughput saturates at 500% aggregated load and at 600% it starts to decline, and with 4 base stations saturation only occurs at 600%. Overall, most terminals and probably cell edge terminals will benefit from CoMP, which is shown in the throughput gain achieved with the input of more base stations in the system.

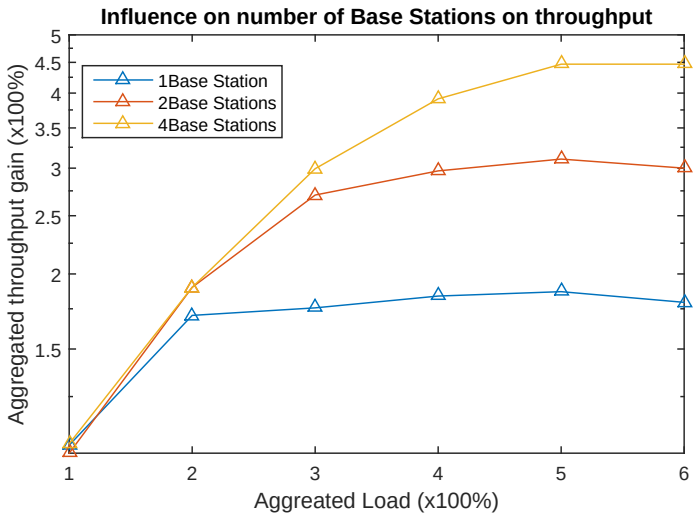


Figure 3.13: effect of Ω on average throughput per number of base stations.

Figure 3.14 depicts the EPUP, defined in section 3.3.1.2 and shows that using CoMP, the energy spent to send one packet can be reduced significantly, and grows slower as the number of base stations in the system increases. Since the number of transmissions is reduced with more base stations, the decrease of EPUP is in accordance to the lesser the energy spent on retransmissions. However, it should be noted that comparing to [28], the power levels and gap used have raised and the transmission power is nearly twofold as the one considered in the mentioned work, thus revealing that the EPUP was actually worsened. Nonetheless it was necessary, in order to attempt reaching a system that provides guarantees for bounds in delay and reliability for URLLC.

Figure 3.15 presents the DAS configuration influence on the maximum service time (values presented are the average of all maxima service times measured per data set). The figure shows that with more base stations, the maximum service time is decreased. Furthermore, it saturates on a higher aggregated load and tends to get to the saturation point slower with more base stations, which can be seen by observing the 2 base stations curve versus the 4 base stations curve. With the decrease in the maximum service time, the epochs become shorter, thus enabling to have more epochs and data transmission in the same time interval.

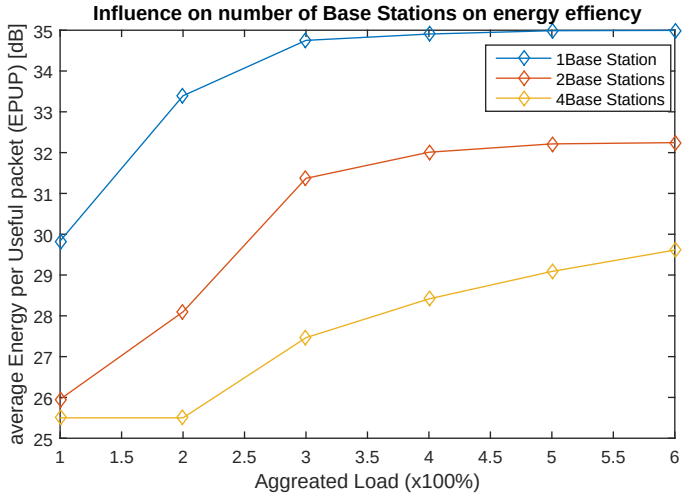


Figure 3.14: effect of Ω on average energy per useful packet.

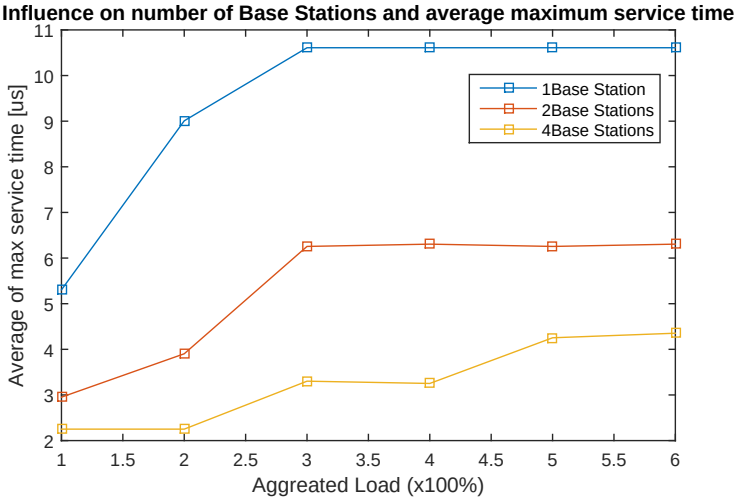


Figure 3.15: effect of Ω on average service time.

3.6.2 DAS effect on number of terminals

To test the effect of the increase of the number of terminals J, a study was made with $P = \{20;50;100;200\}$ MTs and aggregated load $\Omega = 400\%$ for 1, 2 and 4 base stations. $\rho_{ON} = 1$, $A_E = 50$ m^2 and $\beta = 1,..,4$ was considered in these simulations. The $\frac{E_b}{N_0}$ of the transmission power levels vector used, is $Q = \{28,58\}$ dB.

Figure 3.16 shows the behavior of the average delay in function of J. As expected, as J increases so does the average delay. However, with the increasing number of base stations the average delay is lower, which can effectively be seen when comparing the 1 BS and 4 BSs results.

Figure 3.17 depicts the aggregated throughput gain in function of J. Looking at the results, the difference of aggregated throughput gain from 1 BS to 2 BSs is quite high. Another interesting thing to note is that regardless the tests were conducted with

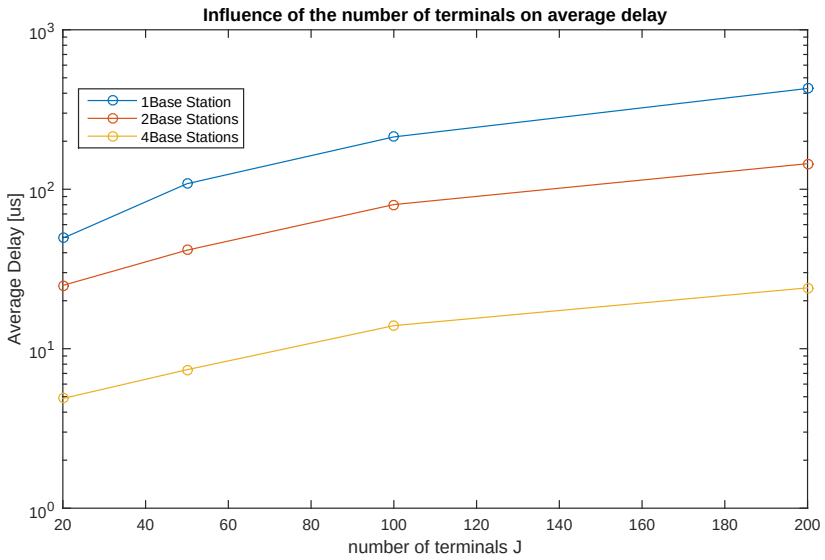


Figure 3.16: effect of J on average delay.

1, 2 or 4 BSs they all show a high throughput gain growth up until 50 MTs. For higher J, throughput starts to decay and eventually reaching a plateau, where doesn't grow or decay. This shows that although the increasing number of base stations improves aggregated throughput gain, having too many MTs transmitting will lead to a plateau in this gain. Since the header also grows linearly with the number of MTs, it becomes a factor that caps this gain when J is large.

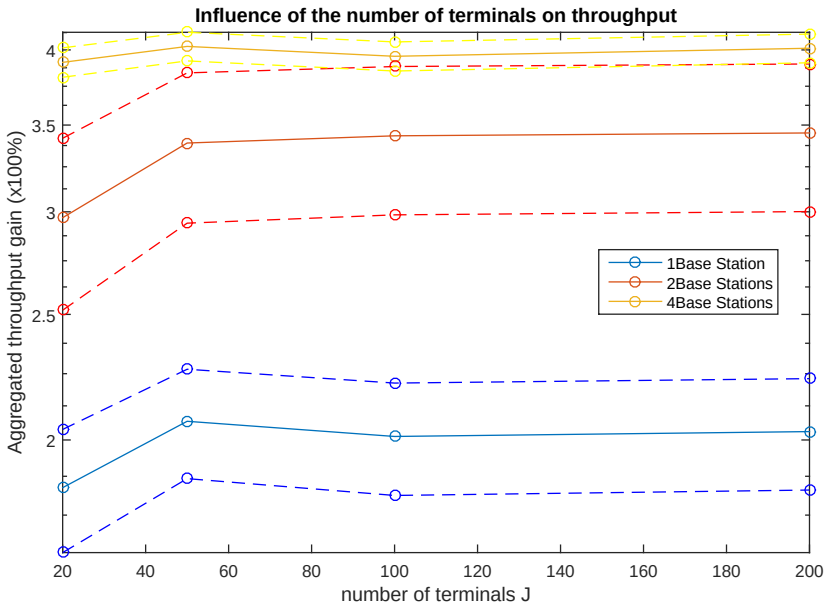


Figure 3.17: effect of J on aggregated throughput with confidence intervals of 95%.

Figure 3.18 presents the results of the EPUP behavior in function of J. The figure

shows the EPUP grows slower with the increasing number of base stations, with a gain of nearly 10 dB from 1 BS to 4 BSs with $P = 200$. Although, as it was mentioned, with more base stations it is spent more energy (the transmission power levels had to be increased). With more BSs the number of retransmissions goes down, hence resulting in the energy saving depicted in the results.

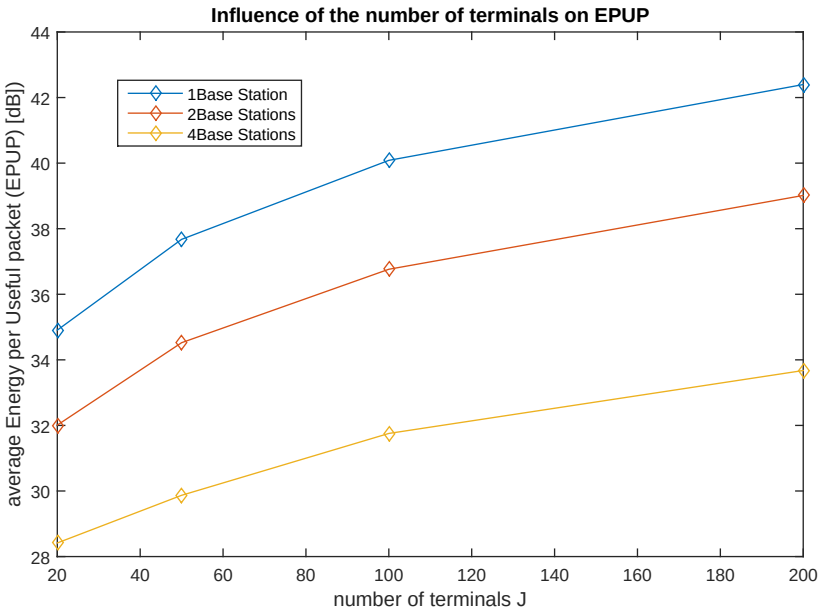


Figure 3.18: effect of J on EPUP.

Figure 3.19 illustrates the results the maximum delay in service time - values presented are the average of all maximum service times measured per data set in relation to J. As J increases, so does the service time, although with the increasing number of base stations, this growth is slower, and a significant gain in the service time delay is obtained, as results show, with $P = 200$, from 1 BS to 4 BSs, the service time delay is cut by little over than fourfold.

3.6.3 Ultra-reliability and low latency suitability analysis

This section discusses the suitability of the M2M H-NDMA protocol with a DAS configuration.

When it comes to provide URLLC services, studying just the average metrics is not enough, since this type of communication has strict requirements, for instance in terms of successful packet rate delivery, which may be as high as $1-10^{-5}$ (or even $1-10^{-9}$). For such demanding restrictions and precisions, metrics must not only be evaluated in terms of averages, but also consider the tail distribution - the tail behavior is inherently related to the tail of random traffic demand, intra/inter cell interference, and users that are at the cell edge power limited, or in deep fade, that needs to be otimized [47].

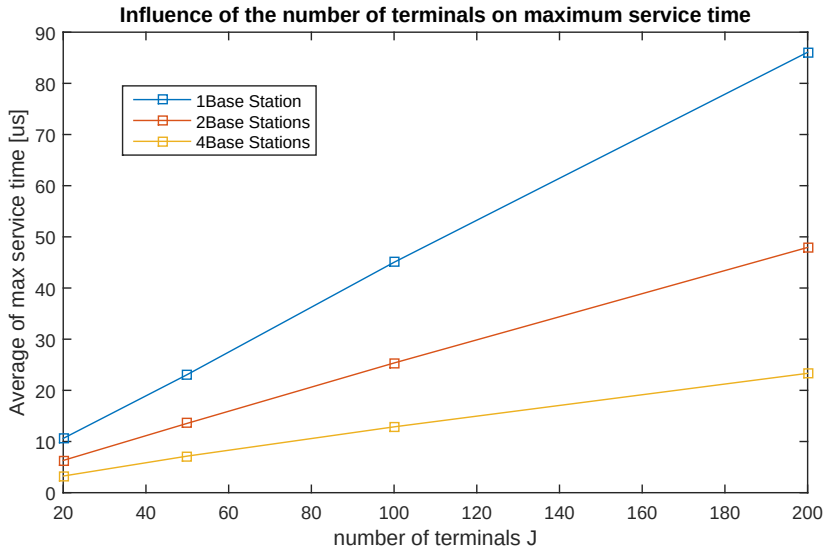


Figure 3.19: effect of J on maximum service time.

To study if the protocol in this dissertation is suitable for providing URLLC services, to evaluate reliability for a degree of precision of 10^{-3} (only 1 packet can fail in 1000), it would require a copious amount of data. Simulations would take very long time, specially for high aggregated loads and a high J. It is not possible to study PER nor delay for this precision. Despite not being able to prove suitability with high degree of precision, studies with lower degree of precision can give an approximate idea about the fit of the protocol to these requirements.

PER was studied using the same data that was used to present the results in sections 3.6.1 and 3.6.2. Results showed no errors when the number of MTs is low, despite the aggregated load. This is a good indicator that it is possible that the protocol could provide URLLC for a small number of MTs, and still handle load peaks, without compromising the reliability bound. The same cannot be said when PER was studied for increasing J and fixed aggregated load. Figure 3.20 shows the results of the average PER for the data sets used in section 3.6.2. Results show that for higher values of J, there are some errors, but with an increasing number of base stations the error reduces, and is under 10^{-3} with 4 BSs. Still, since the precision is low, this is not a good indicator that the protocol can support a high J for URLLC. These errors could be related to the epochs power levels ratio imbalance, having the receiver a hard time separating the signals, resulting in the epoch presenting unsolved MTs in the end of the epoch. This problem could be solved with an APC, by having a initial distribution power level policy and then performing epoch power control.

As for delay, URLLC requirements are of a hard latency of 1ms, which means that the delay must be in the order of microseconds. A queuing delay study was conducted for the data sets presented in sections 3.6.1 and 3.6.2, to see the impact on the Cummulative Distribution Function (CDF) of the delay for J and aggregated load.

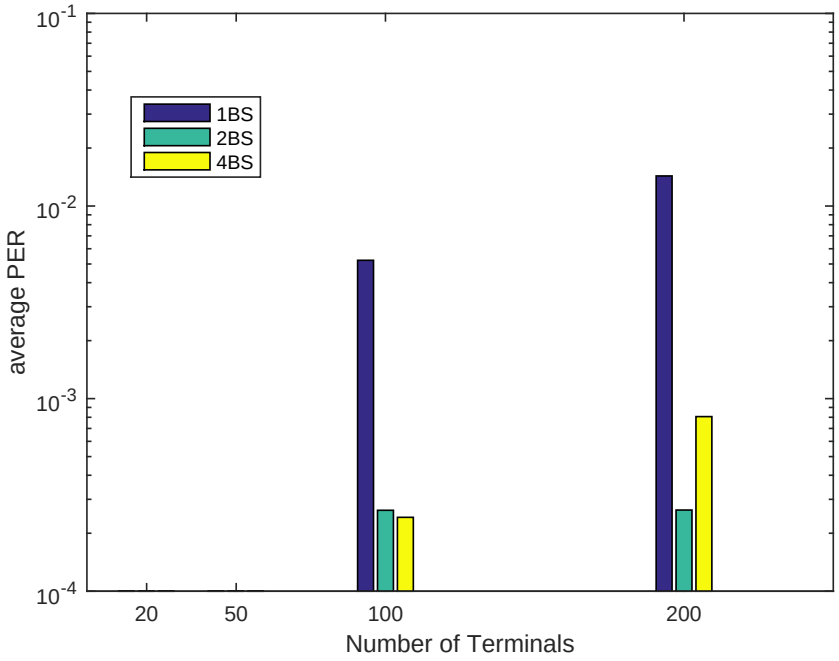


Figure 3.20: effect of J on reliability requirements.

Figures 3.21, 3.22 and 3.23 show the probability of exceeding a certain value of queuing delay in function of the aggregated load. They show that with the increasing number of base stations, the probability of exceeding higher values of delay is lower, and for $P = 20$, with 4 BSs it is possible to attain queuing delays under 50 μs with 500% aggregated load. Hence, for low number of MTs, ultra-low latency might be possible to achieve.

Figures 3.24, 3.25 and 3.26 depict the probability of exceeding a certain value of queuing delay in function of J. The figures show that as J increases, the probability of overrunning a higher queuing delay values increases. Still, with the increase in the number of base stations, the queuing delays are lower, therefore being an enabler for ultra low latency. The delay overrun probability would also be lower for a higher J, but since the header is very long and increases as J does, it will always cap the potential of the queuing delays the protocol could achieve.

All these tests were conducted so far without an APC. With an active power control policy and different initial power levels distribution strategy it might be possible to reduce the delay, and at the same time improve the energy spent, lowering EPUP. The following section will discuss different strategies and present results of the effect of a power control policy.

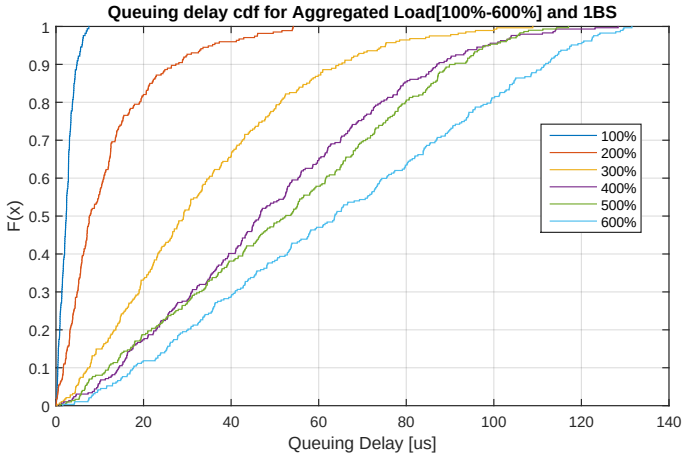


Figure 3.21: Queuing delay for varying Ω and 1BS

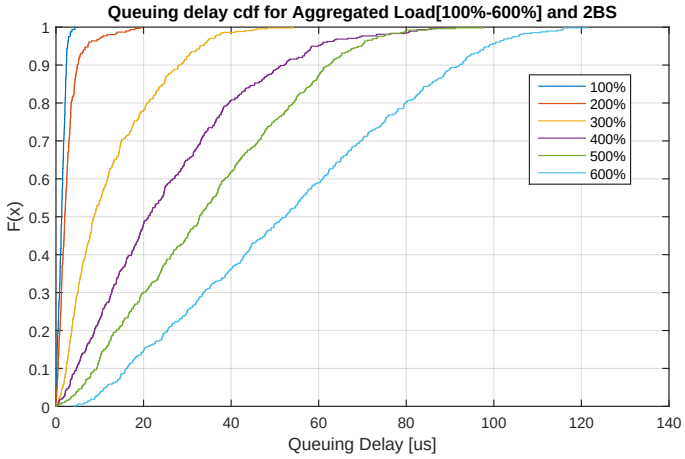


Figure 3.22: Queuing delay for varying Ω and 2BS

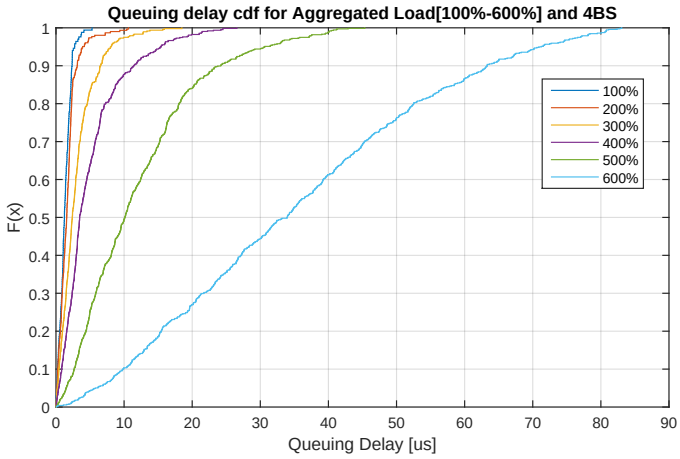


Figure 3.23: Queuing delay for varying Ω and 4BS

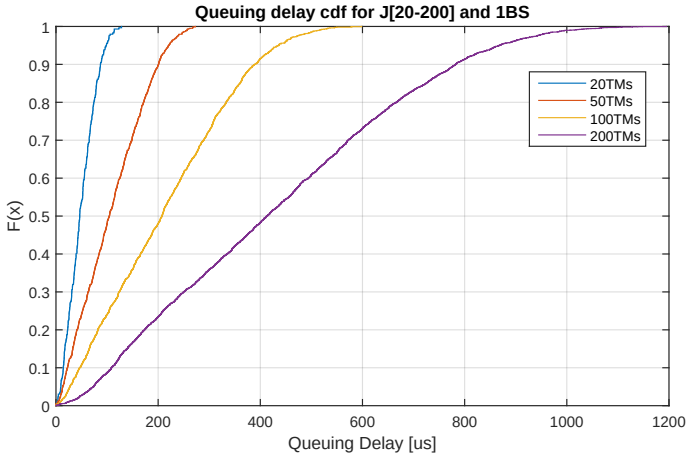


Figure 3.24: Queuing delay for varying J and 1BS

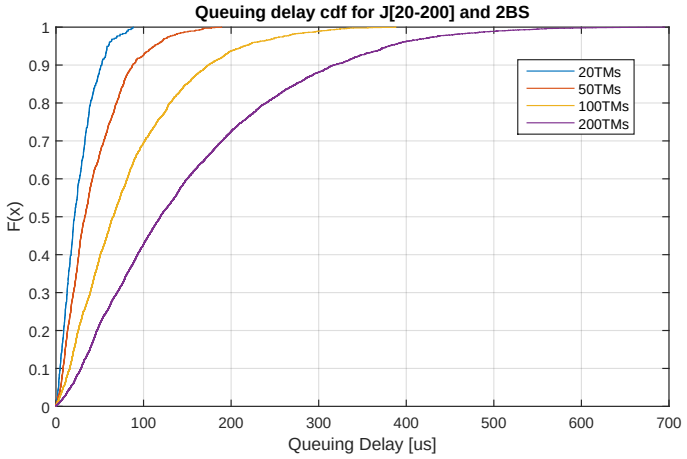


Figure 3.25: Queuing delay for varying J and 2BS

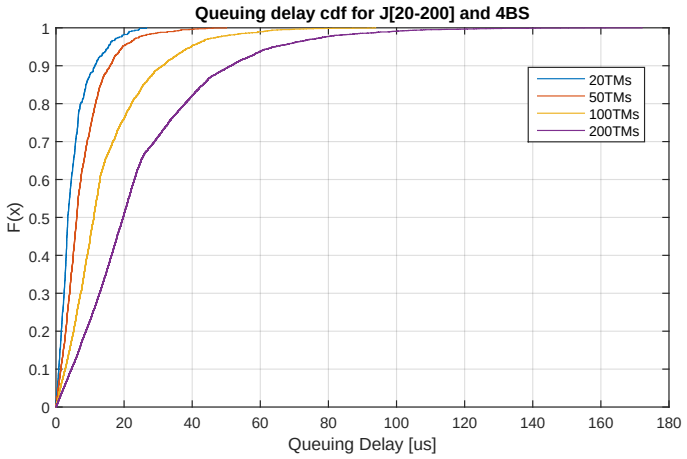


Figure 3.26: Queuing delay for varying J and 4BS

3.7 Impact of power control policies

This section shows the results, the APC implementation, as well as the effect of different initial MTs power levels distributions, that were described in section 3.3.1.2.

To evaluate how the different initial MTs distributions impact the URLLC metrics, a set of simulations were run, one for each initial distribution. The test were done with $P = 200$ MTs and aggregated load $\Omega = 400\%$ for 4 base stations. $\rho_{ON} = 1$, $A_E = 50$ m^2 and $\beta = 4$ were considered in these simulations. The transmission power levels vector used, presented in the form of $\frac{E_b}{N_0}$ is $Q = \{28, 58\}$ dB.

Results showed that as far as PER concerns, all the data sets showed no errors. Figures 3.27 and 3.28 depicts the queuing delay CDF for the different initial distributions. It can be seen that there is little to no difference in the curves of the different data sets. However, when zooming in, it can be seen that the simulation done with the MTs initialized with the lowest power level experienced a worse delay than the other two initial distributions (as expected). The 50/50 initial distribution shows slightly better results in the queuing delay CDF. Still there is not a significant difference between the 3 curves.

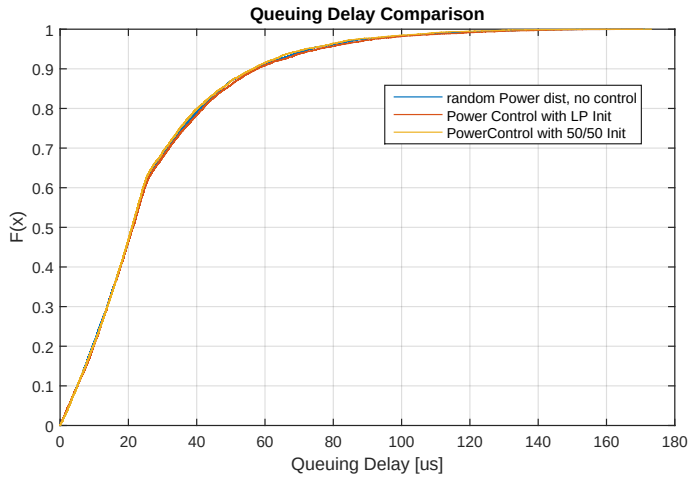


Figure 3.27: CDF of queueing delay for various MTs initial distributions

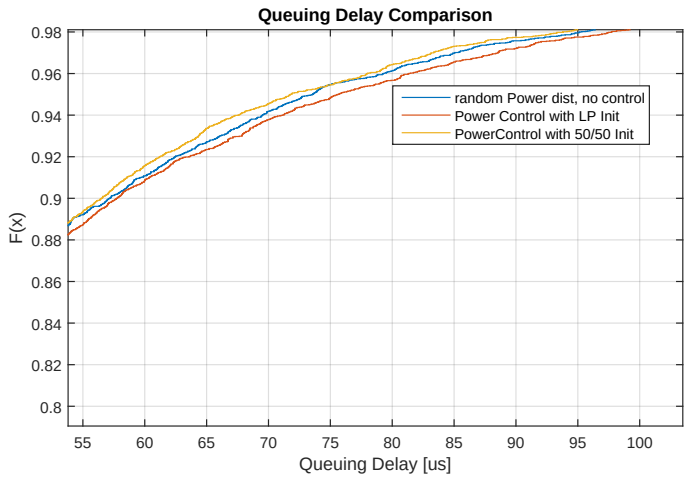


Figure 3.28: Zoomed in section of figure 3.27

To verify just how much the power control algorithm influences the system, a new

metric called *disbalance* was used, that verifies how much did the algorithm actually corrected in every epoch. The *disbalance* is given by

$$disbalance = |HigherRatio - 0.5|, \tag{3.18}$$

where the $HigherRatio$ is the percentage of terminals present in the power level with more MTs in a given epoch.

Figure 3.29 presents the variation of *disbalance* between the 3 initial distributions. It can be seen that after the initial epoch, the difference in the variation is not very significant, since the imbalance hardly ever goes above 10%. To better understand the effect with more accuracy, the table 3.1 presents the standard deviations and variances of the imbalances on the 3 data sets, and table 3.2 presents the same but without the initial transitory samples. The tables show that once the transient is removed, the standard deviations as well as the variances of the *disbalances* are relatively similar, being the random distribution the one presenting the less deviation and variance. From these results, it can be concluded that the power control algorithm did not had a major impact in the overall scenario.

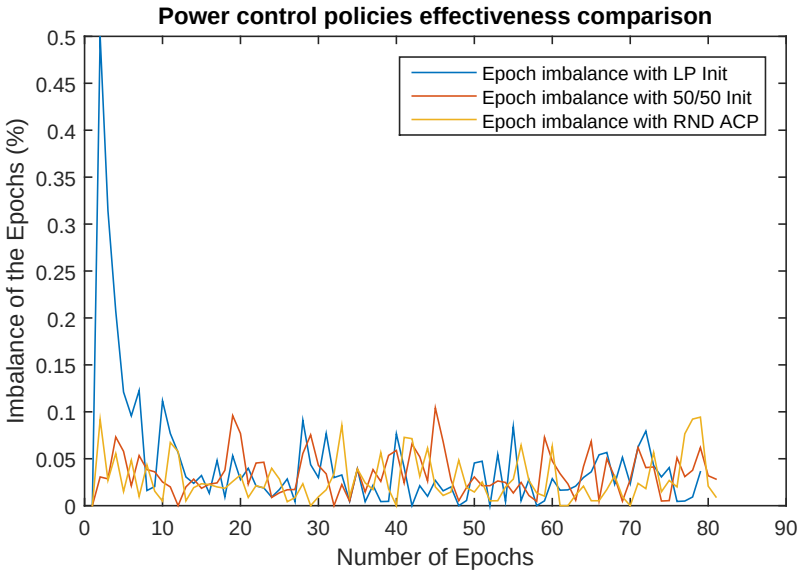


Figure 3.29: Variation of epochs imbalance.

However, the fact that it was not very effective could have been due the load used in the simulations. The higher the aggregated load, the higher are the number of MTs in each epoch transmitting, and once the algorithm runs a couple of times (correcting the transient part during each epoch) at a certain point there won't be anything imbalance to correct, reaching a kind of balance that lasts throughout the end of the simulations. For that reason, a set of simulations were conducted with varying Ω. Parameters for the simulations were $\Omega = \{100\%; 200\%; 300\%; 400\%; 500\%; 600\%\}$ and $P = 200$ MTs for 4 base

Table 3.1: Standard deviations and Variances of the disbalance metric

Initial Distributions	Low Power	50/50	Random
Standard Deviation	0.07	0.02	0.02
Variance	0.005	5e-04	4e-04

Table 3.2: Standard deviations and Variances of the disbalance metric discouting the first 20 epochs

Initial Distributions	Low Power	50/50	Random
Standard Deviation	0.02	0.02	0.02
Variance	5e-04	5e-04	4e-04

Table 3.3: Standard deviations of the disbalance metric with varying load

Initial Distributions / Ω	Low Power	50/50	Random
100%	0.11	0.07	0.07
200%	0.08	0.05	0.04
300%	0.07	0.03	0.03
400%	0.07	0.02	0.02
600%	0.07	0.01	0.01

stations. $\rho_{ON} = 1$ and $A_E = 50$ m^2 were considered for the simulations. The transmission power levels value vector used, presented in the form of $\frac{E_b}{N_0}$ is $Q = \{28, 58\}$ dB.

Figures 3.30, 3.31 and 3.32 show that regarding the queuing delay CDF, there is very little to no difference between the initial power level distributions used. Tables 3.3 and 3.4 show the deviations of the *disbalance* metric, where it can be seen that low power initial distribution is the one that shows the greatest deviation, and 50/50 has a similar deviation to the random distribution.

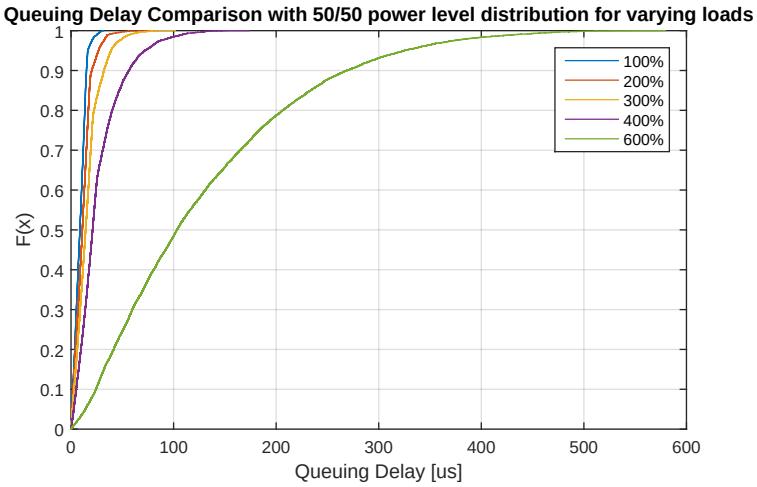


Figure 3.30: Queuing delay CDF for initial distribution of 50/50

Table 3.5 shows the mean and maximum EPUP values for each initial distribution. As it can be seen, the values whether they are average or maximum values, are all close in range, not showing a visible difference. It would be expected that the low power approach

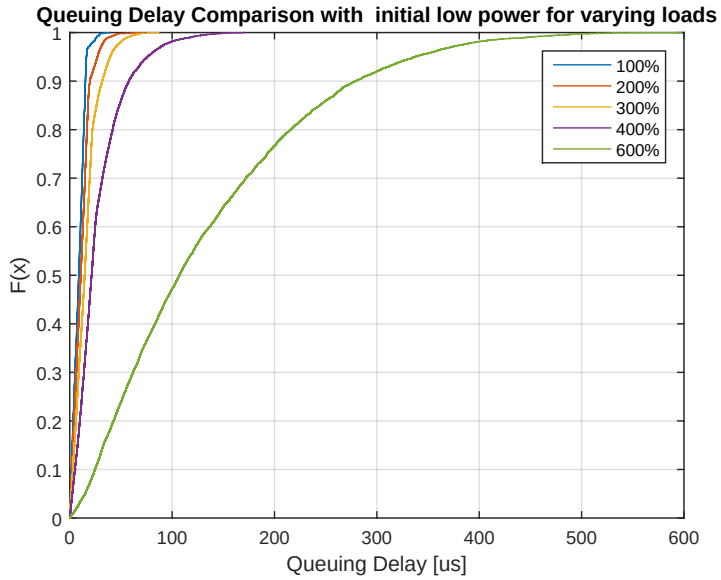


Figure 3.31: Queuing delay CDF for initial distribution of Low Power

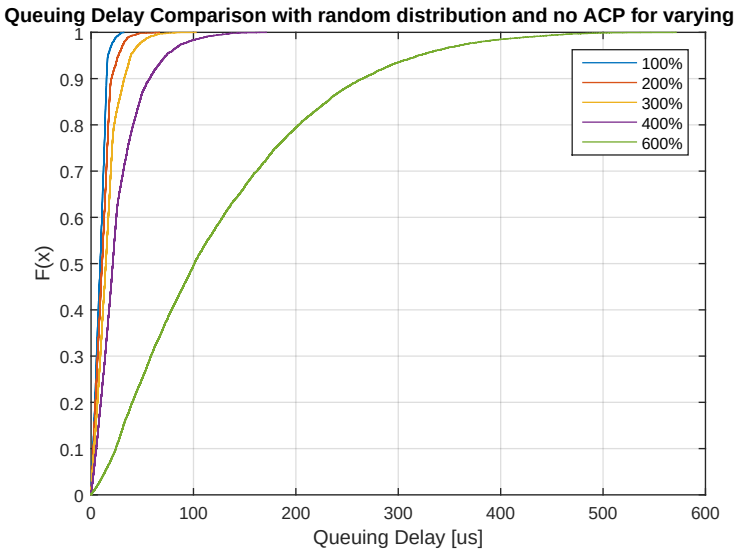


Figure 3.32: Queuing delay cdf for a random initial distribution

Table 3.4: Variances of the disbalance metric with varying load

Initial Distributions / Ω	Low Power	50/50	Random
100%	0.012	0.001	0.001
200%	0.007	0.003	0.002
300%	0.005	0.001	9.9e-04
400%	0.005	5.0e-04	4.2e-04
600%	0.005	8.3e-05	1.7e-04

Table 3.5: Variances of the disbalance metric with varying load

Ω	100%	200%	300%	400%	600%
Low Power mean EPUP	27.41	30.50	32.79	34.39	36.55
Low Power max EPUP	32.68	35.08	37.43	39.01	41.17
50/50 mean EPUP	27.57	30.60	32.53	34.19	36.41
50/50 max EPUP	32.00	35.15	37.17	38.71	40.99
Random mean EPUP	27.08	29.98	32.15	33.90	36.17
Random max EPUP	32.68	35.37	37.18	38.77	40.93

would provide the best results in this field but it is not the case. This can be due to the power control algorithm not working properly. From the delay CDF analysis presented in this section, it can be concluded that there wasn't a meaningful impact on using any of the strategies for power distribution. However, the lack of visible impact can also be due to the preamble overhead the protocol imposes, thus cutting on the possible gains this strategies could achieve, in particular there should have been a visible deviation of when it is used the random power distribution versus the two other strategies. However the values obtained suggest that power was always controlled, which was not the case.

3.8 M2M H-NDMA simulator algorithm

The M2M H-NDMA simulator process is depicted in figure 3.34. At a start of a simulation, the parameters are loaded into the simulator, creating a MATLAB object called system Object, containing all the information about the system.

The system Object is composed by several attributes, some inherited from [28] and others introduced in this dissertation. The representation of the class properties and all the methods are in figure 3.33.

The attributes inherited were the *time* and *maxTime*, variables that represent the current time of the simulator, and the maximum simulator time that is used as a stop condition for the simulator. The *maxTime* is given by $maxTime = P \times \varphi$ where P is the total number of MTs being simulated and φ denotes a number chosen by the user of the simulator. The higher the value, the longer is the simulation time and more packets are generated (in this case, $\varphi = 10$). The *normalized power* is the value in dB from which the power levels of the MTs are computed. *fc* denotes the center frequency, in this case $f_c = 2.5GHz$. *terminal* is the array of objects which contain all the technical and statistical attributes for all terminals and this structure remained unchanged.

As for the attributes introduced or changed in the system Object: the ***NBs*** represents the number of base stations in the system; ***BaseStation*** is the array of objects with all base stations attributes, for example the base stations coordinates; ***powerLevels*** is the vector containing the transmission power values in dB; The ***AssocMap*** contains a matrix, that registers for each terminal what is their associated base station for transmission. This association is done by proximity criteria, being the closest BS to the MT that the terminal

Table 3.6: System Object attributes

Attribute	Description
time	time the simulator is at
P	Number of MTs in the simulator
maxTime	maximum simulator time (used as stop condition for the simulator
normalizedPower	normalized power in dB
fc	centre frequency (GHz)
DistanceMatrix, XiCoefMatrix	distance matrix, with distances of all MTs to all BSs and MTs to MTs. The second matrix has all the channel coefficients values.
terminal	array of objects which contain the properties of all terminals.
NBs	number of base stations in the system
BaseStation	array of objects which contain the properties of all base stations
powerLevels	vector containing the transmission power values in dB
AssocMap	matrix containing information about what MT is associated with which base station (proximity criteria)
PLMap	matrix containing information about all MTs pathloss to all the base stations
TmsInComp	matrix containing information about what MTs are in what CoMP zones

becomes associated to. In the matrix, that is of dimensions of P lines and NBS columns, the association is registered with value 1. All other columns in the same line are 0. The ***PLMap*** is a matrix the same size of ***AssocMap***, containing the pathloss of every MT to every BS in the system, being the pathloss given by the equation in [28],

$$P_L = -(16.9\log_{10} d + 32.8 + 20\log_{10} f_c), \tag{3.19}$$

where d denotes the distance of the MT to the BS.

The ***DistanceMatrix***, is the matrix containing the distances of all MTs to all MTs and also the distances of all MTs to all BSs. The ***XiCoefMatrix*** contains the channel coefficients values representing the difference to the normalized power. Finally, ***TMsInComp*** is the matrix, also of same dimensions as ***AssocMap***, that contains the type of CoMP zone the terminal is. If the terminal is only a COMP 1 zone, the value of the matrix is 999 on the column of the BS that is in the zone of the terminal. Else, depending on the N CoMP zones the MT is, value 1 is set in every column indicating the size of the CoMP zone the MT is and the Base stations involved for that MT CoMP zone. Table 3.6 summarizes the attributes of the system Object.

During the loading of parameters, the base stations coordinates and the terminals attributes are set. Ensuing, the *distanceMatrix* is calculated, and the *PLMap* and *AssocMap* are defined. After these matrices are set, *TmsInComp* matrix is set, creating a map of the existing CoMP Zones in the system and the MTs in each zone.

Once this classification of the MTs per CoMP zone is established, the maximum power transmit constraint is checked. All MTs are tested for the maximum NOMA power level. If they can transmit on that power level, then are also suitable to transmit in the lower level but if they fail the verification, those MTs are constrained to transmitting only on the lowest power level (considering 2 NOMA levels).

Finally, an initial power level distribution is chosen and the *XiMatrix* is computed. Figure 3.36 depicts this process. After all the attributions and computations are done, the simulator enters into a loop. The MTs generate Poisson traffic with a mean Ω_{MT} common for all MTs. Hence, the aggregated load can be defined as

$$\Omega = \Omega_{MT} P. \tag{3.20}$$

At the start of each cycle, the condition *time*>=*maxTime* is checked. If false, the simulator uses the current *time* to update all queues of all terminals and checks if there are any terminals with packets pending. If there aren't, the simulator jumps to the next event time (*time* = *nextEventTime*). Otherwise, the transmission cycle (Epoch Control), runs. The detailed epoch power control algorithm described in section 3.3.1.2, runs inside the Epoch Control block, and is also depicted in figure 3.35.

The receptions are simulated by running several iterations of the IB-DFE receiver for each time slot. To decide which packets get delivered, each reception PER is compared to a uniform random number in the interval $[0,1]$. If the PER is inferior to this number, the packet is assumed to have been successfully delivered to its destination. In every second transmission in a given epoch, the epoch power control algorithm is run, and the *XiMatrix* is updated, except when random power configuration is selected. Lastly, the *time* variable is updated by adding the duration of the control and data packets used in the transmission cycle.

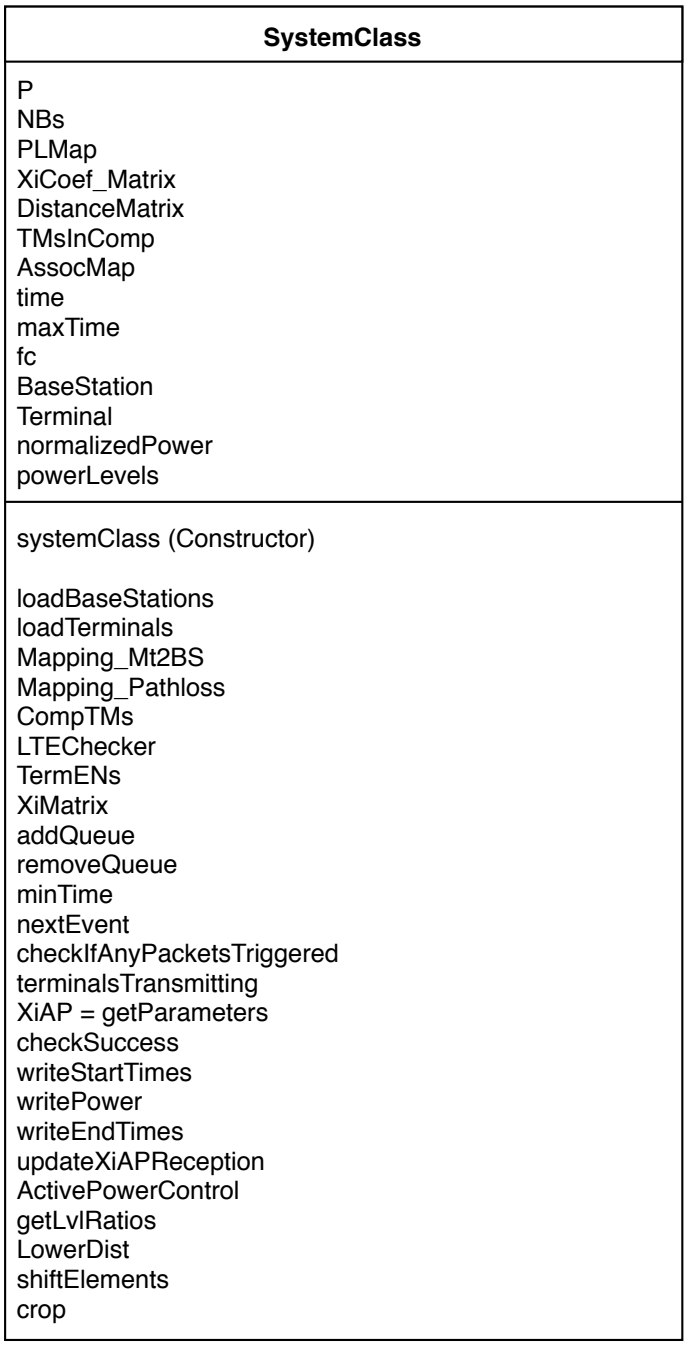


Figure 3.33: Representation of the systemClass and all the methods (bottom rectangle).

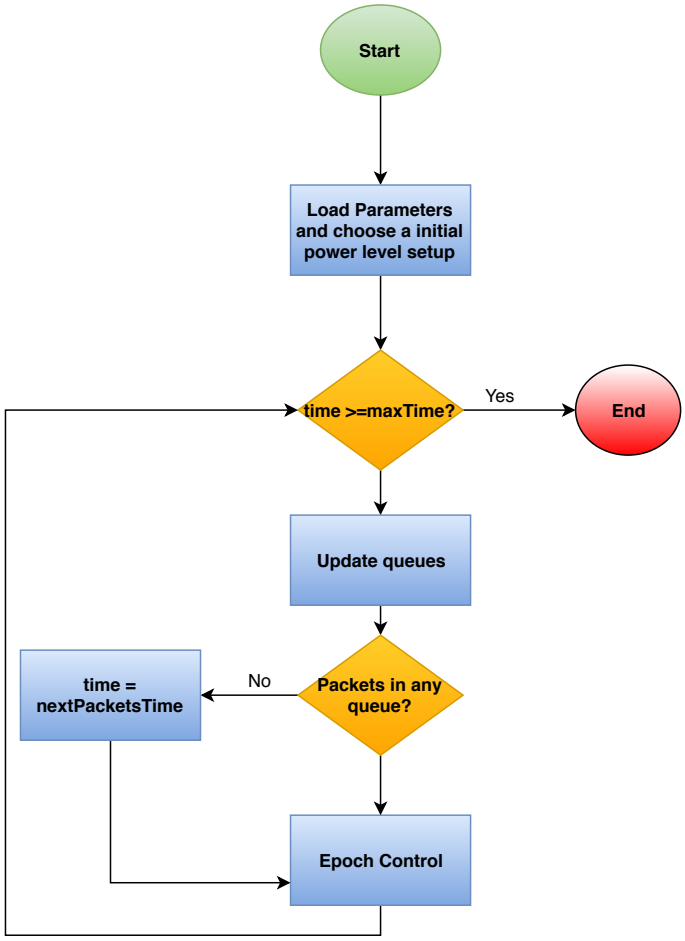


Figure 3.34: Simulator diagram

3.9 Simulator user guide

This section aims to describe how a user should operate the simulator, and where to perform the needed settings to conduct the simulations.

The first thing the user should do, is define how many terminals and base stations he wants in the system, and also the area of the scenario and terminal density. For this configurations, the user should go to the file **TentPoissonGenMts3.m**. An example of a scenario generated is depicted in figure 3.37. After defining the number of terminals, the user should go to the file **TerminalParameters2.m** to generate the various data sets with different aggregated loads. In the file, the desired number of terminals should be inserted, and also how many data slots will the simulation have.

Once all of that data is done, the user should select one data set with the desired load to test, and put them into the simulator folder. The same must be done with the scenario data set.

The user should then open the file **mainScript.m**, where he should set the parameter P and NBS for the number of terminals and base stations being tested, and can change other parameters such as normalized power and directional gain of the antennas to suit

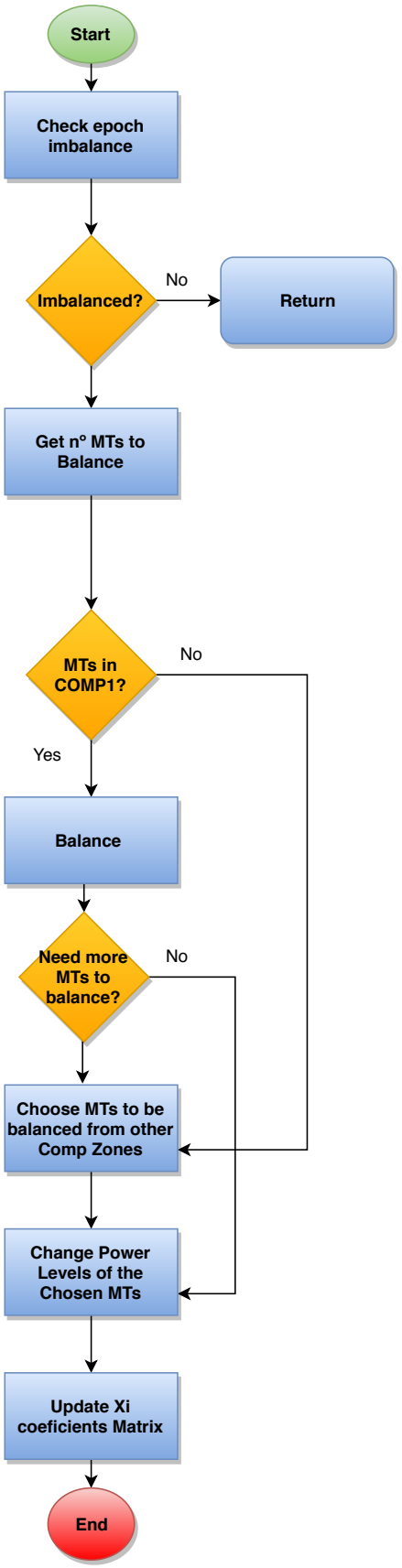


Figure 3.35: Epoch power control algorithm

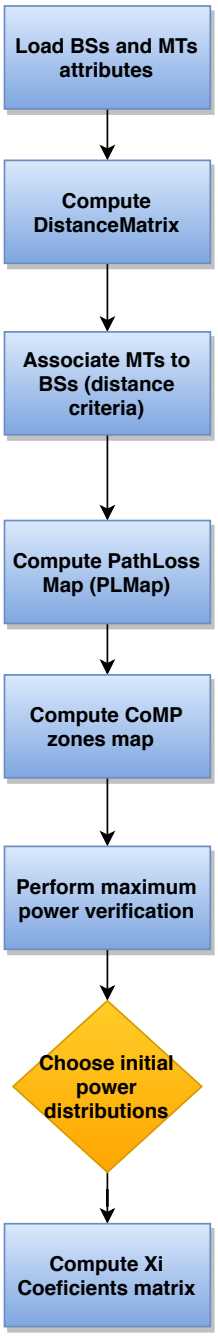


Figure 3.36: Process before the simulator enters into the transmission loop.

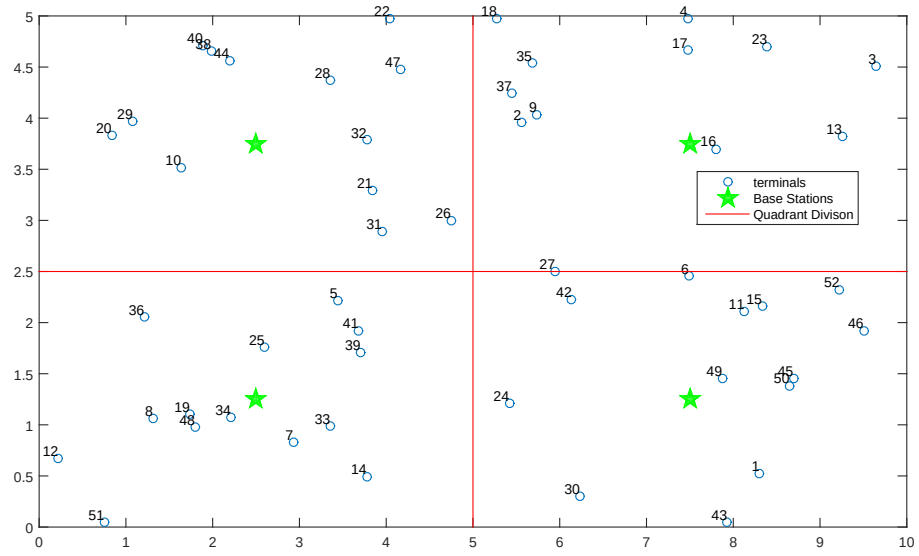


Figure 3.37: Example of a generated scenario.

the user's simulation goal.

Finally, the user should go the **systemClass.m** script, and change the values of the desired power levels to test, and should go to the function *CompTMs*, and select an initial power distribution for the MTs. All that is left to do once everything is done, is simply to run the **mainScript.m**. The simulator will then run according to all the specifications set previously, and in the end will generate an output in the form of a MATLAB data file, that contains the system Object parameters and matrices, and two tables - a *statsTable* with statistics per user and another table, the *logTable* that contains the records of transmissions powers, number of transmissions and the terminals transmitting in each epoch. Figures 3.38 and 3.39 depict an example of the output of these two tables.

3.10 Conclusions

Considering the results obtained in this chapter, it can be concluded that through the spatial densification of the protocol, it is possible to support more terminals and higher loads, achieving lower latencies and improved reliability while also increasing the throughput, although at the expense of increasing the gap between the NOMA power levels. Nonetheless, with the introduction of a DAS, the protocol shows good indications that it may be suitable to provide URLLC services. Power control algorithms must be improved as well as power distribution strategies to lower the EPUP, with the goal of lowering energy expenditure and increasing power saving, which is also a key requirement for URLLC.

statsTable

19x22 cell

	1	2	3	4	5	6	7	8	9
1		'User 1 (HD...	'User 2 (HD...	'User 3 (HD...	'User 4 (HD...	'User 5 (HD...	'User 6 (HD...	'User 7 (HD...	'User 8 (HD
2	'Min Nro. TX'	1	1	2	1	1	2	2	2
3	'Max Nro. TX'	2	2	2	2	2	2	2	2
4	'Average Nr...	1.9444	1.8966	2	1.9677	1.9355	2	2	2
5	'Min Queui...	0.0587	0.6843	0.6375	0.0472	1.0796	0.0595	0.0613	0.3327
6	'Max Queui...	19.4438	15.4134	24.2115	6.1232	16.0048	9.1266	10.9105	8.0736
7	'Average Q...	9.7659	6.1534	11.1707	2.5741	5.9637	3.5406	4.8618	3.4627
8	'Min Servic...	2.2510	2.2510	3.2520	2.2510	2.2510	3.2520	3.2520	3.2520
9	'Max Servic...	3.2520	3.2520	3.2520	3.2520	3.2520	3.2520	3.2520	3.2520
10	'Average Se...	3.1964	3.1484	3.2520	3.2197	3.1874	3.2520	3.2520	3.2520
11	'Successfull...	36	29	34	31	31	33	32	26
12	'Unsuccessf...	34	26	34	30	29	33	32	26
13	'Total Trans...	70	55	68	61	60	66	64	52
14	'Packets Re...	0	0	0	0	0	0	0	0
15	'Dropped P...	0	0	0	0	0	0	0	0
16	'Packet Deli...	0.5143	0.5273	0.5000	0.5082	0.5167	0.5000	0.5000	0.5000
17	'Throughput'	0.2250	0.1813	0.2125	0.1938	0.1938	0.2063	0.2000	0.1625
18	'Total Energy'	47.4510	46.4036	37.3251	46.8533	46.7815	37.1954	37.0618	46.1600
19	'EPUP'	31.8880	31.7796	22.0103	31.9397	31.8679	22.0103	22.0103	32.0103
20									

Figure 3.38: Example of a Stats Table with 20 MTs and 400% load .

	1	2	3	4	5	6	7	8
1	'Epoch Start'	'Epoch End'	'Operating ...	'Terminals T...	'DATA Slots...	'TX Summa...	'Xi'	'PER'
2	0.2188	2.4698	[]	'1'	1	*2x2 cell*	*2x3 cell*	*2x1 cell*
3	2.6895	5.9415	[]	'1,3,5,10,12,...	2	*3x11 cell*	*3x3 cell*	*3x1 cell*
4	6.1613	9.4133	[]	'1,2,3,4,9,10,...	2	*3x14 cell*	*3x3 cell*	*3x1 cell*
5	9.6330	12.8850	[]	'1,2,3,4,5,8,1...	2	*3x15 cell*	*3x3 cell*	*3x1 cell*
6	13.1047	16.3567	[]	'1,2,3,8,11,1...	2	*3x12 cell*	*3x3 cell*	*3x1 cell*
7	16.5765	19.8285	[]	'2,3,4,6,8,9,1...	2	*3x15 cell*	*3x3 cell*	*3x1 cell*
8	20.0483	23.3003	[]	'2,3,4,9,11,1...	2	*3x12 cell*	*3x3 cell*	*3x1 cell*
9	23.5200	26.7720	[]	'1,2,3,4,7,8,1...	2	*3x12 cell*	*3x3 cell*	*3x1 cell*
10	26.9918	30.2438	[]	'1,3,4,5,7,8,9...	2	*3x13 cell*	*3x3 cell*	*3x1 cell*
11	30.4635	33.7155	[]	'1,2,3,4,5,6,8...	2	*3x17 cell*	*3x3 cell*	*3x1 cell*
12	33.9353	37.1873	[]	'1,3,4,5,6,7,8...	2	*3x17 cell*	*3x3 cell*	*3x1 cell*
13	37.4070	40.6590	[]	'1,3,4,5,7,8,9...	2	*3x15 cell*	*3x3 cell*	*3x1 cell*
14	40.8788	44.1307	[]	'1,3,4,5,7,8,1...	2	*3x16 cell*	*3x3 cell*	*3x1 cell*
15	44.3505	47.6025	[]	'1,3,4,5,8,9,1...	2	*3x16 cell*	*3x3 cell*	*3x1 cell*
16	47.8222	51.0742	[]	'1,2,3,4,6,8,9...	2	*3x17 cell*	*3x3 cell*	*3x1 cell*
17	51.2940	54.5460	[]	'1,2,3,6,8,10,...	2	*3x16 cell*	*3x3 cell*	*3x1 cell*
18	54.7657	58.0177	[]	'1,3,4,6,9,11,...	2	*3x12 cell*	*3x3 cell*	*3x1 cell*
19	58.2375	61.4895	[]	'1,3,4,5,6,7,9...	2	*3x17 cell*	*3x3 cell*	*3x1 cell*
20	61.7092	64.9612	[]	'1,3,5,6,7,9,1...	2	*3x16 cell*	*3x3 cell*	*3x1 cell*

Figure 3.39: Example of a Log Table with 20 MTs and 400% load.

Chapter 4

Conclusions

4.1 Final Considerations

This dissertation focused on proposing an adaptation of the M2M H-NDMA protocol to provide URLLC for M2M services with power saving mechanisms. Chapter 2 characterized the existing enablers for URLLC as well as the significant access schemes and existing M2M MAC protocols, focusing on the characterization of the mission critical MACs. Chapter 3 described the new system and scenario considered in this dissertation for the M2M H-NDMA protocol and analyzed the changes needed to be made. With the new scenario of a DAS, it was found that the previous transmission power levels considered in the previous version of the protocol were not adequate. Results showed that for guaranteeing reliability, the power gap should be, for a system up to 4BSs, of 20dB between NOMA power levels. Power control schemes were implemented, in order to try to reduce the delay and the EPUP, to comply to URLLC requirements. Results showed that the various power schemes tested performed similarly, and did not present a major improvement in delay or energy, although being able to guarantee reliability and a reason for it to happen could be due to the preamble overhead of the protocol. Finally, in this dissertation, the evolved M2M-HNDMA detailed algorithm was presented, along with a user guide to assist future users in the use of the simulator.

4.2 Future Work

Future work that could be done in the topic of this dissertation, includes the study of more complex multi-hop scenarios, the re-framing of the packet structure in order to reduce the preamble overhead, and the study providing URLLC services with a hybrid random/scheduled access scheme for M2M networks.

Bibliography

[1] A. Rico-Alvariño, M. Vajapeyam, H. Xu, X. Wang, Y. Blankenship, J. Bergman, T. Tirronen, and E. Yavuz. "An overview of 3GPP enhancements on machine to machine communications." In: *IEEE Communications Magazine* 54.6 (2016), pp. 14–21.

[2] Z. Dawy, W. Saad, A. Ghosh, J. G. Andrews, and E. Yaacoub. "Toward Massive Machine Type cellular Communications." In: *IEEE Wireless Communications* 24.1 (2016), pp. 120–128.

[3] G. Fettweis and S. Alamouti. "5G: Personal mobile internet beyond what cellular did to telephony." In: *IEEE Communications Magazine* 52.2 (2014), pp. 140–145.

[4] NGMN Alliance. "NGMN 5G White Paper." In: *Ngmn* (2015), pp. 1–125. URL: https://www.ngmn.org/5g-white-paper/5g-white-paper.html (visited on 10/24/2017).

[5] C.-p. Li, J. Jiang, W. Chen, T. Ji, and J. Smee. "5G Ultra-Reliable and Low-Latency Systems Design." In: *2017 European Conference on Networks and Communications* (2017), pp. 1–5.

[6] C. Bockelmann, N. Pratas, H. Nikopour, K. Au, T. Svensson, C. Stefanovic, P. Popovski, and A. Dekorsy. "Massive machine-type communications in 5g: Physical and MAC-layer solutions." In: *IEEE Communications Magazine* 54.9 (2016), pp. 59–65.

[7] G. Wunder, P. Jung, M. Kasparick, F. Heinrich, T. Wild, F. Schaich, Y. Chen, S. Brink, I. Gaspar, N. Michailow, A. Festag, L. Mendes, N. Cassiau, D. Ktenas, M. Dryjanski, S. Pietrzyk, B. Eged, P. Vago, and F. Wiedmann. "5GNOW: Non-orthogonal, Asynchronous Waveforms for Future Mobile Applications." In: *IEEE Communications Magazine* 52.2 (2014), pp. 97–105.

[8] P. Agyapong, M. Iwamura, D. Staehle, W. Kiess, and A. Benjebbour. "Design Considerations for a 5G Network Architecture." In: *IEEE Communications Magazine* 52.11 (2014), pp. 65–75.

[9] A. Rajandekar and B. Sikdar. "A survey of MAC layer issues and protocols for machine-to-machine communications." In: *IEEE Internet of Things Journal* 2.2 (2015), pp. 175–186.

[10] X. Xiong, K. Zheng, R. Xu, W. Xiang, and P. Chatzimisios. "Low power wide area machine-to-machine networks: key techniques and prototype." In: *IEEE Communications Magazine* 53.9 (2013), pp. 64–71.

[11] H. Ji, S. Park, J. Yeo, Y. Kim, J. Lee, and B. Shim. "Introduction to Ultra Reliable and Low Latency Communications in 5G." In: (2017). arXiv: 1704.05565.

[12] J. H. Kim, K. J. Choi, K. Lee, and K. S. Kim. "Grant-free Multiple Access for Ultra-Reliable Low-Latency Communications in a Large-Scale Antenna System." In: *The 7th International Conference on Information and Communication Technology Convergence* (2016), pp. 466–470.

[13] M Jamal, B Horia, K Maria, and I Alexandru. "Study of multiple access schemes in 3GPP LTE OFDMA vs. SC-FDMA." In: *Applied Electronics (AE), 2011 International Conference on* (2011), pp. 1–4.

[14] N. Abu-Ali, A. E. M. Taha, M. Salah, and H. Hassanein. "Uplink scheduling in LTE and LTE-advanced: Tutorial, survey and evaluation framework." In: *IEEE Communications Surveys and Tutorials* 16.3 (2014), pp. 1239–1265.

[15] H. Hu, X. Hu, and J. Xiao. "CDMA-based MAC Protocol for Multi-hop Wireless Sensor Networks." In: *2017 2nd IEEE International Conference on Intelligent Transportation Engineering* (2017), pp. 266–271.

[16] X. Wang, T. Ahonen, and J. Nurmi. "Applying CDMA technique to network-on-chip." In: *IEEE Transactions on Very Large Scale Integration (VLSI) Systems* 15.10 (2007), pp. 1091–1100.

[17] Y. Cai, Z. Qin, F. Cui, G. Y. Li, and J. A. McCann. "Modulation and Multiple Access for 5G Networks." In: (2017), pp. 1–31. arXiv: 1702.07673.

[18] J. L. Lu, W. Shu, and M. Y. Wu. "A survey on multipacket reception for wireless random access networks." In: *Journal of Computer Networks and Communications* 2012 (2012).

[19] V. Naware and P. Venkitasubramaniam. "A cross-layer perspective in an uncharted path - Signal processing in random access." In: *IEEE Signal Processing Magazine* 21.5 (2004), pp. 29–39.

[20] E. De Carvalho, E. Bjornson, J. H. Sorensen, P. Popovski, and E. G. Larsson. "Random access protocols for massive MIMO." In: *IEEE Communications Magazine* 55.5 (2017), pp. 216–222.

[21] M. Al-Imari, P. Xiao, M. A. Imran, and R. Tafazolli. "Uplink Non-Orthogonal Multiple Access for 5G Wireless Networks." In: *2014 11th International Symposium on Wireless Communications Systems (ISWCS)* (2014), pp. 781–785.

[22] Y. Liang, X. Li, J. Zhang, and Z. Ding. "Non-Orthogonal Random Access (NORA) for 5G Networks." In: *IEEE Transactions on Wireless Communications* 16.7 (2017), pp. 4817–4831.

[23] Z. Ding, Y. Liu, J. Choi, Q. Sun, M. Elkashlan, I. Chih-Lin, and H. V. Poor. "Application of Non-Orthogonal Multiple Access in LTE and 5G Networks." In: *IEEE Communications Magazine* 55.2 (2017), pp. 185–191.

[24] A. Benjebbour, K. Saito, A. Li, Y. Kishiyama, and T. Nakamura. "Non-orthogonal multiple access (NOMA): Concept, performance evaluation and experimental trials." In: *International Conference on Wireless Networks and Mobile Communications, WINCOM 2015* (2016), pp. 1–6.

[25] Z. Ding, X. Lei, G. K. Karagiannidis, R. Schober, J. Yuan, and V. Bhargava. "A Survey on Non-Orthogonal Multiple Access for 5G Networks: Research Challenges and Future Trends." In: *IEEE Journal on Selected Areas in Communications* 35.10 (2017), pp. 2181–2195.

[26] F. Ganhão, M. Pereira, L. Bernardo, R. Dinis, R. Oliveira, and P. Pinto. "Performance analysis of an hybrid ARQ adaptation of NDMA schemes." In: *IEEE Transactions on Communications* 61.8 (2013), pp. 3304–3317.

[27] R. Samano-Robles and A. Gameiro. "Stability properties of network diversity multiple access with multiple-antenna reception and imperfect collision multiplicity estimation." In: *Journal of Computer Networks and Communications* 2013 (2013).

[28] B. Ramos, L. Bernardo, R. Dinis, R. Oliveira, P. Pinto, and P. Amaral. "Using lightly synchronized MultiPacket Reception in Machine-Type Communication networks." In: *2016 IEEE Globecom Workshops* (2016).

[29] P Suriyachai, U Roedig, and a Scott. "A Survey of MAC Protocols for Mission-Critical Applications in Wireless Sensor Networks." In: *IEEE Communications Surveys & Tutorials* 14.2 (2012), pp. 240 –264.

[30] Z. Zhou, B. Chen, and H. Yu. "Understanding RFID Counting Protocols." In: *IEEE/ACM Transactions on Networking* 24.1 (2016), pp. 312–327.

[31] W. Luo, S. Chen, Y. Qiao, and T. Li. "Missing-tag detection and energy-time tradeoff in large-scale RFID systems with unreliable channels." In: *IEEE/ACM Transactions on Networking* 22.4 (2014), pp. 1079–1091.

[32] Y. Z. Zhao, C. Miao, M. Ma, J. B. Zhang, and C. Leung. "A survey and projection on medium access control protocols for wireless sensor networks." In: *ACM Computing Surveys* 45.1 (2012), pp. 1–37.

[33] H. Wu, C. Zhu, R. J. La, X. Liu, and Y. Zhang. "Fast adaptive S-ALOHA scheme for event-driven machine-to-machine communications." In: *IEEE Vehicular Technology Conference Fall* (2012), pp. 1–5.

[34] D. Tarchi, R. Fantacci, and D. Marabissi. "Proposal of a cognitive based MAC protocol for M2M environments." In: *IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, PIMRC* (2013), pp. 1609–1613.

[35] G. Corrales Madueno, C. Stefanovic, and P. Popovski. "Reliable and Efficient Access for Alarm-Initiated and Regular M2M Traffic in IEEE 802.11ah Systems." In: *IEEE Internet of Things Journal* 3.5 (2016), pp. 673–682.

[36] N. K. Pratas, H. Thomsen, C. Stefanovic, and P. Popovski. "Code-expanded random access for machine-type communications." In: *2012 IEEE Globecom Workshops, GC Wkshps 2012*. 2012, pp. 1681–1686.

[37] Y. Chen and W. Wang. "Machine-to-Machine communication in LTE-A." In: *2010 IEEE 72nd Vehicular Technology Conference - Fall* (2010), pp. 1–4.

[38] G. Wang, X. Zhong, S. Mei, and J. Wang. "An adaptive medium access control mechanism for cellular based machine to machine (M2M) communication." In: *2010 IEEE International Conference on Wireless Information Technology and Systems, ICWITS 2010* (2010), pp. 1–4.

[39] N. Kouzayha, M. Abdallah, J. Costantine, and Z. Dawy. "Sensor applications Energy Efficient IoT Sensor With RF Wake-Up and Addressing Capability." In: *IEEE Sensors Letters* 1.6 (2017), pp. 1–4.

[40] M. Magno and L. Benini. "An Ultra Low Power High Sensitivity Wake-Up Radio Receiver with Addressing Capability." In: *Wireless and Mobile Computing, Networking and Communications (WiMob), 2014 IEEE 10th International Conference on* (2014), pp. 92–99.

[41] L. Gu and J. A. Stankovic. "Radio-triggered wake-up for wireless sensor networks." In: *Real-Time Systems* 29.2-3 (2005), pp. 157–182.

[42] A. H. Sakr and E. Hossain. "Location-Aware cross-tier coordinated multipoint transmission in two-tier cellular networks." In: *IEEE Transactions on Wireless Communications* 13.11 (2014), pp. 6311–6325.

[43] S. M. Islam, N. Avazov, O. A. Dobre, and K. S. Kwak. "Power-Domain Non-Orthogonal Multiple Access (NOMA) in 5G Systems: Potentials and Challenges." In: *IEEE Communications Surveys and Tutorials* 19.2 (2017), pp. 721–742.

[44] P. Marsch and G. Fettweis. "A framework for optimizing the uplink performance of distributed antenna systems under a constrained backhaul." In: *IEEE International Conference on Communications* (2007), pp. 975–979.

[45] F. Ganhão, R. Dinis, L. Bernardo, and R. Oliveira. "Analytical BER and PER performance of frequency-domain diversity combining, multipacket detection and hybrid schemes." In: *IEEE Transactions on Communications* 60.8 (2012), pp. 2353–2362.

[46] N. Benvenuto and S. Tomasin. "Iterative design and detection of a DFE in the frequency domain." In: *IEEE Transactions on Communications* 53.11 (2005), pp. 1867–1875.

[47] M. Bennis, M. Debbah, and H. V. Poor. "Ultra-Reliable and Low-Latency Wireless Communication: Tail, Risk and Scale." In: (2018), pp. 1–26. arXiv: 1801.01270.

Rahim Karim Shamsudin

Licenciado em Ciências da Engenharia Electrotécnica e de Computadores

Protocol for Extreme Low Latency M2M Communication Networks

Dissertação para obtenção do Grau de Mestre em

Engenharia Electrotécnica e de Computadores

Setembro, 2018

FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE NOVA DE LISBOA

Rahim Karim Shamsudin

Licenciado em Ciências da Engenharia Electrotécnica e de Computadores

Protocol for Extreme Low Latency M2M Communication Networks

Dissertação para obtenção do Grau de Mestre em

Engenharia Electrotécnica e de Computadores

Setembro, 2018

FACULDADE DE
CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE NOVA DE LISBOA

Rahim Shamsudin 2018
Protocol for Extreme Low Latency M2M Communication Networks 2018