Università degli Studi di Padova

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE
Corso di Dottorato in Scienza e Tecnologia dell'Informazione

Tesi di Dottorato di Ricerca

On the Support of Massive Machine-to-Machine Traffic in Heterogeneous Networks and Fifth-Generation Cellular Networks

Supporto di Traffico Massivo di Tipo Machine-to-Machine
in Reti Eterogenee e Reti Cellulari di Quinta Generazione

Coordinatore
Ch.mo Prof. Andrea Neviani

Candidato
Marco Centenaro

Supervisore
Ch.mo Prof. Lorenzo Vangelista

Anno Accademico 2016–2017

Alla mia famiglia, per il costante supporto,
e a nonna Rina, mancata troppo presto.
まりかちゃん、ありがとう！

Ignoranti quem portum petat nullus suus ventus est –
Non esiste vento favorevole per il marinaio che non sa dove andare.

Lucius Annaeus Seneca

Abstract

The widespread availability of many emerging services enabled by the Internet of Things (IoT) paradigm passes through the capability to provide long-range connectivity to a massive number of *things*, overcoming the well-known issues of ad-hoc, short-range networks. This scenario entails a lot of challenges, ranging from the concerns about the radio access network efficiency to the threats about the security of IoT networks. In this thesis, we will focus on wireless communication standards for long-range IoT as well as on fundamental research outcomes about IoT networks. After investigating how Machine-Type Communication (MTC) is supported nowadays, we will provide innovative solutions that i) satisfy the requirements in terms of scalability and latency, ii) employ a combination of licensed and license-free frequency bands, and iii) assure energy-efficiency and security.

Sommario

La diffusione capillare di molti servizi emergenti grazie all'*Internet of Things (IoT)* passa attraverso la capacità di fornire connettività senza fili a lungo raggio ad un numero massivo di *cose*, superando le note criticità delle reti ad hoc a corto raggio. Questa visione comporta grandi sfide, a partire dalle preoccupazioni riguardo l'efficienza delle rete di accesso fino alle minacce alla sicurezza delle reti IoT. In questa tesi, ci concentreremo sia sugli standard di comunicazione a lungo raggio per l'IoT sia sulla ricerca di base per le reti IoT. Dopo aver analizzato come vengono supportate le comunicazioni *Machine-to-Machine (M2M)* oggi, forniremo soluzioni innovative le quali i) soddisfano i requisiti in termini di scalabilità e latenza, ii) utilizzano una combinazione di bande di frequenza licenziate e libere e iii) assicurano efficienza energetica e sicurezza.

Contents

Preface

笑顔は生きるエネルギー
The smile is your life force.

Japanese proverb

This thesis is the result of three years of commitment and dedication. I spent the first two years at the Department of Information Engineering (DEI), University of Padova, under the supervision of Prof. Lorenzo Vangelista. During the third year, I had two very important abroad sojourns. I was first a research intern at Nokia Bell Labs Stuttgart, Germany, under the supervision of Dr. Stephan Saur, to study the integration of Internet of Things (IoT) traffic into fifth-generation (5G) cellular networks. Then, I was a visiting researcher at the Yokohama National University, Yokohama, Japan, under the supervision of Prof. Ryuji Kohno, to investigate dependable radio access protocols for Low-Power Wide Area Networks (LPWANs). This thesis was reviewed by Prof. Carlo Fischione (KTH Royal Institute of Technology, Sweden) and Prof. Cedomir Stefanovic (Aalborg University, Denmark).

At the end of this experience, I can say that I am an entirely different person with respect to three years ago. I want to sincerely thank my supervisor Prof. Lorenzo Vangelista for his guidance, both from the professional side and the human side. If my abroad sojourns were fruitful, I due that to my co-supervisors, Dr. Saur and Prof. Kohno: Vielen Dank and ありがとうございました. Let me express gratitude to Nokia and "Fondazione Ing. Aldo Gini" of the University of Padova for funding my experiences in Stuttgart and Yokohama, respectively. Thanks a lot to my laboratory colleagues at DEI, with a particular mention to Dr. Gianluca Caparra, and future doctors Davide Magrin and Michele Polese, with whom I actively collaborated on research projects. A sincere acknowledgement goes to Dr. Andreas Weber of Nokia Bell Labs Stuttgart, Ivano Calabrese and Nicola Bressan of Patavina Technologies.

Let me finally thank the external referees, Prof. Fischione and Prof. Stefanovic, for their valuable comments and suggestions, which helped in improving the overall quality of this thesis.

Padova, January 15, 2018.

Marco Centenaro

笑顔は生きるエネルギー
Il sorriso è la forza della tua vita.

Proverbio giapponese

Questo lavoro è il risultato di tre anni di impegno e dedizione. Ho trascorso i primi due anni del corso di dottorato presso il Dipartimento di Ingegneria dell'Informazione (DEI) dell'Università di Padova, sotto la guida del Prof. Lorenzo Vangelista, mentre nel corso del terzo anno ho avuto la possibilità di intraprendere due interessanti e proficui periodi di formazione all'estero. Il primo periodo si è tenuto presso i Nokia Bell Labs di Stoccarda, Germania, sotto la supervisione del Dott. Stephan Saur, allo scopo di studiare l'integrazione del traffico generato dall'Internet delle cose nella rete cellulare di quinta generazione (5G). In seguito, ho trascorso un periodo di ricerca presso la Yokohama National University (YNU), a Yokohama, Giappone, supervisionato dal Prof. Ryuji Kohno su tematiche riguardanti protocolli di accesso per Low-Power Wide Area Networks (LPWANs). Il presente elaborato è stato revisionato dai professori Carlo Fischione (KTH Royal Institute of Technology, Svezia) e Cedomir Stefanovic (Aalborg University, Danimarca).

Al termine di questo percorso, ritengo di essere profondamente cambiato rispetto a quando iniziai il dottorato di ricerca. Ringrazio sinceramente il mio supervisore Prof. Lorenzo Vangelista per il suo ruolo di guida, sia sotto l'aspetto professionale che umano. Se i due periodi di formazione all'estero sono stati fruttuosi, lo devo ai miei supervisori in loco, Dott. Saur e Prof. Kohno: Vielen Dank, ありがとうございました. Consentitemi di esprimere la mia gratitudine verso Nokia e la Fondazione Ing. Aldo Gini dell'Università di Padova per aver sostenuto economicamente i soggiorni a Stoccarda e Yokohama, rispettivamente. Grazie mille a tutti i colleghi del laboratorio presso il DEI, con particolare riferimento al Dott. Gianluca Caparra e ai futuri dottori Davide Magrin e Michele Polese, con i quali ho collaborato attivamente. Un ringraziamento va anche al Dott. Andreas Weber dei Nokia Bell Labs di Stoccarda, a Ivano Calabrese e Nicola Bressan di Patavina Technologies.

Ringrazio infine i valutatori esterni, Proff. Fischione e Stefanovic, per i validi commenti e suggerimenti che hanno contribuito a migliorare la qualità complessiva dell'elaborato.

Padova, 15 gennaio 2018.

Marco Centenaro

List of Symbols

Statistics

$\mathbb{E}[X]$	expected value of random variable X
$\mathbb{P}[X]$	probability of event X
$f_X(x)$	Probability Distribution Function (PDF) of random variable X
$F_X(x)$	Cumulative Distribution Function (CDF) of random variable X
$F_X^c(x)$	Complementary CDF (CCDF) of random variable X
$\mathcal{N}(\mu, \sigma^2)$	Gaussian random variable with mean μ and variance σ^2
$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{R})$	Gaussian random vector with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{R}$
$I_x(a, b)$	regularized incomplete Beta function of parameters a and b

Linear Algebra

$\boldsymbol{X}$	matrices are denoted by uppercase bold letters
X_{ij}	element in position (i, j) of matrix $\boldsymbol{X}$
$\boldsymbol{X}^*$	complex conjugate of matrix $\boldsymbol{X}$
$\boldsymbol{X}^T$	transpose of matrix $\boldsymbol{X}$
$\boldsymbol{I}_n$	identity matrix of size n
$\mathbf{0}_{m\times n}$	matrix of zeros with m rows and n columns
$\mathbf{1}_{m\times n}$	matrix of ones with m rows and n columns
$\boldsymbol{x}$	vectors are denoted by lowercase bold letters
x_i or $[\boldsymbol{x}]_i$	i-th element of vector $\boldsymbol{x}$
$\boldsymbol{x}^T \cdot \boldsymbol{y} = \sum_\ell x_\ell y_\ell$	inner product between $\boldsymbol{x}$ and $\boldsymbol{y}$
$\|\boldsymbol{x}\| \triangleq (\boldsymbol{x}^T \boldsymbol{x})^{1/2}$	norm of vector $\boldsymbol{x}$
$\|\boldsymbol{x}\|_H$	Hamming weight of binary vector $\boldsymbol{x}$

Set Theory

$\mathcal{A}$	sets are denoted by calligraphic letters
$\|\mathcal{A}\|$	cardinality of set $\mathcal{A}$

Telecommunications

B	bandwidth
N_0	Additive White Gaussian Noise (AWGN) power spectral density
SINR	Signal-to-Interference-plus-Noise-Ratio (SINR)
SNR	Signal-to-Noise Ratio (SNR)
D_C	shadowing decorrelation distance
$\mathcal{S}$	throughput

List of Acronyms

2G	second-generation
3G	third-generation
3GPP	3rd Generation Partnership Project
4G	fourth-generation
5G	fifth-generation
ACB	Access Class Barring
ACK	acknowledgement
ADR	Adaptive Data Rate
AFA	Adaptive Frequency Agility
AG	Access Granted
AGTI	Access Grant Time Interval
API	Application Programming Interface
ARPU	Average Revenue Per User
AWGN	Additive White Gaussian Noise
BLE	Bluetooth Low Energy
BS	Base Station
BSR	Buffer Status Report
CCDF	Complementary CDF
CDF	Cumulative Distribution Function
CDMA	Code Division Multiple Access
CIoT	Cellular IoT
CR	Connection Request
CRC	Cyclic Redundancy Check
C-RNTI	Cell Radio-Network Temporary Identifier

CSCG	Circularly Symmetric Complex Gaussian
CS-MUD	Compressive Sensing-based Multi-User Detection
CSS	Chirp Spread Spectrum
DCI	Downlink Control Information
DL	downlink
DQRAP	Distributed Queuing Random Access Protocol
DSSS	Direct Sequence Spread Spectrum
EAB	Extended Access Barring
ECDF	Empirical CDF
EC-GSM	Extended Coverage GSM
eNB	eNodeB
EPC	Evolved Packet Core
ERP	Effective Radiated Power
ETSI	European Telecommunications Standards Institute
EWL	External Wall Loss
FA	False Alarm
FASA	Fast Adaptive Slotted ALOHA
FBMC	Filter Bank Multi-Carrier
FDM	Frequency Division Multiplexing
FDMA	Frequency Division Multiple Access
F-OFDM	Filtered-OFDM
GFSK	Gaussian Frequency Shift Keying
GLRT	Generalized Likelihood Ratio Test
GNSS	Global Navigation Satellite System
GSM	Global System for Mobile Communications
HARQ	Hybrid Automatic Repeat Request
H2H	Human-to-Human
IEEE	Institute of Electrical and Electronics Engineers
i.i.d.	independent and identically distributed
IMSI	International Mobile Subscriber Identity

IoT	Internet of Things
ISBI	Inter-Service-Band-Interference
ISM	Industrial, Scientific, and Medical
KPI	Key Performance Indicator
LBS	Location-Based Service
LBT	Listen-Before-Talk
LLR	Log-Likelihood Ratio
LTN	Low Throughput Network
LoRa	Long-Range^{TM}
LoRaWAN	Long-Range Wide Area Network^{TM}
LPWAN	Low-Power Wide Area Network
LRT	Likelihood Ratio Test
LTE	Long-Term Evolution
M2M	Machine-to-Machine
MAC	Medium Access Control
MCS	Modulation and Coding Scheme
MD	Missed Detection
MDS	Maximum Delay Spread
MIMO	Multiple-Input-Multiple-Output
ML	Maximum Likelihood
MME	Mobility Management Entity
MMPP	Markov-Modulated Poisson Process
MPR	Multi-Packet Reception
MTC	Machine-Type Communication
MTD	Machine-Type Device
M-TMSI	MME Temporary Mobile Subscriber Identity
MUD	Multi-User Detection
NACK	not-acknowledgement
NB-IoT	Narrowband IoT
NE	Nash Equilibrium

NFC	Near Field Communications
NFV	Network Function Virtualization
NS	Network Server
OFDMA	Orthogonal Frequency Division Multiple Access
PAN	Personal Area Network
OFDM	Orthogonal Frequency Division Multiplexing
PAN	Personal Area Network
PIB	PAN Information Base
PDF	Probability Distribution Function
PDP	Power Delay Profile
PDSCH	Physical Downlink Shared Channel
PHY	Physical Layer
PIB	PAN Information Base
PLE	Path Loss Exponent
PMF	Probability Mass Function
PRACH	Physical Random Access Channel
QoE	Quality of Experience
QoS	Quality of Service
PUCCH	Physical Uplink Control Channel
PUSCH	Physical Uplink Shared Channel
RA	Random Access
RAO	Random Access Opportunity
RAN	Radio Access Network
RAR	Random-Access Response
RB	Resource Block
RE	Resource Element
RF	Radio Frequency
RFID	Radio-Frequency IDentification
RRC	Radio Resource Control
RSSI	Received Signal Strength Indicator

SDN	Software Defined Networking
SIB2	System Information Broadcast 2
SIC	Successive Interference Cancelation
SINR	Signal-to-Interference-plus-Noise-Ratio
SISO	Single-Input-Single-Output
SF	Spreading Factor
SMM	Semi-Markov Model
SNR	Signal-to-Noise Ratio
SPB	Small Packet Block
SOOC	Self-Optimizing Overload Control
SQP	Sequential Quadratic Programming
SR	Scheduling Request
TA	Timing Alignment
TAU	Tracking Area Update
TC-RNTI	Temporary C-RNTI
TDD	Time Division Duplex
TDM	Time Division Multiplexing
TDMA	Time Division Multiple Access
ToA	Time on Air
TTI	Transmission Time Interval
UE	User Equipment
UFMC	Universal-Filtered Multi-Carrier
UF-OFDM	Universal-Filtered OFDM
UL	uplink
UMTS	Universal Mobile Telecommunication System
UNB	Ultra-Narrow Band
UWB	Ultra-Wide Band
VNI	Visual Networking Index
WAN	Wide Area Network
WBAN	Wireless Body Area Network

WLAN Wireless Local Area Network

WSAN Wireless Sensors and Actuators Networks

WSN Wireless Sensor Network

ZC Zadoff-Chu

List of Figures

List of Tables

Introduction

> Every day sees humanity more victorious in the struggle with space and time.
>
> Guglielmo Marconi

As telecommunication technologies continue to evolve rapidly, fueling the growth of service coverage and capacity, new use cases and applications are being identified. Many of these new business areas (e.g., smart metering, in-car satellite navigation, e-health monitoring, smart cities) involve fully-automated communication between devices, without human intervention. This new form of communication is generally referred to as Machine-to-Machine (M2M) Communication, or Machine-Type Communication (MTC), while the involved devices are called Machine-Type Devices (MTDs). Examples of common MTDs are environmental and biomedical sensors, actuators, meters, radio-frequency tags, but also smartphones, tablets, vehicles, cameras, and so on. The number and typology of MTDs are continuously growing, together with the set of M2M applications and services that they enable. As a matter of fact, MTDs are key elements in the emerging Internet of Things (IoT) and Smart City paradigms [1, 2], which are expected to provide solutions to current and future social-economical demands for sensing and monitoring services, as well as for new applications, business models, and industrial sectors, including building and industrial automation, remote and mobile healthcare, elderly assistance, intelligent energy management and smart grids, automotive, smart agriculture, traffic management, and many others [3, 4].

The development of the IoT is an extremely challenging topic, and the debate on how to put it into practice is still open. The discussion is involving all layers of the protocol stack, from the physical transmission up to data representation and service composition [5]. However, the whole IoT architecture rests on the wireless technologies that are used to provide data access to the end devices [6]. For many years, short-range transmission technologies have been considered as a viable way to implement IoT services [7, 8], thus nowadays the most important "de facto" standards in the IoT arena are the following:

1. extremely short-range systems, e.g., Near Field Communications (NFC) enabled devices;

2. short-range passive and active Radio-Frequency IDentification (RFID) systems;

3. systems based on the family of IEEE 802.15.4 standards like ZigBee™, 6LoWPAN, Thread-based systems;

4. Bluetooth-based systems, including Bluetooth Low Energy (BLE);

5. proprietary systems, including Z-Wave™, CSRMesh™, i.e., the Bluetooth mesh by Cambridge Silicon Radio (a company now owned by Qualcomm), EnOcean™;

6. systems mainly based on IEEE 802.11/Wi-Fi™, e.g., those defined by the "AllSeen Alliance"[1] specifications, which explicitly include the gateways, or by the "Open Connectivity Foundation."[2] The AllSeen Alliance is dedicated to the widespread adoption of products, systems, and services that support the IoT with AllJoyn™, a universal development framework [9]. The Open Connectivity Foundation has a similar aim, but different partners [10];

7. wireless solutions based on Ultra-Wide Band (UWB) radio, e.g., the IEEE 802.15.6 standard for Wireless Body Area Network (WBAN) [11, 12].

The vast majority of the connected *things* at the moment is using IEEE 802.15.4-based systems, in particular ZigBee. The most prominent features of these networks are that they operate mainly at 2.4 GHz and optionally in the 868/915 MHz unlicensed Industrial, Scientific, and Medical (ISM) frequency bands, and that the network level connecting these *nodes*[3] uses a mesh topology. The distances between nodes in these kinds of systems range from few meters up to roughly 100 meters, depending on the surrounding environment (presence of walls, obstacles, and so on). Therefore, the IoT connectivity has been characterized so far by

- *Mesh networking.* Multi-hop communication is necessary to extend the network coverage beyond the limited reach of the low-power transmission technology used. Furthermore, the mesh architecture can provide resilience to the failure of some nodes. On the other hand, the maintenance of the mesh network requires non-negligible control traffic, and multi-hop routing generally yields long communication delays, and unequal and unpredictable energy consumption among the devices;

- *Short coverage range – high data rate.* The link level technologies used in these systems tend to privilege the data rate rather than the sensitivity, i.e., in order to recover from the network delays due to the mesh networking, these networks have a relatively high raw link bit rate (e.g., 250 kbit/s), but they are not robust enough to penetrate building walls and other obstacles (even in the 868/915 MHz band). In other words, in the trade-off between rate and sensitivity, rate is usually preferred.

Although these standards are characterized by a very low power consumption, which is a fundamental requirement for many IoT devices, their limited coverage is a major obstacle, in particular when the application scenario involves services that require urban-wide coverage, as in typical Smart City applications [1]. The

[1] `https://www.allseenalliance.org`
[2] `http://www.openconnectivity.org`
[3] *Node* is a term that is frequently used to indicate a connected *thing*, with emphasis on the communication part.

experimentation of some initial Smart Cities services has indeed revealed the limits of the multi-hop short-range paradigm for this type of IoT applications, stressing the need for an access technology able to allow a *place-&-play* type of connectivity, making it possible to connect any device to the IoT by simply placing it in the desired location and switching it on, with no (or minimal) configuration, and without the need for deploying additional devices, such as dedicated gateways or concentrators.

In this perspective, wireless cellular networks may play a fundamental role in the diffusion of IoT, since they are able to provide ubiquitous and transparent coverage [13, 14]. The 3rd Generation Partnership Project (3GPP), which is the standardization body for the most important cellular technologies, is attempting to revamp 2G/GSM to support IoT traffic, implementing the so-called Cellular IoT (CIoT) architecture [15]. On the other side, the latest cellular network standards, e.g., Universal Mobile Telecommunication System (UMTS) and Long-Term Evolution (LTE), were not designed to supply machine-type services to a massive number of devices. In fact, differently from traditional broadband services, IoT communication is expected to generate, in most cases, sporadic transmissions of short packets. At the same time, the potentially huge number of IoT devices asking for connectivity through a single Base Station (BS) would raise new issues related to the signaling and control traffic, which may become the bottleneck of the system. All these aspects make current, fourth-generation (4G) cellular networks unfit to support the envisioned IoT scenarios – in fact the native support of M2M communication is one of the *five disruptive technology directions* for fifth-generation (5G) cellular networks [16].

Nevertheless, from the business point of view, the IoT market is expected to grow exponentially in the very short term, whereas the the standardization process of 5G is still in progress and the first deployments of 5G networks are expected in 2020. Thus, a promising alternative solution, standing in between short-range multi-hop technologies operating in the unlicensed ISM frequency bands, and long-range cellular-based solutions using licensed broadband cellular standards, is provided by the so-called Low-Power Wide Area Networks (LPWANs). This kind of networks exploits sub-GHz, unlicensed frequency bands and is characterized by long-range radio links and star topologies. The end devices are connected to collector nodes, generally referred to as *gateways*, which provide the bridging to the IP world. The architecture of these networks is designed to give wide area coverage and ensure the connectivity also to nodes that are deployed in very harsh environments. On the other hand, a debate is raising in the research community about the effectiveness of LPWANs, in terms of Quality of Service (QoS) and reliability/dependability guarantees.

Table 1 summarizes the variety of wireless solutions that can provide connectivity to things and their main features.

In this thesis, we will consider two of the aforementioned enablers of long-range IoT, i.e., the 5G cellular standard and Long-Range™ (LoRa), which is one of the most prominent LPWAN solutions. Our research will be focusing on the design and performance evaluation of Medium Access Control (MAC) protocols, trying to answer the following question:

> "Under a massive number of packet arrivals from the end nodes, to what extent is the network able to efficiently support these terminals?"

Table 1: Classification of IoT connectivity

	Technology	Bands	Topology
Short range	IEEE 802.15.4, Bluetooth	Unlicensed	Mesh
	IEEE 802.11, IEEE 802.15.6	Unlicensed	Star
Long range	3GPP 4G, 5G	Licensed	Star
	LPWANs	Unlicensed	Star

By abstracting the Physical Layer (PHY), we will mathematically model and evaluate the performance of the radio access protocols for 5G and LoRa, considering the following Key Performance Indicators (KPIs):

- *throughput* (intended as the "number of terminals that successfully transmit uplink (UL) packets," rather than a synonym of "achievable data-rate");
- *outage probability*, that is, the probability that a terminal exceeds the maximum number of allowed transmission attempts, and
- *latency*.

Furthermore, in the second part of this thesis, we will provide innovative ideas on research topics about IoT at large, addressing, in particular, physical-layer security for the IoT and routing protocols for Wireless Sensor Networks (WSNs).

The rest of the thesis is organized as follows.

Part I deals with wireless communication standards for long-range IoT.

- In Chapter 1, we will study contention-free and contention-based radio access protocols to accomodate IoT traffic in 5G networks. The original research contributions of this chapter can be found in Section 1.3, which is taken from [168, 169], in Section 1.4, which is taken from [170], and in Sections 1.5, 1.6, 1.7, which are taken from [171, 172].
- In Chapter 2, we will consider long-range IoT technologies in unlicensed bands, i.e., LPWANs. In particular, we will address the LoRa standard, investigating the capacity and performance of large-scale LoRa networks. The research contributions contained in this chapter are original, and come from [173–175].

Part II deals, instead, with fundamental research about the IoT.

- In Chapter 3, we will address the security issues of IoT networks by providing a novel authentication protocol for IoT terminals, which is based on the estimation of the wireless channels between each end device and a group of trusted anchor nodes. Moreover, we will propose a location verification protocol for IoT terminals, exploiting again the channel estimates of some trusted anchor nodes. The research contributions contained in this chapter are original, and come from [176–178].

- Finally, in Chapter 4, we will investigate the trade-off between the cost of transmitting data (*transport cost*) and the cost of compressing them (*compression cost*) in order to optimize the allocation of flows of data on the wireless links of a WSN. The research contributions contained in this chapter are original, and come from [179].

As a general rule, we inform the reader that the mathematical notation and the acronyms are self-contained in each chapter. The symbols provided in the "List of Symbols" are common to all chapters.

Part I

Wireless Standards for the Internet of Things

Chapter 1

Internet of Things in Licensed Bands

The place-&-play concept calls for terrestrial radio technologies that are capable of providing widespread (ideally) ubiquitous coverage, with extremely low energy consumption, low complexity at the end device, possibly low latency, and minimal cost per bit. The most natural and appealing solution is to include Machine-Type Communication (MTC) in the list of services provided by the existing cellular networks. Indeed, cellular networks have a world-wide established footprint and are able to deal with the challenge of ubiquitous and transparent coverage. Furthermore, the wide-area mobile network access paradigm offers a number of other advantages over local-area distributed approaches, such as higher efficiency, robustness and security, thanks to locally coordinated control, coordinated infrastructure deployment, ease of planning, performance predictability and the possibility of deploying advanced MTC-tailored Physical Layer (PHY) and Medium Access Control (MAC) schemes that shift complexity from Machine-Type Devices (MTDs) to Base Stations (BSs).

Unfortunately, current cellular network technologies will likely be unable to cope with the expected growth of Machine-to-Machine (M2M) services. Indeed, today's standards are designed to provide access to a relatively small number of devices[1] that need to transfer a significant amount of data [180], so that the signaling traffic generated by the management and control plane is basically negligible. M2M services, instead, are generally expected to involve a huge number of devices that generate sporadic transmissions of short packets, making the fourth-generation (4G) cellular network architecture ineffective. For all these reasons, the M2M scenario is considered as a major challenge for the fifth-generation (5G) cellular systems.

In this chapter, we will address the problem of *massive access* in cellular networks. Firstly, we will assess this issue in the Long-Term Evolution (LTE) by means of theoretical evaluations and simulation campaigns. Then, we will propose a MAC protocol for 5G systems to overcome such an issue. We will derive a mathematical model of both 4G and 5G access and use it to compare the performance of the two solutions. The results show that the proposed so-

[1] In the order of the cardinality of the people inside a cell, assuming a one-to-one correspondence between devices and people.

lution provides relevant benefits in terms of signaling overhead and access latency.

The rest of the chapter is organized as follows. In Section 1.1, we will introduce the features of M2M traffic, highlighting the differences with respect to conventional traffic. In Section 1.2, we will thoroughly describe the radio access procedure in LTE, while in Section 1.3 we will assess the massive access problem in current cellular networks. Then, in Section 1.4 we will survey the approaches proposed in literature to tackle the massive access problem, and in Section 1.5 the proposed 5G radio access solution is described. In Sections 1.6 and 1.7 the various radio access solutions for Internet of Things (IoT) traffic in cellular networks are mathematically modeled and their performance is evaluated, respectively. Finally, in Section 1.8, we will draw the conclusions of this chapter.

1.1 M2M Traffic Characterization

MTDs are typically very low-complexity devices, both in the computational and in the Radio Frequency (RF) circuitry, and energy-constrained as well, possibly with capabilities of harvesting energy from the same surrounding environment they are sensing. According to the Cisco® Visual Networking Index (VNI) Forecast [17], a huge growth of the M2M market is expected in the next five years, thus the number of MTDs will increase with an exponential trend. In this section, we will provide an overview of the particular type of traffic generated by this kind of devices, and on the proposed models for it.

After having been predominantly based on voice calls for many years, Human-to-Human (H2H) traffic has recently moved to new Internet-based services, e.g., video streaming, thanks to more powerful devices (smartphones and tablets) and cellular network standards (Universal Mobile Telecommunication System (UMTS) [18] and LTE [19]) specifically designed to provide broadband access to a fairly limited amount of terminals. This "conventional" communication paradigm, however, is completely different from the M2M one. The first insights about the M2M traffic patterns can be found in [20], where real traffic traces are analyzed and a first comparison between M2M traffic and H2H traffic is made. The recorded statistics show that, even though the traffic generated by a single MTD is much smaller than that of a H2H device (e.g., a smartphone), MTDs are much more than the smartphones, and they generate an uplink (UL)-dominant traffic. Moreover, in some cases MTDs activate themselves in a synchronous fashion, e.g., in case of alarms. The traffic session analysis, then, highlights that MTDs are active for less time and M2M sessions occur less frequently than conventional devices. As for the mobility, MTDs (with the exception of tracking devices) are usually less mobile than smartphones.

It is worth observing that nowadays an important part of smartphone-generated traffic related to the *smartapps* (e.g., Facebook, Whatsapp, Line) should be considered MTC, as well: indeed, these applications generate data which are not directly dependent on the actions of the human users. According to [21], the total amount of autonomously generated traffic by this kind of applications per day is more than 4.4 times larger than the total amount of daily voice traffic. However, since the access requirements are mainly determined by the class of the service initiator, with a slight abuse of terminology, in the following

we will use H2H and M2M to refer to human-triggered and machine-triggered services, respectively, whatever the actual nature of the destination.

To sum up, M2M traffic is characterized by

- short packets (e.g., an Ethernet frame of 576 bits),
- long periods between subsequent data transmissions (typically ranging from few tens of minutes to several hours) due to the low duty cycle of the MTDs, and
- UL-dominant communication.

Moreover, the heterogeneous nature of MTDs yields

- M2M traffic with real-time delivery constraints as well as delay-tolerant traffic;
- periodic reporting traffic as well as event-driven reporting traffic;
- partly unsynchronized and partly synchronized access attempts.

Let us describe briefly the possible ways of modeling M2M traffic, which have been described in scientific literature and standards.

1.1.1 M2M Traffic Models Proposed in Literature

Two approaches have been followed by the research community: the first one is based on Markov-Modulated Poisson Process (MMPP) and the second one on Semi-Markov Model (SMM).

Markov-Modulated Poisson Processes

The first idea consists in tuning the time-dependent, packet arrival rate $\lambda_n[t]$ of a Poisson process according to the state $s_n[t]$ of an appropriate Markov chain, with state transition matrix $\boldsymbol{P}$ and stationary probabilities $\boldsymbol{\pi}$, that is, using a MMPP. If I is the number of states of the Markov chain, the global average arrival rate of the MMPP is defined as $\lambda_g = \sum_{i=1}^{I} \lambda_i \pi_i$ [22].

The authors of [23] propose a MMPP-based model for traffic generated by single MTDs. Considering a scenario with N MTDs, multiple MMPP models are coupled to correlate the transitions from "regular reporting" state to "alarm" state. In particular, the N chains describing the behavior of the various MTDs are unidirectionally influenced by a background process Θ (*master process*), which produces samples $\theta[t] \in [0, 1]$. Each MTD $n = 1, \ldots, N$ is assigned a constant parameter $\delta_n \in [0, 1]$ to measure the level of coordination (coordination increases as δ_n approaches 1), therefore, for every MTD, one can define a time dependent parameter $\theta_n[t] = \delta_n \theta[t]$. The time-variant transition matrix of the n-th MTD is finally computed as

$$\boldsymbol{P}_n[t] = \theta_n[t] \times \boldsymbol{P}_{\mathrm{C}} + (1 - \theta_n[t]) \times \boldsymbol{P}_{\mathrm{U}}\,, \tag{1.1}$$

that is, as the convex combination of the globally-known transition matrices $\boldsymbol{P}_{\mathrm{C}}$ and $\boldsymbol{P}_{\mathrm{U}}$, which address the case of perfectly coordinated devices and uncoordinated devices, respectively. In particular, if we consider a simple two-state

Markov model ($I = 2$), where states #1 and #2 represent the *regular* and *alarm* operations, respectively, the two global matrices are defined as

$$\boldsymbol{P}_C = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad \boldsymbol{P}_U = \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix}. \tag{1.2}$$

Note that in the coordinated case an alarm is triggered in one time slot and then the MTD returns to regular operation, while in the uncoordinated case an alarm is never triggered. The global arrival rate λ_g is

$$\lambda_g = \sum_{t=0}^{T} \sum_{n=1}^{N} \sum_{i=1}^{I} \lambda_i \pi_{n,i}[t], \tag{1.3}$$

where T is the desired time horizon. Note that λ_g is not easy to compute, since $\boldsymbol{\pi}_n[t]$ must be computed for every MTD n and time instant t.

This coupled-MMPP model is used to generate arrivals as follows. For each value of t, the transition matrices $\boldsymbol{P}_n[t]$ are generated for all MTDs $n = 1, \ldots, N$. Then, random transitions from $s_n[t-1]$ to $s_n[t]$ are performed for all n, determining the arrival rate $\lambda_n[t]$. Finally, the number of arrivals is generated, together with packet sizes.

Semi-Markov Models

The time-dependence of M2M traffic can also be captured using Markov renewal processes, in particular, SMM, which are particular kinds of renewal reward processes based on a *embedded* Markov chain [24].

The authors of [25] design a four-state SMM to describe the evolution of aggregate M2M traffic. The $S = 4$ states are: *off*, *Periodic Update* (PU), *Event-Driven* (ED), *Payload Exchange* (PE). While the PU and ED states mimic the regular and alarm operations introduced in the previous section, in off and PE states no packets are generated and a larger amount of traffic (e.g., for firmware updates) is exchanged, respectively. The state transition matrix $\boldsymbol{P}$, the distributions of the packet inter-departure time $f_{D,s}(d)$, packet size $f_{Y,s}(y)$, and sojourn time $f_{T,s}(t)$ can be designed to fit the desired MTC traffic pattern for every state $s = 1, \ldots, S$ of the embedded chain. The mean values of inter-departure time $\bar{D}_s$, packet size $\bar{Y}_s$, and sojourn time $\bar{T}_s$ for every state s can be easily computed averaging over the distributions of the random variables. Then, the average data rate of state s is defined as $R_s = \bar{Y}_s / \bar{D}_s$. The stationary probabilities of the embedded Markov chain $\boldsymbol{\pi}^e = [\pi_1, \ldots, \pi_S]$ are computed solving the system of equations $\boldsymbol{\pi}^e = \boldsymbol{\pi}^e \boldsymbol{P}$. Finally, the SMM state probabilities and the total average data rate are

$$\pi_s = \frac{\pi_s^e \bar{T}_s}{\sum_{i=1}^{S} \pi_i^e \bar{T}_i}, \tag{1.4}$$

$$R_{\text{tot}} = \sum_{n=1}^{N} \sum_{s=1}^{S} \pi_{s,n} R_{s,n}, \tag{1.5}$$

respectively, under the assumption that the N MTDs are all equal. The process of packet generation is obtained emulating the behavior of the SMM.

Let us observe that the model proposed in [23] is capable of capturing the behaviour of single MTDs according to their particular spatial correlation index, while the model proposed in [25] is preferable for the simulation of highly populated quasi-synchronized scenarios. Therefore, the former approach is more precise than the latter, since the assumption of homogeneous device is quite strong. However, the computational complexity of the former is much higher than the one required for the latter.

1.1.2 The Contribution of the Standardization Bodies

The 3rd Generation Partnership Project (3GPP), which is the standardization body of cellular networks, proposed several traffic models for M2M traffic in its technical reports.

In [26], a proposal for an aggregate M2M traffic model can be found. Two reference scenarios are considered: the first one deals with uncoordinated traffic (*Model #1*) and the second one with synchronous traffic (*Model #2*). In Model #1, the arrivals are uniformly distributed over a time interval $T = 60$ seconds, while, in Model #2 arrivals follow a Beta distribution of parameters $\alpha = 3$ and $\beta = 4$ over an interval of $T = 10$ s. The traffic patterns of the single MTDs are defined in [27], distinguishing again between *regular reporting* and *triggered reporting* traffic. Two further traffic classes are identified, instead, in [28], i.e., Network Command (NC) and Software Update (SU), to account for downlink (DL) traffic generated by the network to transmit commands and firmware updates.

On the other hand, the Institute of Electrical and Electronics Engineers (IEEE) addresses M2M traffic features in the IEEE 802.16p standard [29]. In [30–32], MTC applications, usage scenarios, and traffic characteristics are identified. In particular, a use case for smart grids is presented in [31]: in this example, the number of metering devices is estimated according to the population and household statistics of urban environments like New York, Washington D.C., and London, and the access rates of various smart metering applications are evaluated, assuming a uniform distribution across the interval of interest.

One may observe that the standardization bodies provide more practical contributions, more focused on real scenarios, and they prefer to adopt simplified M2M traffic models with respect to the ones in the scientific literature.

A summary of the various proposals can be found in Table 1.1.

Table 1.1: Summary and comparison of M2M traffic models

	Ref.	MMPP	SMM	Heuristic	Aggregate	Per-Device	Complexity
Literature	[23]	✓				✓	High
	[25]		✓		✓		High
Standards	[26]			✓	✓		Low
	[27, 28]			✓		✓	Low
	[32]			✓	✓		Low

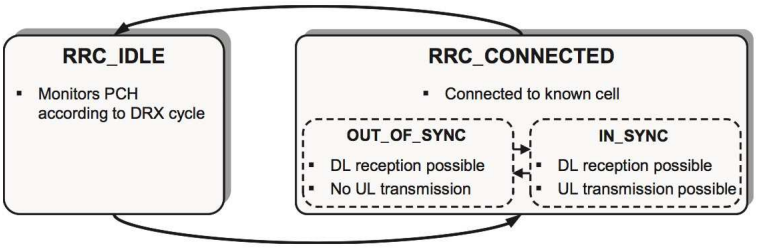


Figure 1.1: RRC state machine, taken from [19]

1.2 Radio Access Procedure in LTE

Let us analyze in detail now the various steps that a device must undergo to "enter" a LTE network, i.e., the radio access protocol of LTE. These steps provide a first piece of traffic (of "control" type) which later in this chapter we will try to minimize for MTDs.

In LTE, a User Equipment (UE) can be in two different states, as shown in Figure 1.1: `RRC_IDLE` or `RRC_CONNECTED` [19]. In `RRC_IDLE`, there is no Radio Resource Control (RRC) context in the Radio Access Network (RAN), that is, the parameters necessary for communication between the terminal and the RAN are not known to both entities, thus the terminal does not belong to a specific cell. Data transfer cannot take place because the terminal sleeps most of the time to reduce energy consumption and UL synchronization is not maintained. In this state, terminals periodically wake up to receive paging messages in the DL.

To start actual data transfer to/from the network, the terminal must switch from `RRC_IDLE` to `RRC_CONNECTED`, establishing a valid RRC configuration. The terminal is then assigned to a cell and is given the Cell Radio-Network Temporary Identifier (C-RNTI). Two substates are assumed in `RRC_CONNECTED` state, i.e., `IN_SYNC` or `OUT_OF_SYNC`, depending on whether the UL is synchronized to the network or not. We recall that, since LTE uses an orthogonal Frequency Division Multiple Access (FDMA)/Time Division Multiple Access (TDMA)-based UL, the synchronization of the various terminals is a mandatory requirement to make all UL transmissions arrive at the BS, also called eNodeB (eNB), at the same time.

Considering the typical traffic pattern generated by a MTD, the time interval between two adjacent packet transmissions is so long that the terminal loses the synchronization with the network. During this time, battery-powered nodes, in order to minimize the energy consumption, go to sleep, turning off the RF circuitry. Thus, after every transmission of a new packet, the terminal switches from `RRC_CONNECTED` to `RRC_IDLE` and needs to switch back again to `RRC_CONNECTED` when it has to send the successive UL packet. To change its state from `RRC_IDLE` to `RRC_CONNECTED`, the MTD has to perform a *radio access procedure*.

The procedure takes place in a dedicated physical channel called Physical Random Access Channel (PRACH) [19], which is multiplexed in time and frequency with the Physical Uplink Shared Channel (PUSCH) and the Physical Uplink Control Channel (PUCCH), as shown in Figure 1.2. The PRACH consists

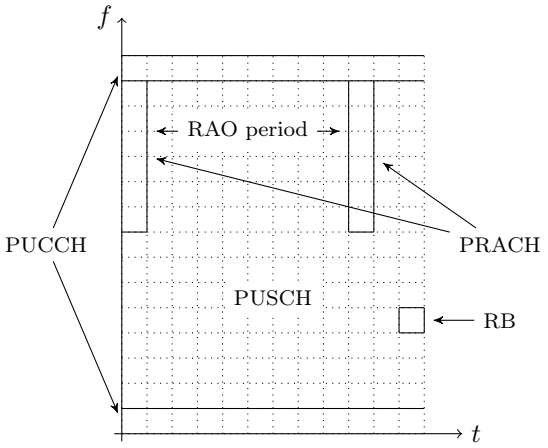


Figure 1.2: LTE uplink physical channels, assuming a bandwidth of 3 MHz ($n_{\mathrm{RB}} = 15$ RBs). The PRACH is used to transmit preambles only, the PUSCH conveys mainly data packets, and the PUCCH is used to transmit signaling information. Note that the PUCCH is instantiated at the edges of the overall available spectrum for two reasons: a) to increase the reliability of the control information by maximizing the frequency diversity and b) not to fragment the UL spectrum in case wide bandwidths are needed [19].

of 6 Resource Blocks (RBs),[2] for an overall bandwidth of 1.08 MHz that is used by each UE to transmit a *preamble*, i.e., a signature composed of a cyclic prefix and a Zadoff-Chu (ZC) sequence that is obtained by shifting a *root sequence*, which is common to all the UEs connected to a certain eNB.[3] Preambles containing different sequences are orthogonal to one another.[4] The periodicity of the Random Access Opportunities (RAOs) δ_{RAO} is variable and is defined by the PRACH Configuration Index, which is broadcast by the eNB on the System Information Broadcast 2 (SIB2) along with the following signaling information:

- `numContentionPreambles`, i.e., the number of preambles for contention-based Random Access (RA)[5];
- `preambleInitialReceivedTargetPower`, i.e., the target power (in dBm) to be reached at the eNB for transmissions on PRACH;
- `powerRampingStep`, i.e., the power ramping step used to increase the transmission power after every failed attempt;
- `preambleTransMax`, i.e., the maximum number of preamble transmission attempts.

[2] A time-frequency physical resource spanning 1 ms (i.e., a subframe or Transmission Time Interval (TTI)) times 180 kHz. It is the minimum-size physical resource of the Orthogonal Frequency Division Multiple Access (OFDMA) grid that can be allocated in LTE.

[3] Actually, 4 different preamble formats (`0` to `3`) are available, with duration from 1 to 3 subframes in order to guarantee the coverage of different cell sizes. In the following, however, we will always refer to format `0`, with preamble duration 1 ms.

[4] For eNBs with a large coverage area there may be more than one root sequence. However the sequences obtained have low cross-correlation [33].

[5] The maximum number of preambles is 64, but the actual number of available signatures for RA purposes is typically lower, usually equal to 54, because some of them are reserved for special purposes, e.g., to trigger a contention-free access procedure during handover.

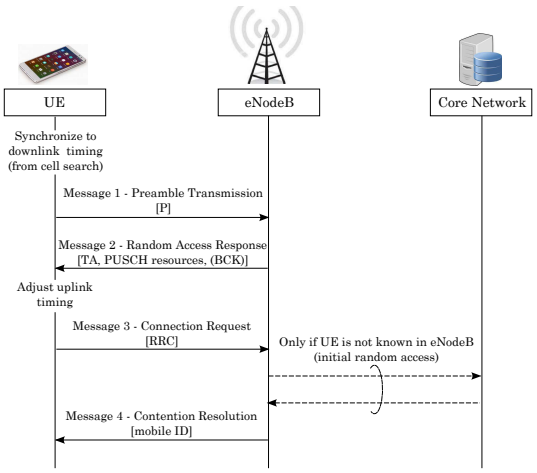


Figure 1.3: Contention-based radio access procedure in LTE

The RA procedure consists in exchanging the following 4 messages (please refer to Figure 1.3).

Preamble Transmission The UE selects a random ZC sequence and transmits it on one of the resources specified by the PRACH Configuration Index. The eNB will detect the sequence by applying a correlator and a peak detector to the received signal [34]. However, since the number of ZC sequences is finite, it may happen that more than one UE select the same sequence, thus incurring in a *collision*. If the colliding UE preambles are received with high enough Signal-to-Noise Ratio (SNR), and are sufficiently spaced apart in time, two energy peaks separated by a time that is longer than the Maximum Delay Spread (MDS) are detected and the eNB will interpret this event as due to a collision (see Figure 1.4). On the other hand, if only one of the colliding preambles is received with high SNR, or the delay of the different preambles is similar,[6] the eNB will not be able to recognize the collision.

We remark that MDS and the preamble detection algorithm are not standardized but left to the eNB vendor. However, according to [35] a missed detection probability lower than 10^{-2} for an SNR value of -14.2 dB and a 2-antenna receiver in Additive White Gaussian Noise (AWGN) channel is required.

Random-Access Response (RAR) The eNB answers to correctly decoded preambles (including those with undetected collision) by sending a RAR message on the Physical Downlink Shared Channel (PDSCH). RAR carries the detected *preamble index*, which corresponds to the sequence sent by the UE, a timing alignment to synchronize the UE to the eNB, a temporary identifier (Temporary C-RNTI (TC-RNTI)), and an UL scheduling grant that specifies the PUSCH resources assigned to the UE to transmit in the

[6] This is typical in Small Cells and Smart Cities scenarios [34]. We will recall this observation in Section 1.5.

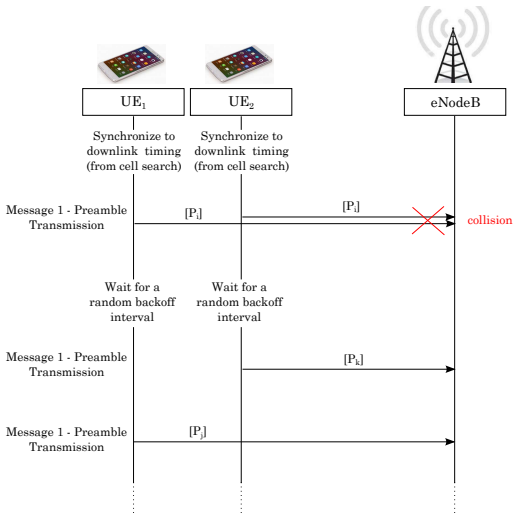


Figure 1.4: Collision event after the preamble transmission

next phase of the RA procedure. If more than one preamble are received, then the RAR multiplexes the responses for all the preambles. If a UE receives a RAR, then it proceeds with the third step; otherwise, it restarts the RA procedure anew (unless it has reached the maximum number of preamble transmission attempts) in the first RAO after a backoff time that is uniformly distributed in the interval [0,`BI`], where `BI` is the Backoff Indicator (denoted by "BCK" in Figure 1.3) carried by the RAR. If the counter of consecutive unsuccessful preamble transmissions exceeds the maximum number of attempts, a RA problem is indicated to the upper layers.

Connection Request (CR) The UE transmits a RRC message containing its core-network terminal identifier in the UL grant resources and starts a Contention Resolution timer. Since, in principle, this message is transmitted in the same manner as scheduled UL data, it exploits a Hybrid Automatic Repeat Request (HARQ) process. However, we remark that the UEs that transmitted the same preamble but whose collision remained undetected will transmit on the same PUSCH resources, colliding again (see Figure 1.5).

Contention Resolution If the eNB correctly receives the RRC message, it replies with an RRC Connection Setup that signals to the UE that the RA phase is successfully completed and with its final identity, i.e., the C-RNTI. Instead, if the Contention Resolution timer expires, the UE repeats the RA procedure from the beginning after a random backoff time. Again, when the number of unsuccessful attempts reaches some specified maximum value, the network is declared unavailable by the UE and an access problem exception is raised to the upper layers.

After successfully completing the four-step RA procedure, further RRC signaling is needed before the UE is finally *connected* to the network. For the sake of simplicity and without loss of generality, we intentionally ignore this

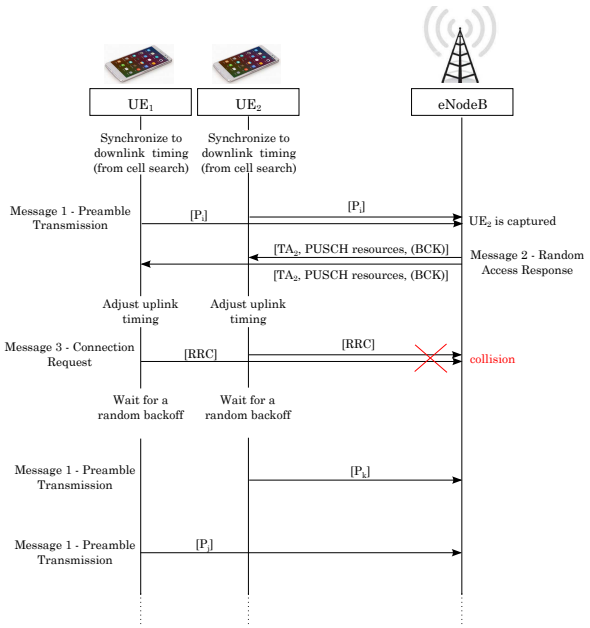


Figure 1.5: Collision event after the CR transmission

phase. At this point, each UE is associated with a contention-free Scheduling Request (SR) opportunity on the PUCCH: it is a reserved resource on which a connected terminal can transmit the request for UL resources. The last two stages remain, then.

Scheduling Request The UE sends a SR message to the eNB, which replies with the indication of some PUSCH resources for data transmission.

Data Transmission The data packet is finally transmitted on the dedicated PUSCH resources.

1.3 Problem Statement

The aforementioned LTE radio access procedure results to be inefficient in a M2M scenario for three distinct reasons:

a) the massive number of preamble transmissions would cause the overload of RA procedure both in UL and DL due to the limited number of available signatures, thus, increasing collision probability, access delay, and failure rate;

b) moreover, additional DL resources would be needed in presence of massive access requests, as the RAR message consists of 56 bits per UE;

c) finally, even assuming that a MTD succeeds in completing the access procedure, the UL payload size is so small that the overall system efficiency would be significantly degraded due to the signaling overhead.

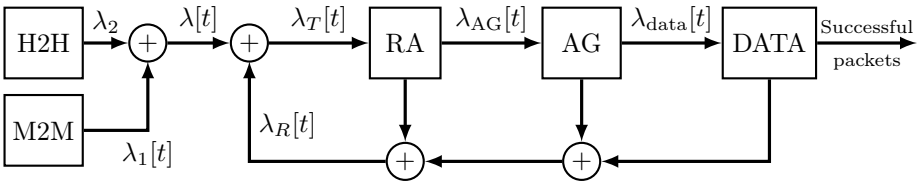


Figure 1.6: Framework for simulations of joint M2M and H2H traffic in cellular networks

Therefore, we can easily foresee that the current LTE radio access procedure does not scale in the presence of massive access attempts by MTDs, resulting in a sharp degradation of the Quality of Service (QoS) of both conventional services and IoT services.

Let us prove this claim, exploiting an easy-to-use framework as well as extensive simulation campaigns.

1.3.1 A Low-Complexity, Simplified Simulation Framework

Let us consider the three phases of the radio access procedure in LTE [36]:

1. RA phase, in which the MTDs contend for UL resources using a slotted-ALOHA-based protocol;

2. Access Granted (AG) phase, where, provided that the RA phase is successful and the requested resources are available, the BS responds to the terminal;

3. the data transmission, in which MTDs transmits the data on the wireless channel, which may introduce errors.

Let us assume that each phase can be fully characterized by a *success probability*. We denote the success probabilities (and the packet arrival rates) of the three phases with P_{RA} (λ_T), P_{AG} (λ_{AG}), and P_{data} (λ_{data}), respectively. Note that both λ_{AG} and λ_{data} are functions of λ_T, since it is $\lambda_{\mathrm{AG}} = \lambda_T P_{\mathrm{RA}}$ and $\lambda_{\mathrm{data}} = \lambda_{\mathrm{AG}} P_{\mathrm{AG}} = \lambda_T P_{\mathrm{RA}} P_{\mathrm{AG}}$. Note also that λ_T is the sum of new transmission attempts (rate λ) and retransmissions (rate λ_R), yielding $\lambda_T = \lambda + \lambda_R$. With a simple computation, recalling that λ_R can be obtained summing the rate of packets blocked at each stage, one obtains

$$\lambda_T = \frac{\lambda}{P_{\mathrm{RA}} P_{\mathrm{AG}} P_{\mathrm{data}}} \,. \tag{1.6}$$

In the following, we will evaluate the impact of joint M2M and H2H traffic on a generic cellular architecture, exploiting the simplified, but effective framework shown in Figure 1.6. Note that the arrival rate of new packets $\lambda[t]$ is time-variant, since it is the sum of the rate of MTC arrivals $\lambda_1[t]$ and the rate of H2H arrivals λ_2, thus

$$\lambda_T[t] = \frac{\lambda_1[t] + \lambda_2}{P_{\mathrm{RA}} P_{\mathrm{AG}} P_{\mathrm{data}}} \,. \tag{1.7}$$

Simulation Setup

We consider a network deployment with conventional UEs as well as MTDs. Regarding the M2M traffic generation, we employ an enhanced version of the model proposed in [23] (see Section 1.1.1). Two possible values of δ_n are defined, $\delta^{(\mathrm{H})} = 0.8$ and $\delta^{(\mathrm{L})} = 0.2$ for the case of high correlation and low correlation, respectively, and three deployment scenarios are addressed:

Scenario 1 all MTDs have high correlation, i.e., $\delta_n = \delta^{(\mathrm{H})}\ \forall n$;

Scenario 2 $\delta_n = \delta^{(\mathrm{H})}$ for a half of the MTDs and $\delta_n = \delta^{(\mathrm{L})}$ for the other half;

Scenario 3 $\delta_n = \delta^{(\mathrm{L})}\ \forall n$.

As for the master process Θ, we are improving the approach of [23] as follows. According to the specifications of the 3GPP's Model #2 (see Section 1.1.2), we define $\theta[t]$ as

$$\theta[t] = \begin{cases} \displaystyle\int_{t \bmod T}^{(t+\tau) \bmod T} f_T(x;\alpha,\beta)\ \mathrm{d}x & \text{if } t \bmod T \neq T-\tau\,, \\ \displaystyle\int_{T-\tau}^{T} f_T(x;\alpha,\beta)\ \mathrm{d}x & \text{otherwise}\,, \end{cases} \tag{1.8}$$

where $f_T(x;\alpha,\beta)$ is the Probability Distribution Function (PDF) of the Beta distribution with $\alpha = 3$ and $\beta = 4$. Note that $\theta[t]$ is periodic of period T. A three-state MMPP is then designed: in addition to *regular* (r) and *alarm* (a) states, already defined in Equation (1.2), the *off* (o) state is defined. Thus, the two global matrices $\boldsymbol{P}_C$ and $\boldsymbol{P}_U$ become

$$\boldsymbol{P}_C = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}, \quad \boldsymbol{P}_U = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}. \tag{1.9}$$

The packet arrival rates for the three states are $\lambda_r = 3.3 \cdot 10^{-3}$ pkt/s/device (i.e., one packet generated every 5 minutes on average, as suggested in [27]), $\lambda_a = 1/\tau$ pkt/s/device, and $\lambda_o = 0$, respectively. The arrival rate of H2H traffic λ_2 is, instead, constant, and computed as

$$\lambda_2 = D \times A \times \zeta\,, \tag{1.10}$$

where D is the population density, A is the coverage area of a cell, and ζ is the average number of calls per hour per person. Setting $D = 10756/\mathrm{km}^2$ (i.e., the population density of New York, USA), $A = \pi \times 0.2^2 = 0.126\,\mathrm{km}^2$ (i.e., a circular coverage area of radius 200 m), and $\zeta = 5/3600$ (i.e., 5 calls per hour per person), then we obtain $\lambda_2 = 1.88\,\mathrm{s}^{-1}$.

We consider a slotted time axis with slot duration τ; the time horizon for the evaluation is Ω. The number of RAOs per slot is

$$L = d \times \ell \times \tau\,, \tag{1.11}$$

where d is the number of preambles and ℓ is the number of RAOs per second per preamble. Thus, according to [26], we can compute P_{RA} as

$$P_{\mathrm{RA}} = e^{-\lambda_T / L}\,. \tag{1.12}$$

Table 1.2: Simulation parameters

Parameter	Value
τ	0.1 s
Ω	60 s
T	10 s
d	54
ℓ	$\{50, 200, 300, 500, 1000\}$
B	$\{1.4, 3, 5, 10, 15, 20\}$ MHz
n_{RB}	$\{6, 15, 25, 50, 75, 100\}$
J	4
λ_r	$3.3 \cdot 10^{-3}$ s^{-1}
λ_a	$1/\tau$ s^{-1}
λ_o	0
λ_2	1.88 s^{-1}

Assuming that n_{RB} RBs are available for a UL bandwidth B, and that every data packet takes $J = 4$ RBs, we define the overall amount of resources for data packets per slot M as

$$M = \frac{n_{\mathrm{RB}}}{J} \times \frac{\tau}{T_{\mathrm{TTI}}}\,, \tag{1.13}$$

where $T_{\mathrm{TTI}} = 1$ ms is the subframe duration.[7] Therefore, as done in Equation (1.12), we define P_{AG} as

$$P_{\mathrm{AG}} = e^{-\lambda_{\mathrm{AG}}/M}\,. \tag{1.14}$$

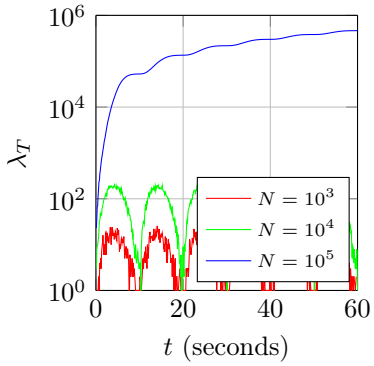
Finally, let us model the wireless channel as an erasure channel, with a constant success probability $P_{\mathrm{data}} = 0.99$.

The complete set of simulation parameters is listed in Table 1.2.

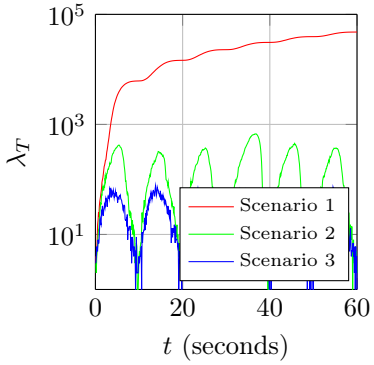
Performance Evaluation

The simulation results are presented in Figure 1.7. Figure 1.7a shows the impact of joint M2M and H2H traffic at the "input" of the cellular network: the total arrival rate λ_T as a function of time is plotted for different amounts of MTDs N, assuming that we are in the Scenario 1, $\ell = 1000$, and $n_{\mathrm{RB}} = 20$. It can be seen that the network is able to support up to 10^4 devices, while if $N = 10^5$ the network becomes unstable. We remark that the pattern of the packet arrivals at the RA phase when the system is stable is due to the periodicity of the master process Θ. Figure 1.7b, instead, shows the impact of spatial correlation of MTDs on the network performance. It can be seen that when the MTDs are highly correlated (Scenario 1) the system is unstable; as the correlation decreases, the network reaches the stability. In Figure 1.7c, we consider $N = 10^4$ MTDs and $n_{\mathrm{RB}} = 20$ RBs, while ℓ varies: it can be seen that if the number of RAOs ℓ is too low, the RA phase becomes the bottleneck of the system. Finally, in Figure 1.7d, for $N = 10^4$ and $\ell = 300$, it is shown that the AG phase becomes the bottleneck when the amount of UL resources is not sufficient to fulfill the incoming packets.

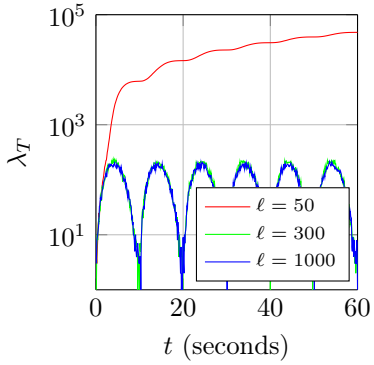
[7] Note that, for the sake of simplicity, we are assuming that the whole bandwidth B is exploited for data transmissions.



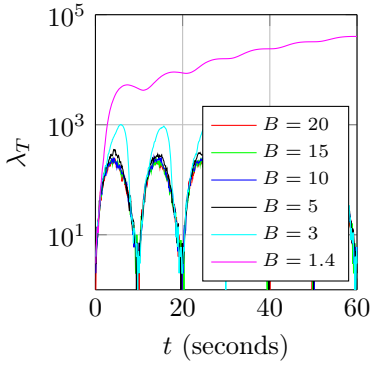
(a) Arrival rate at the RA phase block for different values of N



(b) Impact of spatial correlation on the arrival rate at the RA phase



(c) Arrival rate at the RA phase block for different values of ℓ



(d) Arrival rate at the RA phase block for different values of M

Figure 1.7: Performance evaluation results

1.3.2 Simulation Campaign in ns–3

The previously described framework is a nice tool for a first assessment of the massive access problem in cellular networks, because we can run fast simulations for each value of the various system parameters. However, we want now to confirm our insights using extensive simulations of a realistic Smart City deployment, with the aim of evaluating the delay that a device may undergo while accessing an LTE network in the case of a massive number of access requests (e.g., *event-triggered* accesses). To do so, we resort to one of the most accurate open-source, system-level network simulators, i.e., *network simulator 3* (ns–3, [37]), written in C++, which is particularly suitable to simulate a urban propagation environment. Other open-source simulation platforms are available, e.g., the LTE Vienna Simulator [38], which is based on Matlab, and Omnet++ [39] and LTE-sim [40], both written in C++. However, they cannot be directly used for our purposes. Indeed, the Vienna Simulator is a link-level simulator for the UL, and therefore lacks some of the necessary features to adequately model a network of MTDs, whereas Omnet++ and LTE-sim focus on the higher networking layers through an idealized abstraction of the lower

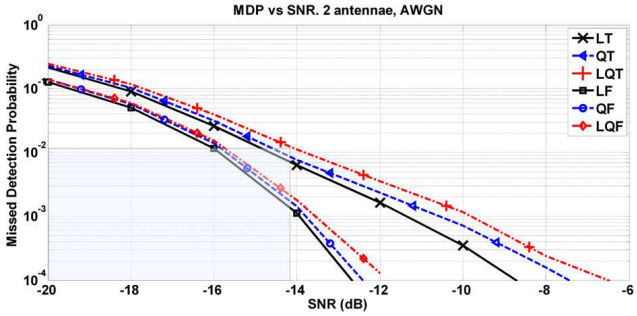


Figure 1.8: Missed detection probability vs SNR performance of various detection algorithms at the eNB receiver, taken from [41]

layers, thus they do not capture the level of detail we need to model the RA procedure performance.

Nevertheless, we found out that the current implementation of the LTE's RA procedure in ns–3 is idealized; therefore, we developed a patch to make the routine suitable to study the impact of M2M traffic in LTE networks in urban scenarios.

LTE Random Access Procedure in ns–3

We refer to version 3.23 of the ns–3 simulator, which uses the LTE-EPC Network simulAtor (LENA) [42] module to simulate the LTE protocol stack and the Evolved Packet Core (EPC) network. In the current implementation of LENA, the preamble is an *ideal* message, i.e., not subject to radio propagation; moreover, CR and Contention Resolution messages are not modeled, thus all collisions are detected and solved at the first step of the RA procedure. Furthermore, we found out that it is not possible for a UE to switch from `RRC_CONNECTED` to `RRC_IDLE` at runtime: LENA allows only to switch from `RRC_IDLE` to `RRC_CONNECTED` once, at the beginning of the simulation.

Therefore, we implemented a more realistic RA procedure, along with the possibility to disconnect UEs from the eNB: the enhanced module is called LENA+.[8] However, to maintain the backward compatibility with the current release, an option has been introduced to use the idealized LENA RA procedure if desired. In the following, we describe in detail the features of LENA+ that were not present in LENA.

PRACH Characterization The PRACH is implemented as a real physical channel, relying on the already developed and tested channel model of ns–3: preambles are now subject to noise and radio propagation, since they are sent on specific time and frequency physical resources, thus the eNB can fail to detect them. Whenever a UE starts the RA procedure, it checks whether it has received SIB2, which carries the PRACH configuration. Then, it chooses a random index drawn uniformly in [0, `numContentionPreambles` - 1] and transmits it in the next RAO. The preamble transmission power (in dBm) is computed according

[8] The source code is available at `https://github.com/signetlabdei/lena-plus`.

Table 1.3: Simulation parameters, taken from [28]

Parameter	Value
Downlink carrier frequency	945 MHz
Uplink carrier frequency	900 MHz
RB bandwidth	180 kHz
Available bandwidth	50 RB
Hexagonal sectors	1
eNBs for each sector	3 (co-located)
eNBs beamwidth (main lobe)	65°
TX power used by eNBs	43 dBm
Max TX power used by MTDs	23 dBm
eNB noise figure	3 dB
MTD noise figure	5 dB
Shadowing	log-normal with $\sigma = 8$
Number of buildings	96
Apartments for each floor	6
Floors for each building	3
MTD speed	0 km/h
Number of MTDs N	{50, 100, 150, 200, 300, 400, 500, 600}
Simulation time $\forall N$	{60, 60, 120, 120, 300, 300, 400, 400} s

Table 1.4: Simulation parameters of LTE PRACH

Parameters	Value
PRACH Configuration Index	1
Preamble format	`0`
Backoff Indicator `BI`	0 ms
`preambleInitialReceivedTargetPower`	−110 dBm
`powerRampingStep`	2 dB
`numContentionPreambles`	54
`preambleTransMax`	∞
Contention resolution timer	32 ms

to [43] as

$$P_{\text{prach}} = \min\{P_{\text{UE,max}}, \texttt{PREAMBLE\_RECEIVED\_TARGET\_POWER} + P_{\text{lc}}\}\,, \quad (1.15)$$

where $P_{\text{UE,max}}$ is the maximum transmit power for a UE, P_{lc} is the estimated path loss. `PREAMBLE_RECEIVED_TARGET_POWER` is given by the MAC layer as [44]

$$\begin{aligned} &\texttt{PREAMBLE\_RECEIVED\_TARGET\_POWER} = \\ &\quad \texttt{preambleInitialReceivedTargetPower} \\ &\quad + \Delta_{\text{preamble}} + (\texttt{PREAMBLE\_TX\_COUNTER} - 1) \\ &\quad \times \texttt{powerRampingStep}\,, \end{aligned} \quad (1.16)$$

where `PREAMBLE_TX_COUNTER` is the number of consecutive preamble transmissions and $\Delta_{\text{preamble}} = 0$ for format `0`. The other parameters are given by the eNB with SIB2, as explained in Section 1.2.

At the eNB side, the SNR is computed for each preamble and a decision on correct or missed detection is made. Among the various eNB detection

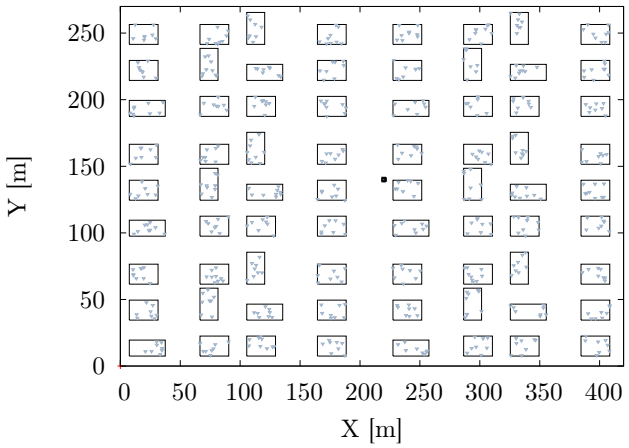


Figure 1.9: A Smart City network deployment example. The rectangles are the buildings, the small triangles are the MTDs, and the black square is the position of the three co-located eNBs.

algorithms compared in [41] (see Figure 1.8), we used a time-domain detector with decimation (denoted with "LT" in the legend of Figure 1.8), which is the simplest algorithm that satisfies the 3GPP requirements described in Section 1.2.

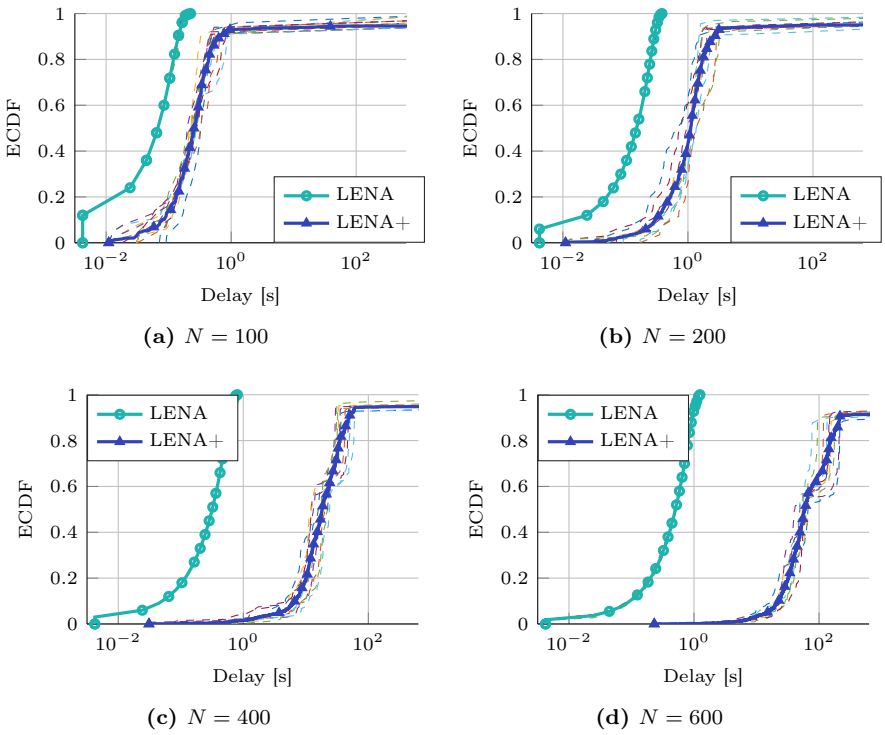
The ns–3 LTE module has also been enhanced enabling the reception of signals which are transmitted on the same time-frequency resources, preventing the exception raised by the default implementation: in this way, we can simulate the real transmission of preambles and, possibly, collision events. If a preamble is correctly received but it is not unique, the collision is detected or not according to a heuristic: since we do not have the Power Delay Profile (PDP) of different users in a system-level simulator as ns–3, as a rule of thumb a collision is detected if

$$\frac{d_{\max} - d_{\min}}{c} > T_{\text{chip}}\,, \tag{1.17}$$

where $d_{\max}$ and $d_{\min}$ are the distances from the eNB of the farthest and closest colliding UE, respectively, c is the speed of light, $T_{\text{chip}} = 1/(2B)$, and $B = 1.08$ MHz is the bandwidth of the PRACH.

The RAR message transmission was already implemented as a message on the PDSCH in the default implementation. For each not-collided preamble or undetected collision, an UL grant on the PUSCH is allocated by the scheduler and added to the RAR response, together with the Backoff Indicator.

CR transmission on granted resources was already implemented in LENA, as well. However, since in the default implementation all the collisions are resolved at the first step, collisions of CRs were not handled, thus the simulator raised an exception. This exception has been handled as follows: a) no capture effect has been considered; b) if two or more CRs collide, they are considered as received with errors, triggering an HARQ retransmission until the maximum number of attempts is reached; after that, the RA procedure starts again. The Contention Resolution timer, that was also not present in LENA, has finally been added.



(a) $N = 100$ **(b)** $N = 200$

(c) $N = 400$ **(d)** $N = 600$

Figure 1.10: ECDFs of access delay for various values of N. The x-axis is expressed in logarithmic scale.

Performance Evaluation

We considered a network scenario based on the specifications of [28]; the main simulation parameters are in Table 1.3, while the PRACH-related parameters can be found in Table 1.4. We refer to an urban environment with a high density of tall buildings; the actual deployment of buildings and MTDs is depicted in Figure 1.9. As for the radio propagation model, we employed the ns–3 Hybrid Buildings Propagation Loss Model, which exploits different propagation models to account for several factors, such as the positions of the UE and the eNB (both indoor, both outdoor, one indoor and the other outdoor), the external wall penetration loss of different types of buildings (i.e., concrete with windows, concrete without windows, stone blocks, wood), and the internal wall penetration loss. We remark that all the MTDs have been placed inside the buildings and their positions are not changed during the simulation, as mandated by [28].

We denote with N the number of MTDs that are trying to simultaneously access the LTE network. For every value of N in Table 1.3, 10 Monte Carlo simulations have been run and the Empirical CDF (ECDF) of the access delays have been produced. Figure 1.10 shows the ECDFs of the delay (in logarithmic scale) for the various values of N, obtained using both the default LENA module and the LENA+ module. We remark that, regarding the LENA+ performance curves, the average ECDF is represented by the solid line and we plotted the ECDFs of the individual Monte Carlo simulations with dashed lines to show

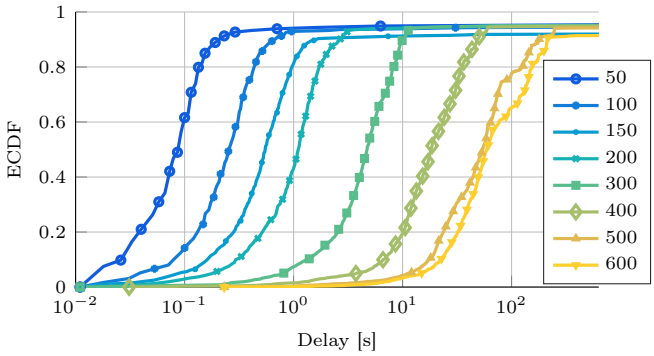


Figure 1.11: Comparison of access delay ECDFs for various values of N. The x-axis is expressed in logarithmic scale.

Table 1.5: Statistics of the access delay experienced by the MTDs that succeeded in completing the access procedure

N	Mean μ [s]	Std. Dev. σ [s]	μ/σ
50	0.235	1.855	0.127
100	0.498	2.608	0.191
150	0.780	2.605	0.300
200	1.481	4.453	0.333
300	5.268	5.359	0.983
400	21.400	15.126	1.415
500	64.234	52.852	1.215
600	77.423	59.256	1.307

the dispersion around the average value. It can be seen that the idealized RA procedure implemented in LENA gives quite unrealistic results, where all the MTDs would succeed in accessing the network in less than 1 s regardless the value of N. The simulations that have been carried out using LENA+, instead, show that, as N grows, the access delay increases, up to hundreds of seconds for most MTDs, which is not acceptable for many delay-constrained Smart City applications, such as alarms. Moreover, using our module, we are able to observe that some UEs (approximately 5% of the total in each simulation) do not succeed in completing the RA procedure during the simulation, despite the unlimited number of transmission attempts allowed (i.e., `preambleTransMax` $= \infty$). This is due to their unfavorable position, e.g., inside buildings which are far away from the eNB. An overall comparison among the average ECDFs for all the values of N is provided in Figure 1.11, which clearly shows that the access delay increases as N grows. As a further insight, we invite the reader to refer to Table 1.5, which contains the average value and standard deviation of the delays of successful MTDs: the statistics confirm the trend.

Finally, Figure 1.12 represents the number of successful RA attempts versus time, for different values of N. It should be noted that, as N increases, the maximum number of successful RA attempts decreases, and is achieved later in time, as a consequence of the higher number of collision events. For $N > 300$, we cannot observe any meaningful peak, denoting that the PRACH is congested

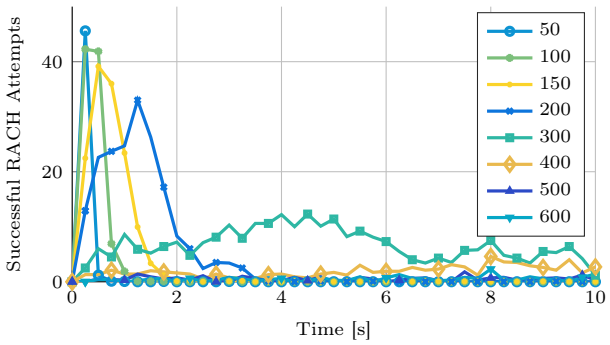


Figure 1.12: Successful RA attempts vs time

and the success probability is very low.

1.4 Related Work on Massive Access

The aforementioned considerations have driven academia, research institutes, industries, and standardization bodies to design amendments/improvements of current standards as well as brand-new solutions to face the challenges posed by M2M services.

1.4.1 Proposed Amendments to the Cellular Standards

Recently, telecommunication providers started to replace second-generation (2G) cellular systems, i.e., Global System for Mobile Communications (GSM), GPRS, and EDGE, and even third-generation (3G) ones (UMTS and HDSPA), with LTE, to provide connectivity to conventional H2H services. In this context, GSM becomes an attractive candidate to support MTC [15], which may exploit the pervasive, worldwide presence of GSM coverage and the empty space left by the migration of the conventional services to 4G networks. However, considering the scarcity of available bands below 6 GHz, and the always growing demand for new wireless services, the refarming of GSM frequency bands is being debated by governments and providers, thus the remaining operational time for GSM is uncertain. In addition, several studies show that the GSM RAN faces serious capacity issues in the presence of the synchronized access of a massive number of MTDs [45, 46]. As a consequence, 3GPP has started enhancing the GSM-related standards to facilitate the support of MTC, tightening the granularity of the transmission resources [46], improving the load control mechanisms and reducing the signaling overhead [47] as well as increasing the network coverage [28]. This technological approach is usually referred to as Cellular IoT (CIoT) or Extended Coverage GSM (EC-GSM) [48].

While these efforts can indeed make GSM a viable solution for massive MTC support in the short/mid term, there are a number of practical considerations and technical limits that will likely prevent GSM from becoming the ultimate access technology for MTDs in the long run. We may think of pushing M2M applications towards UMTS, however, compared to GSM, 3G cellular standards

have a number of disadvantages: they are power-hungry, more complex, and provide a worse (especially indoor) coverage due to the higher higher frequency band employed.

A rather natural option is to resort to the LTE, which is more appealing than UMTS despite the issues shown in Section 1.3. Much work has been done in 3GPP to enhance the support of MTC in LTE. Some enhancements of the radio access protocol were proposed in [49], in order to decrease the RRC signaling in presence of M2M traffic. A smart variant of the default RA procedure is also envisioned in [49]: to shorten the access delay of the default procedure described in Section 1.2, one could think of including the SR into the CR message. If the CR collides or in case of shortage of physical resources on the PUSCH, the MTD must start the RA procedure from scratch.[9]

On the other hand, two entirely new LTE standard amendments were recently introduced.

LTE-M brings new power-saving functionalities, suitable for serving a variety of IoT applications and extend battery life of terminals to 10 years or more, a substantial reduction of device cost and extended coverage [48].

Narrowband IoT (NB-IoT) has a similar goal to LTE-M, but it reduces even more the complexity of LTE-M, offering low-bandwidth data connections at a lower cost [50].

We have to remark, however, that both LTE-M and NB-IoT technologies are still standalone, ad-hoc solutions, which may not be capable of adapting to future (and, in some case, yet unpredictable) network scenarios.

1.4.2 Basic Strategies to Alleviate the PRACH Overload

The PRACH overload problem has been attracting the attention of the scientific community, and many possible methods to improve the LTE's RA procedure have been proposed [51]. Most of these methods provide some form of separation between access requests originated by H2H and M2M services, with the aim of shielding the former from the PRACH overload issues that can be generated by the latter. The various approaches, most of which appeared in 3GPP technical reports (e.g., [26]), differ in the way this separation is enforced. We hence distinguish between "strict" schemes, in which the pool of access resources is split between H2H and M2M, thus achieving perfect isolation between the two types of access requests; and "soft" schemes, where H2H and M2M share the same resources, but with different access probabilities. These two approaches can also be combined, giving rise to "hybrid" schemes.

Strict-Separation Schemes

As mentioned, strict-separation schemes achieve perfect isolation between H2H and M2M access requests by allocating different physical resources to UEs and MTDs. In this category, we can list the following schemes.

[9] In Section 1.6, this protocol variant, denoted as "enhanced 4G," will be mathematically modeled, then in Section 1.7 its performance will be compared with the performance of the default RA procedure in LTE.

Resource Separation The simplest and most immediate way to shield H2H from the risk of access request collisions due to massive MTC requests is to assign orthogonal PRACH resources to H2H and M2M devices. The separation of resources can be done by either splitting the preambles into H2H and MTC groups, or by allocating different RAOs in time and/or frequency to the two categories of terminals [26]. This solution, however, can yield suboptimal performance when the number of resources assigned to each class of devices does not reflect the actual demand.

To be effective, this scheme needs to be coupled with mechanisms to dynamically shift resources among the two classes, according to the respective access request rates. In some scenarios, the network can predict sharp increments of the access load due to MTDs, e.g., using the Self-Optimizing Overload Control (SOOC) scheme proposed in [52] and described later under the "hybrid" category.

Slotted Access This scheme was proposed by 3GPP in [26]. It consists in defining "access cycles" (similar to paging cycles), which contain RAOs dedicated to MTD access requests. Each MTD can only access its dedicated subframes for RA, in a collision-free manner. The reserved RA subframes for each MTD in a cycle are determined from the unique identifier of the devices (namely, the International Mobile Subscriber Identity (IMSI)) and the RA cycle parameter broadcast by the eNB. While this scheme protects H2H devices from MTC, the allocation of dedicated RA subframes to each MTD may yield very long RA cycles and, hence, long access latency, which may not be compatible with the service requirements of delay-constrained M2M applications (e.g., alarms).

Pull-Based Scheme This is a centralized mechanism that allows MTDs to access the PRACH only upon being paged by the eNB [26]. Paging is triggered by the MTC server that is assumed to know in advance when MTDs need to establish a radio link connection, to either send or receive data. The eNB can control the paging taking into account the network load condition, thus preventing PRACH overload. This is already supported by the current specification. The paging message may also include a backoff time for the MTDs, which indicates the time of access from the reception of the paging message. This approach is suitable to manage channel access of MTDs with regular traffic patterns. However, its centralized nature limits the number of MTDs and M2M services that can be managed by a single M2M server. Furthermore, the scheme cannot deal with an unexpected surge of MTD access requests.

Soft-Separation Schemes

In soft-separation schemes there is no neat separation of access resources between M2M and H2H, thus all devices can use the same resources, but with different probabilities: indeed, the separation between MTDs and conventional UEs is achieved in a statistical sense. The main schemes based on this approach are described below.

Backoff Tuning A way to smoothly decrease the rate of channel access requests by MTDs in case of congestion is to assign longer backoff intervals to

MTDs that fail a RA attempt [26]. Although this method can alleviate the contention between H2H and M2M devices in case of peaks of MTD requests, it is not very effective when dealing with stationary MTDs massive access, due to the instability issue that characterizes ALOHA-like access mechanisms.

Access Class Barring The backoff tuning scheme is generalized by the Access Class Barring (ACB) method, which is actually part of LTE specifications. ACB makes it possible to define multiple access classes with different access probabilities [26]. Each class is assigned an access probability factor and a barring timer. The devices belonging to a certain access class are allowed to transmit the preamble in given RAO only if they draw a random number that is lower than the access probability factor. Otherwise, the access is barred and the devices have to wait for a random backoff time, which is determined according to the barring timer of that class, before attempting a new access. The ACB parameters are broadcast by the eNB as part of the system information.

Furthermore, 3GPP proposed the Extended Access Barring (EAB) scheme, which is a method for the network to selectively control access attempts from UEs that can tolerate longer access delays or higher failure probability [26]. These devices will hence be barred in case of overload of the access and/or the core network, without the need to introduce any new access classes. These mechanisms can be used to alleviate the MTD massive access issue by defining a special class for MTDs, with lower access probability factor and/or longer barring timer, or labeling MTDs as EAB devices. However, MTDs with delay-constrained access requirements can be associated to classes with higher access probability and lower barring timer.

ACB mechanisms are quite effective in preventing PRACH overload, but at the cost of longer access delay for MTDs. Moreover, ACB does not solve the access contention problem when many delay-constrained MTDs need to access the channel in a short time interval, as the result of certain events (e.g., alarms triggered by unexpected events, such as failures of the power grid, earthquakes, flooding, and so on). Nonetheless, ACB mechanisms can be combined with other techniques to counteract the PRACH overload due to massive MTD access.

Hybrid Schemes and Other Solutions

Let us discuss now those solutions that cannot be classified as either strict- or soft-separation schemes, since they include aspects from both families or are based on totally different approaches.

Self-Optimizing Overload Control SOOC is a composite scheme presented in [52] to counteract PRACH overload by combining many of the schemes described above, including PRACH resource separation, ACB, and slotted-access schemes. The fundamental feature of the SOOC scheme is the execution of a control loop to collect information for overload monitoring at each RA cycle. Then, based on such data, the eNB adapts the number of subframes for RA in the random access cycles.

More specifically, when a device is not able to get an access grant at the first attempt, it enters the "overloaded control" mode. In this status, the classical p-persistent mechanism is applied in order to regulate access retries for collided terminals. Besides, in order to distinguish between time-tolerant MTDs and

time-sensitive MTDs, two access classes are added to the LTE's ACB scheme for M2M devices (namely, low access priority and high access priority) and different p parameters are set according to the access class of the terminal.

In order to monitor the congestion level of the system, when a terminal receives the RAR message (see Section 1.2), it includes a PRACH overload indicator, which contains the number of RA retries attempted by the device, within the CR. Based on this information, the eNB reacts by dynamically increasing or decreasing the number of PRACH RAOs in the successive cycle in order to maintain a target maximum collision probability for the system. Moreover, in the borderline case when the number of RAOs can not be further increased due to insufficient UL radio resources, the eNB can deny access to low priority MTDs until the overload condition improves.

Unfortunately, although the goal of handling high traffic loads is clear and the proposed scheme surely goes in this direction, [52] only describes SOOC theoretically and no performance results have been presented.

A Modified Random Access Procedure for Fixed-Location MTDs When MTDs are static (e.g., smart meters), the fixed UL Timing Alignment (TA) between the MTDs and the BS can be exploited to reduce the collision probability in the transmission of the CR message in the RA procedure as proposed in [53]. The idea is to add a further step to the conventional LTE RA procedure described in Section 1.2. Upon receiving the RAR, each device should compare the TA value contained in the response with its own TA value: if the two values match, the handshake continues, otherwise the MTD has to retransmit after a random backoff time. Indeed, if the TA received from the eNB does not match the MTD's expected TA, it means that probably the RAR is actually intended for a different MTD that transmitted the same preamble. In this way, the MTD can avoid transmission of the CR, thereby reducing the probability of collision and the access delay.

Bulk MTC Signaling Another possible solution to prevent congestion events may be to enable bulk MTC signaling handling, as stated in [54], where the authors remark the lack of mechanisms to simultaneously handle the overhead generated from a group of MTDs. Indeed, assuming that signaling messages from MTDs are moderately delay tolerant, it may be convenient to minimize the overhead at the eNB by aggregating signaling data coming from MTDs before forwarding them to the core network.

For example, consider the case in which a group of MTDs are triggered to send a Tracking Area Update (TAU) message: the BS could wait a default timeout interval, or until it gathers a sufficient number of signaling messages to forward a single aggregate message towards the Mobility Management Entity (MME). Indeed, since the MTDs are associated to the same MME, the TAU messages will differ only on the MME Temporary Mobile Subscriber Identity (M-TMSI). Considering an average of 30 TAU messages per second, and a message aggregation period of 10 s, 300 TAU messages can be aggregated in a single 1211 bytes message, while individual messages would instead require 4500 bytes. This approach can alleviate the traffic produced by massive channel access across towards the MME, but it does not address the issue of batch MTD transmissions on the access network side.

A Q-Learning Solution The standard RA procedure is basically derived from the classical slotted ALOHA protocol, of which it inherits simplicity and limitations. In particular, the system may drift to an unstable region in the presence of massive M2M traffic. In this context, [55] suggests a solution based on Q-learning to enhance the throughput of RA phase and shield H2H traffic from the performance loss that can be caused by massive M2M access requests. According to the authors, users should be divided in two groups: a learning group containing all MTDs, and a non-learning group composed of H2H devices. MTC uses a virtual frame of RAOs called *M2M-frame*, whose length (in subframes) should be equal to the number of MTDs in the network. Every node keeps a Q-value for each slot of the M2M-frame, which records the transmission history on that slot in consecutive frames. The value is updated after every transmission attempt as

$$Q \leftarrow (1-\alpha) \times Q + \alpha \times r\,, \tag{1.18}$$

where α is the learning rate, and r is the reward, which equals 1 if the transmission is successful or $-k$ otherwise, and k is known as penalty factor and is introduced to mitigate the effect of collisions with H2H devices.

Each MTD will transmit in the slot with the highest Q-value. The performance evaluation shows that, in case of high load from H2H traffic, Q-learning access stabilizes the throughput of RA phase at 35% (approximately the maximum efficiency of slotted ALOHA), as the M2M traffic increases. When the H2H traffic load is low, instead, the proposed solution provides a significant improvement, raising the total RA-phase normalized throughput to 55%. Moreover, delay is reduced and the learning convergence is quickly achieved.

A Game-Theoretic Approach In [56], the problem of H2H and M2M coexistence is formulated in terms of game-theory. In the proposed solution, instead of using a *unique* pool of preambles, different pools are allocated to M2M and H2H users: in particular, there are R_H preambles for H2H, R_M for M2M, and R_B available for both. The i-th MTDs is allowed to extract a random preamble either in the M2M-dedicated pool (action $a_i = M$), in the shared one ($a_i = B$), or to stay silent ($a_i = S$) with a probability distribution that is determined according to the outcome of a game. The game formulation consists of a constant number of H2H users H_B that select preambles from the shared pool, and N MTDs, which are the players of the game. The MTDs play a mixed strategy $\sigma_i(a_i)$, choosing actions M, B or S with probability $p_{i,M}$, $p_{i,B}$ or $1-p_{i,M}-p_{i,B}$, respectively. The preamble transmission has a cost $\lambda \in [0,1]$ (e.g., in terms of energy consumption), yielding the following gains

$$g_i = \begin{cases} 1-\lambda & \text{if transmission is successful;} \\ -\lambda & \text{if transmission fails;} \\ 0 & \text{if transmission is not attempted.} \end{cases} \tag{1.19}$$

Denoting with $P_S(a_i)$ the RA success probability if action a_i, the expected payoff of player i is

$$\mathbb{E}[g_i] = \sum_{a_i \in \{M,B\}} \sigma_i(a_i) \times [P_S(a_i) - \lambda]\,. \tag{1.20}$$

Simulations show that, following a mixed-strategy Nash Equilibrium (NE), every MTD has non-negative utility, and the throughput of both M2M and H2H

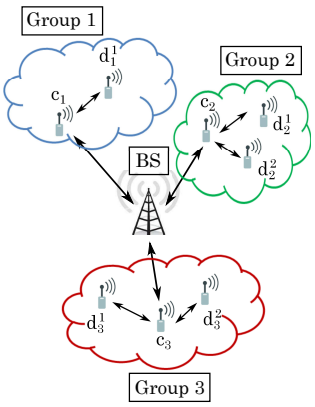


Figure 1.13: Grouping model for MTC proposed in [57]

users is improved with respect to the baseline scheme in the case of overloaded PRACH. Moreover, the authors provide a procedure to estimate the actual number of H2H and M2M devices in real systems, which in practice may have imperfect knowledge of the exact values of H_B and N. The proposed approach is proved to converge quickly and to provide small estimation errors for N.

1.4.3 Enhancement to Energy Efficiency and QoS

Besides the PRACH overload, other relevant challenges like the energy efficiency and the QoS support of M2M services in cellular networks were addressed. We prefer to classify the different techniques according to the methodological nature of the proposed solutions, rather than their specific objectives, which can involve one or more performance indices. Thus, we divided the various solutions in "Clustering Techniques" and "Game Theoretic Approaches;" a machine-learning-based approach is also described.

Clustering Techniques

One possible way to handle massive access to the BS is to appoint a few nodes (called *coordinators* or *cluster-heads*) as relays for the remaining terminals. In this way, the number of access requests to the BS is limited to the number of coordinators. Furthermore, a proper selection of the coordinators can also contribute to decrease the energy consumption of the system by exploiting multi-hop transmissions over high-gain links in place of direct transmissions over poor quality links. The problem now becomes the design of suitable policies for electing the relays and assigning the terminals to the different clusters. In the following we describe some solutions that have been proposed in literature.

Energy-Efficient Clustering of MTDs The authors of [57] proposed a clustering approach to limit the number of simultaneous accesses to the BS and the energy consumption of MTDs. Specifically, the authors consider a scenario with N MTDs, randomly deployed in a single cell centered at the BS, which knows the channel conditions to each terminal. The idea is to group the MTDs

in G clusters and, for each group, select a coordinator that is the only device allowed to communicate with the BS, relaying the communications of the other terminals in its cluster (see Figure 1.13). The total energy consumption of the system can hence be expressed as

$$EC = \sum_{i=1}^{G} \sum_{d_i^j \in G_i \setminus \{c_i\}} \left(\frac{P_t\, L_s}{R(d_i^j, c_i)} + \frac{P_t\, L_s}{R(c_i, BS)} \right), \tag{1.21}$$

where G_i is the set of nodes in the i-th cluster,[10] d_i^j and c_i denote the j-th MTD and the concentrator of cluster i, respectively, $R(x, y)$ is the transmit bit rate from node x to node y with transmit power P_t, while L_s is the length of the packet to be transmitted.

The objective is to minimize (1.21) while keeping the number of groups G below a certain threshold M, hence limiting the maximum number of access requests to the BS and reducing the redundant signaling of M2M services. To this end, the authors of [57] study a number of algorithms that combine different grouping and coordinator selection techniques. The simulation results show that the proposed clustering techniques are effective in reducing the massive access issue and improving the energy efficiency of the MTDs. Indeed, almost all proposed schemes perform better than direct transmission between MTDs and BS in terms of energy consumption and network load. More specifically, in a scenario with randomly distributed MTDs around the BS, the energy consumption tends to decrease with the number G of groups, until it becomes almost constant for $G \geq 10$. On the other hand, when the MTDs distribution over the cell is not uniform (e.g., nodes are concentrated in two or three smaller areas around the BS), the energy consumption is minimized by a lower number of groups.

Another clustering technique to maximize the energy efficiency of MTC has been proposed in [58], considering a cellular network system based on OFDMA, like the LTE. The authors propose to appoint some nodes as coordinators of a certain group of MTDs and use two-hop communication to connect the peripheral nodes to the BS, but, differently from [57], more realistic details are considered. The authors assume that communications between MTDs and coordinators are managed by means of a TDMA scheme, while the coordinators communicate with the BS by using an OFDMA channel. Clustering and coordinator selection are based on an energy-consumption model that accounts for both the energy spent by each terminal to transmit data and some additional energy expenditure due to the circuitry, i.e.,

$$EC' = \sum_{i=1}^{G} \left(\frac{(P_{c_i} + P_{\text{cir}}) D_{c_i}}{R(c_i, BS)} + \sum_{d_i^j \in G_i \setminus \{c_i\}} P_{d_i^j} \times \frac{L_s}{R(d_i^j, c_i)} \right), \tag{1.22}$$

where the notation is as in Equation (1.21), except for P_x that denotes the transmit power of node x, P_{cir} is the fixed circuitry power consumption, and D_{c_i} is the aggregate data received by the coordinator c_i from all the MTDs in its cluster.

[10] The notation $G_i \setminus \{c_i\}$ indicates the set G_i without the element c_i.

The following optimization problem is then formulated:

$$\min_{G,G_i,c_i,P_{c_i}} EC' \tag{1.23a}$$

subject to

$$G \leq N\,, \tag{1.23b}$$

$$P_{d_i^j} = P_t\,, \quad i \in \{1,\ldots,G\}, d_i^j \in G_i \setminus \{c_i\}\,, \tag{1.23c}$$

where N is the total number of MTDs and P_t is a fixed power value. Although this formulation holds under the assumption that all links are subject to flat fading, a similar problem can be defined in the case of frequency selective fading. However, finding the solution of such a problem is very complex, thus the authors of [58] propose a sub-optimal solution that consists in an iterative algorithm that first clusters the MTDs into groups, and then selects the coordinator for each group.

Notably, neither [57] nor [58] account for the energy spent in reception. Furthermore, as for all cluster-based scheme, coordinators are subject to higher power consumption and may fail because of energy depletion before the other nodes. Hence, mechanisms for the rotation of the cluster-head role shall be considered. On the other hand, these countermeasures would require higher costs in terms of signaling and control traffic, which shall also be accounted for.

QoS-Based Clustering In [59–61], clustering is used as an effective solution to manage radio resource assignment to a large population of MTDs with small data transmissions and very disparate QoS requirements: MTDs are grouped into G clusters based on their packet arrival rate ρ and maximum tolerable jitter δ, in such a way that devices in the same group have very similar traffic characteristics and QoS requirements. With this grouping approach, the BS can manage radio resources at a cluster level, rather than per single MTD. As a proof of concept, a simulation-based evaluation is developed in [59] considering the system parameters of LTE and various QoS requirements covering a rather general set of M2M applications: the results show that the proposed grouping scheme can achieve the desired QoS guarantees.

Moreover, in [60], the authors distinguish between deterministic (*hard*) and statistical (*soft*) QoS guarantees to enable more flexibility in the resource assignment at the BS side, and they derive two sufficient conditions to ensure that the QoS constraints of MTDs are satisfied.

The management of the QoS requirements when M2M and H2H applications coexist is addressed in [62]. In this paper, the authors apply clustering to divide the devices in two classes: the "high-priority" class collects all classical H2H users and some delay-sensitive M2M service users, while all the other MTDs are grouped in the "low-priority" class. Then, the authors define the M2M-Aware Scheduling Algorithm (M2MA-SA), which aims at preserving the performance experienced by high-priority services in case of massive presence of low-priority devices. The simulations results confirm that M2MA-SA yields smaller delay, lower outage probability, and higher throughput for H2H users than a basic

scheduling mechanism that does not discriminate between H2H and M2M queues. Of course, this result is obtained by penalizing the performance of M2M users.

Game Theoretic Approaches

An interesting alternative to QoS-based clustering approaches is given by the *distributed matching algorithm* introduced in [63]. Considering a typical single-cell scenario with conventional UEs and MTDs sharing the same physical resources, in this work the radio resources are originally assigned to the UEs only, while MTCs is subordinated to the availability of idle portions of such resources. In particular, each MTD is coupled with a UE and exploits parts of its transmission resources. The coupling between MTDs and UEs results from a distributed algorithm, based on matching theory, which negotiates the associations between UEs and MTDs, according to specific metrics: the final goal is to solve an optimization problem that finds the optimal matching between UEs and MTDs, maximizing the aggregate utilities of all MTDs and UEs. The simulation results show that the proposed algorithm achieves a stable matching state and an average aggregate utility comparable with a classical centralized algorithm, but with lower overhead for the BS.

A Machine-Learning Based Approach

Machine learning techniques are widely applied to protocol design thanks to their ability to deal with very complex systems in a relatively simple and efficient manner. The management of M2M massive access is not an exception, and several studies in this domain have adopted machine learning techniques to address different problems, e.g., the optimal selection of the serving BS for a given MTD. Indeed, one of the expected characteristics of 5G systems is the proliferation of pico- and femto-cells, which will increase the network coverage; as a consequence, each MTD will likely be in the coverage range of multiple BSs that, however, may offer quite disparate QoS, depending on their distance, signal propagation conditions, traffic load, and so on.

In [64], focusing on a LTE network, the authors propose a reinforcement-learning algorithm to enable MTDs to choose the best serving BS. A scenario with just two BSs is considered for the performance evaluation: the simulation results show that the proposed method yields a balanced distribution of the MTDs between the two BSs. Indeed, when the number of nodes that choose one BS increases, the packet delivery delay also increases due to higher congestion at that BS, so that the algorithm will progressively move some devices to the other BS. Moreover, the higher the number of allocated resource blocks by a certain BS, the more the MTDs that select that BS. Finally, it is shown that random BS selection is outperformed by the proposed reinforcement learning algorithm in terms of average packet delivery delay. On the other hand, the algorithm is not capable of promptly reacting to sudden changes of the M2M offered traffic, thus possibly yielding a sub-optimal behavior for relatively long periods.

1.4.4 Clean-Slate Approaches

While the previous studies referred to 4G standards, though with different degrees of abstraction of the system components, a few works have investigated

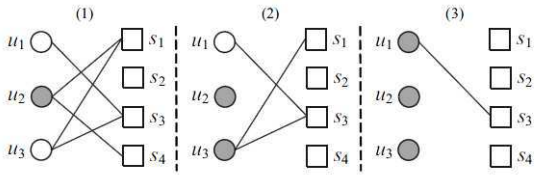


Figure 1.14: Graphical example of the frameless ALOHA protocol

the massive access issues in a *standard-agnostic* manner, with the aim of finding more fundamental results and shading light on the intrinsic performance bounds of these types of systems.

Schemes Based on Slotted ALOHA

The performance of coordinated and uncoordinated transmission strategies for multiple access is analyzed in [65], where it is shown that, for payloads shorter than 1000 bits (which are typical values for MTC), uncoordinated access schemes support more devices than coordinated access mechanisms, because of the lower signaling overhead. A well-known protocol for uncoordinated access is slotted ALOHA. An enhanced version of this protocol, called Fast Adaptive Slotted ALOHA (FASA), is proposed in [66]: taking into account the burstiness of M2M traffic, the knowledge of the idle, successful, or collided state of the previous slots is exploited in order to improve the performance of the access control protocol. In particular, the number of consecutive idle or collision slots is used to estimate the number of active MTDs in the network (the so-called "network status"), enabling a fast update of the transmission probability of the MTDs and, hence, reducing access delays. By means of drift analysis techniques, the authors prove the stability of the FASA protocol when the normalized arrival rate is lower than e^{-1}.

Another improved version of slotted ALOHA, exploiting Successive Interference Cancelation (SIC), called *frameless ALOHA*, is presented in [67]. A simple example illustrating the principle of such SIC-enabled slotted ALOHA is shown in Figure 1.14, which depicts the situation in which $N = 3$ users contend to transmit within the same frame, composed by $M = 4$ slots. Circles and squares represent users and time slots, respectively, while the edges connect the users with the slots in which their respective transmissions take place. All transmissions made by a user in the frame are replicas of the same packet; moreover, every transmission includes pointers to all its replicas. The SIC mechanism works as follows.

1. Slots containing a single transmission (*singleton* slots) are identified and the corresponding transmission resolved. Focusing on Figure 1.14, s_4 is recognized as a singleton slot and the associated packet of user u_2 is hence correctly decoded (see Figure 1.14-1).

2. Using the pointers carried by the decoded packets, all their replicas are removed from the associated slots, i.e., the interference caused by such transmissions is canceled from the aggregate received signal, thus potentially

leading to new singleton slots.[11] In the example of Figure 1.14, the replica sent by u_2 in s_1 is deleted and, as a result, s_1 becomes singleton slot (Figure 1.14-2).

3. The procedure is iterated until either there are no new singleton slots, or all transmissions have been recovered (Figure 1.14-3).

Generalizing this procedure and applying it to the M2M communication context, we can consider a scenario in which N MTDs contend to access the same BS. The protocol assumes that users are slot-synchronized and aware of the start of the contention period, which will be broadcast by the BS. For each slot in the contention period, each active MTD randomly decides whether or not to transmit a replica of the pending packet, according to a predefined probability $p_a = \beta/N$, where β is a parameter that needs to be optimized. After each slot, the BS collects the received compound signal and tries to decode the transmitted packets using the above described SIC procedure. The key feature of the frameless ALOHA proposed in [67] is that the end of the contention period is dynamically determined in order to maximize the throughput. Users that have not successfully delivered their packet by the end of a contention period will keep performing the algorithm in the subsequent contention round. Note that, if the contention period is terminated at the M-th slot and the number of resolved MTDs is N_R, then the instantaneous throughput can be computed as $T_I = N_R/M$. The results presented in [67] show that the proposed algorithm can achieve extremely high throughput and very low loss rate, thus proving the effectiveness and efficiency of the described model in a M2M scenario. The performance of the SIC-based frameless ALOHA scheme can be further improved by considering the *capture effect* [68].

All in all, frameless ALOHA can ideally guarantee high performance in a M2M scenario in terms of throughput. Nonetheless, energy efficiency and complexity aspects have not been considered yet. In particular, the SIC mechanism sets quite high requirements to the BS in terms of storage and processing capabilities. The BS indeed has to store the raw samples of the compound received signal in all the slots of a contention period, and carry out SIC-based signal decoding on many slots in real time. In addition, the frameless ALOHA protocol has a strong impact on MTDs' energy consumption because, for each frame, the devices must transmit a possibly large set of replicas of the same packet to the BS. This aspect is a major issue in M2M communication since, as we already pointed out, many MTDs are constrained by the need to operate for years without any battery replacement/recharge.

Asymptotic Analysis of Massive Access Capacity

In the recent literature [69,70], it was observed that using SIC in combination with Multi-Packet Reception (MPR) capabilities makes it possible to dramatically increase the system throughput even when transmitters are not centrally coordinated. From the simulation results, it can be observed that a BS capable of performing perfect SIC and MPR can theoretically decode an arbitrary large number of simultaneous transmissions by proportionally reducing the per-user

[11] For the sake of simplicity, signal cancellation is assumed to be perfect, i.e., to completely remove the power of the canceled signal without leaving any residual interference.

data rate. Doing this, the aggregate system capacity remains almost constant. Furthermore, it appears that the capacity of the cell depends on the statistical distribution of the signal powers, and the higher the variance, the more effective the SIC. Therefore, combining SIC, MPR, and coded random access techniques, it is possible to support massive access of sporadic transmitters to a common BS, provided that the receiver is capable of performing multi-slot SIC decoding. Once again, however, the analysis has not yet considered aspects related to the energy consumption of the MTDs.

Table 1.6 offers a compound view of the aforementioned solutions, with an indication of the characterizing features and the main targeted performance indices. More specifically, we consider the following aspects.

- *Main challenge*: primary issue addressed by the scheme;
- *3GPP*: the scheme was proposed in a 3GPP technical report and is designed with explicit reference to 3GPP standards (LTE in particular);
- *H2H & M2M*: the scenario assumes the coexistence of H2H and M2M services;
- *Performance indices*: the scheme is designed to improve the following figures of merit: i) minimization of access delay; ii) minimization of energy consumption of MTDs; iii) maximization of access probability/throughput of UEs and/or MTDs.

Table 1.6: Comparison of the approaches proposed in literature

Solution	Main Challenge	3GPP	H2H & M2M	Performance Indices		
				Delay	Energy Efficiency	Access Probability
Resource Separation [26]	PRACH overload	✓	✓			
Slotted Access [26]	PRACH overload	✓				✓
Pull-based Scheme [26]	PRACH overload	✓		✓		✓
Backoff Tuning [26]	PRACH overload	✓	✓			✓
Access Class Barring [26]	PRACH overload	✓	✓*	✓		
SOOC [52]	PRACH overload			✓		✓
RA for Fixed-Location [53]	PRACH overload			✓		✓
Bulk Signaling [54]	PRACH overload				✓	✓
Q-learning [55]	PRACH overload		✓	✓		✓
Game Theoretic Scheme [56]	PRACH overload		✓		✓	✓
Energy-Efficient Clustering [57, 58]	Energy consumption				✓	
QoS-Based Clustering [59–61]	QoS for M2M			✓		
M2M-Aware Scheduling [62]	QoS for H2H&M2M		✓	✓		✓
Matching Theory Scheme [63]	QoS for H2H&M2M		✓			✓
Reinforcement Learning [64]	BS selection			✓		✓
Clean Slate Approaches [67, 68, 70]	Massive access					✓

NOTES: "Delay" is intended as the time from the first transmission attempt until the successful conclusion of the access procedure. (*): the scheme can support H2H and M2M separation, though it is not specifically designed for this purpose.

1.5 Proposed Radio Access Protocol for 5G

We will now introduce a novel radio access protocol for sporadic transmissions of small-data packets, which are suitable for 5G networks because it provides a resource-efficient packet delivery by exploiting a *connection-less* approach. The core of our work resides in the derivation of an analytical framework to evaluate the performance of both 4G and 5G radio access protocols. The final goal is the comparison between the aforementioned solutions employing both our analytical framework and computer simulations. The performance evaluation results show the benefits of the protocols envisioned for 5G in terms of signaling overhead and access latency.

Let us remark that the fundamental drawback of many innovative solutions described in Section 1.4 is that they require to modify the current cellular network air interface (or, especially for those approaches presented in Section 1.4.4, even to add a *brand-new* air interface, separated from the current one): therefore, the implementation of such solutions would be a serious issue. Moreover, the additional air interface could not be used for other services in case IoT traffic is not present, thus wasting the allocated spectrum. The aim of our work, instead, is to provide a *unified* air interface for 5G which is able to integrate broadband services and IoT services at the same time, ensuring the backward compatibility with current cellular standards like LTE. With this flexible design, radio resources can be dynamically allocated for those services that actually need them. Simultaneous support of multiple services by sub-band wise optimization of PHY parameters like subcarrier spacing or pulse shape is discussed in [71] and [72], respectively. An additional filtering of the sub-bands can mitigate Inter-Service-Band-Interference (ISBI) [73].

In this section, we first present the PHY specifications, then the proposed solution is introduced in two variants, i.e., the *One-Stage* protocol and the *Two-Stage* protocol [74]. Possible feedback formats are discussed and, finally, a comparison with LTE is provided. Without loss of generality, in the following we assume perfect synchronization of all UL transmissions at the eNB.[12]

1.5.1 Physical Layer Design

Let us refer to Figure 1.15 and consider a multi-carrier transmission system, based on an OFDMA, consisting of elementary resource units called Resource Elements (REs), equivalent to one subcarrier and one time symbol (OFDM symbol). A group of REs over S subcarriers and T symbols forms a Resource Block (RB). In the following we assume that a RB spans a period of one subframe, also called Transmission Time Interval (TTI), of duration T_{TTI}. We remark that such a PHY design is implemented by the latest cellular network technologies like LTE.

Without loss of generality, we define a Small Packet Block (SPB) as a group of J RBs stacked in frequency.[13] The time duration of a SPB is still one subframe, as for the RB. In every subframe, M SPBs are available, where M_{SR} SPBs are dedicated for scheduling requests and M_{D} SPBs for actual data transmission,[14]

[12] We invite the reader to refer to Section 1.5.5 for the motivation of this assumption.

[13] We recalled that the concept of SPB was already envisioned in Section 1.3.1.

[14] As an alternative to this Frequency Division Multiplexing (FDM) scheme, which complies

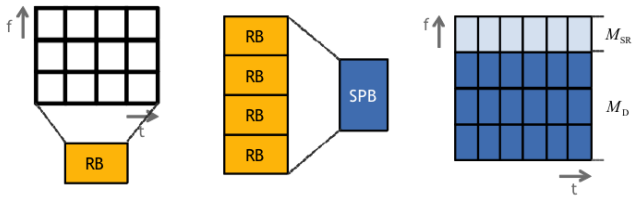


Figure 1.15: Proposed OFDM structure. The white boxes denote the REs, the yellow ones the RBs, and the blue ones the SPBs.

in such a way that

$$M = M_{\mathrm{SR}} + M_{\mathrm{D}} \,. \tag{1.24}$$

In the proposed solution a SR is represented by a code sequence, i.e., a signature, similar to an LTE preamble, that is mapped on the radio resources. If we assume that R orthogonal codes can be distinguished per RB, a total amount of RJM_{SR} SRs per TTI can be detected at the eNB side. There is not a one-to-one mapping between SRs and data SPBs, but we rather allow for an *over-provisioning* of SRs, i.e., we assume that the number of signatures is greater than the actually available data resources. For this reason, the over-provisioning parameter N is introduced, so that

$$R \times J \times M_{\mathrm{SR}} \simeq N \times M_{\mathrm{D}} \,, \tag{1.25}$$

under the assumption (without loss of generality) that N is an integer value in this derivation and in the following of the chapter. Note that parameter N may equivalently be defined as the ratio between the aggregate number of SRs and the amount of data SPBs. Substituting $M_{\mathrm{D}} = M - M_{\mathrm{SR}}$ from (1.24) in (1.25), the value of M_{SR} can be determined as a function of N, yielding

$$M_{\mathrm{SR}} = \left\lceil \frac{NM}{RJ+N} \right\rceil \,. \tag{1.26}$$

Equivalently, N can be obtained as a function of M_{SR} as follows:

$$N = \left\lfloor \frac{RJM_{\mathrm{SR}}}{M - M_{\mathrm{SR}}} \right\rfloor \,. \tag{1.27}$$

We remark that $N \geq 1$, since at least one signatures should be associated to every data SPB. A graphical representation of the SR mapping into RBs and of the over-provisioning factor N with respect to one data SPB is provided in Figure 1.16.

1.5.2 One-Stage Protocol

The first variant of the proposed solution is called *One-Stage* protocol. A graphical representation of the protocol can be found in Figure 1.17. The protocol consists of three steps.

to the current resource allocation scheme in LTE, a Time Division Multiplexing (TDM) scheme can be applied in the case of a very small system bandwidth. However, as access delay is one key performance indicator, in the following of the section we will mainly focus on FDM. A hybrid FDM/TDM scheme will be proposed in Section 1.5.4.

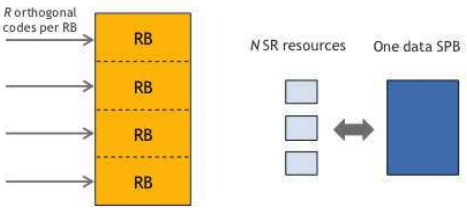


Figure 1.16: Mapping of SR resources in RBs and over-provisioning factor N compared to data SPBs

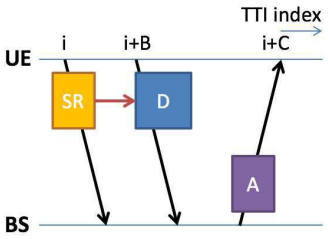


Figure 1.17: One-Stage protocol

1. The MTD sends a SR with index $r = 1, \ldots, NM_{\mathrm{D}}$ that points uniquely to one specific payload resource.

2. The MTD sends its data packet utilizing the SPB corresponding to the chosen SR, either in the same subframe or in one of the subsequent subframes.

3. If the data transmission is successful, then the eNB acknowledges the packet; otherwise a not-acknowledgement (NACK) message is sent to the MTD.

The transmission of the SR in step 1 is utilized for *activity detection*. Although the reservation of extra RBs for SRs is not mandatory if the One-Stage protocol operates in a standalone system, we remark that step 1 becomes necessary in a multi-service interface that has to support both collision-free and contention-based data transmissions. Furthermore, the SR can implicitly hold some extra information like the size of the SPB or the used Modulation and Coding Scheme (MCS) for the data. Finally, we observe that the predefined mapping between SRs and data SPBs could be disadvantageous in the presence of frequency selective fading. In case the MTD has some channel knowledge based previous transmission attempts, which is a valid assumption if the propagation conditions are static, the MTD can avoid SRs pointing to data resources in a deep fade.

This radio access solution is much faster than LTE in case the transmission is successful, and it requires significantly less DL feedback. There are some disadvantages, though: the high collision probability reduces the throughput and the coexistence with scheduled transmissions may be difficult to handle. For these reasons, this solution is envisaged for very small packets and low traffic load.

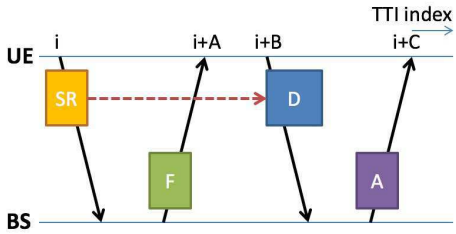


Figure 1.18: Two-Stage protocol

1.5.3 Two-Stage Protocol

The benefit of a high over-provisioning factor N resides primarily in a reduction of the probability that more than one MTDs select the same preamble index to send a SR. Nevertheless, this positive effect is not exploited in the One-Stage protocol. Indeed, even though the terminals pick different preambles, if their SRs point to the same data SPB, we cannot avoid the collision in step 2 and all collided data packets are lost. Furthermore, to provide a higher value of N we have to increase M_{SR} and, consequently, decrease M_{D}. Therefore, the best option for the One-Stage approach is to minimize the value of N.

On the other hand, we may take advantage of over-provisioning as follows. As one variant of the previous protocol let us assume that the eNB is able to reject part of the detected SRs in order to prevent packet collisions on data SPBs.[15] This second proposed approach is called *Two-Stage* protocol and its operation is graphically explained in Figure 1.18. The protocol consists of four steps.

1. The MTD chooses a random index $r = 1, \ldots, NM_{\mathrm{D}}$ and sends a SR to the eNB, requesting UL radio resources.

2. The eNB sends information related to the assignment of a data SPB as feedback, i.e., which SPB to use and in which subframe to use it. In case the SR is not acknowledged, the MTD randomly selects a new index r and sends a new SR after a random time offset $\beta \in [\beta_{\min}\ ,\ \beta_{\max}]$.

3. The MTD sends data utilizing the assigned data SPB.

4. The eNB acknowledges the data transmission if received successfully. In case the data SPB is not acknowledged, the MTD restarts the procedure from step 1. The available number of SR transmissions are restricted to Θ in order to avoid the overload of the system.

This second radio access solution is promising because, using a high over-provisioning factor, it reduces the collision probability and, therefore, increases the throughput. Of course, collisions on data resources still happen if more than one MTD choose the same preamble index in step 1. On the other hand, it requires an additional delay with respect to the One-Stage protocol due to the

[15] Alternatively, over-provisioning of SRs could be exploited through Multi-User Detection (MUD) techniques. If the eNB is aware that one data SPB is utilized by multiple UEs, it may apply, e.g., SIC. This aspects will be part of our future work.

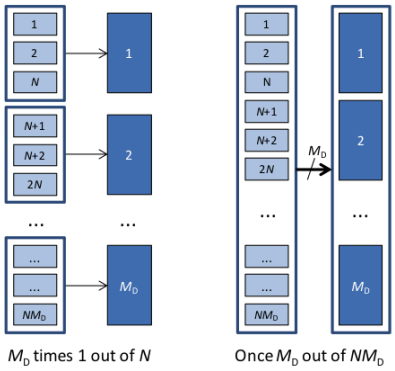


Figure 1.19: Difference between tagged data SPBs (left hand side) and pooled data SPBs (right hand side)

feedback required after the SR transmission. For these reasons, this solution is envisaged for bigger packets and high traffic load.

We remark that a dynamic usage of the two protocols is possible, according to the traffic load.

Physical Resources Grouping

We can further customize the Two-Stage approach by splitting the total amount of available data SPBs M_{D} into K distinct groups, where K is such that $1 \leq K \leq M_{\mathrm{D}}$. This splitting can be either fixed per specification or can be adapted dynamically by the eNB per DL control channel according to the current traffic situation. For the sake of simplicity, in the following we assume that every group comprises the same number of SPBs M_{D}/K and has the same arrival rate of new users. Consequently, $NM_{\mathrm{D}}/K \simeq RJM_{\mathrm{SR}}/K$ SRs are associated to every group, and each MTD sends a SR that is associated with the targeted group. Two special cases are noteworthy:

1. the *tagged data* case, in which $K = M_{\mathrm{D}}$, i.e., each group consists of exactly one SPB,
2. and the *pooled data* case, in which, instead, $K = 1$, i.e., all SPBs belong to one single group.

A graphical representation of these extreme cases can be found in Figure 1.19. In the first special case, each SR points exactly to one single data resource. As a consequence, the required feedback from the eNB is minimized (just acknowledgement (ACK) or NACK must be indicated, since the data SPB is already fixed), however, the scheduling is not flexible. With the transmission of the SR, it is already clear which SPB will be utilized for the data packet later on. As for frequency selective channels, similar consideration to the One-Stage case can be done. In the pooled case, instead, there is no predefined tagging between SRs and data SPBs. Consequently, the eNB has full scheduling flexibility, i.e., the eNB can distribute the full set of SPBs among the received SRs. This comes along with the drawback of an increased feedback effort in the DL: for each acknowledged SR the eNB has to indicate the assigned data SPB or vice versa.

The physical resource grouping allows for a differentiation of service types, device classes, packet sizes, or transport formats in a real and complex communication system. As an example, $K = 2$ groups can be configured, one group for high priority services (e.g., fire alarms), which consists of a big number of SPBs for a comparably low number of MTDs, and, vice versa, a second group for low priority services (e.g., air temperature measurement), which consists of only few SPBs for many MTDs. Consequently, the high priority service will experience a significantly lower collision probability and a higher success rate.

1.5.4 Feedback Formats

As DL signaling efficiency concerns, we provide some consideration about the feedback format for the proposed protocols, focusing both on the case of a fixed time relation and a relaxed time relation between the steps of the protocol.

Fixed Time Relation

Let us assume that a fixed time relation exists between the steps of the Two-Stage protocol, e.g., A, B, C TTIs as depicted in Figure 1.18. As a consequence, the SR feedback from the eNB consists only of the particular SPB index that is assigned to each request. The following feedback formats for the SR ACK are proposed.

- *Option 1*: For every SPB the acknowledged SR is indicated. Since for every SPB we have to identify the SR we acknowledge within the corresponding group, we need a binary vector of length

$$F_1 = M_{\mathrm{D}} \left\lceil \log_2 \left(\frac{N M_{\mathrm{D}}}{K} \right) \right\rceil \,[\mathrm{bit}]\,. \tag{1.28}$$

  Note that the number of required bits for this feedback format is very low, but the MTD must look for its SR index in the selected group (thus, making M_{D}/K searches).

- *Option 2*: For every SR the assigned data SPB is indicated. Since for every SR we have to identify the assigned SPB within the corresponding group, we need a binary vector of length

$$F_2 = N M_{\mathrm{D}} \left\lceil \log_2 \left(\frac{M_{\mathrm{D}}}{K} + 1 \right) \right\rceil \,[\mathrm{bit}]\,. \tag{1.29}$$

  Note that the $+1$ accounts for an additional codeword for the NACK. Moreover, in the case of tagged data resources ($K = M_{\mathrm{D}}$) it is $F_2 = N M_{\mathrm{D}}$. We remark that this kind of option is larger in terms of bits, but the UE now does not need to search for the SR it sent.

Relaxed Time Relation

A performance gain is expected if the constraint of fixed time scheduling is relaxed, i.e., if we allow to delay a data packet transmission from an entirely occupied subframe to a later one. Two approaches for a fully flexible data SPB scheduling are proposed.

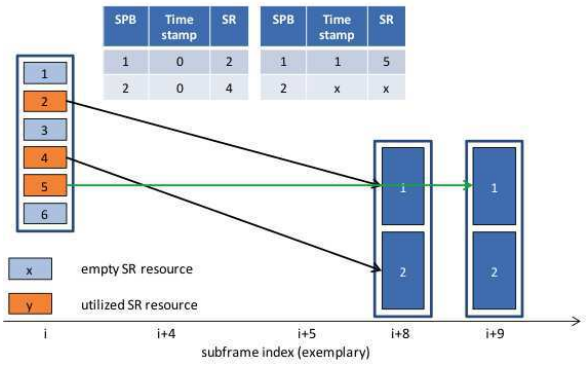


Figure 1.20: Example of time stamp feedback with $M_{\mathrm{D}}/K = 2$, $N = 3$, $W = 2$, $A = 4$, and $B = 8$

Feedback with Time Stamp Considering a window of W subframes, we assume that the feedback comprises, in addition to the assigned SPB, a subframe index $w = 0, \ldots, W-1$ indicating the additional delay that has to be added to the minimal offset between the reception of SR feedback and the data transmission. Note that the fixed time relation is a particular case of the relaxed time relation with $W = 1$. Under this assumption, the length of feedback messages are

$$F_1^{(\mathrm{TS})} = M_{\mathrm{D}} \left\lceil \log_2 \left(\frac{WNM_{\mathrm{D}}}{K} \right) \right\rceil \text{ [bit]} \tag{1.30}$$

using Option 1 and

$$F_2^{(\mathrm{TS})} = NM_{\mathrm{D}} \left\lceil \log_2 \left(\frac{WM_{\mathrm{D}}}{K} + 1 \right) \right\rceil \text{ [bit]} \tag{1.31}$$

using Option 2. A graphical example of time stamp feedback is provided in Figure 1.20, assuming that $M_{\mathrm{D}}/K = 2$, $N = 3$, $W = 2$, $A = 4$, and $B = 8$. In subframe i, 3 MTDs choose indices $r_1 = 2$, $r_2 = 4$, and $r_3 = 5$, respectively, and send their SRs. The third MTD, after the default delay of $A = 4$ TTIs, reads the feedback information, but does not find the acknowledgement of $r_3 = 5$. Since $W = 2$, the MTD is allowed to look for its SR again in subframe $i + A + 1 = i + 5$. As it finds its SR in the feedback together with $w = 1$, it starts the data transmission on SPB number 1 in subframe $i + B + w = i + 9$. Note that, if the third MTD had found $w = 0$, it would have not interpreted the feedback in subframe $i + 5$ as for itself, but as the acknowledgement for another UE that sent the same SR in subframe $i + 1$.

Queueing-Based Feedback A promising approach to reduce the number of required feedback bits in the Two-Stage protocol with pooled resources consists in the Distributed Queuing Random Access Protocol (DQRAP) [75]. Many DQRAP-based protocols have been designed for wireless communications, e.g., for Wireless Local Area Network (WLAN) [76]. The drawback of this approaches is, however, that the eNB needs to be able to detect the collision of MTDs utilizing the same SR resource. This would require a huge additional complexity at the eNB receiver side. Therefore, a simplified queueing scheme using a single

Table 1.7: Mapping between pointer $P = Q + L$ and time-frequency position of assigned SPBs for data transmission

	$i+B$	$i+B+1$	$i+B+2$
$m=1$	$P=1$	$P=M_{\mathrm{D}}+1$	$P=2M_{\mathrm{D}}+1$
$m=2$	$P=2$	$P=M_{\mathrm{D}}+2$	$P=2M_{\mathrm{D}}+2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$m=M_{\mathrm{D}}$	$P=M_{\mathrm{D}}$	$P=2M_{\mathrm{D}}$	$P=3M_{\mathrm{D}}$

queue instead of two, as in DQRAP, has been designed. The queueing-based feedback consists of the length Q of queue $\mathcal{Q}$, accounting for the number of terminals that have already been acknowledged and are waiting to transmit, and of binary vector V, indicating for every SR whether it is active ($v_\ell = 1$) or not ($v_\ell = 0$), where v_ℓ is the ℓ-th value in vector V. Upon receiving this kind of feedback, a UE that chooses the SR index $r = X$ computes pointer P as follows:

$$P = Q + \sum_{\ell=1}^{X} v_\ell = Q + L\,, \tag{1.32}$$

where $L \triangleq \sum_{\ell=1}^{X} v_\ell$. The assigned TTI index t and SPB index m are derived from P according to Table 1.7, where it is assumed that the minimal timing offset between SR and data transmission is of B subframes. It can be seen that $t = i + B + \lfloor (P-1)/M_{\mathrm{D}} \rfloor$, while $m = [(P-1) \mod M_{\mathrm{D}}] + 1$.

An example of queueing-based feedback is provided in Figure 1.21, where we assume $M_{\mathrm{D}} = 24$, $K = 1$, and $B = 8$. We assume that a MTD chooses index $r = 17$ and sends the SR to the eNB. In subframe $i + A$ the MTD receives as feedback information $Q = 18$ and the vector V that is shown in the figure. It computes $L = 8$ and $P = 18 + 8 = 26$ and realizes that it has been scheduled in subframe $i + B + 1 = 9$ in SPB $m = 2$.

Note that the feedback message length using the queueing-based feedback is $F_Q = NM_{\mathrm{D}} + \lceil \log_2 Q \rceil$ [bit], where NM_{D} is the length of vector V. However, this feedback scheme can be generalized to $K > 1$ as well, sending K different values of Q and the vector V. Therefore, we can generalize the feedback length as follows:

$$F_Q = K\left(\frac{NM_{\mathrm{D}}}{K} + \lceil \log_2 Q \rceil\right) = NM_{\mathrm{D}} + K\lceil \log_2 Q \rceil\,[\mathrm{bit}]\,. \tag{1.33}$$

Feedback Comparison

A comparison of all the proposed feedback formats is provided in Table 1.8, assuming $M_{\mathrm{D}} = 24$, $N = 3$, and $W = 9$. To provide a fair comparison between the approaches with a relaxed time relation, we must provide a conversion between the window size W and the queue length Q. Assuming the *worst* case for the queueing-based approach, in which the generic MTD points the last data SPB, it is:

$$W \geq \frac{\max\{P\}}{\frac{M_{\mathrm{D}}}{K}} = \frac{Q + \frac{M_{\mathrm{D}}}{K}}{\frac{M_{\mathrm{D}}}{K}} = \frac{K}{M_{\mathrm{D}}}Q + 1 \tag{1.34}$$

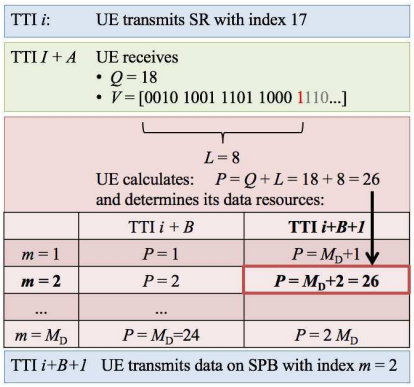


Figure 1.21: Example of queueing-based feedback, assuming $M_{\mathrm{D}} = 24$ and $K = 1$

Table 1.8: Feedback format lengths for the Two-Stage protocol, assuming $M_{\mathrm{D}} = 24$, $N = 3$, and $W = 9$

Feedback format	General formula [bit]	2-stage tagged ($Q = 8$)	2-stage pooled ($Q = 192$)
Option 1	$M_{\mathrm{D}} \left\lceil \log_2 \left(\frac{N M_{\mathrm{D}}}{K} \right) \right\rceil$	**48**	**168**
Option 2	$N M_{\mathrm{D}} \left\lceil \log_2 \left(\frac{M_{\mathrm{D}}}{K} + 1 \right) \right\rceil$	72	360
Option 1 with TS	$M_{\mathrm{D}} \left\lceil \log_2 \left(\frac{W N M_{\mathrm{D}}}{K} \right) \right\rceil$	**120**	240
Option 2 with TS	$N M_{\mathrm{D}} \left\lceil \log_2 \left(\frac{W M_{\mathrm{D}}}{K} + 1 \right) \right\rceil$	288	576
Queueing based	$N M_{\mathrm{D}} + K \lceil \log_2 Q \rceil$	144	**80**

yielding

$$W = \left\lceil \frac{K}{M_{\mathrm{D}}} Q + 1 \right\rceil \quad \text{and} \quad Q = \left\lfloor (W-1) \frac{M_{\mathrm{D}}}{K} \right\rfloor . \tag{1.35}$$

We infer that in the case of tagged resources the most efficient feedback format is option 1 with time stamp, while in case of pooled resources the most efficient one is the queueing-based feedback. It must be taken into account, however, that option 1 with time stamp forces the MTD to look for its SR in the feedback message, i.e., it is more computationally demanding with respect to the other formats.

1.5.5 Comparison with LTE

The proposed radio access solution provides many advantages with respect to LTE. Firstly, it introduces a *contention-based* transmission paradigm in which a MTD is allowed to send UL data without undergoing the LTE four-step access procedure for collision resolution, thus reducing the signaling overhead, in particular the DL feedback, decreasing the packet delivery delay, and allowing for a significantly higher number of simultaneously active MTDs per radio cell. Collision resolution is achieved through retransmissions after a random backoff time, which is sufficient for most delay-tolerant applications. The proposed

Table 1.9: Comparison of feedback lengths after SR transmission for LTE and the proposed protocol. For the tagged case and the pooled case, the option 1 with time stamp and the queueing-based feedback formats are considered, respectively.

	LTE	1-stage	2-stage tagged ($Q = 8$)	2-stage pooled ($Q = 192$)
General formula [bit]	$M_{\mathrm{D}}(8 + 48)$	0	$M_{\mathrm{D}}\lceil \log_2(WN) \rceil$	$NM_{\mathrm{D}} + \lceil \log_2 Q \rceil$
$M_{\mathrm{D}} = 24$, $N = 3$, $W = 9$	**1344**	0	120	80

feedback formats, indeed, are *broadcast* messages, while the LTE RAR consists of individual messages, since time offsets for each MTD are included and resources for Message 3 are not pre-configured. In particular, according to [44], every RAR requires one octet for the header and six octets for the data, i.e., 56 bits per SPB. Moreover, the configuration of PDSCH, which carries the RAR messages, requires a Downlink Control Information (DCI) Format 1A message of 25 bits [77]. It is worth noticing that PHY overhead should be accounted also for the proposed feedback format, but 5G PHY specifications are not yet defined. However, assuming that 5G physical channels will be similar to LTE ones and considering that in LTE the MAC overhead is predominant, we neglect the PHY overhead contribution and compare just the MAC layer overhead. As shown in Table 1.9, a number of 24 SRs would result in an aggregate RAR size of 1344 bits including headers, thus the proposed feedback formats provide an improvement of more than one order of magnitude with respect to LTE.

Secondly, the proposed protocols can be easily combined with the *connectionless* concept [78]. As machine-type traffic is characterized by *sporadic infrequent* transmissions of small packets, the *connection-oriented* paradigm of LTE is highly inefficient. Apart from the four-step RA protocol itself, a cascade of signaling messages has to be exchanged between the MTD and the network before the MTD is in `RRC_CONNECTED`, `IN_SYNC` state and data transmission is possible. We aim to avoid this overhead and include source and destination addresses directly in the SPB without prior connection setup, as recommended in [79]. Apart from a more efficient usage of radio resources, the connectionless concept helps to reduce energy consumption in the MTD, mainly due to a much shorter on-time of the radio module in the MTD compared to LTE. This helps to achieve a clearly longer battery lifetime: sensor networks, indeed, typically aim for a MTD battery lifetime of several years, but LTE has not been designed for that purpose. We remark that other solutions were proposed in literature in this direction [80, 81], but their focus is rather on a mere re-engineering of LTE RRC procedures to support M2M traffic, whereas our contribution offers an additional degree of freedom with respect to RRC states. Also, we will complement these concepts with a mathematical framework that can be generally applied for various configurations of radio access protocols.

Finally, we remark that the protocol implementation is not an issue, because current LTE physical channel specifications may be partly reused, but also selectively complemented by novel concepts like Universal-Filtered OFDM (UF-OFDM), also known as Universal-Filtered Multi-Carrier (UFMC), that

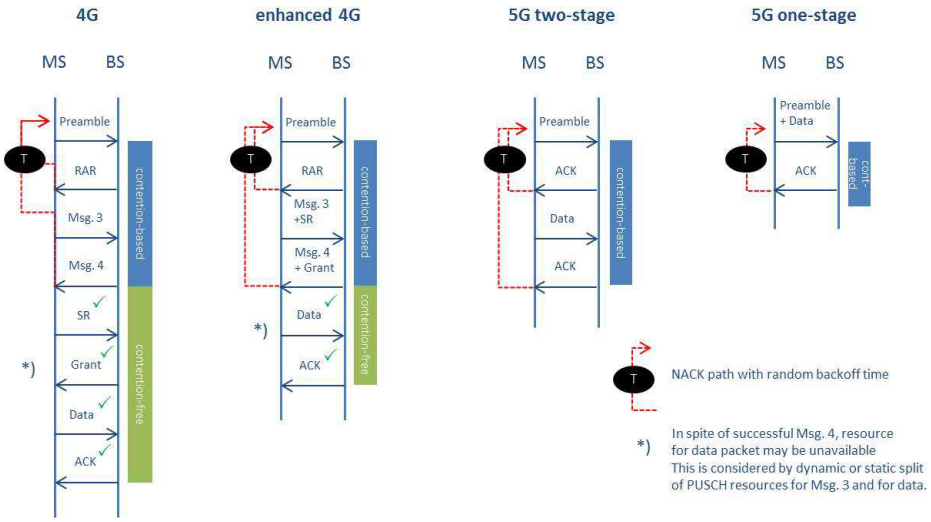


Figure 1.22: Comparison between 4G and 5G radio access solutions

allows for relaxed synchronization for data transmission without severe performance degradation [82]. Filter Bank Multi-Carrier (FBMC) [83] and Filtered-OFDM (F-OFDM) [84] are alternative waveform concepts addressing the same problem: the common approach is the application of filtering to minimize the mutual interference between users caused by imperfect synchronization of UL signals.

1.6 Mathematical Models

In this section, we will propose a mathematical characterization of the radio access protocols depicted in Figure 1.22 (4G, enhanced 4G, Two-Stage 5G, and One-Stage 5G) using an analytical approach similar to [85]. We assume that the arrival process of new packets at the system follows a Poisson distribution of rate λ (expressed in packets per second). The overall arrival rate at the system, denoted with λ_T, is obtained summing new transmission attempts and retransmissions, i.e., $\lambda_T = \lambda + \lambda_R$. It is assumed also that the time interval between two RAOs, denoted as δ_{RAO}, is equal to one TTI, since in every subframe M_{SR} SPBs are available for scheduling requests.

Let us recall that the eNB may not detect a collision in step 1 due to the capture effect of the channel or because the collided UE are not separable in terms of Power Delay Profile (PDP) [34]. For this reason, in practical systems the detection of collided preambles is often not considered, thus in the following of the analysis we assume that the eNB is *not* able to detect a collision event at step 1.

1.6.1 Model of the One-Stage Protocol

From the perspective of a generic device in a set of j nodes, each of which randomly chooses one resource out of n available resources, the probability that

another contender node selects the same resource is

$$q(j,n) \triangleq 1 - \left(1 - \frac{1}{n}\right)^{j-1} . \tag{1.36}$$

Let us define, then, the one-shot[16] *failure probability* p_f as the average of $q(j,n)$, with $n = M_{\mathrm{D}}$, over the Poisson distribution of j overall arrivals at the system in one RAO:

$$p_f = \mathbb{E}_j[q(j, M_{\mathrm{D}})] = \sum_{j=1}^{+\infty} \left[1 - \left(1 - \frac{1}{M_{\mathrm{D}}}\right)^{j-1}\right] e^{-\Delta} \frac{\Delta^j}{j!} \leq 1 - \left(1 - \frac{1}{M_{\mathrm{D}}}\right)^{\Delta - 1} , \tag{1.37}$$

where $\mathbb{E}[\cdot]$ denotes the expected value and $\Delta \triangleq \lambda_T \delta_{\mathrm{RAO}}$. We remark that the inequality holds for the Jensen's inequality and p_f is a function of λ_T.

The impact of multiple transmission attempts can be evaluated as presented in [85] by exploiting the Bianchi's model [86]. Recalling that Θ denotes the maximum number of transmission attempts, it can be proved that the *outage probability*, i.e., the probability of exceeding the maximum number of transmission attempts, of the radio access protocol is given by

$$p_{\mathrm{outage}} = p_f^{\Theta} . \tag{1.38}$$

The *average number of transmission attempts* is

$$\bar{\theta} = \sum_{\theta=1}^{\Theta} \theta \times \mathbb{P}[\theta \text{ tx}] = \sum_{\theta=1}^{\Theta-1} \theta p_f^{\theta-1}(1 - p_f) + \Theta p_f^{\Theta-1} = \frac{1 - p_f^{\Theta}}{1 - p_f} , \tag{1.39}$$

where $\mathbb{P}[\theta \text{ tx}]$ is the probability that a packet undergoes θ transmission attempts. If we count only successfully delivered packets the mean number of transmission attempts becomes

$$\bar{\theta}_{\mathrm{ACK}} = \sum_{\theta=1}^{\Theta} \theta \mathbb{P}[\theta \text{ tx}|\text{pkt ok}] = \sum_{\theta=1}^{\Theta} \theta \frac{p_f^{\theta-1}(1 - p_f)}{1 - p_{\mathrm{outage}}} = \frac{1 - (\Theta + 1)p_f^{\Theta} + \Theta p_f^{\Theta+1}}{(1 - p_f)(1 - p_f^{\Theta})} , \tag{1.40}$$

where $\mathbb{P}[\theta \text{ tx}|\text{pkt ok}]$ is the probability that a packet undergoes θ transmission attempts given that it is successfully delivered.

Finally, the value of λ_T can be determined solving the following fixed-point equation:

$$\lambda_T = \bar{\theta} \times \lambda = \frac{1 - p_f^{\Theta}}{1 - p_f} \times \lambda . \tag{1.41}$$

Note that if $\Theta = 1$ then $\lambda_T = \lambda$. The *throughput*, defined as the number of successful data packets per overall number of SPBs, can then be computed as

$$\mathcal{S} = \lambda \times (1 - p_{\mathrm{outage}}) . \tag{1.42}$$

[16] Allowing just one transmission attempt.

1.6.2 Model of the Two-Stage Protocol with Pooled Resources

Let us split the analysis of the protocol in two phases: the *preamble transmission* phase and the *data transmission* phase. The one-shot failure probability in this case is defined as follows:

$$p_f \triangleq 1 - (1 - p_c)(1 - p_d^A)\,, \tag{1.43}$$

where p_c is the *collision probability* in the preamble transmission phase and p_d is the *dropping probability* during the access granted phase.

The collision probability can be computed similarly to Equation (1.37), simply considering now the number of preambles in place of the amount of data SPBs. Therefore, we obtain

$$p_c = \mathbb{E}_j[q(j, NM_{\mathrm{D}})] \leq 1 - \left(1 - \frac{1}{NM_{\mathrm{D}}}\right)^{\Delta - 1}. \tag{1.44}$$

The data transmission phase, instead, is modeled as a *queueing system* in which the customers, i.e., the successful SRs, are *impatient* customers [87]. In particular, we are interested in evaluating the *long-term fraction of users who are lost*, that is, the dropping probability of the queue. Let us denote the arrival rate at the queue, the queue service rate, the number of servers, and the maximum waiting time with Λ, μ, m, and τ, respectively. The dropping probability for a M/M/m queue is defined as

$$p_d(\Lambda, m\mu, \tau) \triangleq \frac{(1-\rho)\rho\Omega}{1 - \rho^2\Omega}\,, \tag{1.45}$$

where $\rho = \Lambda/(m\mu)$ and $\Omega = e^{-m\mu(1-\rho)\tau}$. The system should be modeled as a M/D/m queue with impatient costumers, but no closed-form expression is known for this kind of queues. Nevertheless, according to [88], the expression of dropping probability for M/M/m queues is an excellent approximation for M/G/m queueing systems, including M/D/m queueing systems.

In the Two-Stage protocol, the arrival rate at the queue, denoted by λ_A, is the number of activated preambles (both collided and not) per time unit and it can be computed as follows. Let us define the random variable X as the number of users selecting the same preamble index. Since the average number of arrivals per preamble per subframe is $\alpha = \Delta/(NM_{\mathrm{D}})$, X is distributed according to a Poisson distribution of parameter α. We denote with ω the probability of preamble activation (a function of parameter α),

$$\omega(\alpha) = 1 - \mathbb{P}[X = 0] = 1 - e^{-\alpha}\,, \tag{1.46}$$

and assume that preamble activations are independent from each other. Then, we can compute the average arrival rate at the access granted λ_A as

$$\lambda_A = \frac{NM_{\mathrm{D}}}{\delta_{\mathrm{RAO}}} \times \omega(\alpha)\,, \tag{1.47}$$

since we model the number of activated preambles as a binomial random variable of parameters NM_{D} and $\omega(\alpha)$. The service rate of the access granted phase and the number of servers are

$$\mu_D = \frac{1}{T_{\mathrm{TTI}}} \quad \text{and} \quad m_D = M_{\mathrm{D}}\,, \tag{1.48}$$

respectively. The maximum waiting time of a SR is

$$\tau_q = W \times T_{\mathrm{TTI}}\,. \tag{1.49}$$

Finally, the access granted dropping probability is evaluated as

$$p_d^A = p_d(\lambda_A, m_D\mu_D, \tau_q)\,. \tag{1.50}$$

The impact of multiple transmission attempts can be evaluated as done in the One-Stage scheme, using formulas (1.38), (1.39), and (1.41) to evaluate the outage probability, the average number of transmissions, and the aggregate arrival rate λ_T, respectively. The system throughput can be computed as done in Equation (1.42).

1.6.3 Model of the Two-Stage Protocol with Grouped Resources

In the case of grouped resources, the access granted phase is characterized as a system with K parallel queues. We recall that, for the sake of simplicity, we assume that the groups have exactly the same number of dedicated resources. The rate of activated preambles at the generic queue k is

$$\lambda_A^{[k]} = \frac{NM_{\mathrm{D}}/K}{\delta_{\mathrm{RAO}}} \times \omega(\alpha) = \frac{1}{K} \times \lambda_A \qquad \forall k = 1, \ldots, K\,. \tag{1.51}$$

The number of servers of a single queue becomes

$$m_D^{[k]} = \frac{M_{\mathrm{D}}}{K} = \frac{1}{K} \times m_D \qquad \forall k = 1, \ldots, K\,, \tag{1.52}$$

while the service rate μ_{D} remains unchanged. The dropping probability at the access grant phase is obtained using (1.50), but considering now the expressions of $\lambda_A^{[k]}$ and $m^{[k]}$ in place of λ_A and m_D, respectively.

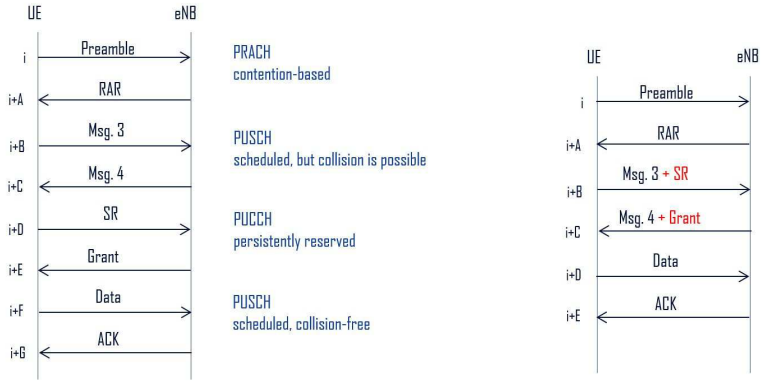
1.6.4 Model of LTE Radio Access for Small Packet Traffic

To provide a quantitative comparison between the proposed protocols for 5G and the current cellular standard, the LTE RA procedure has been tailored to small packet traffic and modeled following the steps shown in Figure 1.23: we will consider both the default RA procedure described in Section 1.2 (see Figure 1.23a) and the protocol variant proposed by 3GPP in [49] and described in Section 1.4.1 (see Figure 1.23b).

Let us recall that initially each UE is in `RRC_IDLE` to minimize the energy consumption between two subsequent packet transmissions: indeed, the interarrival time of packets is typically much longer than the `RRC_CONNECTION_RELEASE` timer. Therefore, the UE has to switch to `RRC_CONNECTED` state through the LTE RA procedure. We will neglect the aspects related to the RRC connection setup, in order to focus only on the core of the LTE radio access procedure.

We assume that n_{RB} RBs are available for PRACH, PUCCH, and PUSCH. For the sake of a fair comparison, we assume that every MTD requests a grant of fixed size[17] of J RBs, i.e., a data SPB, to send UL data on PUSCH. In the

[17] Under this assumption, the UE does not need to send its Buffer Status Report (BSR), since the dimension of the data to transmit are already fixed; thus, in the following analysis we will not allocate resources for BSR messages.



(a) Default radio access procedure in LTE **(b)** Enhanced 4G radio access protocol

Figure 1.23: Timing of LTE radio access protocols. RRC connection setup is neglected.

following, we describe the setup of the various physical channels and derive the model of LTE radio access procedure.

PRACH The PRACH takes $n_{\mathrm{PRACH}} = 6$ RBs, stacked in frequency [19]. We set δ_{RAO} to one TTI, thus the PRACH is instantiated in every subframe. We recall that, in the entire pool of 64 Zadoff-Chu orthogonal sequences, $d = 54$ signatures are used for contention-based access and the remaining ten for contention-free access. Therefore, the number of distinguishable preambles per PRACH RB per subframe is

$$R_{\mathrm{PRACH}} = \frac{d}{n_{\mathrm{PRACH}}} = 9\,. \tag{1.53}$$

PUCCH We denote with n_{PUCCH} the number of RBs dedicated to PUCCH. This quantity must be a multiple of two, because the PUCCH is always instantiated at the opposite sides of the UL bandwidth [19]. Since PUCCH format 1 is dedicated to SRs, the maximum number of UEs that can be accomodated on PUCCH is given by

$$N_{\mathrm{PUCCH}}^{\max} = R_{\mathrm{PUCCH}} \times n_{\mathrm{PUCCH}} \times T_{\mathrm{SR}}\,, \tag{1.54}$$

where R_{PUCCH} is the number of orthogonal codes distinguishable per PUCCH RB and T_{SR} is SR periodicity.

PUSCH The PUSCH resources are used to accomodate both CRs and data SPBs. Since the RB is the smallest resource that can be allocated in LTE, we assume that a CR message occupies exactly one RB, thus n_{CR} RBs are allocated every subframe for CR messages. Parameter n_{CR} should be upper bounded by the maximum number of UL grants that a RAR message can carry in a TTI, i.e., $n_{\mathrm{CR}} \leq 3$, but we relax this constraint assuming that the entire DL bandwidth is dedicated to small packet traffic. The number of RBs for PUSCH is then

$$n_{\mathrm{PUSCH}} = n_{\mathrm{CR}} + M_{\mathrm{D}}^{\mathrm{LTE}} \times J\,, \tag{1.55}$$

where $M_{\mathrm{D}}^{\mathrm{LTE}}$ is the number of data SPBs for LTE. We remark that it is $M_{\mathrm{D}}^{\mathrm{LTE}} \leq M_{\mathrm{D}}$, because the resources on PUSCH must be shared between CRs and data SPBs.

The values of n_{PUCCH}, n_{PUCCH}, and n_{PUSCH} are such that

$$n_{\mathrm{PRACH}} + n_{\mathrm{PUCCH}} + n_{\mathrm{PUSCH}} = n_{\mathrm{RB}} \,. \tag{1.56}$$

4G Radio Access Protocol Model Let us focus on the default RA procedure in LTE (see Figure 1.23a). The preamble collision probability is

$$p_c^{\mathrm{LTE}} = \mathbb{E}_j[q(j,d)] \leq 1 - \left(1 - \frac{1}{d}\right)^{\Delta - 1} \,. \tag{1.57}$$

As done for the Two-Stage approach, we exploit the theory of queues with impatient customers to model the CR step. The arrival rate, service rate, and number of servers in this case are

$$\lambda_A^{\mathrm{LTE}} = \frac{d}{\delta_{\mathrm{RAO}}} \times \omega \left(\frac{\Delta}{d}\right), \; \mu_A = \frac{1}{T_{\mathrm{TTI}}} \text{ , and } m_A = n_{\mathrm{CR}} \,, \tag{1.58}$$

respectively, while the maximum waiting time is equal to the RAR window size, i.e., $\tau_A = W_{\mathrm{RAR}} \times T_{\mathrm{TTI}}$. Therefore, the drop probability of the CR phase is

$$p_d^{\mathrm{CR}} = p_d(\lambda_A^{\mathrm{LTE}}, m_A \mu_A, \tau_A) \,. \tag{1.59}$$

On the other hand, the data transmission takes place only if there are enough resources available. This step can be modeled as a queue with impatient customers, as well. The arrival rate is given by the number of packets that succeeded in getting a grant for the CR, i.e.,

$$\lambda_D = \lambda_S \times \left(1 - p_d^{\mathrm{CR}}\right) \,. \tag{1.60}$$

While the service rate μ_D is as defined in Equation (1.48), the number of servers is $m_D^{\mathrm{LTE}} = M_{\mathrm{D}}^{\mathrm{LTE}}$. The maximum waiting time is given by the SR periodicity, i.e.,

$$\tau_D = T_{\mathrm{SR}} \times T_{\mathrm{TTI}} \,. \tag{1.61}$$

Therefore, the drop probability of the data phase is

$$p_d^D = p_d(\lambda_D, m_D^{\mathrm{LTE}} \mu_D, \tau_D) \,. \tag{1.62}$$

The one-shot failure probability of the overall RA procedure is

$$p_f^{\mathrm{LTE}} = 1 - \left(1 - p_c^{\mathrm{LTE}}\right)\left(1 - p_d^{\mathrm{CR}}\right)\left(1 - p_d^D\right) \tag{1.63}$$

and the outage probability is

$$p_{\mathrm{outage}}^{\mathrm{LTE}} = \left(p_f^{\mathrm{LTE}}\right)^{\Theta} \,. \tag{1.64}$$

The average number of preamble transmission attempts and the aggregate arrival rate can be computed using Equations (1.39) and (1.41). Finally, the throughput of the overall system is defined as

$$\mathcal{S}_{\mathrm{LTE}} = \lambda \times (1 - p_{\mathrm{outage}}^{\mathrm{LTE}}) \,. \tag{1.65}$$

Table 1.10: System parameters for the performance evaluation

Variable	Value
Bandwidth	20 MHz
Orthogonal Frequency Division Multiplexing (OFDM) subcarriers	1200
Subcarrier spacing	15 kHz
T	14
S	12
n_{RB}	100
n_{PUCCH}	4
n_{PRACH}	6
R_{PRACH}	9
R_{PUCCH}	18
n_{CR}	30
$M_{\mathrm{D}}^{\mathrm{LTE}}$	15
J	4
$M = M_{\mathrm{SR}} + M_{\mathrm{D}}$	24
R	9
T_{TTI}	1 ms
δ_{RAO}	1 ms

Enhanced 4G Radio Access Protocol Model The protocol variant in Figure 1.23b [49, Section 5.3.3] consists in sending the SR along with the CR. In the Contention Resolution message, if the process is successful, the eNB indicates the UL resources on which the MTD has to transmit the data packet.

The mathematical model of this radio access protocol is exactly the same as the default RA procedure, but for maximum waiting time at the data transmission stage τ_D. Indeed, in this case it depends on the Contention Resolution timer length $W_{\mathrm{resolution}}$, that is

$$\tau_D = \frac{W_{\mathrm{resolution}}}{8} \times T_{\mathrm{TTI}}\,, \tag{1.66}$$

where we need to divide $W_{\mathrm{resolution}}$ by a factor 8 to account for fixed pattern of the HARQ process involved in the transmission of the CR.

1.7 Performance Evaluation

In this section, the performance of the proposed 5G radio access protocols for IoT traffic is evaluated and compared with the LTE RA procedure. The analytical results will be compared with the computer simulation results. Finally, a discussion of the results is provided.

1.7.1 Performance Metrics and Evaluation Assumptions

The system performance is evaluated in *ideal* conditions, i.e., assuming an error-free channel. Moreover, if two UEs in the same cell use the same resource

Table 1.11: Protocol parameters for the performance evaluation

Variable	Value
Θ	4
A	0 (1-stage) 3 (2-stage) 3 (LTE)
B	0 (1-stage) $A+1=4$ (2-stage) $A+6=9$ (LTE)
C	3 (1-stage) $B+3=7$ (2-stage) $B+8=17$ (LTE)
D	$C+4=21$
E	$D+4=25$
F	$E+4=29$
G	$F+4=33$
$\beta_{\min}$	0 ms
$\beta_{\max}$	10 ms
T_{wake}	0.5 ms
W_{RAR}	1
$W_{\text{resolution}}$	8
T_{SR}	1

(data or SR resource) both transmissions are lost. We will compare the three system performance metrics:

1. the *throughput*;
2. the *outage probability*;
3. and the *average transmission delay*.

In particular, the average transmission delay is defined as the period between the generation of a new data packet and the reception of final ACK. Under the assumption of independent transmission attempts, the delay $\mathcal{D}$ can be computed as

$$\begin{aligned}\mathcal{D} &= \sum_{\theta=1}^{\Theta}[T_{\text{TX}} + (\theta-1)\times(T_{\text{RETX}}+\bar{\beta})]\times\mathbb{P}[\theta \text{ tx attempts}|\text{final ACK}] \\ &= \sum_{\theta=1}^{\Theta}[T_{\text{TX}} + (\theta-1)\times(T_{\text{RETX}}+\bar{\beta})]\times\frac{p_f^{\theta-1}(1-p_f)}{1-p_f^{\Theta}} \\ &= T_{\text{TX}} + (T_{\text{RETX}}+\bar{\beta})\times\frac{p_f-\Theta p_f^{\Theta}+(\Theta-1)p_f^{\Theta+1}}{(1-p_f^{\Theta})(1-p_f)},\end{aligned} \tag{1.67}$$

where T_{TX} is the average delay of a *successful* transmission attempt, T_{RETX} is the average delay of a *unsuccessful* transmission attempt, and $\bar{\beta}=(\beta_{\max}-\beta_{\min})/2$ is the average backoff time between subsequent transmission attempts. The details about the computation of T_{TX} and T_{RETX} can be found in Section 1.7.3.

Table 1.12: Physical channels sizes for the performance evaluation of the LTE radio access protocols, assuming $n_{\mathrm{RB}} = 100$

n_{PRACH}	n_{PUCCH}	4G		Enhanced 4G	
		n_{CR}	M_{D}	n_{CR}	M_{D}
6	4	54	9	54	10
		30	15	30	16
		26	16	26	17

In the following, analytical results and computer simulation results are compared considering the system parameters and the protocol parameters that can be found in Tables 1.10 and 1.11, respectively. It is worth noticing that the PHY layer parameters for LTE and 5G are the same, e.g., $R = R_{\mathrm{PRACH}}$, to provide a fair comparison. Moreover, the number of SPBs M has been obtained subtracting the RBs dedicated to PUCCH from the overall number of RBs, i.e., $M = (n_{\mathrm{RB}} - n_{\mathrm{PUCCH}})/J$.

As for the timing parameters, in LTE the processing time at the MTD side between the reception of the RAR and the CR transmission takes 5 TTIs, i.e, $B = A + 6$. In 5G, instead, the processing time at the MTD side between the reception of the ACK of the SR and the data transmission can be minimized due to the optimized feedback design described in Section 1.5.4, thus we assume that $B = A + 1$. Moreover, a mean waiting time between the MTD wake-up and the beginning of the next TTI T_{wake} of half TTI is considered. We neglect, instead, the possible offsets between UL and DL frame, the propagation time, as well as the time required for the wake-up process of the device and the delay introduced by an additional final ACK from the application layer, which could be quantified in a few additional milliseconds on aggregate. Also, we remark that our definition of delay includes the time between data transmission and reception of the ACK, i.e., the duration $C - B$. However, in practice, as soon as the data is successfully decoded at the eNB, it may be already forwarded to its final destination. Thus, the duration $C - B$ will not extend the overall end-to-end delay of the service.

1.7.2 Pure Protocol Performance

A preliminary study has been made to evaluate the performance of the 4G radio access protocols, comparing the default protocol with the protocol variant. Then, we made four comparative to test the performance of the proposed 5G protocols, varying N, K, W, and R, respectively.

Performance of LTE Radio Access Protocols

We made a first performance evaluation of two 4G radio access protocols in order to optimize the system parameters for the comparison with respect to the 5G solutions. We considered the parameters listed in Table 1.12: note that we assume that in the enhanced 4G mode we can omit the PUCCH, thus increasing the PUSCH size of one extra data SPBs. Moreover, not all the possible values of n_{CR} are allowed. For the protocol variant case, n_{CR} should not exceed the maximum number of preambles d, that is, $n_{\mathrm{CR}} \leq d$. For the default LTE

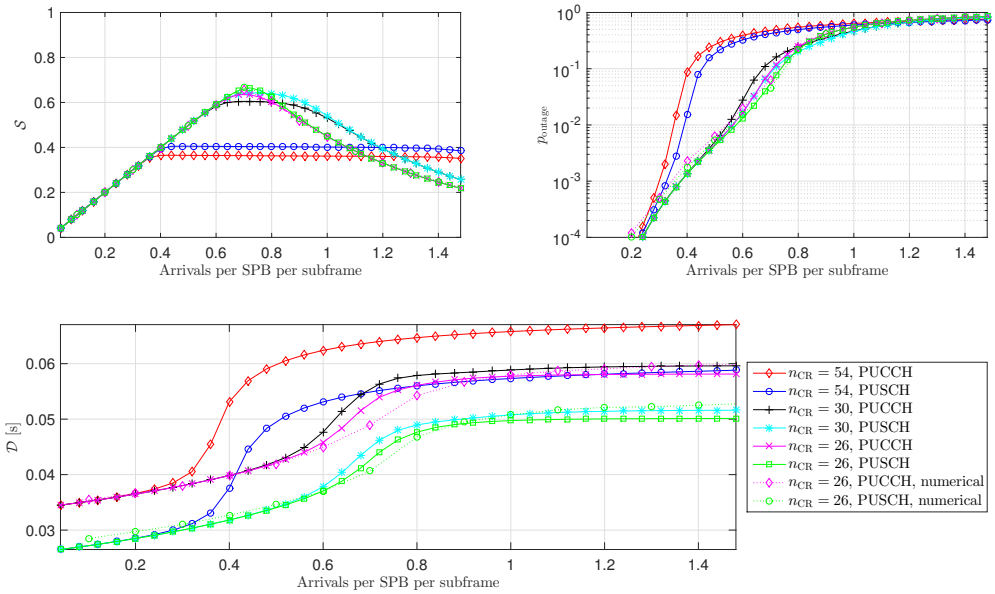


Figure 1.24: Impact of n_{CR} on the default LTE procedure and the protocol variant (denoted with "PUCCH" and "PUSCH" in the legend, respectively) when $n_{\mathrm{PUCCH}} = 4$. The analytical results and the simulation results are represented in solid lines and dashed lines, respectively.

operation, instead, in order not to cause a shortage of PUCCH signatures, we should ensure also that

$$n_{\mathrm{CR}} \times T_{\mathrm{SR}} \leq N_{\mathrm{PUCCH}}^{\max} = R_{\mathrm{PUCCH}} \times n_{\mathrm{PUCCH}} \times T_{\mathrm{SR}} \,. \tag{1.68}$$

Therefore, the possible values of n_{CR} are those that satisfy the following inequality:

$$n_{\mathrm{CR}} \leq \min\{R_{\mathrm{PUCCH}} \times n_{\mathrm{PUCCH}}, d\} \,. \tag{1.69}$$

The graphical results can be found in Figure 1.24; for a pair of curves the comparison between the theoretical results and computer simulation results is provided to assess the correctness of the model. We can observe that the enhanced 4G protocol provides a slightly higher throughput and a faster delivery for successful packets than the default approach. Let us remark that the higher the value of n_{CR}, the lower M_{D} becomes, thus the data transmission phase will be the bottleneck of the system. On the other hand, the lower the value of n_{CR}, the lower the number of MTDs that access the network becomes: the bottleneck of the system will be the second stage, i.e., the CR transmission. Therefore, a tradeoff must be found according to the traffic conditions.

For the sake of a fair comparison with respect to the 5G system, we will consider the default RA protocol (taking into account the control channel), with $n_{\mathrm{CR}} = 30$ (see Table 1.10), since for such value of n_{CR} the system throughput is maximized.

Impact of Over-Provisioning Factor N

The graphical results can be found in Figure 1.25, where we assume $M_{\mathrm{SR}} \in \{1, 2, 4\}$, tagged data resources, i.e., $K = M_{\mathrm{D}}$, and no windowing ($W = 1$).

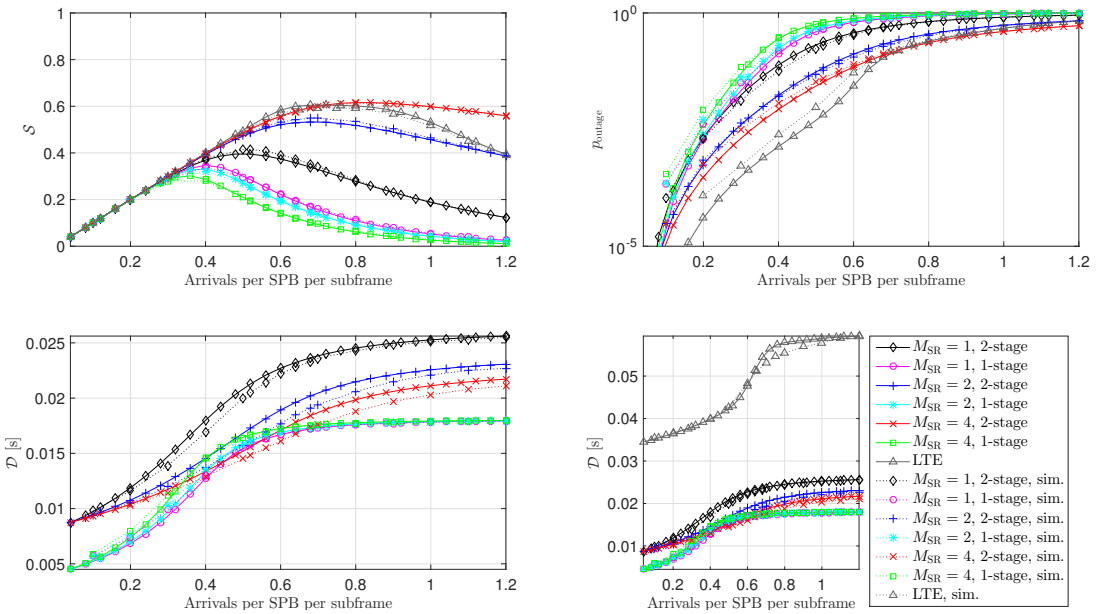


Figure 1.25: Impact of over-provisioning factor N, i.e., the trade-off between SR resources M_{SR} and tagged data resources M_{D}. M_{SR} is varied in $\{1, 2, 4\}$, $K = M_{\text{D}}$ and no windowing is allowed ($W = 1$). The analytical results and the simulation results are represented in solid lines and dashed lines, respectively. LTE parameters according to Tables 1.10 and 1.11.

The solid lines and dashed lines denote the results of the theoretical model and the numerical evaluation, respectively. It can be seen that the Two-Stage protocol with tagged resources outperforms the One-Stage protocol in terms of throughput, failure probability, and outage probability, while the One-Stage protocol provides a faster delivery for successful packets if the arrival rate is sufficiently low. For high arrival rates, indeed, the outage probability of the One-Stage protocol approaches one, meaning that very few packets are successfully delivered. Moreover, as M_{SR} increases, the One-Stage protocol performance is degraded, while the Two-Stage protocol performance improves, as expected. Finally, we want to remark that the theoretical curves and the empirical curves nicely overlap in terms of throughput, failure probability, outage probability, and average number of transmission attempts of successful packets. The greatest difference is in the delay plots, where the gap between the theoretical evaluation and the empirical evaluation in the Two-Stage protocol is due to the assumption of statistical independence between the two stages of the transmission as well as among successive transmission attempts in the theoretical model.

The gains of 5G protocols over LTE mainly regard the delay, because of the additional signaling exchange shown in Figure 1.23a. Please also note that the impact of MUD is not yet included in this analysis. A further increase of throughput is expected if a sophisticated receiver could decode at least one or more packets in spite of a collision. This aspect will be investigated as future work.

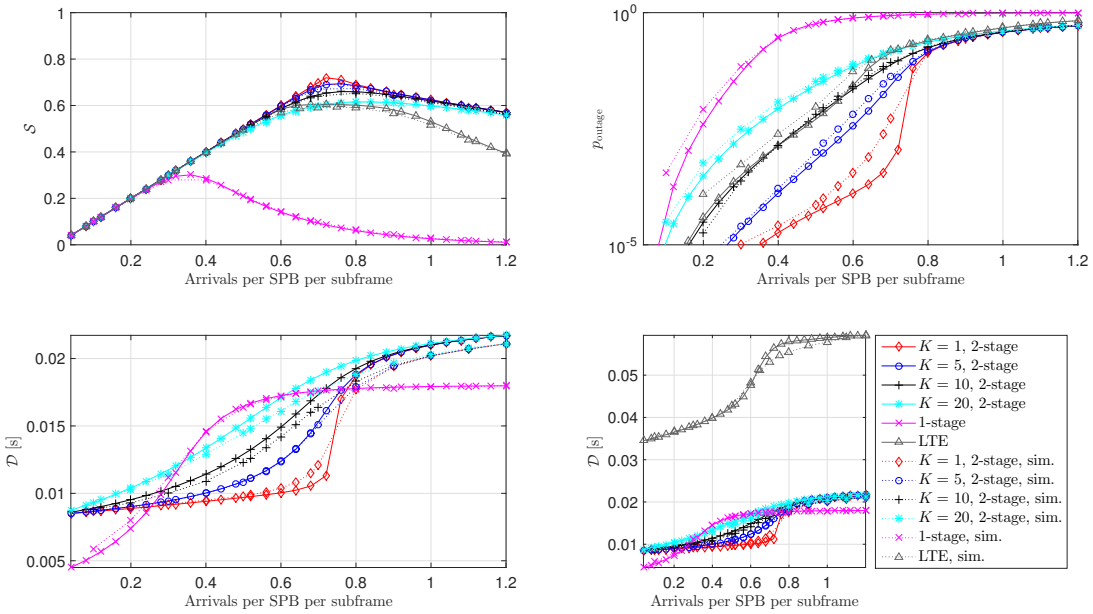


Figure 1.26: Impact of grouping with parameter K for $M_{\text{SR}} = 4$ without windowing ($W = 1$). The analytical results and the simulation results are represented in solid lines and dashed lines, respectively.

Impact of the Number of Groups K

The graphical results can be found in Figure 1.26, where we fix $M_{\text{SR}} = 4$, $W = 1$, and let K vary in the set $\{1, 5, 10, 20\}$. We observe that the increase in the number of groups K degrades the performance of the Two-Stage protocol; therefore, the pooled version is more efficient than the tagged version due to the enhanced flexibility for the assignment of data SPBs at the eNB. However, as already discussed, the benefit of the tagged variant is the smaller DL feedback size. Moreover, grouping may be needed for efficient service differentiation and prioritization.

Impact of Window Size W

The graphical results can be found in Figure 1.27, where it is $M_{\text{SR}} = 4$, $K = M_{\text{D}} = 20$, and $W \in \{1, 2, 4, 8\}$. The time flexibility results to be beneficial if associated with tagged data SPBs. Indeed, it can be seen that an increase in the window size W boosts the performance of the Two-Stage protocol. The time window W does not provide an additional benefit if the pooled variant of the Two-Stage protocol is applied. The reason is that the potential of increased flexibility is already fully exploited in frequency direction as explained above. Thus, tagged resources combined with time windowing (see Figure 1.27, $W = 8$) can be seen as equivalent solution to pooled resources (see Figure 1.26, $K = 1$).

Impact of PHY

Finally, we investigate the impact of the number of detectable preamble sequences per RB R. An increase of R is equivalent to a higher over-provisioning

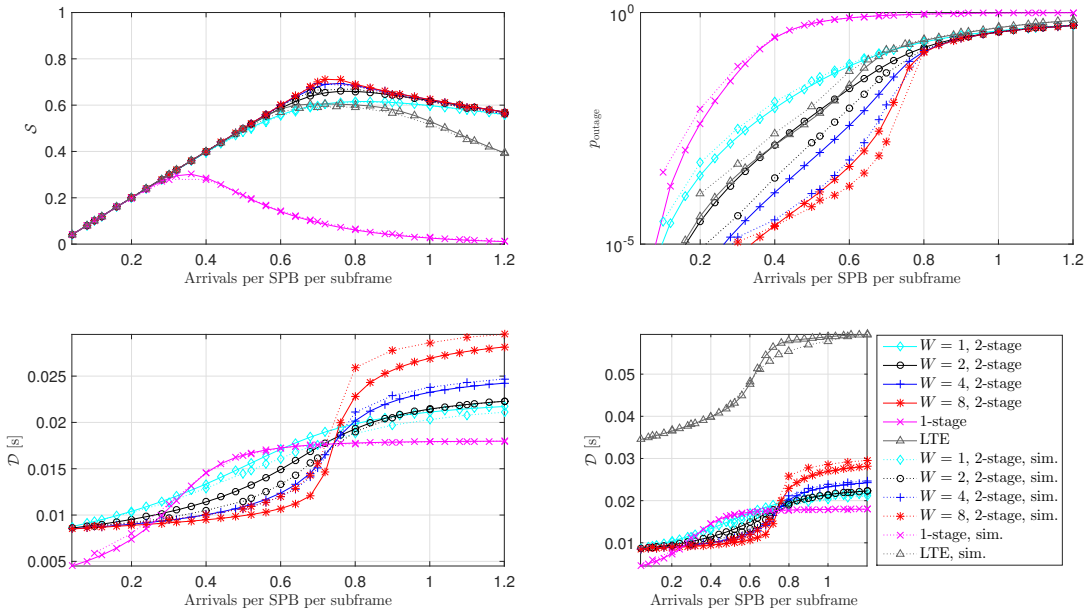


Figure 1.27: Impact of windowing with parameter W for $M_{\text{SR}} = 4$ and $K = M_{\text{D}}$. The analytical results and the simulation results are represented in solid lines and dashed lines, respectively.

factor N, however without the need to reserve a larger portion of the radio resources for SRs, i.e., in contrast to Figure 1.25 we keep the values for M_{SR} and M_{D} constant. With a novel preamble sequence design like the one introduced in [89] at least a duplication of the number of preamble sequences can be achieved, i.e., $R = 18$ instead of $R = 9$. The comparison is shown in Figure 1.28. Obviously, all performance metrics clearly benefit from the higher number of preambles due to a significantly smaller collision probability. The dynamic increase of the over-provisioning factor N by switching from $R = 9$ to $R = 18$ is an important means to mitigate a congestion through sporadic massive access of arrivals. The drawback, however, is a slightly increased probability for missed detections and false alarms of SRs. We remind that we assume perfect preamble detection capabilities throughout this chapter.

1.7.3 On the Delay Computation

In this section the computation of the average delay of a successful transmission attempt, denoted with T_{TX}, and of a failed transmission attempt, denoted with T_{RETX}, is provided.

One-Stage Protocol

In the One-Stage protocol T_{TX} is simply given by

$$T_{\text{TX}} = T_{\text{wake}} + [(i + C) - i + 1] \times T_{\text{TTI}} = T_{\text{wake}} + (C + 1) \times T_{\text{TTI}} \qquad (1.70)$$

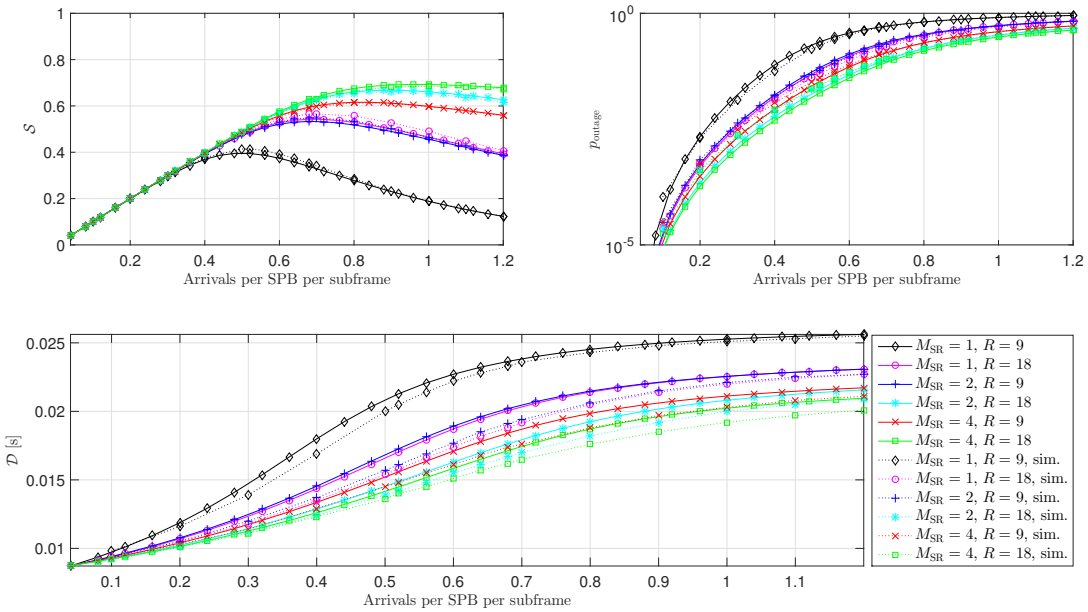


Figure 1.28: Impact of PHY design, according to parameter R, for $M_{\text{SR}} = \{1, 2, 4\}$, $K = M_{\text{D}}$, and $W = 1$

and T_{RETX} is equal to the average transmission time without the wake-up time, i.e.,

$$T_{\text{RETX}} = T_{\text{TX}} - T_{\text{wake}} = (C + 1) \times T_{\text{TTI}} \,. \tag{1.71}$$

Two-Stage Protocol

In the Two-Stage protocol without windowing, i.e., $W = 1$, T_{TX} is expressed as in Equation (1.70). For window sizes W such that $W > 1$, instead, we must account for the average delay introduced by the window W. This can be done exploiting the theory of queues with impatient customers [90]. The relationship between queue dropping probability p_d, worst case average wait time τ, and average waiting time $\bar{W}_{\text{wait}}$ is defined as

$$p_d = \frac{\bar{W}_{\text{wait}}}{\tau} \,, \tag{1.72}$$

thus the average waiting time is computed as

$$\bar{W}_{\text{wait}} = p_d \times \tau \,. \tag{1.73}$$

In the case of the Two-Stage approach, we have to plug in the values of p_d^A in Equation (1.50) and τ_q.

The successful transmission interval duration, then, can be derived as

$$T_{\text{TX}} = T_{\text{wake}} + (C + \bar{W}_{\text{wait}} + 1) \times T_{\text{TTI}} \,, \tag{1.74}$$

while T_{RETX} is obtained averaging between the delays introduced if a failure occurs after the preamble transmission or after the data transmission, i.e.,

$$T_{\text{RETX}} = p_d^A \times (B + 1) \times T_{\text{TTI}} + \left(1 - p_d^A\right) \times (T_{\text{TX}} - T_{\text{wake}}) \,. \tag{1.75}$$

4G

In order to make the packet transmission as fast as possible, we set the LTE protocol parameters to their minimum values, i.e., $W_{\mathrm{RAR}} = 1$, $W_{\mathrm{resolution}} = 8$, and $T_{\mathrm{SR}} = 1$, as stated in Table 1.11. As a consequence, no time flexibility is allowed. The successful transmission attempt duration is

$$T_{\mathrm{TX}} = T_{\mathrm{wake}} + (G+1) \times T_{\mathrm{TTI}}\,, \tag{1.76}$$

while the retransmission time is

$$T_{\mathrm{RETX}} = [p_d^A \times (A+1) + (1-p_d^A) \times (C+1)] \times T_{\mathrm{TTI}}\,. \tag{1.77}$$

Enhanced 4G

The successful transmission duration for the protocol variant is shorter, since

$$T_{\mathrm{TX}} = T_{\mathrm{wake}} + (E+1) \times \mathrm{TTI}\,, \tag{1.78}$$

while T_{RETX} is defined as in Equation (1.77).

1.8 Conclusions and Ways Forward

In this chapter, we have addressed the issue of massive M2M radio access in cellular networks. We first identified the features of IoT traffic, highlighting the differences with respect to the conventional H2H traffic, and surveyed the M2M traffic models proposed by the research community and the standardization bodies. Then, we showed by means of a simplified framework as well as extensive simulation campaigns that the 4G cellular network reveals to be inefficient if put under strain by a huge amount of terminals, due to the very high signaling. A lot of solutions to this issue problem have been proposed in literature, but most of them are not of practical use in real network architectures.

Therefore, we proposed a novel radio access protocol for 5G systems to support sporadic small UL data traffic, aiming at minimal signaling overhead and scalability with respect to the number of IoT devices per radio cell. The main characteristic of our proposals consists in the transmission of data already in the first or in the second stage of the communication: in contrast to LTE, the two protocols are *contention-based* and eventually *connectionless*, i.e., there is no collision resolution mechanism and connection setup and release are not required before and after data transmission.

We derived a mathematical model of both 4G and 5G radio access protocols and compared them in terms of throughput, outage probability, and delivery delay; these models have been validated by system-level simulations. The envisioned solutions for 5G provide substantial advantages with respect to LTE, especially in the reduction of the latency and signaling overhead.

As for future work on this topic, we are planning to exploit MUD techniques in the protocol design and investigate their impact on the system performance [181].

As a closing remark on the massive access topic, let us observe that, in addition to innovative radio access protocol designs, other approaches and technologies for 5G may help in supporting M2M services. In the following, we will quickly comment on some of these techniques.

Massive MIMO This technique consists in equipping the BS with much more antennas than the number of cheap single-antenna devices: in this way, the channels to different devices would be quasi-orthogonal, thus making it possible to increase the spectral efficiency by using simple spatial multiplexing/demultiplexing procedures [91, 92]. Therefore, massive Multiple-Input-Multiple-Output (MIMO) can dramatically enlarge the number of simultaneous transmissions that can be successfully received by a (powerful) BS, without burdening the peripheral nodes. These characteristics, in principle, make massive MIMO attractive for supporting MTC. The limit of such approach is that enabling *massive* MIMO for a *massive* number of MTDs may require an exceedingly large number of antennas at the BS, which can be practically infeasible. Furthermore, despite the huge interest in massive MIMO, there is still much to be learned, in particular regarding the propagation and cost effectiveness, so that the actual performance and feasibility of this technique are still under investigation.

Small Cells One possible solution for dealing with the increase of the number of devices in hot spot areas is the densification of the network by employing small cells [93]. Small cells are indeed deployed to reduce the distance between devices and access points, thus enabling higher bit rates (or lower transmit power and interference), while also improving the spatial reuse. In an M2M setting, however, the focus is not on high transmit rate, but rather on reliable and ubiquitous connectivity. MTDs are in fact expected to be spread across wide areas, also where human-generated broadband traffic may be light, e.g., along highways/road/railroads or in agricultural areas. Hence, providing access to MTDs will require uniform and ubiquitous coverage, which is not economically sustainable by using microcells. Even in case of a high concentration of MTDs in relatively small areas, the Average Revenue Per User (ARPU) of MTD-based services is likely lower compared to conventional services, thus not justifying the deployment of small cells for the sake of MTD-coverage only. Finally, the densification of the network does not impact the signaling overhead at the PHY layer, which is inefficient due to the MTC characteristics.

mmWave Communication After years of striving to squeeze more spectral efficiency from the crowded bandwidth used by current microwave cellular systems, the huge bandwidth available at mmWave frequencies, from 3 to 300 GHz, represents an irresistible attraction for 5G systems. Although the signal propagation at these frequencies is not yet thoroughly understood, the measurements reported in [94] indicate that transmission can occur even in the absence of line of sight, though with a much higher path loss exponent. In combination with large antenna arrays, mmWave communication can make it possible to reach huge bitrates over short distances. However, the sensitivity to blockage, the rapid power decay with distance, and the higher power requirements of mmWave communications make this technology less attractive for MTDs that, instead, need long-range, low-power, and low bitrate connections.

Virtualization Software Defined Networking (SDN) and Network Function Virtualization (NFV) are two emerging paradigms that basically consist in abstracting low-level network functionalities to enable a much more

flexible management of the network resources and a better and adaptive support of different types of services [95]. The accomplishment of these concepts would make it possible to differentiate the services offered to the different traffic flows and to dynamically instantiate network elements where and when needed. Ideally, these mechanisms shall deliver the illusion of "infinite capacity," giving to each application exactly the resources it needs to achieve the desired Quality of Experience (QoE). This vision is extremely appealing for what regards the support of massive M2M traffic, in that SDN can naturally provide separation between M2M and H2H traffic, while guaranteeing the desired QoS levels to each type of flow (both at the access network and across the core network). Moreover, the fine-grained and per-flow resource allocation paradigm enabled by SDN will result in a better utilization of the network resources, thus contributing to alleviate the massive access problem. NFV, on the other hand, can be used to dynamically shape the network architecture according to the traffic requirements. For example, NFV can instruct network elements in a certain area to act as concentrators to collect MTDs data, or as relays to extend the coverage range, or even as additional BSs to satisfy temporary peaks of access requests. While this virtualization principle can bring a disruptive change in the architectural design of next generation communication systems, major research efforts are still required to turn this vision into reality.

Chapter 2

Internet of Things in Unlicensed Bands

The standardization process of the upcoming fifth-generation (5G) standard is proceeding at a fast pace, however, the market growth of Internet of Things (IoT) is increasing at a way faster pace. This exponential growth opens the ground to a wide range of new technologies providing ready-to-use solutions for companies and individuals who want to exploit the potential of IoT as soon as possible. This group of technologies is commonly referred to as Low-Power Wide Area Networks (LPWANs). In this chapter, we will first give an introductory overview about LPWANs on unlicensed bands at large, then we will focus on one of the most prominent these technologies: Long-Range^{TM} (LoRa).

The rest of the chapter is organized as follows. In Section 2.1, we introduce the communication paradigm of LPWANs, highlighting the differences with respect to short-range wireless solutions as well as cellular systems. The most important LPWAN technologies are surveyed in Section 2.2, with special reference to LoRa, whose performance is first assessed with some experimental results in Section 2.3. Then, by means of simulation campaigns, we evaluate the performance of LoRa in large-scale IoT networks, focusing both on the uplink (UL) capacity (see Section 2.4) and downlink (DL) capacity (see Section 2.5). Finally, in Section 2.6 we draw the conclusions about this topic and outline possible future extensions.

2.1 A New Paradigm: Long-Range IoT Communications in Unlicensed Bands

Most LPWANs operate in the unlicensed ISM bands centered at 2.4 GHz, 868/915 MHz, 433 MHz, and 169 MHz, depending on the region of operation.[1] The radio emitters operating in these frequency bands are commonly referred to as "Short Range Devices," a rather generic term that suggests the idea of coverage

[1] A further set of bands that is suitable for the implementation of LPWANs is the TV White Space (TVWS) spectrum. These frequencies are made available for unlicensed users when the spectrum in not being used by licensed services. The most prominent LPWAN technology that jointly exploits ISM bands and TVWS bands is Weightless (`www.weightless.org`).

ranges of few meters, which was indeed the case for the previous ISM wireless systems. Nonetheless, the ERC Recommendation 70-03 [96] specifies that *"The term* Short Range Device *(SRD) is intended to cover the radio transmitters which provide either uni-directional or bi-directional communication which have low capability of causing interference to other radio equipment."* Therefore, there is no explicit mention of the actual coverage range of such technologies, but only of the interference caused.

LPWAN solutions are indeed examples of "short range devices" with cellular-like coverage ranges, in the order of 10-15 km in rural areas, and 2-5 km in urban areas. This is possible thanks to a radically new physical layer design, aimed at very high receiver sensitivity. For example, while the nominal sensitivity of ZigBee^{TM}and Bluetooth receivers is about -125 dBm and -90 dBm, respectively, the typical sensitivity of a LPWAN receiver is around -150 dBm (see Section 2.2).

The downside of these long-range connections is the low data rate, which usually ranges from few hundred to few thousand bit/s, significantly lower than the bitrates supported by the actual "short-range" technologies, e.g., 250 kbit/s in ZigBee^{TM}and 1-2 Mbit/s in Bluetooth. However, because of the signaling overhead and the multi-hop packet forwarding method, the actual flow-level throughput provided by such short-range technologies may actually be significantly lower than the nominal link-level bitrate. For example, in [97] it is reported that a 6LoWPAN network based on a mesh topology using an IEEE 802.15.4 physical layer, with a nominal link level bit-rate of 250 kbit/s, reaches a unicast throughput of about 0.8 kbit/s and a multicast throughput lower than 1.5 kbit/s.

While such low bitrates are clearly unsatisfactory for most common data-hungry network applications, many Smart City and IoT services are expected to generate a completely different pattern of traffic, characterized by sporadic and intermittent transmissions of very small packets, typically in the order of few hundred bytes, for monitoring and metering applications, remote switching or control of equipment, and so on [28]. Furthermore, many of these applications are rather tolerant to delays and packet losses and, hence, are suitable for the connectivity service provided by LPWANs.

Another important characteristic of LPWANs is that the *things*, i.e., the end devices, are connected directly to one (or more) gateway with a single-hop link, very similar to a classic cellular network topology. This greatly simplifies the coverage of large areas, even nation-wide, by re-using the existing infrastructure of the cellular networks. For example, LoRa systems are being deployed by telecommunication operators like Orange and Bouygues Telecom in France, by Swisscom in Switzerland, and by KPN in the Netherlands, while Sigfox has already deployed a nation-wide access network for M2M and IoT devices in many central European countries, from Portugal to France. Furthermore, the star topology of LPWANs makes it possible to have greater control on the connection latency, thus potentially enabling the support of interactive applications that require predictable response times such as, for example, the remote control of street lights in a large city, the operation of barriers to limited-access streets, the intelligent control of traffic lights, and so on. Besides the access network, the similarity between LPWANs and legacy cellular systems further extends to the bridging of the technology-specific wireless access to the IP-based packet switching core network.

Therefore, LPWANs inherit the basic aspects of the legacy cellular systems

architecture that, however, is stripped off its most advanced features, such as the management of user mobility and resource scheduling. The combination of the simple but effective topology of cellular systems with a much lighter management plane makes the LPWAN approach particularly suitable to support services with relatively low Average Revenue Per User, such as those envisioned in Smart City scenarios.

A clear evidence of the appeal of LPWAN technologies in the IoT arena is given by the ever increasing number of products and applications that rely on these technologies for communication. For example, Sensing Labs[2] produces sensors for telemetry and metering to enable smart building applications. Enevo[3] uses wireless devices to monitor the fill-level of waste containers. Sayme[4] provides a street lighting remote management system that increases energy efficiency and reduces maintenance expenses. Turbo Technologies[5] designed a wireless geomagnetic detector for smart parking purposes. Elmar[6] is implementing a smart grid network across the entire island of Aruba. Finally, Mueller Systems[7] developed a communication network that fully automates the management of water resources.

2.2 A Review of LPWAN Technologies

In the following, we quickly overview four of today's most prominent technologies for LPWANs, namely DASH7^{TM}, Sigfox^{TM}, Ingenu^{TM}, and LoRa.[8] In particular, we will describe in greater detail the LoRa technology, which is gaining more and more momentum, and whose specifications are publicly available, thus making it possible to appreciate some of the technical choices that characterize LPWAN solutions.

2.2.1 Dash7

DASH7^{TM} is one of the very first LPWAN technologies [99, 100], and it paved the way to a number of LPWAN solutions. In contrast to the other three technologies, which are proprietary solutions, Dash7 originates from the ISO/IEC 18000-7 standard for active Radio-Frequency IDentification (RFID) tags [101] on the Industrial, Scientific, and Medical (ISM) 433 MHz band.

The DASH7 Alliance[9] is in charge of maintaining the technology, which, other than the band aroung 433 MHz, supports other two, sub-GHz frequency bands (868 and 915 MHz) using a 2-Gaussian Frequency Shift Keying (GFSK) modulation scheme. The communication is bi-directional and the default network topology is a tree topology, where the gateway is the root of the network and the endpoint devices are the leaves, and a third type of devices called "sub-controllers" can be used to extend the network coverage. A star topology can also be implemented, as well as tag-to-tag communication.

[2] `http://sensing-labs.com`
[3] `http://www.enevo.com`
[4] `http://www.sayme.es`
[5] `http://www.turboes.com/english/`
[6] `https://www.elmar.aw`
[7] `http://www.muellersystems.com`
[8] For a exhaustive survey on LPWANs, we invite the reader to refer to [98].
[9] `www.dash7-alliance.org`

Security is also addressed in DASH7, and is similar to the security of IEEE 802.15.4 standard, using AES-CBC for authentication and AES-CCM for authentication and encryption.

2.2.2 Sigfox

Sigfox,[10] the first proprietary LPWAN technology proposed in the IoT market, was founded in 2009 and has been growing very fast since then, recently becoming part of the European Telecommunications Standards Institute (ETSI) specifications for Low Throughput Networks (LTNs).

The Sigfox physical layer employs an Ultra-Narrow Band (UNB) wireless modulation. The first releases of the technology only supported uni-directional UL communication, i.e., from the device towards the aggregator; however, bi-directional communication is now supported. Sigfox claims that each gateway can handle up to a million connected objects, with a coverage area of 30-50 km in rural areas and 3-10 km in urban areas.

Regarding the security aspects of Sigfox networks, as a general approach, Sigfox focuses on the network security itself, leaving the payload security mechanisms to the end users, both at the transmitting side, i.e., the Sigfox node, and at the receiving side, i.e., the applications linked to the Sigfox Cloud via Application Programming Interfaces (APIs) or callback functions.

2.2.3 Ingenu

An emerging star in the landscape of LPWANs is Ingenu, a trademark of On-Ramp Wireless, a company headquartered in San Diego (USA).[11] On-Ramp Wireless has been pioneering the IEEE 802.15.4k standard [102]. The company developed and owns the rights of the patented technology called Random Phase Multiple Access (RPMA®) [103], which is deployed in different networks. Conversely to the other LPWAN solutions, this technology works in the 2.4 GHz band but, thanks to a robust physical layer design, can still operate over long-range wireless links and under the most challenging RF environments.

From a security point of view, the RPMA technology offers six state-of-the-art guarantees: (i) mutual authentication; (ii) message integrity and replay protection; (iii) message confidentiality; (iv) device anonymity; (v) authentic firmware upgrades, and (vi) secure multicasts.

2.2.4 The LoRa System

LoRa is a new physical layer LPWAN solution, which has been designed and patented by Semtech Corporation that also manufactures the chipsets [104].

LoRa PHY

With the term "LoRa," we specifically refer to the proprietary Physical Layer (PHY) modulation, which is a derivative of Chirp Spread Spectrum (CSS) [105]. Traditional CSS has been innovated by Semtech Corporation in order to ensure the phase continuity between different chirp symbols in the preamble part of

[10] `http://www.sigfox.com`
[11] `http://www.onrampwireless.com`

the physical layer packet, thus enabling a simpler and more accurate timing and frequency synchronization, without requiring expensive components to generate a stable local clock in the LoRa node.

The technology employs a spreading technique, according to which a symbol is encoded in a *longer sequence of bits*, thus reducing the signal to noise and interference ratio required at the receiver for correct reception, without changing the frequency bandwidth of the wireless signal. The length of the spreading code is equal to $2^{\mathtt{SF}}$, where $\mathtt{SF}$ is the value of a tunable parameter called Spreading Factor (SF) in the LoRa jargon, that can be varied from 7 up to 12, thus making it possible to provide variable data rates, giving the possibility to trade throughput for coverage range, or link robustness, or energy consumption [106]. We remark that the higher the SF, the longer and more reliable the packet transmission will be. The raw bitrate of LoRa can be expressed as [106]

$$R_b = \frac{1}{T_b} = \frac{4}{4+\mathtt{CR}} \times \frac{1}{T_c} = \mathtt{SF} \times \frac{4}{4+\mathtt{CR}} \times \frac{1}{T_s} = \mathtt{SF} \times \frac{\frac{4}{4+\mathtt{CR}}}{\frac{2^{\mathtt{SF}}}{B}}\,, \qquad (2.1)$$

where $\mathtt{CR}$ is the channel coding rate, and $B = 125$ kHz is the typical LoRa bandwidth. Note that $T_s = 2^{\mathtt{SF}}/B$ is the LoRa symbol time.

The system works mainly in the 902-928 MHz band in the US and in the 863-870 MHz band in Europe, but can also operate in the lower ISM bands at 433 MHz and 169 MHz. According to the regulation in [107], the radio emitters are required to adopt duty cycled transmission (0.1%, 1%, or 10%, depending on the sub-band), or the so-called Listen-Before-Talk (LBT) Adaptive Frequency Agility (AFA) technique, a sort of carrier sense mechanism used to prevent severe interference among devices operating in the same band. LoRa (as well as Sigfox) mainly uses the duty-cycled transmission option [108], which limits the rate at which the end device can actually generate messages. However, by supporting multiple channels, LoRa makes it possible for an end node to engage in longer data exchange procedures by changing carrier frequency, while respecting the duty cycle limit in each channel.

Three different categories of ISM sub-bands are distinguished in Europe (see [96]):

1. h1.4 (868-868.6 MHz), with maximum 36 seconds per hour Time on Air (ToA), 1% duty cycle to be shared between all sub-channels in each sub-band, and a Effective Radiated Power (ERP) limit of 14 dBm;

2. h1.5 (868.7-869.2 MHz), with maximum 3.6 seconds per hour ToA, 0.1% duty cycle, and a ERP limit of 14 dBm;

3. h1.6: (869.4-869.65 MHz) with maximum 360 seconds per hour ToA, 10% duty cycle, and a ERP limit of 27 dBm.

Table 2.1 proposes a lineup of 6 LoRa channels according to ETSI constraints on European ISM bands. We want to remark that each end device is allowed to transmit on channels belonging to different sub-bands in order to increase the aggregate ToA as long as the duty cycle limit in each sub-band is respected.

Note that channel #6 falls in the h1.6 band, for which a 10%-duty cycled transmission and a much higher transmit power (27 dBm vs the standard 14 dBm) are allowed. Therefore, this channel can be exploited for communications of longer messages over larger distances.

Table 2.1: Channel lineup for LoRa according to ETSI regulations (t_i is the ToA in channel i)

#	Carrier f	B	Time per hour	ToA	Max ERP	Regime
1	868.1 MHz	125 kHz	$t_1 + t_2 + t_3 \leq 36$	1%	25 mW (14 dBm)	h1.4
2	868.3 MHz	125 kHz	$t_1 + t_2 + t_3 \leq 36$	1%	25 mW (14 dBm)	h1.4
3	868.5 MHz	125 kHz	$t_1 + t_2 + t_3 \leq 36$	1%	25 mW (14 dBm)	h1.4
4	868.85 MHz	125 kHz	$t_4 + t_5 \leq 3.6$	0.1%	25 mW (14 dBm)	h1.5
5	869.05 MHz	125 kHz	$t_4 + t_5 \leq 3.6$	0.1%	25 mW (14 dBm)	h1.5
6	869.525 MHz	125 kHz	$t_6 \leq 360$	10%	500 mW (27 dBm)	h1.6

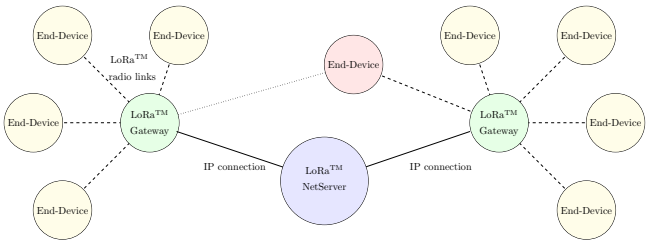


Figure 2.1: LoRa system architecture

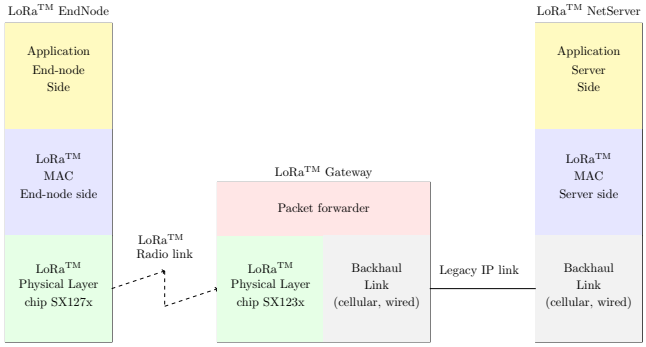


Figure 2.2: LoRa protocol architecture

LoRaWAN™

While the PHY layer of LoRa is proprietary, the rest of the protocol stack, known as Long-Range Wide Area Network™ (LoRaWAN), is kept open, and its development is carried out by the LoRa Alliance,[12] led by IBM, Actility, Semtech, and Microchip.

As exemplified in Figure 2.1, the LoRa network is typically laid out in a *star-of-stars* topology, where the end devices are connected via a single-hop LoRa link to one or many gateways that, in turn, are connected to a common Network Server (NetServer) via standard IP protocols.

The gateways relay messages between the end devices and the Network Server (NS) according to the protocol architecture represented in Figure 2.2.

[12] `https://www.lora-alliance.org/`

Unlike in standard cellular network systems, however, the end devices are not required to associate to a certain gateway to get access to the network, but only to the NS. The gateways act as a sort of relay/bridge and simply forward to their associated NS all successfully decoded messages sent by any end device, after adding some information regarding the quality of the reception. The NetServer is hence in charge of filtering duplicate and unwanted packets, and of replying to the end devices by choosing one of the in-range gateways, according to some criterion (e.g., best radio connectivity). The gateways are thus totally transparent to the end devices, which are logically connected directly to the NS. Note that current full-fledged LoRa gateways allow the parallel processing of up to 9 LoRa channels, where a channel is identified by a specific sub-band and SF.

This access mode greatly simplifies the management of the network access for the end nodes, moving all the complexity to the NS. Furthermore, the end nodes can freely move across cells served by different gateways without generating any additional signaling traffic in the access network, nor in the core network. Finally, we observe that increasing the number of gateways that serve a certain end device will increase the reliability of its connection to the NS, which may be interesting for critical applications.

LoRa device classes

A distinguishing feature of the LoRa network is that it envisages three classes of end devices, named *Class A* (for *All*), *Class B* (for *Beacon*) and *Class C* (for *Continuously listening*), each associated to a different operating mode [108].

Class A defines the default functional mode of LoRa networks, and must be mandatorily supported by all LoRa end devices. Class-A end devices initiate UL transmissions in a totally asynchronous manner. As shown in Figure 2.3, after each UL transmission, the end device will open (at least) two reception windows, waiting for any command or data packet returned by the NS. The first window is opened on the same channel as the UL frame, while the second window is opened on a different sub-band (previously agreed upon with the NS) in order to increase the resilience against channel fluctuations. End devices of Class A are mainly intended for monitoring applications, where the data produced by the end devices have to be collected by a control station.

Class B has been introduced to decouple UL and DL transmissions. Class-B end devices, indeed, synchronize with the NS by means of beacon packets which are broadcast by Class-B gateways and can hence receive DL data or command packets in specific time windows, irrespective of the UL traffic. Therefore, Class B is intended for end devices that need to receive commands from a remote controller, e.g., switches or actuators.

Class C is defined for end devices without (strict) energy constraints (e.g., connected to the power grid), which can hence keep the receive window always open (except when transmitting).

LoRa MAC

According to the LoRaWAN specifications [108], the Medium Access Control (MAC) protocol is basically an ALOHA protocol. A description of the protocol

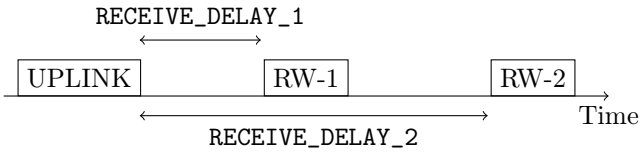


Figure 2.3: Class-A MAC protocol in LoRaWAN. The UL packet transmission starts in a totally asynchronous manner. After the end of the UL phase, two receive windows (denoted as "RW-1" and "RW-2") are opened after a delay of `RECEIVE_DELAY_1` and `RECEIVE_DELAY_2`, respectively.

can be found in [108]. Overall, the LoRa MAC has been designed attempting to mimic as much as possible the interface of the IEEE 802.15.4 MAC towards the higher layers. The objective is to simplify the accommodation, on top of the LoRa MAC, of the major protocols now running on top of the IEEE 802.15.4 MAC, such as 6LoWPAN and CoAP. A clear analogy is the *authentication* mechanism, which is taken directly from the IEEE 802.15.4 standard using the 4-octet Message Integrity Code.

LoRa IP Connectivity

LoRaWAN employs the IEEE 64-bit Extended Unique Identifier (EUI) to automatically associate IPv6 addresses to the LoRa nodes. Therefore, IPv6 and 6LoWPAN protocols can be deployed on LoRaWAN networks, thus enabling transparent interoperability with the IP-based world.

Security in LoRa

Security aspects are taken into account in the LoRaWAN specifications as well [108]. Several layers of encryption are employed, using (i) a Unique Network key to ensure security at the network layer; (ii) a Unique Application key to ensure end-to-end security at the application layer, and finally (iii) a Device-specific key to secure the joining of a node to the network.

In Table 2.2, a comparison between the aforementioned LPWAN radio technologies in terms of coverage range, frequency bands, data rate is provided.

Table 2.2: Comparison between LPWAN radio technologies

	Dash7	Sigfox	Ingenu	LoRa
Coverage range (km)	< 5	rural: 30-50 urban: 3-10	≈ 15	rural: 10-15 urban: 3–5
Frequency bands (MHz)	various, sub-GHz	868 or 902	2400	various, sub-GHz
ISM band	✓	✓	✓	✓
Bi-directional link	✓	✓	✗	✓
Data rate (kbps)	9.6-167	0.1	0.01-8	0.3-37.5
Nodes per gateway	Unknown	$\approx 10^6$	$\approx 10^4$	$\approx 10^4$

2.3 Some Experimental Results

In this section, we corroborate the arguments of the previous sections by reporting some observations based on some initial deployments of LoRa networks.

2.3.1 A LoRa Deployment Test

A LoRa private network has been installed by Patavina Technologies s.r.l. in a large and tall building (19 floors) in Northern Italy, for a proof of concept of the capabilities of the LoRa network. The objective is to monitor and control the temperature and the humidity of the different rooms, with the aim of reducing the costs related to heating, ventilation, and air conditioning. To this end, different wireless and wired communication technologies (including powerline communication) had been tried, but these solutions were mostly unsatisfactory, requiring the installation of repeaters and gateways in basically every floor to guarantee mesh connectivity and access to the IP backbone. Instead, the LoRa technology has made it possible to provide the service by installing a single gateway on the ninth floor and placing 32 nodes all over the building, at least one per floor. The installation included the integration of the NS with a monitoring application and with the databases already in use. At the time of writing, the installation has been flawlessly running for two years and is being considered as the preferred technology for the actual implementation of the energy saving program in many other buildings.

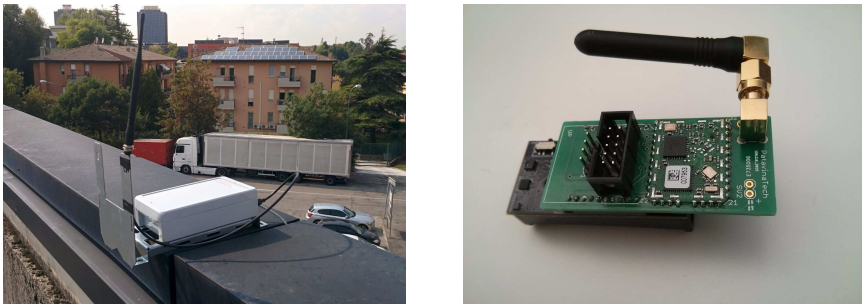
We want to remark that the LoRa network connectivity has been put under strain placing the nodes in elevators and in other places known to be challenging for radio connectivity. All the stress tests have been successfully passed. The envisioned next step is to install a gateway on an elevated site to serve multiple buildings in the neighborhood.

This proof of concept is particularly relevant as it provides, on the one side, interesting insights on how pertinent and practical the LPWAN paradigm is for a Smart City scenario and, on the other side, some intuition from the economical point of view. Indeed, though extremely limited in its extent, the positive experience gained in the proof-of-concept installation of the LoRa system in a building bodes well to the extension of the service to other public and private buildings, realizing at the same time an infrastructure for other Smart City services. According to Analysis Mason 2014 data, the number of LPWAN smart building connections is projected to be 0.8 billions by 2023 [109] and, according to the McKinsey Global Institute analysis, the potential economic impact of IoT in 2025 for Homes and Cities is between $1.1 and $2.0 trillion [110]. Thus, LPWAN solutions appear to have both the technical and the commercial capabilities to become the game changer in the Smart City scenario.

2.3.2 LoRa Coverage Analysis

One of the most debated aspects of LPWAN is the actual coverage range. This is crucial for a correct estimation of the costs for city-wide coverage, which clearly have an important impact on the Capital Expenditure of the service providers.

To gain insight in this respect, we carried out a coverage experimental test of LoRa networks in the city of Padova, Italy. The deployment area consists



(a) **(b)**

Figure 2.4: Experimental setup to assess LoRa coverage. Figure 2.4a depicts the gateway, i.e., a Kerlink LoRa IoT station model 0X80400AC (www.kerlink.fr), while Figure 2.4b shows a Patavina Technologies node mounting an IMST iM880A-L LoRa module.

in an urban environment, which resembles a typical commercial area of a big Italian city, crossed by a busy six-lane, two-way street, with office buildings and shopping malls (up to 5-6 floors) on both sides, and intersections with secondary roads regulated by traffic lights and roundabouts. The aim of the experiment was to assess the "worst case" coverage of the technology, to have a conservative estimate of the number of gateways required to cover the whole city. To this end, we placed a gateway with no antenna gain at the the top of a two storey building, without antenna elevation, in an area where taller buildings are present.

Figure 2.4 shows the experimental setup, while Figure 2.5 shows the results of the test. It can be seen that, in such harsh propagation conditions, the LoRa technology was able to cover a cell of about 2 km of radius. However, the connection at the cell edge is guaranteed only when using the lowest bit rate (i.e., the longest spreading sequence, which provides maximum robustness), with low margin for possible interference or link budget changes. For this reason, we assumed a nominal coverage range of 1.2 km, a value that ensures a reasonable margin (about 14 dB) to interference and link budget variations due, e.g., to fading phenomena.

Using this parameter, we attempted a rough coverage planning for the city of Padova, which extends over an area of about 100 square kilometers. The resulting plan is shown in Figure 2.6, from which we observe that, with the considered conservative coverage range estimate, the coverage of the entire municipality can be reached with a total of 30 gateways, which is less than half the number of sites deployed by one of the major cellular operators in Italy to provide mobile cellular access over the same area.

Finally, we observe that Padova municipality accounts for about 200,000 inhabitants. Considering 30 gateways to cover the city, we get about 7,000 inhabitants per gateway. The current LoRa gateway technology claims the capability of serving 15,000 nodes per gateway, which accounts for about 2 *things* per person. Considering that the next generation of gateways is expected to triple the capacity (by using multiple directional antennas), in the long term we can expect that a basic coverage of the city may grant up to 6-7 *things* per person, on average, which seems to be adequate for most Smart City applications. Any

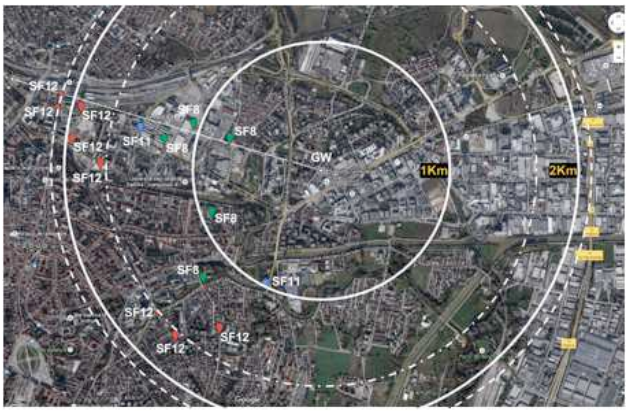


Figure 2.5: LoRa system single cell coverage in Padova, Italy. Colored dots represent some of the measurement spots, which are associated with the minimum SF required for robust communication. The dash-dotted circle and the dashed circle enclose the coverage edge area, where communication is only possible at the minimum transmit rate (i.e., using $\texttt{SF} = 12$).

further increase in the traffic demand can be addressed by installing additional gateways, a solution similar to densification in cellular networks.

2.4 Performance of LoRa Networks in Presence of Massive Uplink Traffic

A debate is going on about the effective performance of LPWANs, in order to understand whether they are a viable solution for the deployment of IoTs networks. In this section, we aim at evaluating the performance LoRa in a typical urban scenario: to do so, we implemented a new LoRa module in one of the most accurate open-source system-level network simulators, ns–3 [37].

The related work on this topic is quite limited, since the interest of the research community in the relatively new technologies of LPWANs started growing lately. In [111], the authors provide an exhaustive technical analysis of the LoRa modulation system, comparing it with other LPWAN solutions. In [112], instead, some field trials of LoRa end nodes are carried out in a urban environment and computer simulations of the LoRa MAC layer procedures are run to evaluate the throughput of a LoRa network.

However, comprehensive and accurate system-level simulations of LoRa networks that consider a number of end devices which are deployed in a realistic urban propagation scenario, with streets and buildings, are still missing: it is exactly in this context that our work is presenting novel and interesting results demonstrating that a LoRaWAN provides a higher throughput than a typical ALOHA-based scheme, thanks to the new access scheme it employs. Furthermore, we show that the LoRaWAN network can scale well, since a higher number of gateways increases considerably the coverage and reliability of the UL. Finally, we demonstrate via a simulation campaign that a packet success rate above 95% is achieved when a gateway is serving a number of devices in the order of 10^4,

Figure 2.6: Coverage plan using LoRa system for the city of Padova, Italy

as claimed by the proponents of LoRaWAN [113].

The rest of the section is structured as follows. In Section 2.4.1 and Section 2.4.2 we describe the link-level assumptions andthe system-level assumptions, respectively. Then, in Section 2.4.3 the performance evaluation results are shown and discussed.

2.4.1 Link-Level Assumptions

As done in other system-level simulation tools, e.g., the Vienna Long-Term Evolution (LTE) simulator [114], our system-level simulations are based on two models:

- the *link measurement model* aims at estimating the signal strength at the receiver side and is used to abstract the effects of propagation on signal strength as well as to average out small-scale fading and similar effects;
- the *link performance model* determines the probability of correctly receiving a packet at reduced complexity, given the previously computed link strength, the amount of interferers and other system-level effects.

Link Measurement Model

Given a transmitter-receiver pair, the link measurement model aims at estimating the strength of the signal at the receiver side. Let us denote the transmitter and receiver antenna gain with G_{tx} and G_{rc}, respectively, and the

transmit power with P_{tx}. Then, the received power is expressed as

$$P_{\text{rx}} = \frac{P_{\text{tx}} G_{\text{tx}} G_{\text{rc}}}{L} e^{\xi}\,, \tag{2.2}$$

where L is the path loss and e^{ξ} is the lognormal shadowing component, i.e., $\xi \sim \mathcal{N}(0, \sigma^2)$. In the logarithmic domain, Equation (2.2) becomes

$$P_{\text{rx}}^{\text{dB}} = P_{\text{tx}}^{\text{dB}} + G_{\text{tx}}^{\text{dB}} + G_{\text{rc}}^{\text{dB}} - L^{\text{dB}} + 4.34\xi\,. \tag{2.3}$$

The path loss L^{dB} consists of two contributions: the *propagation loss*, which depends on the distance between transmitter and receiver, and the *building penetration loss*, due to the wall attenuation, thus

$$L^{\text{dB}} = L_{\text{propagation}}^{\text{dB}} + L_{\text{buildings}}^{\text{dB}}\,. \tag{2.4}$$

Propagation Loss Model According to [115], the propagation loss (also called *external* path loss) is computed as

$$\begin{aligned} L_{\text{propagation}}^{\text{dB}} = {} & 40(1 - 4 \times 10^{-3} \times h) \log_{10} R|_{\text{km}} \\ & - 18 \log_{10} h|_{\text{m}} + 21 \log_{10} f|_{\text{MHz}} + 80\,, \end{aligned} \tag{2.5}$$

where $h \in [0, 50]$ m is the gateway antenna elevation, measured from the average rooftop level. We want to remark that the antenna elevation has a massive impact in the performance of the system [116]. Assuming $f = 868$ MHz and $h = 15$ m, it follows

$$L_{\text{propagation}}^{\text{dB}} = 120.5 + 37.6 \log_{10}(R|_{\text{km}})\,. \tag{2.6}$$

Building Penetration Loss Model In order to model the losses that are caused by external and internal walls of buildings, we resort to the model described in [28]. The overall building penetration loss $L_{\text{buildings}}^{\text{dB}}$ is the sum of the following three contributions:

1. External Wall Loss (EWL);
2. the internal wall loss; and
3. the losses caused by floors and ceilings.

We omit the details about how each one of these contributions is modeled, inviting the reader to refer to [28].

Correlated Shadowing Generation Many studies on shadowing in wireless networks can be found in the literature. In particular, [117] provides a structured synthesis of the existing literature on correlation in wireless shadowing. Two kinds of correlation are usually considered [118].

1. If a transmitter sends a message to a receiver, we expect that the amount of shadowing experienced by the receiver is correlated with the shadowing affecting any other device that is "close" to it. This correlation is a function of the distance separating the two devices, and is usually modeled with an exponential function [119].

Table 2.3: Gateway sensitivity to different SFs

Spreading Factor Index	Sensitivity [dBm]
7	−130.0
8	−132.5
9	−135.0
10	−137.5
11	−140.0
12	−142.5

2. If two devices which are close to each other transmit, we expect their shadowing values to be correlated at the receiver side. This effect is described as site-to-site cross-correlation in [118].

The most common correlation model is a decaying exponential of distance (distance-only model, [117, Section VI-B]). Denoting the distance between end nodes i and j with d_{ij}, the shadowing correlation is

$$\rho_{ij}(d_{ij}) = e^{-d_{ij}/D_C} , \tag{2.7}$$

where $D_C > 0$ is a tunable parameter called decorrelation distance.

As for the implementation of correlated shadowing components, the most common way to generate shadowing maps (i.e., 2D functions that describe shadowing at each point in the map) exploits Cholesky factorization [117]. To reduce the computational effort required to produce the maps, [120] proposes an alternative method. However, to simulate a urban scenario with tens of thousand of nodes, as envisioned for a LoRa network, we resort to the heuristic approach proposed by [121]. Assuming a shadowing decorrelation distance $D_C = 110$ m [28], we generate a regular grid in which each square has a side length of D_C and draw an independent Gaussian random variable at each vertex of the grid. To calculate the shadowing values of nodes which are not exactly placed on a vertex of the grid, we interpolate the values of the grid using an exponential covariance function as explained in [121]. This captures correctly the first one of the two aspects of the shadowing correlation that we listed above. In order to also express the fact that a receiver "sees" two correlated shadowing values from neighbouring devices, we make use of the same shadowing map for every point belonging to the same square in the grid.

Link Performance Model

The link performance model aims at abstracting the real implementation of the physical layer transmission chain and at making interference computations more manageable. It is based on a model of the gateway receiver and a couple of look-up tables that are used to model the aspects of sensitivity and interference.

Receiver Sensitivity Let us denote with S_i the sensitivity of the gateway receiver for $\texttt{SF} = i$. The gateway (UL) sensitivity is summarized in Table 2.3 [113].

For each value in Table 2.3, we need to factor in the gain of the receiver antenna G_{rc} (improving the reception in general). It can be seen that an increase

of SF yields a better sensitivity, with regular steps of 2.5 dB. In case of DL transmissions, since the sensitivity of an end device is assumed to be worse than the sensitivity of a gateway, we introduce an offset of 3 dB and, once again, we have to factor in the antenna gain.

Any received packet with $\texttt{SF} = i$ whose power is below the threshold S_i cannot be detected by the gateway; if, instead, the received power is above S_i, then it can be detected. In this case, we also assume that the receiver will lock on the incoming signal and start receiving the packet.

A further assumption regards the received power of the packet, which is computed thanks to Equation (2.3) and assumed to be constant for the whole duration of the reception. This implies that when a packet is received with a high enough power to start being detected, it will be detectable (i.e., above the sensitivity) for the rest of the time it takes to be completely received.

SINR Matrix Since our objective is to simulate the behaviour of a standalone LoRaWAN network, we assume that interference only comes from other LoRa transmissions. By making this assumption, we can leverage the (partial) orthogonality property of different SFs to model whether a packet survives interference from other LoRa transmissions or not. Let us introduce the following (relative) Signal-to-Interference-plus-Noise-Ratio (SINR) threshold matrix [111]:

$$\boldsymbol{\Theta} = \begin{bmatrix} 6 & -16 & -18 & -19 & -19 & -20 \\ -24 & 6 & -20 & -22 & -22 & -22 \\ -27 & -27 & 6 & -23 & -25 & -25 \\ -30 & -30 & -30 & 6 & -26 & -28 \\ -33 & -33 & -33 & -33 & 6 & -29 \\ -36 & -36 & -36 & -36 & -36 & 6 \end{bmatrix} . \tag{2.8}$$

The element Θ_{ij} is the SINR margin (in dB units) that a packet sent with $\texttt{SF} = i$ must have in order to be decoded if the interfering packet has $\texttt{SF} = j$. We remark that, in presence of multiple interfering packets, we need to satisfy the margin conditions with all the interfering packets, summing the received power values for each SF. Therefore, referring to the Single-Input-Single-Output (SISO) case [114, Section III-C3], we recall the general definition of SINR:

$$\mathsf{SINR} = \frac{P_{\mathrm{rc},0}}{N_0 B + \sum_{\ell=1}^{N_{\mathrm{int}}} P_{\mathrm{rc},\ell}} , \tag{2.9}$$

where $P_{\mathrm{rc},0}$ is the power of the packet under consideration, N_0 is the Additive White Gaussian Noise (AWGN) power spectral density, N_{int} is the number of interfering packets, and $P_{\mathrm{rc},\ell}$ is the power of the ℓ-th interfering packet. Focusing on an end device using $\texttt{SF} = i$ and a set $\mathcal{I}_j$ of interferers using $\texttt{SF} = j$, we define

$$\mathsf{SINR}_{ij} = \frac{P_{\mathrm{rc},0}}{N_0 B + \sum_{\ell \in \mathcal{I}_j} P_{\mathrm{rc},\ell}} . \tag{2.10}$$

Therefore, a packet with $\texttt{SF} = i$ is correctly decoded if, for every j (i.e., for every set $\mathcal{I}_j$ of interfering packets with the same SF), the following inequality holds:

$$\mathsf{SINR}_{ij}^{\mathrm{dB}} > \Theta_{ij} . \tag{2.11}$$

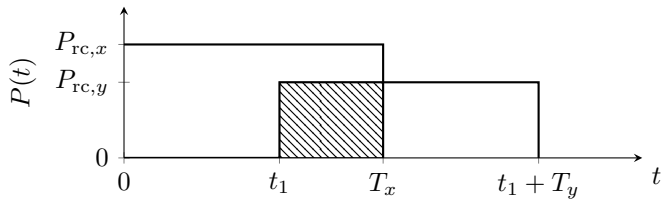


Figure 2.7: *Power equalization* of colliding packets. The highlighted energy is spread on the duration of the packet.

A further remark must be made. The elements in matrix $\mathbf{\Theta}$ are calculated assuming that the two packets are perfectly overlapping. However, in the general case, packets are not perfectly synchronized. Because of this, we must *equalize* the interfering power value for the computation of the SINR. Consider the situation illustrated in Figure 2.7, in which a packet with $\texttt{SF} = x$ is received at time $t = 0$ and whose transmission lasts T_x. A packet with $\texttt{SF} = y$ is received at time $t = t_1$ and its transmission lasts T_y. The energy of packet x is $E_x = P_{\mathrm{rc},x} T_x$, while the interfering energy is $E_y^{\mathrm{interf}} = P_{\mathrm{rc},y}(T_x - t_1)$. Therefore, the *equalized* interfering power is:

$$P_{\mathrm{rc},y}^{\mathrm{interf}} = \frac{E_{\mathrm{rc},y}^{\mathrm{interf}}}{T_x} = \frac{P_{\mathrm{rc},y}(T_x - t_1)}{T_x} = P_{\mathrm{rc},y}\left(1 - \frac{t_1}{T_x}\right). \tag{2.12}$$

Similarly to the example above, we assume that, in general, the interfering energy for any reciprocal position of the useful signal and the interfering signal can be "spread out" on the signal in order to then compute the SINR using Equation (2.10). Denoting with t_{ol} the period of time during which the interferer is overlapping with the useful signal, the general formula becomes:

$$P_{\mathrm{rc},y}^{\mathrm{interf}} = \frac{P_{\mathrm{rc},y} \times t_{\mathrm{ol}}}{T_x}. \tag{2.13}$$

This assumption is justified by the fact that the underlying channel code employed by the modulation makes use of an interleaver: even if the interference is concentrated on a few consecutive symbols, we can assume that a good interleaver will spread it out and eventually correct the errors caused by the interferer.

Moreover, thanks to the channel coding technique used by the LoRa modulation standard, we can also assume that we will always correctly receive a packet that is above sensitivity and survived interference, due to the fact that the curves of the bit error rate versus SINR decline very sharply as SINR grows above the thresholds reported in matrix $\mathbf{\Theta}$ in Equation (2.8).

Characterization of the Gateway Receiver We assume that a single LoRa gateway is capable of emulating 8 receivers working in parallel, as explained in [113]. These 8 *receive paths* are connected to the same antenna, and have the following characteristics.

- The center frequency of each receive path can be individually configured.
- Any SF can be received without prior configuration on any receive path.

- When more than one receive path is listening to the same channel, we assume that they can manage in parallel as many packets as the number of listening receive paths. The packets may even have the same SF. In other words, if there are multiple receive paths on the same frequency and a packet arrives, only one receive path "locks" on the incoming signal, leaving the other ones free to sense more incoming packets.

- If a packet arrives at a certain LoRa channel and there are no available receive paths listening, the packet is lost.

2.4.2 System-Level Assumptions

In the following of this section, we explicitly refer to a LoRa Class-A network, where transmissions are always initiated by the end devices, in a totally asynchronous manner. For this purpose, the end node may choose at random one channel. One of the parameters of the system is the *report periodicity* τ: in our scenario, every end device is assigned a random initial reporting delay, after which the node generates a new packet every τ seconds. In this subsection, no DL transmissions, i.e., messages from the gateways to the end devices, are considered. This limitation will be removed in a future version of this work. Anyway, we do not consider it a heavy limitation, since we expect most of the traffic in a LPWAN to be UL.

Spreading Factor Assignment

At the beginning of the simulation, each device is assigned a SF as follows. We first calculate the power level that each gateway would receive from the end device. Then, we pick the gateway with the highest received power and set the SF based on that value. The assignment is done according to the gateway sensitivity: we assign the end device the lowest SF that would still be above the gateway sensitivity. Note that, due to the shadowing and the presence of buildings, the closest gateway to a device may not always be the one that receives the highest power from that device. As an example, suppose that the best gateway for a device receives a power of -137 dBm. In this case, considering the sensitivity values contained in Table 2.3, it can be seen that $\texttt{SF} = 9$ would be too low, while we can receive the end device packets if they are sent using $\texttt{SF} \in \{10, 11, 12\}$. Since we are interested, in general, in minimizing the ToA, we set the end device to use $\texttt{SF} = 10$. An example of SF assignment can be found in Figure 2.8.

Channel Lineup

LoRaWAN dictates the use of at least three mandatory channels at center frequencies 868.1, 868.3, and 868.5 MHz in the European ISM frequency bands. When sending a packet, the end node picks one of these three channels at random. In our simulations featuring multiple LoRa channels, thus we have decided to rely on the following, fixed allocation of the gateway's 8 receive paths.

- Since there are 3 channels in the h1.4 sub-band (with 1% duty cycle) that are used for UL communication, we will allocate 3 receive paths to the first LoRa channel, 3 to the second, and 2 to the third one.

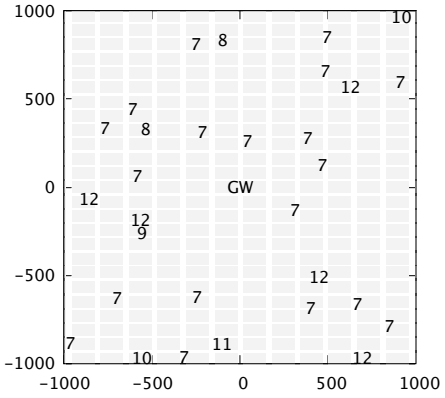


Figure 2.8: An example of random distribution of nodes around a gateway, denoted by "GW." The gray rectangles represent the buildings, while the numbers denote the SFs assigned to the various end devices and their position on the map. The distances are expressed in meters.

- We assume the h1.6 sub-band will be used exclusively for DL communication, thus no gateway receive path will be allocated to this channel.

2.4.3 Performance Evaluation

Leveraging the ns–3 simulator module we developed, we have been able to evaluate various performance metrics of a LoRa network. Several tests were conducted in order to estimate throughput, packet error probability, and gateway coverage.

Throughput Performance

The first simulation campaign aimed at evaluating the throughput $\mathcal{S}$ as a function of the offered traffic Λ. The network scenario is characterized by a single central gateway and N end devices, uniformly distributed in a circular space around it of radius $r = 7500$ m. This particular radius value was chosen because r is the maximum distance at which the gateway and an end device using $\texttt{SF} = 12$ are able to communicate above sensitivity, considering the propagation loss only. The simulations have been performed on a single LoRa channel, and the gateway has only one receive path enabled for all simulations measuring throughput. Since we are interested in evaluating the utmost performance of the LoRa modulation, no duty cycle restrictions are enabled at this stage.

For the throughput computation, we suppose that the device $i = 1, \ldots, N$ generates every τ_i seconds a packet which occupies the channel for $T_{\text{pkt},i}$ seconds in order to be transmitted. We compute the network offered traffic as described in [122]:

$$\Lambda = \sum_{i=1}^{N} \frac{T_{\text{pkt},i}}{\tau_i} \,. \tag{2.14}$$

For a given value of Λ, throughput $\mathcal{S}$ is obtained as

$$\mathcal{S} = \Lambda \times P_{\text{succ}} \,, \tag{2.15}$$

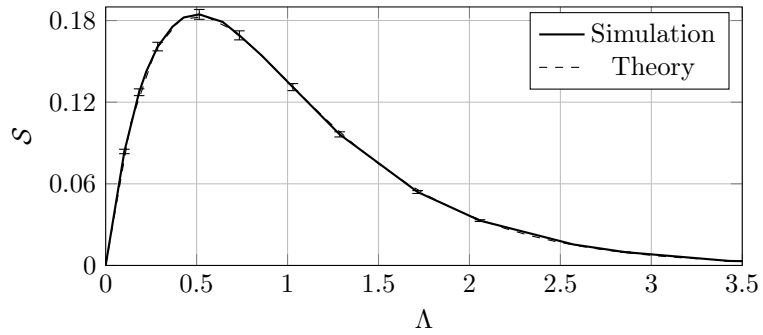


Figure 2.9: Throughput versus offered traffic for $\texttt{SF} = 7$ and ideal packet collisions

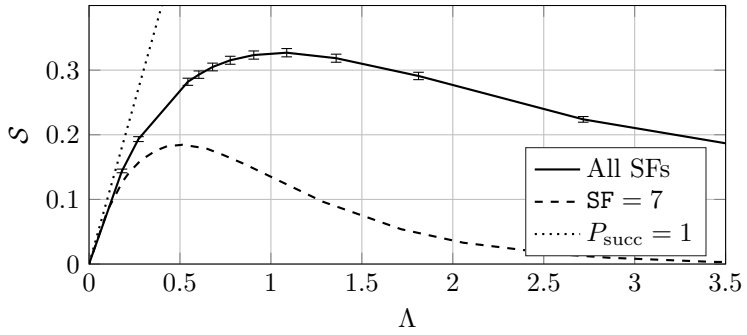


Figure 2.10: Throughput performance of a LoRa network with real wireless channel (solid line) and ideal channel conditions (dashed line)

where the probability of success of a given packet P_{succ} is the ratio between the number of successfully received packets and the total number of sent packets.

As a first validation of our simulator we expect, under ideal channel conditions, the shape of the throughput curve to be that of a typical ALOHA network. We assume then ideal channel conditions and that overlapped packets are always collided and, consequently, lost. Turning off the link measurement model, all end devices transmit with $\texttt{SF} = 7$ and all packets are received with the same power by the central gateway. As expected, the performance result of this test, shown in Figure 2.9, mimics the theoretical ALOHA throughput trend [122], where $P_{\text{succ}} = e^{-2\Lambda}$, thus $\mathcal{S}$ is maximized by $\Lambda^\star = 0.5$, yielding $\mathcal{S}^\star \simeq 18\%$. For all figures featuring the throughput metric, 95% confidence intervals are also shown.

After the validation, we evaluated the impact of real wireless links using the proposed link measurement model: indeed, the presence of a real channel motivates the usage of all possible SFs. The simulation results in Figure 2.10 show a large throughput increase with respect to the previous case.

We also studied the impact of $\texttt{SF} = 12$ transmissions on the performance of the LoRa network. The simulation results shown in Figure 2.11 demonstrate that excluding end devices with the highest SF is beneficial when the system load is high, because the collisions with other end device transmissions are reduced, thus the success probability increases. This is in line with the mandate by the LoRa Alliance to exclude from public networks the end nodes that can transmit

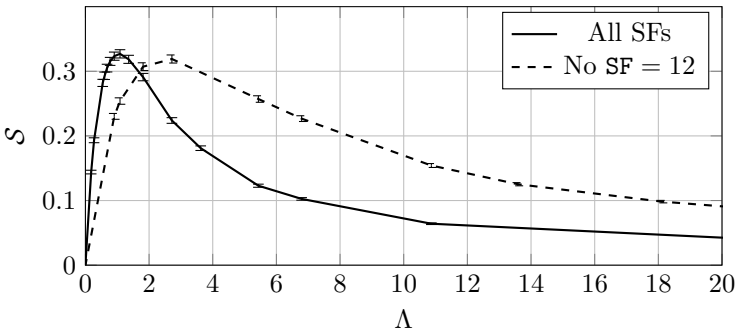


Figure 2.11: Throughput performance with and without SF = 12 (solid and dashed line, respectively)

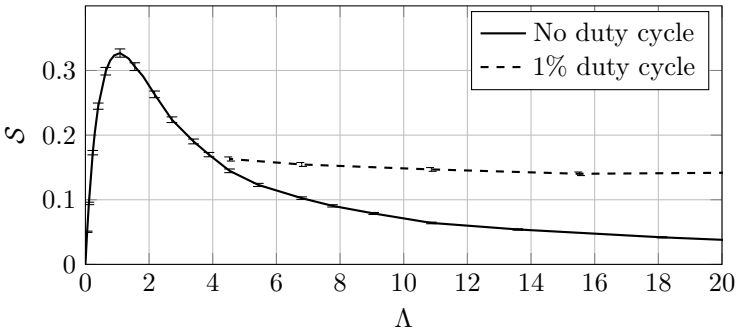


Figure 2.12: Effect of duty cycle limitations on throughput

only at SF = 12.

Finally, we investigated the impact of duty cycled transmissions at the end device. Figure 2.12 shows that a duty cycle of 1% is beneficial for the system because it limits traffic and hence collisions, thus providing a higher throughput than the system without duty cycle restrictions.

Success Probability Performance

The second simulation campaign aimed at estimating the probability of successfully receiving a packet in a LoRa network. Since we are interested in the performance of real networks, this simulation scenario features 18 gateways that are placed in an hexagonal grid around a central gateway. In these simulations each gateway will cover a radius of 1.5 km, thus the area in which we place end devices is a circle of radius 7.5 km centered on the central gateway. This allows us to simulate inter-cell interference besides intra-cell interference. Even though the simulation features 19 gateways, we are interested only in the devices belonging to the area that is covered by the central gateway, so the collected data regard packets that were generated inside this region of interest. To add realism to the simulation, the entire area features buildings (whose dimensions and distance follow the layout of Manhattan) and generation of correlated shadowing is enabled. If a device is randomly assigned to an area occupied by a building, that device will be marked as "indoor" and transmissions involving it will suffer

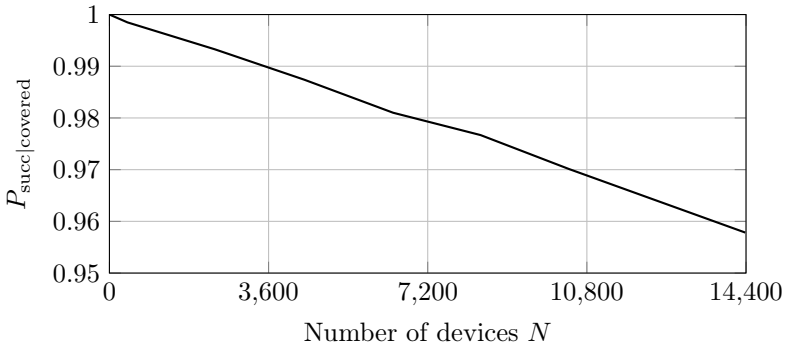


Figure 2.13: Packet success probability (covered nodes only) as a function of the total number of end devices in the coverage area of the central gateway

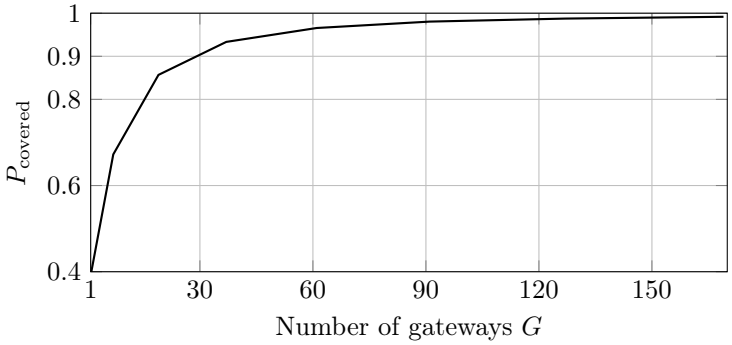


Figure 2.14: Coverage probability for a node as a function of the number of gateways covering a circular area of radius 7.5 km

from building penetration losses. As for the traffic generation, we refer to the Mobile Autonomous Reporting model for periodic reports described in [28]. Also the size of the application level payload is randomized, following a Pareto distribution as described in [28] with payload size in the $[10, 30]$ bytes range.

Figure 2.13 shows the packet success probability as a function of the number of end devices in the central gateway coverage area. This probability ignores packets that arrived at the central gateway under sensitivity (because of heavy building loss or shadowing), thus the decreasing trend of the success probability is to ascribe only to interference or to the unavailability of adequate reception paths at the gateway. In particular, we noticed that, on average, the 20% of the end devices cannot connect to the gateway due to particularly unfavourable channel conditions. Nevertheless, these nodes remain active and cause interference to neighbouring nodes. The trend appears to be linear with the number of devices in the network, with a success probability around 97% for a network with 10^4 end devices. This is coherent with Semtech's claim that a gateway is able to support a network of around 10^4 nodes [113].

Gateway Coverage Assessment

In the final simulation campaign we study how increasing the number of gateways G that serve a fixed amount of end devices can enhance the reliability of their connections to the Network Server. This aspect is particularly interesting for critical applications, where the packet reception (by *any* gateway) is crucial. We simulated a circular urban scenario of radius 7.5 km, where end devices are served by an increasing number of gateways deployed in a hexagonal grid setup. The results shown in Figure 2.14 state that, in order to achieve a reliability above 90%, we should deploy gateways in such a way that every gateway covers 6 km^2 or, equivalently, a radius of 1200 m around it.

A consequence of the densification of gateways is that the number of end nodes with $\texttt{SF} > 7$ decreases, thus increasing the number of collisions between packets having the same SF (and thus worse SINR isolation, see (2.8)). In a real LoRa network, the Adaptive Data Rate (ADR) mechanism should be able to keep the network in a state where SF orthogonality can still be leveraged to increase throughput.

2.5 Performance of LoRa Networks in Presence of Massive Downlink Traffic

While for a wide variety of IoT services a best-effort packet transmission policy is acceptable, for some other IoT applications the end node may need to know whether the packet was correctly received by the gateway or not, in order to implement some higher Quality of Service (QoS) requirements. Thus, the scope of this section is to evaluate the impact of the DL feedback on LoRa networks, to investigate the degree of dependability a LoRa network can achieve. In particular, we will focus on the MAC protocol implemented by LoRa, providing an abstraction of the PHY, with the aim of evaluating the *ideal performance* of the access protocol, that is, without considering the side effects of the actual positions of the various end nodes and the consequently different wireless channel propagation conditions.

The interest on the DL performance of LoRa is growing only recently, thus the related work is quite limited. One of the first insights about the limitations of LoRa DL was given in fact in [112]. In [123], more accurate computer simulations were carried out to emulate realistic LoRa deployments and study the network capacity limits; however, the authors implemented a simplified version of the MAC protocol, e.g., considering a single transmission attempt from the terminals. More complete simulation tools and extensive campaigns are proposed by [124] and [125], confirming the previous insights and highlighting further aspects. However, despite all the aforementioned papers provide extremely useful intuitions about the *effects* of DL feedback on LoRa networks, they lack a comprehensive analysis of the *reasons* for these effects. In this section, instead, we aim at investigating how the UL traffic composition and the different parameters are intertwined and how they influence the network behaviour. The findings of this work make it possible to tune the network parameters according to the target QoS and Key Performance Indicators (KPIs) of a given IoT service.

The rest of the section is structured as follows. In Section 2.5.1, we describe

the simulation setup, while in Section 2.5.2 the performance evaluation results are shown and discussed.

2.5.1 Simulation Setup

Let us introduce the assumptions we made and the simulation setup we considered to evaluate the performance of LoRa in ideal channel conditions.

Types of Data Packets

As for the MAC protocol, we consider the most-widely adopted operation mode (Class A), which is shown in Figure 2.3. DL messages can be received only after an UL transmission, since two *receive windows* are opened by the node: the first window is opened on the same channel as the node's UL communication, while the second window is opened on a different sub-band previously agreed with the NS. Note that the delays to open both the receive windows are referred to the end of the UL transmission. The receive window duration is

$$\mathsf{RW} = T_{\text{preamble}} + W\,, \tag{2.16}$$

where T_{preamble} is the preamble duration of the DL packet (see Equation (2.17)) and $W \geq 0$ is the excess receive window size we introduce to enable some time flexibility at the NS side.[13]

Two categories of data packets are defined [108]: a) *unconfirmed-data* packets and b) *confirmed-data* packets. The typical best-effort operation of LoRa networks evaluated in the previous section falls in the former category: packets are just sent in the UL without any guarantee on the successful delivery. The latter class, instead, may be used by IoT applications whose traffic requires to be delivered to the NS with a certain reliability guarantee: if the acknowledgement (ACK) of the UL packet is not received, the end device is allowed to transmit the packet up to `NbTrans` times, with a backoff time of `ACK_TIMEOUT` between adjacent attempts.

A wide range of parameters is defined in LoRaWAN to implement the MAC. We invite the reader to refer to Table 2.4 to find the list of parameters of interest for this work and their values. In addition to already mentioned parameters, we define the following ones:

- `PL`: payload size;
- `IH`: flag to disable headers;
- `DE`: flag to enable low data rate optimisation;
- `CRC`: flag to enable Cyclic Redundancy Check (CRC) of the payload.

Assumptions

We consider a Class-A LoRa network, where new packets are generated by the end devices at a rate λ following a Poisson arrival process. Each new packet is given a certain *priority*:

[13] We remark that the terminal will continue to listen to the DL channel only if it is able to detect the DL packet preamble.

Table 2.4: MAC parameters

Parameter	Value
SF	$\{7, 8, 9, 10, 11, 12\}$
PL	15 bytes for UL 1 byte for DL
IH	0
CR	1
DE	0
CRC	1 for UL 0 for DL
RECEIVE_DELAY_1	1 s
RECEIVE_DELAY_2	2 s
ACK_TIMEOUT	1 s
NbTrans	$\{1, \ldots, 15\}$

1. *high-priority* packets require to acknowledge the outcome of the data transmission (i.e., they come from confirmed-data terminals);
2. *low-priority* packets do not require any ACK (i.e., they come from unconfirmed-data terminals).

The fraction of high-priority packets is $p_h \in [0, 1]$, while $p_l = 1 - p_h$ denotes the fraction of low-priority end devices. Moreover, we assume that every new packet is assigned at random

1. one of the 6 available SFs and
2. one out of $n_{\text{ch}} = 3$ channels for the UL transmissions, as mandated by LoRaWAN.

In particular, we will assume that the n_{ch} channels are allocated on *different* sub-bands[14] with a typical 1% duty cycle constraint mandated by the European regulations [96]; we recall that, other than for UL transmissions, these channels will accommodate the DL feedback in the first receive window. Moreover, we assume that the second receive window is opened on the h1.6 sub-band, which instead has a 10% duty cycle limit [96].

Therefore, two kinds of packets are sent through the network: UL data packets and DL ACKs. The ToA T of both is

$$T = T_{\text{preamble}} + T_{\text{payload}} = [(n_{\text{preamble}} + 4.25) + n_{\text{payload}}] \times T_s \,, \tag{2.17}$$

where n_{payload} is [126]

$$n_{\text{payload}} = 8 + \max\left(\left\lceil \frac{8 \times \texttt{PL} - 4 \times \texttt{SF} + 28 + 16 \times \texttt{CRC} - 20 \times \texttt{IH}}{4 \times (\texttt{SF} - 2 \times \texttt{DE})} \right\rceil (\texttt{CR} + 4), 0 \right) . \tag{2.19}$$

We invite the reader to refer to Table 2.5 for the ToA of UL and DL packets (denoted by T_{pkt} and T_{ack}, respectively) for the various SF values. Regarding the PHY abstraction, we assume ideal channel conditions: two (or more) packets

[14] Thus, the duty cycle limit is not aggregated, rather per-channel.

Table 2.5: UL and DL packet ToA for all SF values

SF	R_b [kpbs]	T_{pkt} [s]	T_{ack} [s]
7	5470	0.0463	0.0259
8	3215	0.0927	0.0517
9	1760	0.1649	0.0829
10	980	0.3297	0.1659
11	440	0.5775	0.3318
12	250	1.1551	0.6636

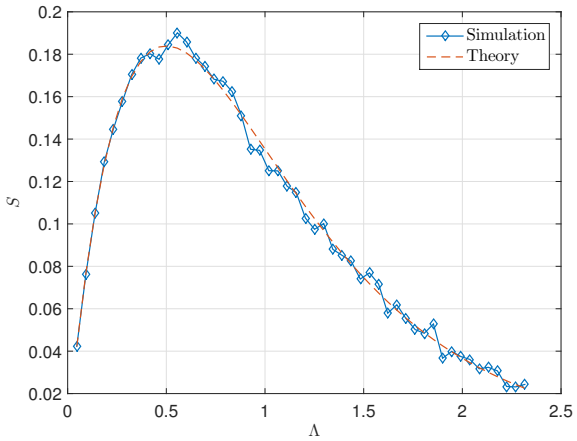


Figure 2.15: Comparison between simulation results with only low-priority traffic ($p_h = 0$) and theoretical performance of pure ALOHA protocol

collide if and only if their transmissions overlap in time *and* they use the same SF. If the packets overlap in time but use different values of SF, then they are using distinct collision domains (because the transmissions with different SFs can be well approximated as orthogonal), thus they are successful.

We assume that G gateways are available, and they are able to reach all end devices. In our simulations, we will assume that the gateways are provided with ideal reception capability, i.e., with infinite receive paths, and that each gateway has one antenna to transmit. We want to remark that the gateways have to comply with the regulations about duty cycle as well. The DL feedback is transmitted by the gateways using the same SF of the UL packet.

As for the timing relations, we set a receive window excess length W greater than zero to enable some scheduling flexibility at the NS side. Note that, in general, W is such that

$$W \geq 0\,, \tag{2.20a}$$

$$W < \texttt{RECEIVE\_DELAY\_2} - \texttt{RECEIVE\_DELAY\_1} - T_{\text{preamble}}\,. \tag{2.20b}$$

Validation of the Simulation Setup

To validate the simulation setup, we resort again to the ALOHA protocol theory [122]. Denoting with $T_{\text{pkt},i}$ and τ_i the packet duration and the packet

Table 2.6: Simulation parameters

Parameter	Value
λ	$[0, 30]$ packets per UL channel per second
n_{ch}	3
G	$\{1, \ldots, 5\}$
`duty_cycle_1`	1%
`duty_cycle_2`	10%
p_h	$[0.05, 0.3]$
W	0.5

periodicity of node i, the offered traffic Λ of a pool of N end devices is

$$\Lambda = \sum_{i=1}^{N} \frac{T_{\text{pkt},i}}{\tau_i} = T_{\text{pkt}} \times \frac{N}{\tau} = T_{\text{pkt}} \times \lambda \,, \tag{2.21}$$

under the assumptions that all nodes have the same packet periodicity and use the same SF (thus, having the same transmission duration). The throughput $\mathcal{S}$ is defined as in Equation (2.15). The comparison between the simulation results with $p_h = 0$ (only low-priority traffic, implementing exactly the ALOHA protocol) and `SF` $= 7$ for all packets and the theoretical result is shown in Figure 2.15: the two curves are nicely overlapping.

2.5.2 Performance Evaluation

Let us now show and discuss the simulation results. We developed an event-driven simulator in Matlab$^{\text{TM}}$ to emulate the LoRa network scenario described in Section 2.5.1, using the parameters summarized in Table 2.6. We considered two KPIs:

1. the *throughput* $\mathcal{S}$, defined as the number of packets that are correctly delivered, and
2. the *outage probability* (p_{outage}), that is, the probability that a generic packet is not correctly delivered to destination within `NbTrans` transmission attempts.

We studied how the various parameters impact on the performance of a LoRa network in the following studies.

Preliminary Considerations About the Trade-off Between G and n_{ch}

The number of gateways G and the number of channels n_{ch} are closely related due to the presence of the duty cycle constraints. Indeed, if we use just one gateway and one frequency band, then it is very likely that we will exceed the duty cycle threshold given by the regulations. On the other hand, either increasing the number of gateways while keeping fixed the number of channels or vice versa, it is possible to limit the effect of the duty cycle, which is always calculated per device and per channel.

In our simulations, we fixed $n_{\text{ch}} = 3$ (as mandated by LoRaWAN) and tuned the value of G to reduce the effect of the duty cycle on the network. Indeed, while

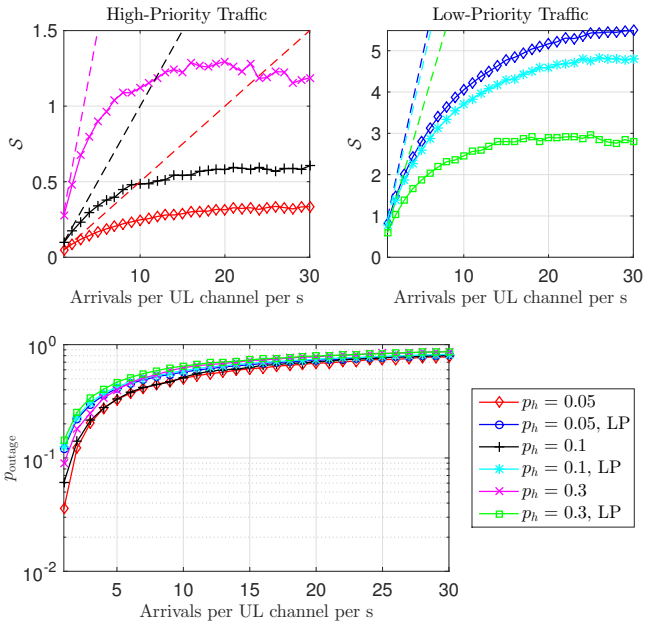


Figure 2.16: Impact of p_h using all SFs, $G = 5$, and `NbTrans` $= 3$. The low-priority traffic curves are denoted by "LP"; the remaining curves are for high-priority traffic. The dashed lines indicate the maximum achievable throughput.

the randomisation of the device operating sub-band allows the given gateways to exploit different duty cycles in different sub-bands, in case of high traffic load it is necessary to increase the value of G anyway. On the other hand, parameter G should be kept as low as possible in order to reduce the collision rate of DL messages.

A third degree of freedom is given by the number of SFs we enable. Indeed, if we use all SFs, the collision probability of packets is reduced. Therefore, more packets arrive at the NS, increasing the rate of DL transmissions and consequently the ToA spent by the gateways.

Impact of p_h

Figure 2.16 shows the results obtained by letting the fraction of high-priority packets p_h vary. Clearly, the high-priority packets can achieve lower outage probabilities than low-priority ones. On the other hand, it is apparent that a higher value of p_h impacts a lot the network performance. Not only the high-priority traffic throughput decreases with respect to the maximum achievable throughput, but also the unconfirmed-data users suffer an important throughput reduction.

Impact of `SF`

We first analyse how end devices with different SFs behave when all SFs are enabled. We invite the reader to refer to Figure 2.17, which reports the outage probability of the various SFs. It can be seen that, as expected, the lower the

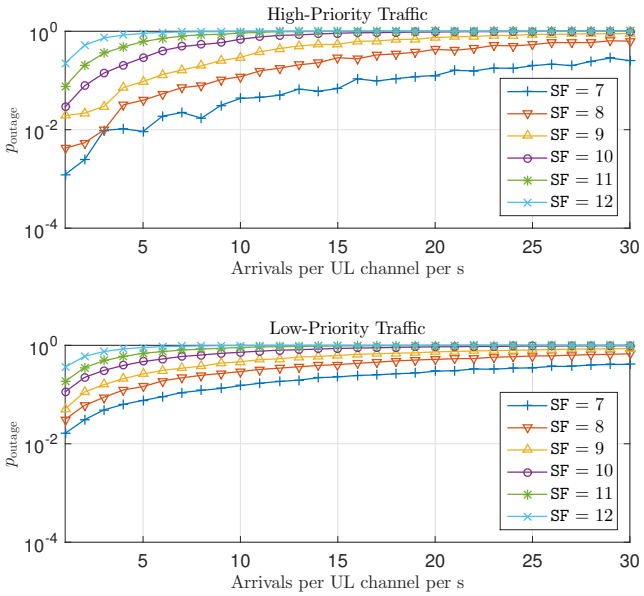


Figure 2.17: Performance of packets with different SFs when $p_h = 0.1$, $G = 5$, and `NbTrans` $= 3$

value of `SF`, the lower the outage probability, both in the case of high-priority and low-priority traffic.

Moreover, Figure 2.18 shows the SF impact if we assume that all end nodes in the network use the same SF sequence. We can observe that the network performance is drastically degraded if we increase the SF, since for every unit increase in the SF the ToA doubles (see Table 2.5).

We want to remark that the random assignment of SFs is clearly a sub-optimal choice. Indeed, the ADR mechanisms of LoRa [108] networks basically is needed to reduce the number of end devices with high SFs, since they are extremely harmful for the overall network performance.

Impact of `NbTrans`

Finally, Figure 2.19 depicts the performance for a variable number of transmission attempts when all end devices are using `SF` $= 7$. As expected, a higher value of `NbTrans` yields a lower outage probability for high-priority users and increases the peak throughput, while degrading the performance of low-priority users. However, too many transmission attempts (see `NbTrans` $= 15$ vs `NbTrans` $= 9$) would results in a detrimental effect, even worsening the problem. We also note that it is not convenient to enable only one transmission attempt for high-priority traffic: the outage probability is higher than the low-priority traffic due to the collisions of DL messages.

What's more, when the network becomes unstable, in the case of a higher value of `NbTrans` the throughput performance drops faster with respect to the case in which we allow fewer transmission attempts.

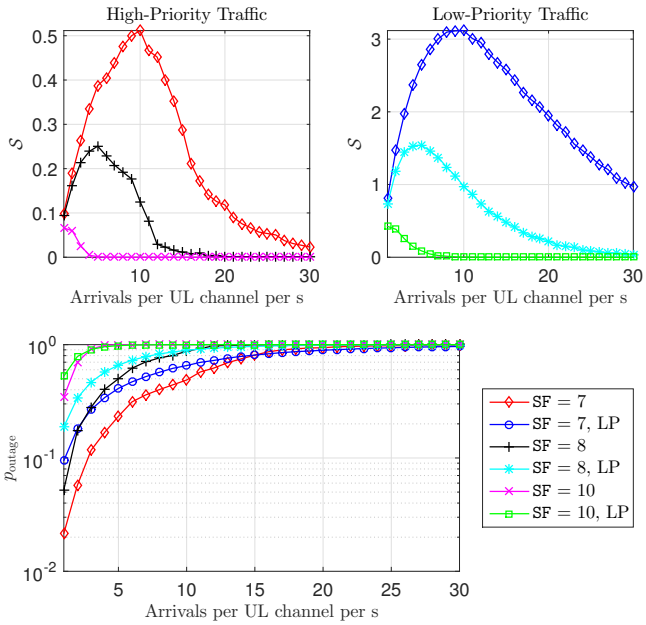


Figure 2.18: Impact of `SF` when $p_h = 0.1$, $G = 1$, and `NbTrans` $= 3$. The low-priority traffic curves are denoted by "LP"; the remaining curves are for high-priority traffic.

2.6 Conclusions and Ways Forward

In this chapter, we described the emerging LPWAN paradigm for IoT connectivity. These solutions are based on long-range radio links, in the order of tens of kilometers, and a star network topology. Therefore, LPWANs are inherently different from usual IoT architectures, which are, instead, typically characterized by short-range links and mesh topology. After a brief introduction of the most prominent LPWAN technologies, we focused on one of them: LoRa.

The LoRa experimental trials showed that the LPWAN paradigm has the potential to complement current IoT standards as an enabler of Smart City applications, benefiting from long-range links. Then, we implemented a system-level simulator in ns–3 to simulate a whole LoRa network consisting of tens of thousands of end devices: the simulation results show that the LoRaWAN access protocol provides a higher throughput with respect to a basic ALOHA scheme, thanks to the partial orthogonality between its SFs. Moreover, we proved that the LoRaWAN architecture can scale well, mainly due to the fact that an increase in the number of gateways enhances the coverage and reliability of the UL, as well. Finally, we assessed the performance in terms of dependability of a LoRa network under a massive number of packet arrivals, part of which require to be acknowledged by the NS. To do so, we implemented an event-driven simulator in Matlab to emulate the MAC protocol defined by LoRaWAN, while abstracting the PHY implementation. The simulation results showed that the performance of LoRa is severely impacted if the fraction of end devices of confirmed-data type grows. On the other hand, if we fix a relatively low fraction of high-priority users, increasing the number of transmission attempts yields a higher peak throughput, but the situation gets worse in case of network instability.

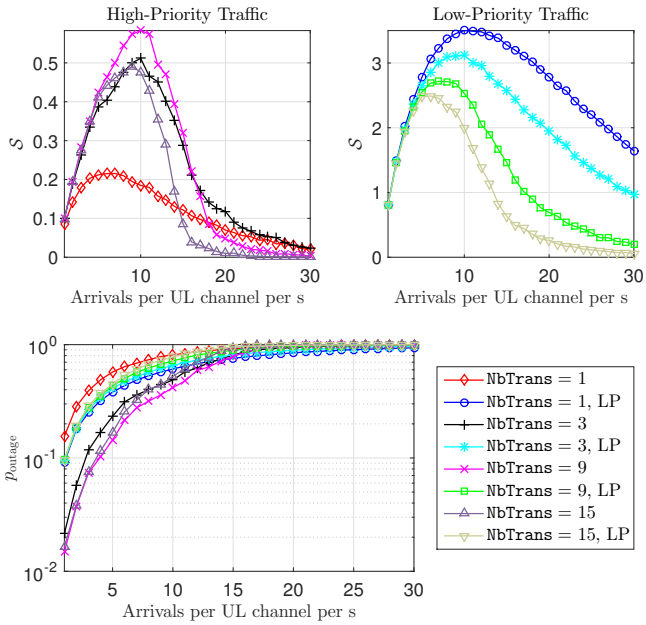


Figure 2.19: Impact of `NbTrans` when `SF` $= 7$, $G = 1$, and $p_h = 0.1$. The low-priority traffic curves are denoted by "LP"; the remaining curves are for high-priority traffic.

As future work, we plan to extend our ns–3 and Matlab simulators, adding more functionalities and implementing strategies that can boost the radio access performance in LoRa.

Part II

Fundamental Research on the Internet of Things

Chapter 3

Physical Layer Security for the Internet of Things

In this first chapter about fundamental research on the Internet of Things (IoT), we will tackle one of the most relevant (and often neglected) aspects in this field, that is, the security of IoT networks. Since the IoT *objects*, usually referred to as Machine-Type Devices (MTDs), are envisioned to operate with minimal human intervention, many concerns regarding the security of IoT networks are raising.

Therefore, in Section 3.1 we will focus on the *authentication* of IoT terminals, i.e., the problem of determining whether a message has been truly transmitted by a "honest" IoT device. We will design an efficient (in terms of energy expenditure and signaling overhead) authentication protocol to ensure that no malicious MTD is transmitting messages in place of a legitimate MTD.

Moreover, due to the growing importance of localization in many IoT applications (e.g., asset monitoring and tracking), there is room to study techniques for *location verification*, which ensure that the position of the transmitting MTD is correct: this particular security aspect is addressed in Section 3.2.

Finally, in Section 3.3 we will draw the conclusions of our work.

3.1 An Efficient Authentication Protocol

Message authentication is an important feature of communications systems, and it will become more and more important as devices will autonomously communicate without much user intervention in IoT scenarios. Moreover, in this pervasive context, MTDs will include very simple computational capabilities and have limited energy budgets. Therefore, new authentication schemes that do not require heavy exchange of keys and use of security protocols may be extremely useful.

Currently, authentication mechanisms are already deployed in Wireless Sensors and Actuators Networks (WSAN) and operate at the Medium Access Control (MAC) or higher layers using cryptographic approaches. For example, the IEEE 802.15.4 standard encompasses the extension of counter mode encryption and cipher block chaining message authentication code (CCM*) algorithm [127, 128]. Like conventional CCM [129], it encrypts and adds a signature

(MIC – Message Integrity Code) to the input data. In addition, the CCM* has the possibility to encrypt only or to add the signature to the input data without encryption only. It must be pointed out that, in this context, some PAN Information Base (PIB) attributes (in particular, the keys) must be properly configured. Here is where a higher layer cooperation or some other mechanisms are needed. Actually, in some implementations, the MAC security tables (including the keys) are set at software compile time and cannot be changed dynamically at runtime.

A possible solution to ease authentication is provided by *physical layer authentication* mechanisms, where the features of the channel over which transmission occurs are exploited. As described in [130], the physical layer authentication mechanisms can be divided into two categories: one using keys for the authentication and the other not requiring keys.

Key-based authentication schemes have been extensively studied in the '80s [131, 132]. In [133], message authentication is interpreted as a hypothesis testing problem, thus extending the previous scenarios. More recently, the presence of noise in the authentication procedure (still based on keys) has been considered in [134,135]. Recent studies have been focused on joint channel and authentication coding [136]. The problem of key-based authentication schemes, however, is that a shared key is needed, having in turn the problem of generating and managing the key, which can be particularly difficult for non-controlled devices with constrained computation capabilities.

Therefore, we focus on *key-less authentication*, which only relies on the characteristics of the channel over which the communication occurs. We will investigate solutions that exploit the features of wireless transmissions and can well integrate other authentication procedures implemented in the higher layers. In particular, the random nature of the wireless channel depending on the spatial position of the transmitter or relatively long time periods (a few seconds) can be exploited to distinguish nodes that are placed in different locations. At the same time, we will also take advantage of the time-invariant nature of the wireless channel for static nodes over short time intervals (up to a hundred millisecond).

In this section, we will exploit the fact that the legitimate receiver knows the channel, while the eavesdropper sees another channel due to a different position with respect to the legitimate receiver. Thus, the receiver can perform a two-stage authentication: it first estimates the channel using a message that has been authenticated by some other means; then, for forthcoming messages, it checks if the channel is the same of the first transmission. The attacker can suitably process the transmitted signal in order to let the receiver estimate a different channel, and various deterministic attack strategies have been considered in [137, 138]. A statistical attack strategy has been investigated in [139], while in [140] it has been proved that secure authentication is possible when the legitimate transmitter has a noisy channel to the receiver whose behavior cannot be completely simulated by the attacker.

Let us consider the map of a planar IoT in Figure 3.1, in which a specific *concentrator node* c is collecting data from many *source nodes* through wireless links. In particular, we assume a *star* topology, in which the source nodes communicate with the concentrator node. Moreover, a number of *anchor nodes* are available to cooperate with the concentrator and are considered as trusted. Anchor nodes overhear the communication for authentication purposes, but may

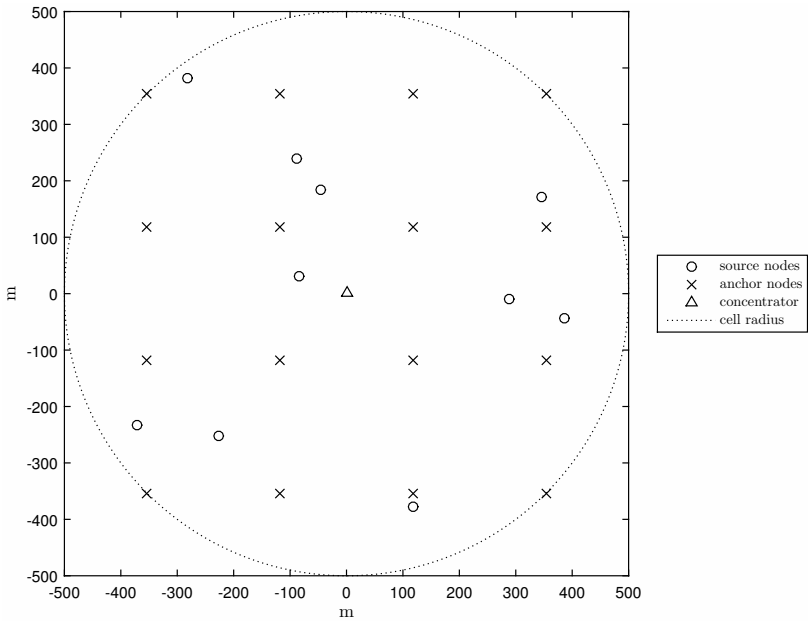


Figure 3.1: Map of a planar IoT including source nodes (i.e., the circles), anchor nodes (i.e., the crosses) and the concentrator node `c` (i.e., the triangle). The dotted line indicates the coverage radius of the cell.

act as relays as well, if needed. We want to remark that we could easily fit also a *tree* topology, in which source nodes communicate only with the anchor nodes and anchor nodes communicate with the concentrator, in order to fit existing short-range IoT standards. For example, the Zigbee$^{\text{TM}}$ standard provides a *coordinator node* (which plays the role of our concentrator), assisted by many routers (which play the role of our anchor nodes) and end devices (which play the role of our source nodes). The IEEE 802.15.4 standard provides a *first Personal Area Network (PAN) coordinator* (which plays the role of our concentrator), assisted by many PAN coordinators (which play the role of our anchor nodes) and devices (which play the role of our source nodes).

Consider that the generic source node `s` is transmitting. We will assume that the first transmission between `s` and `c` is authenticated by some initial pairing procedure, e.g., authentication at high layers by previously shared secrets, or manual pairing by the user checking the consistency of transmitted and received messages. The core of this section is how to authenticate forthcoming packets. For broadband transmissions, a first option is that the concentrator node `c` compares the estimate of channel impulse response of the newly received message with that of the initial transmission. An attacking node trying to impersonate a source node but transmitting from another location will in general have a different channel impulse response to `c` and therefore it will not be authenticated. The wider the bandwidth of the signal is, the higher the precision of this authentication process will be. For narrowband transmissions, which are typical of an IoT scenario, instead, this authentication process is highly imprecise. In this case, indeed, the wireless channel between two devices is characterized by an attenuation which depends (mainly) on the distance between the two devices (*path loss*). However,

we can exploit the availability of anchor nodes (whose authenticity is guaranteed) spread over the area of the source nodes, that estimate the channel gains of ongoing transmissions: by considering the aggregate estimates of the anchors, it is possible to obtain a precise authentication of the messages.

A second problem that we address in this section is the energy balance of the anchor nodes. In fact, we envision that in the IoT scenario the anchor nodes may be battery-powered devices, that are possibly recharged by some renewable energy source (e.g., by solar cells). Therefore, the number of transmissions that can be authenticated is limited. However, since multiple source-concentrator pairs will be communicating over the IoT area, with different distances to the various anchor nodes, we study the possibility of activating different anchor nodes according to the transmitting source, thus making the authentication process more efficient.

On the other hand, the signaling traffic required to perform the authentication procedure must be taken into account, as well. Indeed, from the point of view of a network administrator, an intensive exchange of control messages between the anchor nodes and the concentrator containing the channel estimates to perform the authentication may put the network under strain, eventually causing the collapse of the network itself. Therefore, we design strategies which are efficient from the point of view of the signaling and we find a trade-off between energy efficiency and the amount of network traffic needed to perform the authentication of the data packets.

The rest of this section is organized as follows. In Sections 3.1.1 and 3.1.2, we introduce the reference scenario of our proposal and the attacker model, respectively. Then, in Section 3.1.3 the authentication protocol is proposed. In Section 3.1.4, the criteria for the selection of anchor nodes are addressed and in Section 3.1.5 three possible optimization problems of anchor nodes' usage are derived. In Section 3.1.6, the performance of the proposed physical layer authentication framework and of the envisioned techniques to optimize the energy consumption of anchor nodes is evaluated. Then, the concept of signaling-efficient anchor node selection is introduced (Section 3.1.7), as well as the trade-off between energy efficiency and signaling efficiency (Section 3.1.8). Distributed strategies for the anchor node selection are provided in Section 3.1.9, and a final overall comparison of the various optimization approaches in given in Section 3.1.10.

3.1.1 Reference Scenario

We consider an Cellular IoT (CIoT) scenario with M legitimate sources, N anchor nodes (with indices $i = 1, \ldots, N$), and one concentrator node $\mathtt{c}$. The concentrator $\mathtt{c}$ gathers the reports of the anchors to decide whether a received packet actually comes from a legitimate source.[1] In the IoT scenario, where MTDs transmit at a low rate, we assume that the communication channel between each source $\mathtt{s}$ and each anchor node i is narrowband and can be represented by a single complex coefficient. In particular, the complex channel gain (including path loss and fading) from source node $\mathtt{s}$ to anchor node $i = 1, \ldots, N$ is a random

[1] In principle, we can combine the channel estimates of the concentrator with those of the anchor nodes to increase the performance of the authentication procedure. However, without loss of generality, in the following we will not consider the channel estimate of the concentrator.

variable denoted by $h_i(\mathtt{s})$, and the channel power gain is

$$\mathbb{E}[|h_i(\mathtt{s})|^2] = \lambda_i\,, \qquad i = 1, \ldots, N\,, \tag{3.1}$$

where λ_i is the path loss value of the link between anchor node i and the source node $\mathtt{s}$. We collect the gains of all links to sources $\mathtt{s}$ into the vector

$$\boldsymbol{h}(\mathtt{s}) = [h_1(\mathtt{s}), \ldots, h_N(\mathtt{s})]\,. \tag{3.2}$$

We assume that each link has uncorrelated gains, i.e.,

$$\mathbb{E}[h_i(\mathtt{s})h_j^*(\mathtt{s})] = 0\,, \qquad \forall i \neq j, \tag{3.3}$$

where $(\cdot)^*$ denotes the complex conjugate operator.

We assume that the communication between the anchor nodes and the concentrator node $\mathtt{c}$ is secure, either by some additional communication feature or because these nodes are connected to $\mathtt{c}$ through a wired link (e.g., optical fiber). Moreover, the anchor nodes have a limited energy budget: we assume that, when involved in the authentication of a MTD, an anchor node consumes a fixed amount of energy so that each anchor node is able to perform at most Q message authentications.

3.1.2 Attacker Model

The concentrator $\mathtt{c}$ aims at establishing whether a message reporting as sender the source node $\mathtt{s}$ is actually coming from $\mathtt{s}$. On the other hand, an attacker node $\mathtt{a}$ aims at having his message accepted by the concentrator node as coming from legitimate source $\mathtt{s}$. For this purpose, the attacker can pre-process the message in order to induce a certain channel to each of the anchor nodes. In particular, as a worst case we assume that the attacker is equipped with multiple antennas, thus being able to induce any channel to each anchor nodes.

Let $\boldsymbol{g} = [g_1, \ldots, g_N]$ be the vector containing the forged channel gains from the attacker to the N anchor nodes. We assume that the attacker has only a partial knowledge of the channel gains from the source to the anchor nodes. Let $\boldsymbol{z} = [z_1, \ldots, z_N]$ be the random vector of the N observations available to the attacker (e.g., the channel gains from $\mathtt{s}$ to the N anchor nodes) and we assume that each observation is correlated only with the channel from $\mathtt{s}$ to a anchor node i, thus

$$\frac{\mathbb{E}[z_i h_i^*(\mathtt{s})]}{\mathbb{E}[|z_i|^2]} = \rho\,, \qquad i = 1, \ldots, N\,, \tag{3.4}$$

$$\mathbb{E}[z_i h_j^*(\mathtt{s})] = \mathbb{E}[z_i z_j^*] = 0\,, \qquad \forall i \neq j. \tag{3.5}$$

Note that (3.4) establishes that the correlation coefficient is the same for all anchor nodes and (3.5) yields that only source-anchor and attacker-anchor channels relative to the same anchor are correlated. However, we will assume that the attacker does neither know vector $\boldsymbol{h}(\mathtt{s})$ nor has access to its estimate. This is reasonable since the only way to have access to this estimate is to replace the anchor nodes.[2] The only knowledge available to the attacker on channel $\boldsymbol{h}(\mathtt{s})$ is its joint statistics with the observations $\boldsymbol{z}$. Finally, legitimate nodes have no clue of the presence of the attacker, and in particular they do not know neither $\boldsymbol{z}$ nor $\boldsymbol{g}$.

[2]It could be achieved only by placing spoofing nodes very close to each anchor node in order to estimate the same channel.

3.1.3 Proposed Authentication Protocol

The envisioned authentication procedure operates into two phases.

Phase 1 In the initial phase, the anchor nodes receive a message coming from source $\mathtt{s}$ that has been authenticated by some other methods (e.g., by a key-based authentication procedure), thus they estimate the channel gain vector $\hat{\boldsymbol{h}}^{(0)}(\mathtt{s})$ (which by reciprocity is assumed as that from $\mathtt{s}$ to the anchor nodes) and report this estimate to $\mathtt{c}$.

Phase 2 Upon receiving the k-th message reportedly coming from source $\mathtt{s}$, a sub-set of anchor nodes estimates the channel and reports such estimate to $\mathtt{c}$. In particular, let $\boldsymbol{c}(\mathtt{s},k)$ be a N-size column binary vector denoting the *configuration* of anchor nodes that authenticate the k-th message, i.e.,

$$[\boldsymbol{c}(\mathtt{s},k)]_i = \begin{cases} 1 & \text{if anchor } i \text{ is active in the authentication,} \\ 0 & \text{otherwise.} \end{cases} \tag{3.6}$$

Note that if $[\boldsymbol{c}(\mathtt{s},k)]_i = 1\ \forall i$, then we are employing all anchors. The concentrator obtains from the active anchor nodes the estimated channel gain vector

$$\hat{\boldsymbol{h}}^{(k)}(\mathtt{s}) = [\hat{h}_1^{(k)}(\mathtt{s}),\ldots,\hat{h}_N^{(k)}(\mathtt{s})]\,, \qquad k>0\,, \tag{3.7}$$

being non-zero only for corresponding entries of $\boldsymbol{c}(\mathtt{s},k)$ equal to one. If the actual transmitter is $\mathtt{s}$, then

$$\hat{h}_i^{(k)}(\mathtt{s})[\boldsymbol{c}(\mathtt{s},k)]_i \approx h_i(\mathtt{s})[\boldsymbol{c}(\mathtt{s},k)]_i, \qquad i=1,\ldots,N. \tag{3.8}$$

On the other hand, if the attacker is transmitting then

$$\hat{h}_i^{(k)}(\mathtt{s})[\boldsymbol{c}(\mathtt{s},k)]_i \approx g_i(\mathtt{s})[\boldsymbol{c}(\mathtt{s},k)]_i, \qquad i=1,\ldots,N. \tag{3.9}$$

The authentication performed by $\mathtt{c}$ on $\hat{\boldsymbol{h}}^{(k)}(\mathtt{s})$ must discern between two hypotheses:

- $\mathcal{H}_0$: packet k comes from $\mathtt{s}$,
- $\mathcal{H}_1$: packet k has been transmitted by the attacker $\mathtt{a}$.

The decision between the two hypotheses is taken by comparing estimates $\hat{\boldsymbol{h}}^{(k)}(\mathtt{s})$, $k>0$ with estimates $\hat{\boldsymbol{h}}^{(0)}(\mathtt{s})$. In the following, we will assume that the channel realization in two subsequent phases is subject to different fading (but still correlated), while the path loss remains constant, under the assumption that sources do not move between the two phases. Once a packet is deemed as authentic, the original estimate $\hat{\boldsymbol{h}}^{(0)}(\mathtt{s})$ is updated exploiting the newer one $\hat{\boldsymbol{h}}^{(k)}(\mathtt{s})$ in order to track channel variations over time. Moreover, we assume Additive White Gaussian Noise (AWGN).

Decision Process When the transmission is not performed by $\mathtt{s}$, we expect that the channel estimates of Phase 2 significantly differ from those of Phase 1. However, even when the transmission is actually performed by $\mathtt{s}$, the estimates in the two phases may differ due to occurred channel variations, noise and interference. Therefore, the decision process is prone to two well known types of errors:

- False Alarms (FAs), occurring when a legitimate packet is deemed as not being transmitted by $\mathtt{s}$, and
- Missed Detections (MDs), occurring when the impersonation attack succeeds, and the message coming from $\mathtt{a}$ is accepted as authentic.

The quality of the detection process is determined by the probabilities of these two events; note that in general a lower value of one yields a higher value of the other. The detection procedure that, for a given FA probability, minimizes the MD probability is the Likelihood Ratio Test (LRT). However, this approach requires the knowledge of the statistics of the channel of the attacker. Moreover, if $\mathtt{a}$ is able to forge the channels to the anchor nodes, the LRT technique requires the knowledge of the attacking strategy, i.e., vector $\boldsymbol{g}$. Since it is unrealistic to have such a knowledge, the LRT must be dropped in favor of the Generalized Likelihood Ratio Test (GLRT) [141], in which the knowledge of $\boldsymbol{g}$ is replaced by its Maximum Likelihood (ML) estimate $\hat{\boldsymbol{h}}^{(k)}(\mathtt{s})$.

This test works as follows. Let us assume for simplicity that all N observations from the anchor network are available, and denote with $f_{\hat{\boldsymbol{h}}^{(k)}(\mathtt{s})|\mathcal{H}_0}(\boldsymbol{a})$ the Probability Distribution Function (PDF) of $\hat{\boldsymbol{h}}^{(k)}(\mathtt{s})$ under hypothesis $\mathcal{H}_0$. Similarly, let $f_{\hat{\boldsymbol{h}}^{(k)}(\mathtt{s})|\mathcal{H}_1,\boldsymbol{g}}(\boldsymbol{a}|\boldsymbol{b})$ be the PDF of $\hat{\boldsymbol{h}}^{(k)}(\mathtt{s})$ under hypothesis $\mathcal{H}_1$ and given that $\boldsymbol{g}=\boldsymbol{b}$. The Log-Likelihood Ratio (LLR) of the estimated channel $\hat{\boldsymbol{h}}^{(k)}(\mathtt{s})$ is defined as[3]

$$\log \frac{f_{\hat{\boldsymbol{h}}^{(k)}(\mathtt{s})|\mathcal{H}_1,\boldsymbol{g}}(\hat{\boldsymbol{h}}^{(k)}(\mathtt{s})|\hat{\boldsymbol{h}}^{(k)}(\mathtt{s}))}{f_{\hat{\boldsymbol{h}}^{(k)}|\mathcal{H}_0}(\hat{\boldsymbol{h}}^{(k)}(\mathtt{s}))} \propto \frac{2}{\sigma^2}\|\hat{\boldsymbol{h}}^{(k)}(\mathtt{s})-\hat{\boldsymbol{h}}^{(0)}(\mathtt{s})\|^2 \triangleq \Psi\,, \tag{3.10}$$

where $\sigma^2=\mathrm{E}[\|\hat{\boldsymbol{h}}^{(k)}(\mathtt{s})-\hat{\boldsymbol{h}}^{(0)}(\mathtt{s})\|^2]$. According to the GLRT, the authenticity is established by comparing the LLR in Equation (3.10) (or its proportional variant Ψ) with a threshold θ as follows:

$$\text{if } \Psi \le \theta, \text{ then decide for } \mathcal{H}_0\,; \tag{3.11a}$$

$$\text{else if } \Psi > \theta, \text{ then decide for } \mathcal{H}_1\,. \tag{3.11b}$$

We note from Equation (3.10) that Ψ is a random variable depending both on the estimate accuracy and on the transmitting node (either $\mathtt{s}$ or $\mathtt{a}$). In particular, conditioned on $\mathcal{H}_0$ and for any realization of $\boldsymbol{h}(\mathtt{s})$, as shown in [138], Ψ is a central chi-squared distributed random variable with $2N$ degrees of freedom, yielding the FA probability

$$P_{\mathrm{FA}} = \mathbb{P}[\Psi > \theta|\mathcal{H}_0] = 1 - F_{2N,0}(\theta)\,, \tag{3.12}$$

[3]We use log for the natural (base-e) logarithm.

where $F_{n,y}(x)$ is the Cumulative Distribution Function (CDF) of a non-central chi-squared random variable with n degrees of freedom and non-centrality parameter y. On the other hand, conditioned on $\mathcal{H}_1$, specific realizations of $\boldsymbol{h}(\mathtt{s})$ and the forged vector $\boldsymbol{g}$, Ψ is a non-central chi-squared distributed random variable with $2N$ degrees of freedom and non-centrality parameter

$$\beta = \frac{2}{\sigma^2}\|\boldsymbol{g} - \boldsymbol{h}(\mathtt{s})\|^2 , \tag{3.13}$$

yielding the MD probability, i.e., the probability that the case in (3.11a) is verified when $\mathtt{a}$ is transmitting,

$$P_{\mathrm{MD}}(\boldsymbol{h}(\mathtt{s}), \boldsymbol{g}) = \mathbb{P}[\Psi \leq \theta | \mathcal{H}_1, \boldsymbol{h}(\mathtt{s}), \boldsymbol{g}] = F_{2N,\beta}(\theta) . \tag{3.14}$$

We observe that the MD probability depends on the attack channel vector $\boldsymbol{g}$, which is random because it depends on the attacker observations and on its attack strategy; therefore, for a probabilistic attack strategy, the average MD probability over the attack distribution is [138]

$$P_{\mathrm{MD}}(\boldsymbol{h}(\mathtt{s})) = \int_0^\infty F_{2N,x}(\theta) f_{\beta|\boldsymbol{h}(\mathtt{s})}(x|\boldsymbol{h}(\mathtt{s}))\, \mathrm{d}x . \tag{3.15}$$

For instance, if Rayleigh fading is assumed, and $\boldsymbol{h}(\mathtt{s})$ and $\boldsymbol{z}$ jointly Circularly Symmetric Complex Gaussian (CSCG) vectors, the optimal attack – both in the maximum MD probability sense of [138] and in the minimum divergence sense of [139] – is itself jointly CSCG with $\boldsymbol{h}(\mathtt{s})$ and $\boldsymbol{z}$ and can be written as

$$\boldsymbol{g} = \boldsymbol{\Xi}\boldsymbol{z} = \boldsymbol{\Omega}\boldsymbol{h}(\mathtt{s}) + \boldsymbol{\epsilon} , \tag{3.16}$$

with $\boldsymbol{\Xi}$ and $\boldsymbol{\Omega}$ complex matrices, and $\boldsymbol{\epsilon}$ a zero mean CSCG vector independent of $\boldsymbol{h}(\mathtt{s})$. Under the assumption of (3.5), both the matrices $\boldsymbol{\Xi}$ and $\boldsymbol{\Omega}$ in (3.16), as well as the covariance matrices of $\boldsymbol{z}$ and $\boldsymbol{\epsilon}$ are diagonal (see [138, App. A] or [139, Sect. V]), thus we can write

$$g_i = \xi_i z_i = \omega_i h_i(\mathtt{s}) + \epsilon_i \tag{3.17}$$

where

$$\omega_i = \xi_i \rho_{z_i h_i(\mathtt{s})} \sigma_{z_i} / \sigma_{h_i(\mathtt{s})} , \tag{3.18}$$

$$\sigma^2_{\epsilon_i} = |\xi_i|^2 \sigma^2_{z_i} (1 - |\rho_{z_i h_i(\mathtt{s})}|^2) , \tag{3.19}$$

while $\sigma_{h_i(\mathtt{s})}$ and σ_{z_i} represent the standard deviations of $h_i(\mathtt{s})$ and z_i, respectively.

It is worth assessing the average MD probability when the optimal attack is performed and the channel $\boldsymbol{h}(\mathtt{s})$ is Gaussian distributed with independent and identically distributed (i.i.d.) entries, i.e., $P_{\mathrm{MD}} = \mathrm{P}[\Psi \leq \theta | \mathcal{H}_1]$. This measure is relevant when the sequence of transmitted messages is long enough to span a significant portion of the channel fading statistics[4]. In the case of N independent observations, $\beta = 2\sum_i |(1-\omega_i)h_i(\mathtt{s}) + \epsilon_i|^2/\sigma^2$ becomes the sum of N independent exponentially distributed random variables, each with mean [138, App. A]

$$\frac{1}{\zeta_i} = \frac{2}{\sigma^2}(|1 - b_i|^2 \lambda_i + \sigma^2_{\epsilon_i}) = \frac{2}{\sigma^2}(1 - |\rho_{z_i h_i(\mathtt{s})}|^2)\lambda_i . \tag{3.20}$$

[4]We remark that fading is independent on each phase, thus the MD probability is averaged over the fading. The case of constant fading over the phases can be addressed by a similar approach but leads to hardly tractable expressions.

Then, under the simplifying assumption[5] that the ζ_i are all distinct, the average MD probability is

$$P_{\mathrm{MD}} = 2\sum_{i=1}^{N} \zeta_i \left(\prod_{j \neq i} \frac{1}{1-\zeta_i/\zeta_j} \right) \left[\sum_{m=0}^{\infty} \frac{\bar{\gamma}(N+m;\theta/2)}{(2\zeta_i+1)^{m+1}} \right], \tag{3.21}$$

where $\bar{\gamma}(r;a) = \frac{1}{\Gamma(r)} \int_0^a x^{r-1} e^{-x}\, \mathrm{d}x$ denotes the normalized lower incomplete Gamma function. Observe that, since $F_{2N,x}(\theta)$ is a decreasing function of x for every θ, and the CDF of β is a decreasing function of each λ_i once $\rho_{z_i h_i(\mathtt{s})} = \rho$ is kept fixed, P_{MD} is itself a decreasing function of each λ_i for a given ρ. In other words, better legitimate channel gains yield a lower probability of confusing an attacker as a legitimate source.

Finally, let us remark that, if we do not exploit the whole anchor network, rather we consider only a sub-set of estimates from a certain configuration of anchors $\boldsymbol{c}(\mathtt{s},k)$, the FA and MD probability expressions become, respectively,

$$P_{\mathrm{FA}} = 1 - F_{2L(\mathtt{s},k),0}(\theta)\,, \tag{3.22}$$

$$P_{\mathrm{MD}} = 2\sum_{i=1}^{N} [\boldsymbol{c}(\mathtt{s},k)]_i \zeta_i \left[\prod_{j \neq i} \frac{1}{(1-\zeta_i/\zeta_j)^{[\boldsymbol{c}(\mathtt{s},k)]_i}} \right] \left[\sum_{m=0}^{\infty} \frac{\bar{\gamma}(L(\mathtt{s},k)+m;\theta/2)}{(2\zeta_i+1)^{m+1}} \right], \tag{3.23}$$

where $L(\mathtt{s},k) = \|\boldsymbol{c}(\mathtt{s},k)\|_H$ is the Hamming weight of vector $\boldsymbol{c}(\mathtt{s},k)$, i.e., the number of active anchor nodes in the selected configuration.

3.1.4 Anchor Node Selection Criteria

In this context, it is reasonable to assume that most of the energy cost of the anchor nodes comes from their transmission of the authentication packets to the concentrator node $\mathtt{c}$. Therefore, while on the one hand we would like to exploit as many anchor nodes as possible to decrease the MD probability, on the other hand, in a scenario in which the anchor nodes are battery-powered devices, it is important to optimize their usage. Thus, the envisioned optimization procedure must, at the same time, ensure an accurate message authentication, and on the other hand reduce the power consumption of the anchor nodes. Our objective is to minimize the utilization of the trusted anchor nodes in the network for authentication purposes, when there are M sources in the system, by using the configurations (i.e., sub-sets) of anchors.

Let us observe first that, with N anchor nodes, $2^N - 1$ configurations are possible for source $\mathtt{s}$, and let us denote the ℓ-th available configuration with the binary vector $\boldsymbol{c}_\ell(\mathtt{s})$, of Hamming weight $L_\ell(\mathtt{s})$. For example, if $N = 4$, vector $\boldsymbol{c}_8(\mathtt{s}) = [1\,0\,0\,0]^T$ denotes the eighth configuration for the authentication of node $\mathtt{s}$, where anchor #1 is active, while anchors #2, #3, and #4 are not active. Not all configurations are suitable to authenticate $\mathtt{s}$, however: the selected anchor

[5] This assumption in made only for the sake of obtaining a more compact expression in (3.21). If it does not hold, the PDF of β can be derived with a slight complication as described in [142].

nodes must satisfy some performance constraints on the FA and MD probabilities of the authentication process. Denoting the target values with $P^{\star}_{\mathrm{FA}}$ and $P^{\star}_{\mathrm{MD}}$, one should require

$$P_{\mathrm{FA}} \leq P^{\star}_{\mathrm{FA}} \quad \text{and} \quad P_{\mathrm{MD}} \leq P^{\star}_{\mathrm{MD}} \tag{3.24}$$

for all transmitting sources. Those configurations that satisfy the constraints in Equation (3.24) are denoted as *admissible configurations*.

Moreover, since we aim at selecting admissible configurations that yield longer network lifespan (which is related to the anchor node usage), we observe that if configuration $\boldsymbol{c}_\ell(\mathtt{s})$ is admissible and anchor node i is not active ($[\boldsymbol{c}_\ell(\mathtt{s})]_i = 0$), the configuration $\boldsymbol{c}'(\mathtt{s})$ obtained by activating node i ($[\boldsymbol{c}'(\mathtt{s})]_i = 1$, and $[\boldsymbol{c}'(\mathtt{s})]_j = [\boldsymbol{c}_\ell(\mathtt{s})]_j \ \forall j \neq i$) yields additional power consumption while still being admissible. Therefore, the newly obtained configuration $\boldsymbol{c}'(\mathtt{s})$ is worse than the original one $\boldsymbol{c}_\ell(\mathtt{s})$ in terms of energy consumption. Then, we want to consider only *efficient* admissible configurations, i.e., the admissible configurations with a minimal set of active nodes.

Let $a_{\mathtt{s}}$ be the number of efficient admissible configurations for source node $\mathtt{s}$, and denote each configuration as $\boldsymbol{c}_\ell(\mathtt{s})$, $\ell = 1, \ldots, a_{\mathtt{s}}$. We can collect all efficient admissible configurations into the $N \times A$ binary matrix

$$\boldsymbol{C} = [\ \boldsymbol{c}_1(1)\ \cdots\ \boldsymbol{c}_{a_1}(1)\ \cdots\ \boldsymbol{c}_1(M)\ \cdots\ \boldsymbol{c}_{a_M}(M)\ ]\ , \tag{3.25}$$

where

$$A = \sum_{m=1}^{M} a_m \tag{3.26}$$

is the total number of efficient admissible configurations. Note that, thanks to the definition of efficient admissible configuration, A is much smaller than the total amount of possible configurations, i.e., $M \times (2^N - 1)$.

Now, we observe that the activity of each source node is random; therefore, even if we deterministically select a configuration to authenticate each user, the utilization of the anchor nodes is a random variable. In order to further balance the usage of the anchor nodes, we propose to randomize the choice of the configuration for the authentication of each source position, by

1. assigning a probability distribution to the admissible configurations and
2. randomly and independently selecting the configuration to be employed for each authentication of that source according to the assigned distribution.

In this way, the minimization of the usage of the anchor nodes consists in finding a suitable probability distribution for the admissible configurations used for the verification of each source position. Let $\pi_\ell(\mathtt{s})$ be the probability (or, equivalently, the fraction of times) that configuration $\boldsymbol{c}_\ell(\mathtt{s})$ is used and let us stack these probabilities into the A-size column vector

$$\boldsymbol{\pi} = [\ \pi_1(1)\ \cdots\ \pi_{a_1}(1)\ \cdots\ \pi_1(M)\ \cdots\ \pi_{a_M}(M)\ ]^T\ , \tag{3.27}$$

where $(\cdot)^T$ denotes the transpose operator.

The proposed authentication protocol works as follows: when a packet (presumably) coming from source node $\mathtt{s}$ is received by the concentrator node $\mathtt{c}$, the latter draws a configuration according to $\boldsymbol{\pi}$ and sends it in broadcast to the anchor nodes, triggering the participation of selected anchors. Then, the authentication process continues as described in Section 3.1.3.

Anchor Network Lifespan

As a metric to guide the choice of the probabilities and to assess the performance of the proposed methods, we consider the *anchor network lifespan*, defined as the smallest number of authentication processes after which at least one anchor node runs out of power. The underlying assumption is that if any anchor node is no longer available, there will be some source position for which no efficient admissible configuration exists, therefore, no reliable authentication can be performed.[6]

Since the choice of the configuration is random, the network lifespan is a random variable, too. In order to derive its statistical description, let us denote with $y_i(k)$, $i = 1, \ldots, N$, $k \geq 1$, the number of message authentications performed by anchor node i out of the first k authentications performed by the network. We recall that each anchor can perform up to Q authentications before running out of energy. The CDF of the random lifespan $\mathcal{L}$ of the anchor network can be written as

$$F_{\mathcal{L}}(k) \triangleq \mathbb{P}[\mathcal{L} \leq k] = \mathbb{P}[\max_i y_i(k) > Q]\,. \tag{3.28}$$

The evaluation of the above expression requires the joint distribution of $y_i(k)$, which is fairly complicated by the correlations introduced by the specific set of efficient admissible configurations. However, denoting with u_i the probability of using anchor node i, one can easily see that the marginal distribution of each $y_i(k)$ is binomial with parameters (k, u_i). Then, we can upper and lower bound the CDF in (3.28) as

$$\max_i \mathbb{P}[y_i(k) > Q] \leq \mathbb{P}[\max_i y_i(k) > Q] \leq \sum_i \mathbb{P}[y_i(k) > Q] \tag{3.29}$$

and hence

$$I_{\max_i u_i}(Q+1, k-Q) \leq F_{\mathcal{L}}(k) \leq \sum_i I_{u_i}(Q+1, k-Q)\,, \tag{3.30}$$

where $I_x(a,b) \triangleq B(x;a,b)/B(1,a,b)$ is the regularized incomplete beta function, $B(x;a,b) = \int_0^x t^{a-1}(1-t)^{b-1}\,\mathrm{d}t$.

Observe that the term on the left is the CDF of a *negative binomial* random variable, thus we can obtain an upper bound on the expected network lifespan by integrating the Complementary CDF (CCDF) as follows:

$$\mathbb{E}[\mathcal{L}] \triangleq \sum_k [1 - F_{\mathcal{L}}(k)] \leq \sum_k [1 - I_{\max_i u_i}(Q+1, k-Q)] = \frac{Q+1}{\max_i u_i}\,, \tag{3.31}$$

where the last equality is given by the mean of the negative binomial random variable.

However, the bounds in (3.30) and (3.31) may be rather loose, as will be seen in Section 3.1.6, so we will also resort to the approximation of $F_{\mathcal{L}}(k)$ that can be obtained by neglecting the statistical dependence among $y_1(k), \ldots, y_N(k)$, that is

$$F_{\mathcal{L}}(k) \simeq 1 - \prod_{n=1}^{N} \mathbb{P}[y_i(k) > Q] = 1 - \prod_{n=1}^{N} [1 - I_{u_i}(Q+1, k-Q)]\,. \tag{3.32}$$

[6] If by taking off some node there still exists an efficient admissible configuration, the definition of the network lifespan can be easily modified and the derivations of this section are easily adjusted (see Section 3.1.5).

Although not justified, the above approximation is seen to be quite good from the numerical results in Section 3.1.6.

3.1.5 Configuration Probability Optimization

We now provide three possible methods to compute vector $\boldsymbol{\pi}$. First, let $\boldsymbol{u}$ be a N-size column vector with i-th entry u_i. Assume that ϕ_m is the probability that source node $m \in \{1, \dots, M\}$ is transmitting, therefore $\sum_{m=1}^{M} \phi_m = 1$. Let us define the $A \times A$ diagonal matrix $\boldsymbol{\Phi}$ that weights the admissible configurations by the probabilities that the corresponding transmitter is active, i.e.,

$$\boldsymbol{\Phi} = \begin{bmatrix} \phi_1 \cdot \boldsymbol{I}_{a_1} & 0 & 0 & 0 \\ 0 & \phi_2 \cdot \boldsymbol{I}_{a_2} & 0 & 0 \\ & & \ddots & \\ 0 & 0 & & 0 \\ 0 & 0 & 0 & \phi_M \cdot \boldsymbol{I}_{a_M} \end{bmatrix}, \tag{3.33}$$

where $\boldsymbol{I}_n$ is the identity matrix of size $n \times n$. Then, $\boldsymbol{u}$ can be written as

$$\boldsymbol{u} = \boldsymbol{C}\boldsymbol{\Phi}\boldsymbol{\pi}\,. \tag{3.34}$$

In most cases, however, it is reasonable to assume that each source is transmitting with the same probability $\phi_m = 1/M \ \forall m$; in this case, we have

$$\boldsymbol{u} = \frac{1}{M}\boldsymbol{C}\boldsymbol{\pi}\,. \tag{3.35}$$

Upper Bound Maximization

Since the lower bound on the expected lifespan (3.31) is inversely proportional to the maximum value of u_i, a first approach for the optimization of $\boldsymbol{\pi}$ is the minimization of $\max_i u_i$, under the constraint that only efficient admissible configurations are used each time. The optimization problem can then be written as follows:

$$\min_{\boldsymbol{\pi}} \max_{i} u_i \tag{3.36a}$$

subject to (3.27), (3.35), and

$$0 \le \pi_\ell(m) \le 1\,, \quad \ell = 1, \dots, a_m\,,\ m = 1, \dots, M\,, \tag{3.36b}$$

$$\sum_{\ell=1}^{a_m} \pi_\ell(m) = 1\,, \quad m = 1, \dots, M\,. \tag{3.36c}$$

We remark that the constraint (3.36c) ensures that for each source node that needs to be authenticated there is always a configuration that can be used. If by taking off some node there still exists an efficient admissible configuration, the optimization problem can be easily fixed by ignoring the index of that node in the maximization. The proposed min-max problem (3.36) can be solved as a linear programming problem

$$\min_{\boldsymbol{\pi}, t} t \tag{3.37a}$$

subject to (3.27), (3.35), (3.36b), (3.36c), and

$$\frac{1}{M}\boldsymbol{C}\boldsymbol{\pi} \leq t\mathbf{1}_{N\times 1}\,, \tag{3.37b}$$

where $\mathbf{1}_{N\times 1}$ is an N-size column vector with entries all equal to 1.

We must observe that by solving the min-max problem we are maximizing an upper bound on the average anchor network lifetime, or its CCDF; however, maximizing the bound does not necessary correspond to maximize the bounded value (average or CCDF). Therefore, we explore two other possible methods to choose the configuration probability, based on the minimization of the variance of u_i or on the minimization of their power, respectively.

Minimum Variance Optimization

As we will see in the following of this section, in most cases the optimized node utilizations u_i are almost constant, i.e., $u_i \approx u_j\ \forall i,j$. This is also intuitive if one thinks that, starting from a feasible solution, we can reduce the utilization of the most used node by increasing the probability of efficient admissible configurations that do not contain that node, thus increasing the utilization of other anchor nodes. Therefore, let us choose $\boldsymbol{\pi}$ in order to minimize the variance of u_i:

$$f(\boldsymbol{\pi}) = \sum_{i=1}^{N}\left(u_i - \frac{1}{N}\sum_{j=1}^{N}u_j\right)^2\,. \tag{3.38}$$

Denoting by $\mathbf{1}_{N\times N}$ the $N\times N$ matrix containing all entries equal to 1, $f(\boldsymbol{\pi})$ can be expressed in matrix form as follows:

$$f(\boldsymbol{\pi}) = ||\boldsymbol{C}\boldsymbol{\pi} - \frac{1}{N}\mathbf{1}_{N\times N}\boldsymbol{C}\boldsymbol{\pi}||^2 = ||\boldsymbol{C}\boldsymbol{A}\boldsymbol{\pi}||^2 = \boldsymbol{\pi}^T\boldsymbol{C}^T\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{C}\boldsymbol{\pi} = \boldsymbol{\pi}^T\boldsymbol{C}^T\boldsymbol{A}\boldsymbol{C}\boldsymbol{\pi}\,, \tag{3.39}$$

where $\boldsymbol{A} \triangleq \boldsymbol{I}_N - \frac{1}{N}\mathbf{1}_{N\times N}$ with $\boldsymbol{I}_N$ the $N\times N$ identity matrix, is a symmetric and idempotent matrix. The problem of minimizing (3.38) can now be written as

$$\min_{\boldsymbol{\pi}} \boldsymbol{\pi}^T\boldsymbol{C}^T\boldsymbol{A}\boldsymbol{C}\boldsymbol{\pi} \tag{3.40}$$

subject to (3.27), (3.35), (3.36b), and (3.36c). Note that the objective function of problem (3.40) is convex and constraints (3.27), (3.35), (3.36b), and (3.36c) are affine transformations. The optimization problem is convex and can be solved using well-known techniques such as the interior point method.

Least Squares Optimization

With the minimum variance optimization, we aim at making all utilization probabilities similar to each other, however, we do not explicitly minimize the average node utilization probabilities. Therefore, as a third optimization method, we consider the minimization of the sum of the square probabilities of utilization across the anchor nodes, i.e.,

$$\min_{\boldsymbol{\pi}} \sum_{i=1}^{N} u_i^2 = \min_{\boldsymbol{\pi}} \boldsymbol{\pi}^T\boldsymbol{C}^T\boldsymbol{C}\boldsymbol{\pi} \tag{3.41}$$

subject to (3.27), (3.35), (3.36b), and (3.36c). Also in this case the convexity of the objective function and the affine nature of the constraints make the problem easily solvable.

3.1.6 Baseline Authentication Protocol vs Energy-Efficient Anchor Selection: Performance Comparison

We will now evaluate the performance of the proposed authentication procedures, in particular, we will compare the performance of the baseline authentication protocol (using all the available anchor nodes) with that of the proposed algorithms to compute $\boldsymbol{\pi}$. We consider a CIoT scenario, with a single cell having a circular shape of radius 500 m. Anchor nodes are placed on a regularly-spaced grid inside the circle, and we consider various densities of the anchor nodes, namely 4, 9, 16, or 25 nodes in the circle. Transmissions are performed with unitary power. The deterministic component of the wireless channel, i.e., the path loss, is computed as (in dB)

$$[\lambda(\eta, d)]_{\mathrm{dB}} \triangleq -10\eta \log_{10}\left(\frac{4\pi d}{\Lambda}\right), \tag{3.42}$$

where η is the Path Loss Exponent (PLE), d is the distance between the transmitter and the receiver, Λ is the wavelength, defined as the ratio between the speed of light and the carrier frequency f. We assume that $f = 900$ MHz, which is the typical carrier frequency value considered in the context of CIoT [28]. η is in $[2, 3]$, which is a reasonable assumption for the radio-wave propagation in an urban scenario: we recall that as η increases, the propagation environment becomes harsher. Finally, we set as target FA probability $P_{\mathrm{FA}}^{\star} = 10^{-4}$ for the authentication method.

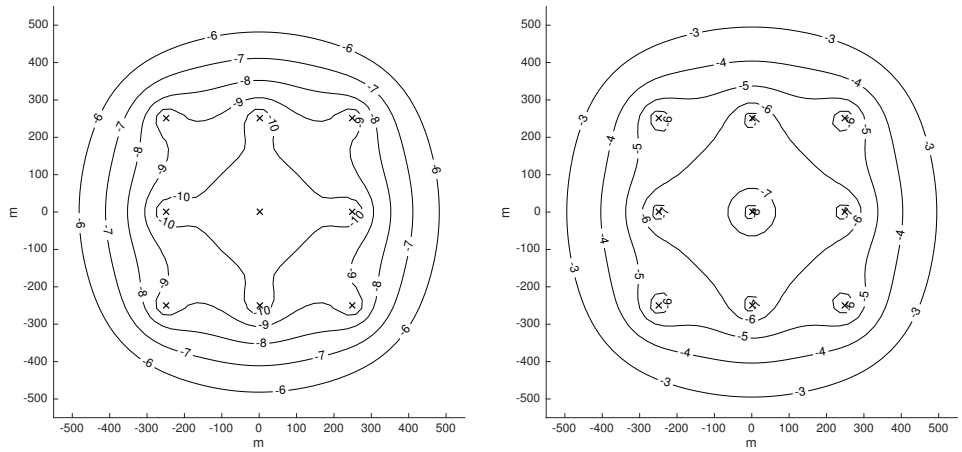
Missed Detection Probability of the Baseline Protocol

Figure 3.2 shows (in log scale) the average (with respect to noise and channel realization) MD probability as a function of the legitimate source node position, with $N = 9$ anchor nodes and a correlation factor for the attacker node $\rho \in \{0.1, 0.5\}$. The PLE is set to $\eta = 2$. The Signal-to-Noise Ratio (SNR), defined as the average (over fading) power ratio for a sensor-anchor distance of 250 meters, is 15 dB. We observe that the positions at the center of the circle provide a lower MD probability, since the average channel gain sensed by the anchor nodes is higher than for external positions, especially as the source node moves to the circle border.

We also investigate the impact of a different number of anchor nodes N and PLE η on the performance of the proposed authentication protocol. Figure 3.3 shows the CCDF of the MD probability, considering different fading and source node position realizations. Note that P_{MD} decreases as the number of anchor nodes increases. On the other hand, a higher PLE leads to an increase of MD probability as the signal power received at the anchor nodes is decreased.

Enhancement of Anchor Network Lifespan: a Comparison

We now assess the performance of the proposed authentication method and the various approaches for the choice of the configuration probabilities in terms



(a) $\rho = 0.1$ **(b)** $\rho = 0.5$

Figure 3.2: Logarithm of the MD probability as a function of legitimate source node position for two values of ρ, $N = 9$, $\mathsf{SNR} = 15$ dB at a distance of 250 m

of the anchor network lifespan. We focus on the case with $N = 9$ anchor nodes, $M = 10$ source nodes, $\eta = 2$, $\rho = 0.1$, and $\mathsf{SNR} = 30$ dB.

Figures 3.4, 3.5, and 3.6 show the CDF of the network lifespan $\mathcal{L}$ for the min-max, the minimum variance and the minimum power methods introduced in Section 3.1.5. We report the CDF of the empirical lifespan obtained by Monte Carlo simulation, together with the upper and lower bond of the CDF (see (3.30)). Moreover, we report the approximation of the CDF obtained by assuming independent anchor node usage (see (3.32)), indicated with label "independent" in the figures. It can be observed that the lower bound has a quite loose performance, while both the independent approximation and the upper bound are quite close to the empirical CDF. When comparing the various optimization methods, we observe that they perform approximately the same. In order to better assess the differences among the methods, Figure 3.7 reports the empirical CDF for the various optimization methods. We observe that the min-max optimization provides the highest anchor network lifespan, while the minimum variance and minimum power approaches have different behaviors at different outage probabilities.

We have just found out that the best optimization strategy consists in solving the min-max problem given by (3.36). Let us show now the effective anchor usage of a given deployment of source nodes.

Example Let us consider the parameters reported in Table 3.1 and the IoT network deployment of Figure 3.1. For a single realization of the source nodes deployment, Figure 3.8 shows the anchor node utilization probabilities $u_i \ \forall i$ after optimization. As expected, we observe that all anchor nodes are used on average with a similar probability. Moreover, if we compare it with the case in which all anchors are always used ($u_i = 1 \ \forall i$) we note a sharp decrease of the anchor node utilization probability to 0.12, yielding ten-times the network lifespan.

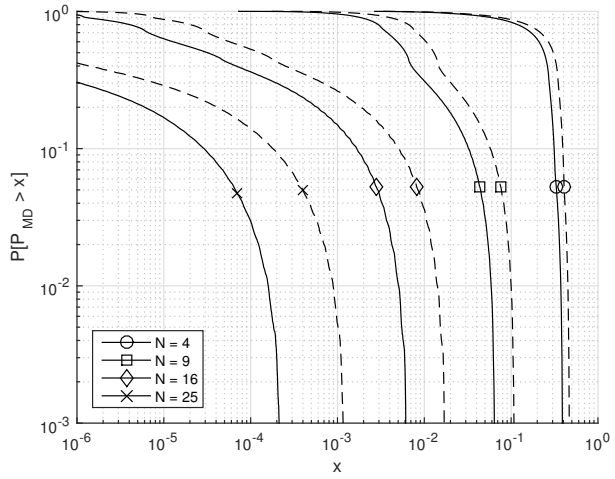


Figure 3.3: CCDF of MD probability for various amounts of anchor nodes N and PLE values $\eta = 2$ (solid lines) and $\eta = 2.5$ (dashed lines)

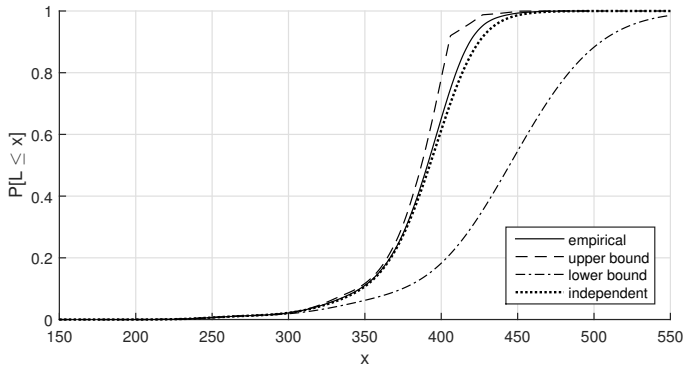


Figure 3.4: Empirical CDF and bounds of the anchor network lifespan $\mathcal{L}$ using the configuration probability vector $\boldsymbol{\pi}$ obtained solving the min-max problem

3.1.7 Signaling-Efficient Anchor Selection

By solving the min-max problem (3.36), we aim at distributing the burden of the authentication procedure between all the anchor nodes, so that their utilization is as similar as possible and, therefore, the lifespan of the authenticating network is maximized. However, to achieve this objective, the optimization procedure may employ configurations involving multiple anchor nodes, i.e., configurations with a high Hamming weight, resulting in a heavy amount of signaling traffic. In the following of this section, we will provide a method to minimize the Hamming weight $L_\ell(m)$ $\forall \ell, m$ of the admissible configurations, regardless the overall utilization probability of the anchor nodes.

Recalling that anchor nodes having a better channel can yield a lower MD probability, for each source node we order the anchor nodes with decreasing channel gain, and we run the ordered list adding anchor nodes to the configuration until we find a configuration that satisfies the target FA probability. With this

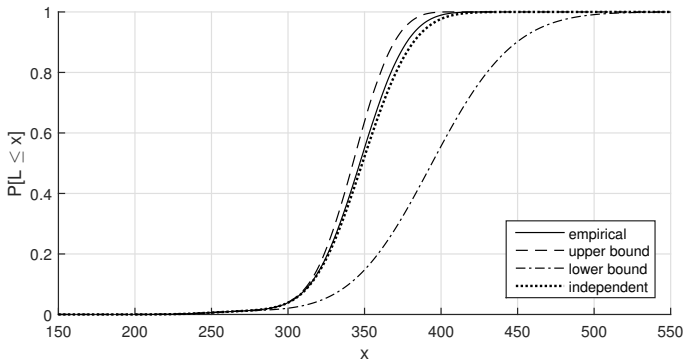


Figure 3.5: Empirical CDF and bounds of the anchor network lifespan $\mathcal{L}$ using the configuration probability $\boldsymbol{\pi}$ obtained solving the minimum variance problem

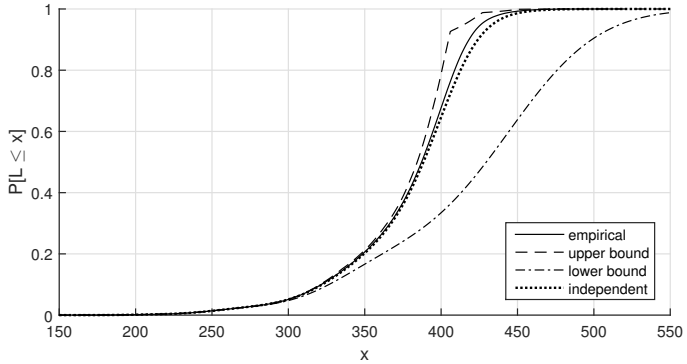


Figure 3.6: Empirical CDF and bounds of the anchor network lifespan $\mathcal{L}$ using the configuration probability $\boldsymbol{\pi}$ obtained solving the least squares problem

approach, no min-max algorithm must be solved and we have a single broadcast message from the concentrator node c to indicate the selected configuration, i.e., we minimize the signaling traffic due to authentication purposes. Note that this SNR-based anchor node selection does not take into account the energy consumption of the anchors. Indeed, while the minimum number of nodes to achieve authentication is used, it may occur that, e.g., in case the source node distribution is not uniform, some anchor nodes are often activated while others are never activated, thus limiting the anchor network lifespan. Therefore, this technique does not maximize the network lifespan in general and is suboptimal with respect to the solution of (3.36) from the point of view of the energy consumption.

Example Let us consider the network scenario with parameters reported in Table 3.1 and deployment of Figure 3.1. Figure 3.9 shows the anchor node utilization probabilities $u_i \ \forall i$ when the SNR-based anchor selection method is used. It can be seen that the utilization is not equal among the anchor nodes: some of them are never used, while others remain active most of the time to authenticate source nodes. The maximum anchor node utilization probability is

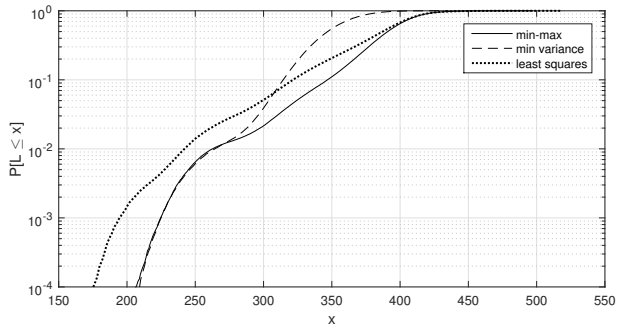


Figure 3.7: Empirical CDF of the anchor network lifespan $\mathcal{L}$ for the various optimization methods

Table 3.1: Simulation parameters for the reference scenario

Parameter	Value
f	900 MHz
η	2
ρ	0.5
Cell radius	500 m
SNR	30 dB (at 250 m)
N	16
M	10
Δ	0.1
$P^{\star}_{\mathrm{FA}}$	10^{-4}
$P^{\star}_{\mathrm{MD}}$	10^{-4}

about 0.27, providing a network lifespan which is 45% of the maximum lifespan obtained solving problem (3.36). Still, when compared to the case in which all anchor nodes are always used for authentication, we have a four-times longer network lifespan.

3.1.8 A Trade-Off Between Energy Efficiency and Signaling Efficiency

A possible way to reduce the amount of authentication overhead and, at the same time, to balance the anchor node utilization consists in limiting the Hamming weight $L_{\ell}(m)\ \forall \ell, m$ of the admissible configurations in matrix $\boldsymbol{C}$. Two distinct approaches can be identified:

- put a *hard* constraint on the Hamming weight of the configurations, modifying the computation of matrix $\boldsymbol{C}$ by upper bounding the Hamming weight of every configuration;

- set a *soft* constraint on the Hamming weight of the configurations, modifying the min-max problem formulation in order to upper bound the average Hamming weight of the admissible configurations.

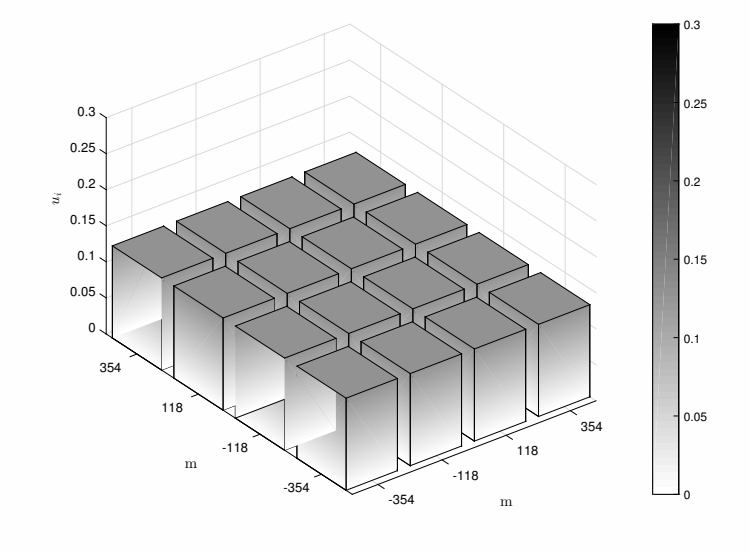


Figure 3.8: Anchor utilization probabilities u_i $\forall i$, solving the min-max problem, as for the IoT deployment in Figure 3.1

Min-Max Problem with Hard Constraint

This strategy consists in modifying the construction of matrix $\boldsymbol{C}$ so that it contains only admissible configurations $\boldsymbol{c}_\ell(m)$ with Hamming weight $L_\ell(m) \leq K$, where

$$K \geq \max_m \min_\ell L_\ell(m) \tag{3.43}$$

is the minimum feasible Hamming weight for an efficient admissible configuration, yielding a *reduced* configuration matrix $\boldsymbol{C}'$. We remark that a particular value K_m for every source m can also be set, provided that $K_m \geq \min_\ell L_\ell(m)$ $\forall m$. The problem formulation is the following:

$$\min_{\boldsymbol{\pi}} \max_i u_i \tag{3.44a}$$

subject to (3.27),

$$\boldsymbol{u} = \boldsymbol{C}'\boldsymbol{\Phi}\boldsymbol{\pi} \tag{3.44b}$$

and (3.36b), (3.36c).

Note that the SNR-based anchor selection policy presented in Section 3.1.7 is a special case of problem (3.44), where a single configuration of anchor nodes is allowed for each source node $\mathbf{s}$. Also in this case the problem can be linearized for an efficient solution.

Example Let us consider the network scenario with parameters reported in Table 3.1 and deployment of Figure 3.1. Figure 3.10 shows the anchor node utilization probabilities u_i $\forall i$ when the min-max problem with hard constraints (3.44) is solved, imposing $K_m = \min_\ell L_\ell(m)$ $\forall m$. We observe that in this deployment example we achieve a similar utilization probability among the anchor nodes with respect to the original min-max strategy in (3.36). However, the maximum usage is slightly higher than the min-max.

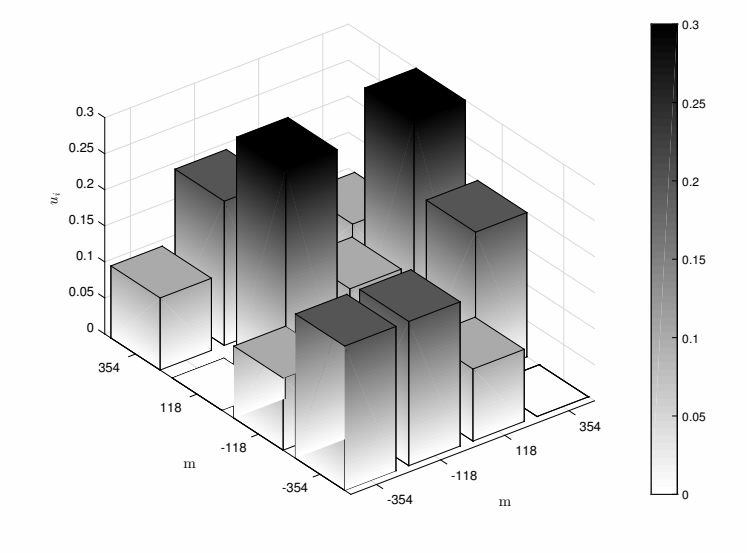


Figure 3.9: Anchor node utilization probabilities u_i $\forall i$, employing the SNR-based approach, as for the IoT deployment in Figure 3.1

Min-Max Problem with Soft Constraint

This approach consists in imposing a constraint on the *average* Hamming weight of the admissible configurations. Note that now we are shaping the structure of the vector of the admissible configuration probabilities $\boldsymbol{\pi}$, rather than the intrinsic structure of matrix $\boldsymbol{C}$, as done in problem (3.44). The proposed optimization problem is the following:

$$\min_{\boldsymbol{\pi}} \max_{i} u_i \tag{3.45a}$$

subject to (3.36b), (3.36c), (3.35), and

$$\sum_{\ell=1}^{a_m} L_\ell(m)\pi_\ell(m) \leq K_m\,, \qquad m = 1, \ldots, M\,. \tag{3.45b}$$

We denote with Δ the fixed lag that exceeds the minimum Hamming weight $\min_\ell L_\ell(m)$ $\forall m$. Note that if we impose that $\Delta = 0$, then the problems (3.45) and (3.44) provide the same solution.

Example For the usual network scenario with parameters reported in Table 3.1 and deployment of Figure 3.1, Figure 3.11 shows the anchor node utilization probability u_i obtained solving the optimization problem (3.45) for a single realization of the source nodes deployment, with $\Delta = 0.1$. It can be seen that this method provides a more flat utilization probability among the anchor nodes with respect to the solution of problem (3.44). Moreover, note that in this case the maximum utilization probability is greater than the maximum utilization probability of problem (3.36) and lower than the utilization probability of problem (3.44).

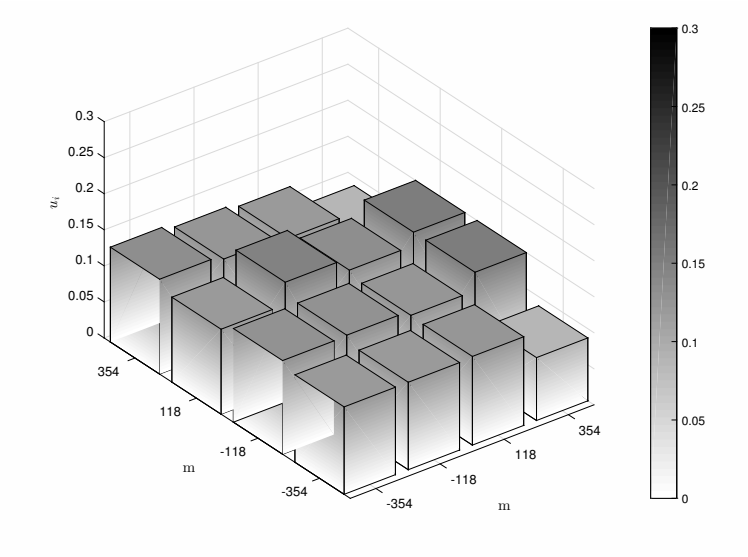


Figure 3.10: Anchor node utilization probabilities u_i $\forall i$, solving problem (3.44), as for the IoT deployment in Figure 3.1

3.1.9 Distributed Anchor Node Selection

The optimization procedures of Section 3.1.5 are centralized because

- the concentrator node c collects all the channel estimates of Phase 1 to build the matrix $\boldsymbol{C}$;
- the concentrator node c solves the optimization problem;
- for each packet transmission the concentrator node c sends N control messages, one per anchor node, indicating which configuration is selected[7].

Therefore, the centralized procedure requires intense exchange of control messages from the concentrator c to the anchor nodes.

We will propose now decentralized solutions that have reduced message exchange requirements are considered.

Distributed Configuration Selection

A first way of distributing the centralized procedure consists in eliminating the transmission of the control message from the concentrator node to the anchor nodes. In this solution, we assume that the messages from the anchor nodes to the concentrator node can be overheard by all anchor nodes, e.g., because they are transmitted over the wireless medium or because they all go through a wired bus. To this end, we must consider a preliminary operation (performed at the end of Phase 1), in which the concentrator node c sends the admissible configurations matrix $\boldsymbol{C}$ and the optimal configurations usage probability vector $\boldsymbol{\pi}$ in multicast to all anchor nodes.

[7]When the same control message can reach more nodes simultaneously (e.g., in a wireless broadcast scenario), the concentrator c can send a single message indicating the configuration.

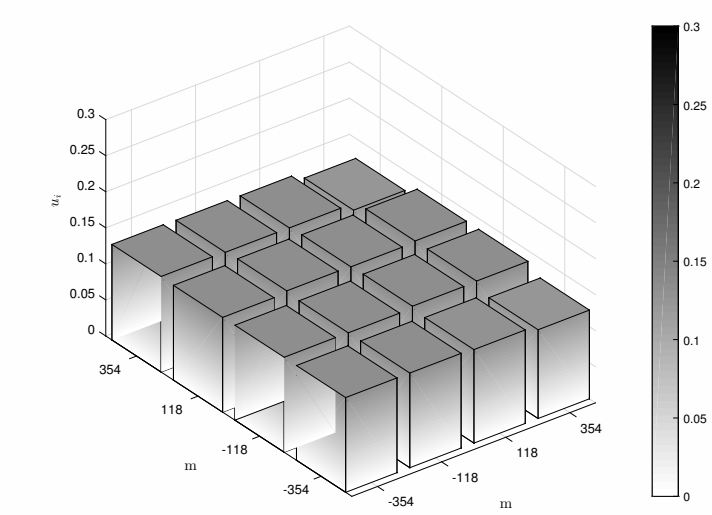


Figure 3.11: Anchor node utilization probabilities u_i $\forall i$, solving problem (3.45), as for the IoT deployment in Figure 3.1

Then, in Phase 2 upon the transmission of each data packet, a round-robin procedure is used to select the desired configuration in a distributed fashion. In this procedure, anchor nodes have each a fixed time slot assigned in which they can provide their channel estimate to the concentrator. The first anchor node, i.e., the anchor node that owns the first slot, provides the feedback with probability u_1. The other anchor nodes in the meantime listen to the control channel and are able to detect whether anchor node 1 transmits or not. Anchor node 2 then selects among the admissible configurations those that match the initial behavior of node 1, and decides whether to transmit or not according to the probability of transmission conditioned on the transmission of anchor node 1. The third anchor node overhears what happens in the two previous slots and again transmits with a probability that is determined by the subset of admissible configurations that have been identified by the behavior of the two anchor nodes.

In general, to authenticate node m, at slot τ, define $\mathcal{C}_\tau$ as the set of configurations that are compatible with the transmissions performed in the previous $\tau - 1$ slots and that have a non-zero probability of being used. Moreover, let $\mathcal{R}_\tau$ be the set of configurations in $\mathcal{C}_\tau$ in which node i is active, i.e.,

$$\mathcal{R}_\tau = \{\ell \in \mathcal{C}_\tau : [c_\ell(m)]_i = 1\}\,. \tag{3.46}$$

Then, anchor node i will transmit with probability

$$p_{\text{tx}} = \frac{\sum_{\ell \in \mathcal{R}_\tau} \pi_\ell(m)}{\sum_{\ell \in \mathcal{C}_\tau} \pi_\ell(m)}\,. \tag{3.47}$$

Example A simple example is now provided to better understand the proposed approach. Let us consider the configuration matrix $\boldsymbol{C}$ and the optimal configuration usage probability vector $\boldsymbol{\pi}$ in (3.48), where $N = 5$, $a_m = 3$ and

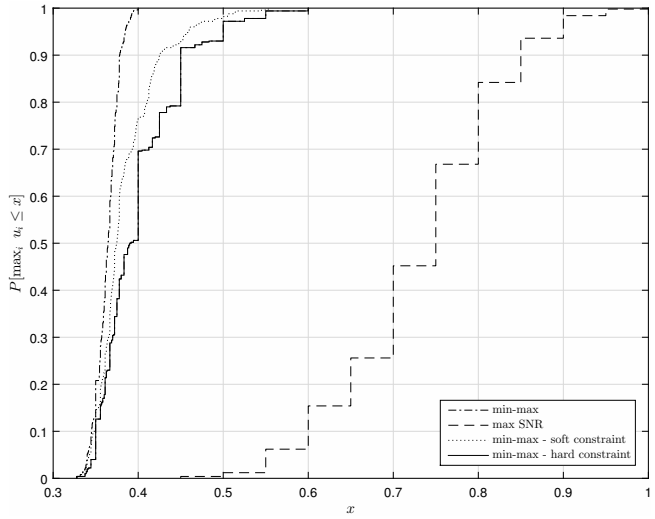


Figure 3.12: CDF of the anchor node utilization probability in the proposed authentication methods

the section of the matrix $\boldsymbol{C}$ relative to the source $m = 3$ is

$$\boldsymbol{C} = \begin{bmatrix} & 0 & 0 & 0 & \\ & 1 & 0 & 0 & \\ \cdots & 0 & 1 & 1 & \cdots \\ & 0 & 1 & 0 & \\ & 0 & 0 & 1 & \end{bmatrix}, \qquad \boldsymbol{\pi} = \begin{bmatrix} \vdots \\ 0.8 \\ 0.1 \\ 0.1 \\ \vdots \end{bmatrix} \tag{3.48}$$

where clearly $\boldsymbol{\pi}$ contains zeros in all entries pertaining to the same source, except the three highlighted in (3.48).

When the transmitter sends the packet, anchor node 1 knows it must not participate. Therefore, slot 1 will remain empty. Then, in the next slot, anchor node 2 participates with probability 0.8. About node 2, observe that it can authenticate the source by itself and hence the only considered configuration that includes node 2 is $\boldsymbol{c}_1(3) = [0\,1\,0\,0\,0]^T$. Therefore, if anchor node 2 participates, no other node will participate. Otherwise, in the following slot, node 3 is required to participate (since it is active in all remaining configurations with non-zero probability) and, therefore, it sends its report to c. In slot 4 anchor node 4 collaborates with probability 0.5 and finally, node 5 remains silent if node 4 transmits, otherwise it sends its report to the concentrator c.

Distributed SNR-Based Anchor Node Selection

The SNR-based anchor node selection described in Section 3.1.7 could be partially distributed by avoiding the reporting of the channel gains in Phase 1 when not used in the selected configuration. In particular, the proposed SNR-based distributed algorithm works as follows. In Phase 1, each anchor node estimates the channel, however without immediately transmitting it to the concentrator c. Instead, anchor node i waits a time $w(|h_i^{(0)}(m)|)$, which is a

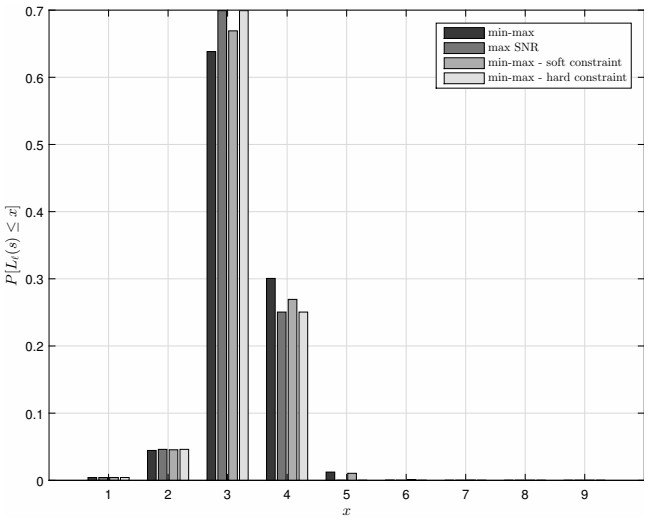


Figure 3.13: PMF of the number of anchor nodes involved in a source node authentication

decreasing function of the estimated SNR, thus for anchor nodes having a higher SNR the forwarding of the channel estimate to node c will be faster. When the anchor node c has obtained an admissible configuration, it sends a broadcast message to stop the forwarding from the anchor nodes.

Now, let $F^c(x) = \mathrm{P}[|h_i^{(0)}(m)| \geq x] = 1 - F(x)$ be the CCDF of the SNR over source node position statistics and fading statistics. Then, we set the waiting time as

$$w(|h_i^{(0)}(m)|) = F^c(|h_i^{(0)}(m)|) \cdot T_0 = [1 - F(|h_i^{(0)}(m)|)] \cdot T_0 \,, \tag{3.49}$$

where T_0 is a constant chosen in order to minimize authentication packet collisions. The choice of the waiting time according to (3.49) ensures a uniform distribution of the transmissions within the interval T_0, thus minimizing the duration of the authentication procedure.

However, the collection of channel gains in Phase 1 can take longer than the fully centralized approach, where a simple round Robin approach is used to all gain values. Moreover, possible collisions of control packets transmitted by the anchor nodes must be handled.

3.1.10 Final Performance Comparison

Let us finally compare the performance of the energy-efficient anchor selection (using the min-max problem (3.36)), the signaling-efficient anchor selection, and the possible trade-offs between the two. For the comparison we used the parameters in Table 3.2. As for the characterization of problems (3.44) and (3.45), we remark that only minimum Hamming weight configurations, i.e., $K_m = \min_\ell L_\ell(m)\ \forall m$ are considered.

Figure 3.12 shows the CDF of the maximum anchor node utilization probability, $\max_i u_i$. It can be seen that the min-max-based optimization problems (3.36), (3.44), and (3.45) provide a much lower utilization probability. In particular,

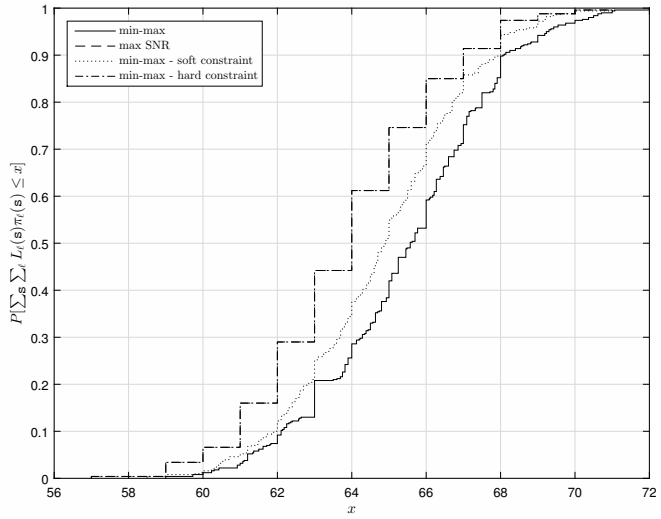


Figure 3.14: CDF of the energy consumption per authentication round

Table 3.2: Simulation parameters for the performance evaluation

Parameter	Value
f	900 MHz
η	2
ρ	0.1
Cell radius	500 m
SNR	25 dB (at 250 m)
N	9
M	20
Δ	0.1

as expected, the solution of problem (3.45) is in-between the solutions of problems (3.36) and (3.45). On the contrary, the SNR-algorithm is clearly the worst solution in terms of anchor nodes lifespan.

Figure 3.13, instead, depicts the PDF of the *number of anchor nodes involved in the authentication of a source node*. We note that the SNR-based policy and problem (3.44) have approximately the same behavior, using less reports from the anchor nodes with respect to problems (3.36) and (3.45), thus being signaling-efficient methods. Moreover, it is clear that the intermediate approach with soft constraint (3.45) represents a trade-off between the min-max approach of (3.36) and the intermediate approach with hard constraint (3.44).

Finally, Figure 3.14 depicts the CDF of the *average number of reports used in a whole authentication round*, i.e., for the authentication of all the M source nodes. This metric is obtained as $\sum_m \sum_\ell L_\ell(m) \cdot \pi_\ell(m)$. We note that the SNR-based policy and problem (3.44) employ the minimum number of anchor nodes; therefore, the two curves are overlapped. On the contrary, the min-max problem (3.36) tends to use more anchor nodes in performing the authentication, while the soft approach (3.45) is once again in-between.

3.2 Energy-Efficient Location Verification

In the context of the IoT, localization of terminals is a desirable feature that can enable an entire class of services commonly referred to as Location-Based Services (LBSs) [8] that are specifically related to the position of the users. Such applications may vary from the tracking of assets and goods in logistics, to services in which the user is charged by the provider based on his position, e.g., road tolling.

Ensuring the correctness of tags' positions is of a fundamental importance for the effectiveness of these applications. Moreover, with the increasing adoption of LBSs, the motivation to attack such systems in order to obtain an economic or a competitive advantage grows. Therefore, appropriate security mechanisms for location verification should be designed to provide assurance of the terminal position against forgery attempts, i.e., the spoofing attacks [143, 144].

3.2.1 Related Work

The Received Signal Strength Indicator (RSSI) can be seen as a ranging measure between the transmitter and the receiver, thus converting RSSI into a distance measure and performing triangulation provides a low-cost and low-complexity localization procedure for IoT terminals. However, it has been shown [145] that RSSI does not provide accurate results. One could argue that for precise localization it is sufficient to equip IoT terminals with a Global Navigation Satellite System (GNSS) module, but in most cases this option is prevented by costs and energy availability constraints. Moreover, GNSS signals are typically not available in indoor or deep urban environments and may themselves be subject to spoofing attacks [146].

Several solutions have been proposed by the research community to improve the precision of RSSI-based positioning, and most of them are based on the presence of some location-aware nodes (called anchor nodes) that are deployed in the area of interest. For example, in [147], a distributed localization scheme with location verification for Wireless Sensor Networks (WSNs) is proposed, and the anchor nodes' positions are preset. In [148–150] multiple-step procedures are proposed to estimate the users' locations in Wireless Local Area Networks (WLANs): after collecting RSSI measurements from anchors to build a probabilistic map of the area, signals coming from generic nodes are matched in real-time to the most likely recorded positions.

As for the security aspects of positioning, several results can be found in the literature about location verification algorithms for RSSI-based localization, as well. Most solutions rely on the availability of trusted nodes [151, 152], although approaches that do not require them are also available [153, 154]. Rather than proposing methods based on an explicit exchange of messages for location verification purposes, some authors exploit the intrinsic characteristics of the wireless channel between source nodes and anchor nodes. In [155], the shadowing components estimated by the anchor nodes are employed to authenticate the position of terminals in fifth-generation (5G) cellular networks. A formal performance analysis of wireless location verification systems based on multiple trusted anchors under correlated shadowing conditions is provided by [156].

In this section, we focus on an IoT context where anchor nodes are available

for location verification purposes, but we aim at optimizing their utilization for a judicious use of resources. Indeed, existing literature typically assumes that all anchor nodes are always active for location verification. However, in some scenarios, e.g., when the anchors have a limited energy capacity because they are battery-powered, it is important to use each them only when necessary, in order to extend the network lifespan. Considering the location verification procedure proposed by [156] as a building block, we design an optimization problem to derive the activation policies of the anchor nodes, as done in Section 3.1. The motivation behind our work is to minimize the energy expenditure and the network overhead of the location verification algorithm, given a target performance in terms of FA and MD probabilities. We want to remark that the proposed framework may be applied in two prominent network architectures for IoT: 1) CIoT [50] and 2) Long-Range$^{\text{TM}}$ (LoRa) (see Chapter 2). In the former case, our algorithm minimizes the energy consumption of the anchor nodes, i.e., the distributed antennas. In the latter, our solutions allows to minimize the overhead traffic coming from the various gateways towards the NetServer. Moreover, another interesting application of the proposed framework is the cross-check of positions sporadically obtained by the GNSS module, thus increasing their resilience against attacks.

The rest of the section is organized as follows. In Section 3.2.2 we describe the localization verification framework, while the performance evaluation results are discussed in Section 3.2.3.

3.2.2 System Model

We consider an IoT network, as shown in Figure 3.15, in which various legitimate end nodes transmit data to a unique concentrator, which is assisted by N anchor nodes deployed in the area of interest. We assume that the signals coming from the end nodes can be received by all anchors.[8] Moreover, we assume that an attacker node $\mathtt{a}$ is present in the area at position $\boldsymbol{x}_{\mathtt{a}}$: it aims at transmitting messages to the concentrator by pretending he is located at a different position $\boldsymbol{x}_{\mathtt{s}}$ of his choice. Note that vectors $\boldsymbol{x}_{\mathtt{a}}$ and $\boldsymbol{x}_{\mathtt{s}}$ are two-dimensional coordinates on a plane.

The column vector of the average received power at the N anchor nodes when $\mathtt{a}$ is transmitting is denoted as $\boldsymbol{v} = [v_1, \ldots, v_N]^T$. The average received power vector at the N anchors when $\mathtt{s}$ is transmitting is denoted as $\boldsymbol{u} = [u_1, \ldots, u_N]^T$.

In logarithmic (dB) scale, the i-th entries of $\boldsymbol{u}$ and $\boldsymbol{v}$ are, respectively,

$$u_i = p - 10\eta \log_{10}\left(d_{\mathtt{s},i}\right) \tag{3.50}$$

and

$$v_i = p - 10\eta \log_{10}\left(d_{\mathtt{a},i}\right) , \tag{3.51}$$

where p is the reference received power at a unitary (1 m) distance, η is the path loss exponent, and $d_{\mathtt{s},i}$ ($d_{\mathtt{a},i}$) is the distance between $\mathtt{s}$ ($\mathtt{a}$) and the i-th anchor node. Note that $\boldsymbol{u}$ ($\boldsymbol{v}$) depends only on the path loss between the real (pretended) location and the various anchors. The shadowing vector $\boldsymbol{\xi} = [\xi_1, \ldots, \xi_N]^T$ is a

[8] We remark that, in the case of CIoT, the anchor nodes represent the remote radio heads (RRHs) of the base station, while in a LoRa network they can be associated to the gateways that forward packets to the NetServer.

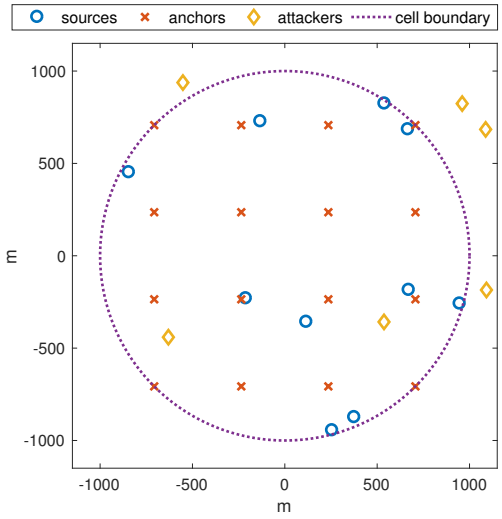


Figure 3.15: Network deployment example, where $N = 16$ anchors are deployed in a square grid. The diamonds indicates the worst case attacker positions.

Gaussian vector with zero mean and $N \times N$ covariance matrix $\boldsymbol{R}$. The entries of $\boldsymbol{R}$ are the spatial correlations between the various anchor pairs; considering anchors i and j at distance d_{ij}, we have

$$R_{ij} = \sigma_s^2 \exp\left(-\frac{d_{ij}}{D_C} \ln 2\right), \tag{3.52}$$

where σ_s^2 is the shadowing power in dB and D_C is the decorrelation distance.

The location verification is an hypothesis testing problem between the following two alternatives:

1. $\mathcal{H}_0$: the signal comes from a legitimate transmitter $\mathtt{s}$, actually located at position $\boldsymbol{x}_\mathtt{s}$. In this case, the received power vector at the N anchors is $\boldsymbol{y} = \boldsymbol{u} + \boldsymbol{\xi}$, distributed as

$$\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{u}, \boldsymbol{R}). \tag{3.53}$$

2. $\mathcal{H}_1$: the attacker $\mathtt{a}$ is transmitting from a position which is at least at distance r from the legitimate one, i.e.,

$$||\boldsymbol{x}_\mathtt{a} - \boldsymbol{x}_\mathtt{s}|| \geq r. \tag{3.54}$$

Assuming that the attacker can alter its transmit power p_x, we have that the received power vector becomes $\boldsymbol{y} = p_x \mathbf{1}_{N\times 1} + \boldsymbol{v} + \boldsymbol{\xi}$ and is distributed as

$$\boldsymbol{y} \sim \mathcal{N}(p_x \mathbf{1}_{N\times 1} + \boldsymbol{v}, \boldsymbol{R}). \tag{3.55}$$

The decision rule is based on the log-likelihood ratio, i.e., the logarithm of the ratio between the PDFs given the two hypothesis

$$\Psi \simeq \log \frac{f(\boldsymbol{y}|\mathcal{H}_1)}{f(\boldsymbol{y}|\mathcal{H}_0)}. \tag{3.56}$$

Given a decision threshold θ, the concentrator decides on the hypothesis testing problem by checking whether $\Psi \gtrless \theta$. If $\Psi \geq \theta$, we decide for hypothesis $\mathcal{H}_1$, otherwise we decide for $\mathcal{H}_0$.

Attack Strategy We assume that the attacker is able to optimize its transmit power p_x and actual position $\boldsymbol{x}_{\mathtt{a}}$. In particular, in order to maximize the probability of attack success, power and position are chosen to minimize the Kullback-Leibler (KL) divergence between the conditional distributions of random vectors $\boldsymbol{y}$ under the two hypotheses. Since $\boldsymbol{y}$ is Gaussian distributed in both cases, we are considering the KL divergence between two multivariate Gaussian distributions with different means, which is given by

$$\phi(p_x, \boldsymbol{x}_{\mathtt{a}}) = \frac{1}{2}(p_x \mathbf{1}_{N\times 1} + \boldsymbol{v} - \boldsymbol{u})^T \boldsymbol{R}^{-1}(p_x \mathbf{1}_{N\times 1} + \boldsymbol{v} - \boldsymbol{u})\,. \tag{3.57}$$

Under the assumption that the attacker minimizes the KL, the received power vector becomes

$$\boldsymbol{y} = \boldsymbol{w}^\star + \boldsymbol{\xi}\,, \tag{3.58}$$

where

$$\boldsymbol{w}^\star = \frac{(\boldsymbol{u} - \boldsymbol{v}^\star)^T \boldsymbol{R}^{-1}\mathbf{1}_{N\times 1}}{\mathbf{1}_{N\times 1}^T \boldsymbol{R}^{-1}\mathbf{1}_{N\times 1}}\mathbf{1}_{N\times 1} + \boldsymbol{v}^\star\,, \tag{3.59}$$

and $\boldsymbol{v}^\star$ is obtained by plugging the optimal attacker position into the definition of $\boldsymbol{v}$.

FA and MD Probabilities With the test (3.56) and under the assumptions (3.57)-(3.59), the FA probability can be derived as [156]

$$P_{\mathrm{FA}} = \mathcal{Q}\left[\frac{\ln\lambda + \frac{1}{2}(\boldsymbol{w}^\star - \boldsymbol{u})^T \boldsymbol{R}^{-1}(\boldsymbol{w}^\star - \boldsymbol{u})}{\sqrt{(\boldsymbol{w}^\star - \boldsymbol{u})^T \boldsymbol{R}^{-1}(\boldsymbol{w}^\star - \boldsymbol{u})}}\right], \tag{3.60}$$

while the MD probability for the optimal attack is

$$P_{\mathrm{MD}} = 1 - \mathcal{Q}\left[\frac{\ln\lambda - \frac{1}{2}(\boldsymbol{w}^\star - \boldsymbol{u})^T \boldsymbol{R}^{-1}(\boldsymbol{w}^\star - \boldsymbol{u})}{\sqrt{(\boldsymbol{w}^\star - \boldsymbol{u})^T \boldsymbol{R}^{-1}(\boldsymbol{w}^\star - \boldsymbol{u})}}\right] \tag{3.61}$$

with $\mathcal{Q}(z) = (1/\sqrt{2\pi}) \int_z^{+\infty} \exp(-z^2/2)\ \mathrm{d}z$.

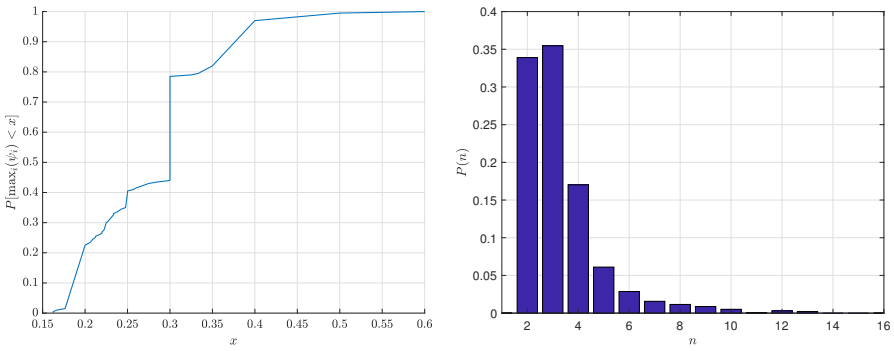
Anchor Node Usage Optimization To optimize the usage of the anchor network, we resort to the energy-efficient optimization framework described in Section 3.1.5. In particular, we will solve the min-max problem (3.36) to obtain the probability vector of anchor configurations $\boldsymbol{\pi}$.

3.2.3 Performance Evaluation

We evaluated the proposed scheme considering typical physical layer parameters for CIoT [28]. The anchor nodes are deployed in a regular squared grid inscribed in the cell circular area; the legitimate nodes, instead, are deployed randomly and the attacker positions are optimized as discussed in Section 4.1. The attacker position is optimized for each examined configuration in order to

Table 3.3: Simulation parameters for the reference scenario

Parameter	Value
f	900 MHz
η	3.5
D_C	110 m
σ_s^2	8 dB
Cell radius	1000 m
p	-10 dB (at 1 m)
N	16
M	10
r	400 m
$P_{\text{FA}}^{\star}$	5×10^{-2}
$P_{\text{MD}}^{\star}$	5×10^{-2}
Realizations	200



(a) **(b)**

Figure 3.16: Performance evaluation results. On the left-hand side, the CDF of the maximum usage of the anchor nodes. On the right-hand side, the histogram of the average number of anchor nodes used to verify the position of a single source node.

induce the optimal attack strategy. The generation of the M legitimate sources (and of the corresponding attackers) is repeated multiple times to average the results over many realizations of the scenario. All the simulation parameters are summarized in Table 3.3. In the following, we describe the results that assess the performance of the proposed approach.

Maximum Utilization Figure 3.16a shows the empirical CDF of the maximum utilization in the various realizations of the scenario. The CDF was obtained by recording the maximum utilization $\max_i \psi_i$ obtained after solving the proposed optimization problem for every deployment of the source nodes. Note that the anchor node $i^{\star}$ with the highest utilization may be different in each realization. We also recall that the upper bound on the value of $\max_i \psi_i$ is achieved when the mostly used anchor is employed in the verification of all the terminal locations, thus $\max_i \psi_i \leq 1$. Thus, this plot shows that the proposed optimized activation policy provides a reduction of the maximum utilization of at least 60%, since in $x = 0.4$ the CDF of the maximum usage is 97%.

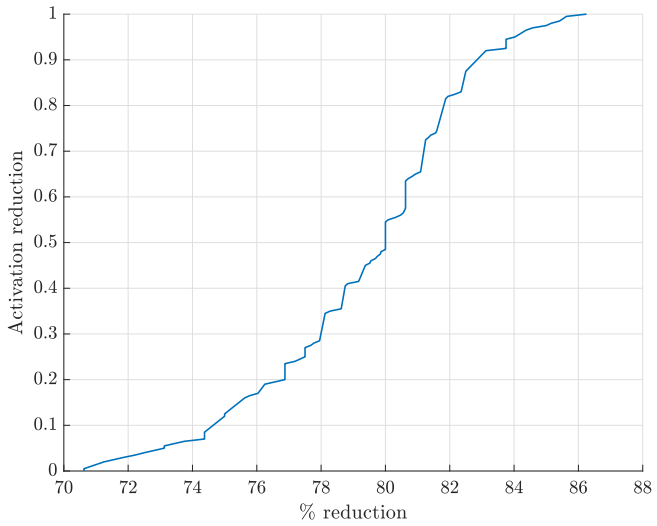


Figure 3.17: CDF of the percentage saving in anchor utilization with respect to the full utilization

Configuration Weight Distribution Another interesting result is the empirical Probability Mass Function (PMF) of the number of anchor nodes involved in the location verification of a generic source node, which is reported in Figure 3.16b. The PMF is obtained recording the utilization of the configurations with Hamming weight $n = 2, \ldots, 16$ in the various realizations. This result shows that, in most cases, the location verification procedure requires configurations that comprise 2 or 3 anchors, which is a major reduction with respect to employing all the $N = 16$ anchors.

Utilization Reduction Finally, in Figure 3.17 we plot the empirical CDF of the percentage reduction in the utilization achieved by the proposed activation policy with respect to the case in which all the anchor nodes are employed. Note that, contrary to Figure 3.16a where we focused only on the anchor with the maximum utilization, in this plot we consider the entire set of anchors. From this graph, it can be seen that we can reduce the number of times in which a generic anchor is active between 70% and 86%.

3.3 Conclusions

Due to the pervasive nature of devices and their impact on our daily life, the future IoT needs new low-cost and efficient techniques to improve security.

In the first part of this chapter, we addressed the problem of authentication in a Cellular IoT (CIoT) scenario by exploiting the fading of the wireless links between the device to be authenticated and a set of trusted anchor nodes. We first derived the False Alarm (FA) and Missed Detection (MD) probabilities of the authentication detection process; then, we addressed the problem of optimizing the use of the anchor nodes for optimization purposes, with the

objective of maximizing the anchor network lifespan while ensuring target FA and MD probabilities. Bounds on the Cumulative Distribution Function (CDF) and average anchor network lifespan have been derived and three strategies that optimize the anchor nodes utilizations have been proposed. The numerical results showed that the three methods provide approximately the same performance in terms of message authentication, while the method that maximizes the upper bound on the average anchor network lifespan achieves the highest anchor network lifespan. Finally, we proposed a signaling-efficient technique to manage the anchor network utilization, and two possible trade-offs between energy-efficiency and signaling-efficiency.

In the second part of this chapter, instead, we focused on the aspects regarding the security of localization algorithms. In particular, we proposed an energy-efficient physical layer location verification method for IoT networks in which the concentrator node is assisted by anchor nodes, whose activation rate is minimized while guaranteeing a given performance in terms of FA and MD probabilities.

Chapter 4

Joint Optimization of Lossy Compression and Transport in Wireless Sensor Networks

Energy-efficient data dissemination in distributed Wireless Sensor Networks (WSNs) is a key requirement, and many papers have appeared in the past few years on this subject [157]. A large body of work exists on in-network data aggregation [158, 159], lossless [160] and lossy compression [161], as reducing the number of bits to be transmitted entails a smaller energy consumption for communication.

In this chapter, we jointly address the problems of lossy data compression at the WSN sources and of constructing efficient routes toward the WSN data gathering point (the sink). This amounts to exploring the fundamental tradeoffs between rate-distortion at the sources and the cost associated with transporting this information, exploiting the best possible flow allocation scheme. To the best of our knowledge, a complete solution to this problem is still lacking and our present analysis represents an initial step toward it.

Among many other papers on compression and energy efficient routing, the following ones are of particular interest and are closely related to our current work. The authors of [162] construct a data gathering tree in such a way that the sum of the computation and communication costs is minimized using lossless compression. Their goal is to tune the complexity associated with compression based on energy availability, by adjusting the degree of compression at the sources while jointly routing data. They propose a simple greedy approximation to minimal Steiner tree routing and prove that it provides good average performance in general topologies. Paper [163] addresses the joint routing and compression problem by applying Lyapunov optimization theory. The objective is to dynamically decide whether and at which nodes to compress, devising centralized and distributed schemes. The Lyapunov drift is used to maintain the nodes' backlog queues stable over time and the routing topology is given. In [164], the authors exploit spatial correlation to aggregate data, constructing energy-efficient data aggregation trees. Paper [165] analyzes a simple theoretical model of data gathering networks with data compression where each node can preprocess the gathered data before sending it to the base station. The focus is

on efficient compression and transmission scheduling over single-hop networks.

As we mentioned above, here we study the problem of joint compression and transport. With this we refer to the inherent tradeoffs that are associated with data compression at the sources and energy consumption for the transport of such compressed data using the most efficient routing paths in terms of energy consumption and transport capacity. Our main concern is to understand how much processing has to be performed at the sources, exploiting some lossy compression algorithm, so that the compressed information is efficiently disseminated through a given network graph where compression and routing are jointly evaluated.

The distinctive aspects of our present work are that: a1) we consider lossy compression at the sources, i.e., we can trade the signal representation accuracy for the number of bits to transmit, and a2) the network topology is given but routing paths are not, so the flow allocation problem (paths and data rate on each link) is jointly solved with a1). One of the network setups that we are interested in consists of topologies where a bottleneck link inherently exists a few hops away from the sources, no matter the routing path, and some action (e.g., compression) has to be taken to reduce the amount of traffic that flows over it, as otherwise the data transport problem would be infeasible, leading to high packet drop rates. In this case, the presence of the bottleneck must be somehow backpropagated to the sources to allow for data compression, while jointly obtaining an efficient routing topology.

This leads to a *multi-objective* optimization problem that entails the joint minimization of o1) a compression cost and o2) a transport cost. Assuming a generic lossy compression algorithm at the sources, which may be source-specific, objective o1 consists of minimizing a compression cost, while at the same time never exceeding certain source-specific maximum distortion levels. To this end, we design rate-distortion functions that accurately match those of practical data compression algorithms [161]. Objective o2 entails the minimization of a transport cost, which is modeled as the energy consumption associated with the transmission over the selected links.

The multi-objective optimization framework that we present in this chapter keeps the above facts into account, providing a solution to the considered compression and transport problem, i.e., finding the corresponding Pareto curve in terms of distortion and transport costs. The framework allows the optimization of the network under capacity and distortion constraints, while accounting for general cost models. Our approach is an initial step toward the joint optimization of compression (source processing) and transport and can be extended in several ways.

To summarize, in this chapter:

- we formulate a joint rate-distortion and routing (flow allocation) problem in distributed WSNs. This problem is posed as a convex optimization program that we solve to obtain the optimal flows for the whole WSN.

- We account for realistic rate-distortion curves, capitalizing on the work in [161], and we model the routing problem considering a simplified but meaningful approach to capture the scheduling of channel access resources at the medium access layer.

- We discuss the results of our problem in two selected network scenarios,

quantifying the mutual dependence between compression at the sources and network flow allocation, and discussing relevant tradeoffs.

- We discuss possible extensions that will be considered in our future work.

The remainder of the chapter is structured as follows. In Section 4.1 we present the system model, treating the WSN network as a graph and detailing how flows are scheduled at the medium access layer. In Section 4.2 we detail the optimization problem, which is then solved in Section 4.3, where we evaluate the performance of two selected network scenarios and discuss the relevant tradeoffs between flow scheduling and compression. Our conclusions and possible extensions of the presented analysis are discussed in Section 4.4.

4.1 System Model

We represent the WSN using a directed graph $\mathcal{G} = (\mathcal{N}, \mathcal{L})$ [166], where $\mathcal{N}$ is the set of nodes, with $|\mathcal{N}| = N$, and $\mathcal{L}$ is the set of wireless links between node pairs with $|\mathcal{L}| = L$. The positions of the sensors on the plane and the network connections are assumed to be fixed.

4.1.1 Set of Nodes Characterization

Set $\mathcal{N}$ is partitioned as $\mathcal{N} = \mathcal{S} \cup \mathcal{R} \cup \mathcal{D}$, where $\mathcal{S}$, $\mathcal{R}$, and $\mathcal{D}$ are the disjoint sets of source nodes, relay nodes, and destination nodes, respectively. Source nodes gather data from the surrounding environment and can compress this data through lossy source coding techniques before injecting it into the transport network. The compression process reduces the bitrate that goes into the WSN, but introduces some distortion in the reconstruction of the data at the destination. To capture this tradeoff, every source node s is characterized by a monotonically decreasing *rate-distortion function* $D_s : \mathbb{R} \mapsto [0, 1]$, that takes the transmitted data rate as input and outputs the information distortion. Moreover, we assume that a *distortion threshold* Δ_s^{thr} exists: if the reconstruction error exceeds Δ_s^{thr}, the signal generated by s is no longer useful for the final destination. Relay nodes, instead, are just packet forwarders: they neither generate new data nor perform data aggregation.[1] Finally, since a common destination for all source nodes is usually assumed in WSNs, $\mathcal{D}$ is a singleton consisting of the *sink*, i.e., $\mathcal{D} = \{d\}$. Therefore, set $\mathcal{N}$ is expressed as

$$\mathcal{N} = \{s_1, \ldots, s_S\} \cup \{r_1, \ldots, r_R\} \cup \{d\}\,. \tag{4.1}$$

Note that the cardinalities of the three partitions of network devices are S, R, and $D = 1$, respectively, so that $S + R + 1 = N$.

4.1.2 Set of Edges Characterization

We label the edges of the graph with natural numbers, thus, $\mathcal{L} = \{1, \ldots, L\}$. We denote by $\mathcal{E} = \{1, \ldots, E\} \subseteq \mathcal{L}$ the subset of edges that are linked to the

[1] This simplifying assumption is introduced in this first study to highlight more clearly some behaviors, and the extension of this framework to a more general model is part of our future work.

source nodes; assuming that each source is attached to a single router, the cardinality of $\mathcal{E}$ is equal to that of $\mathcal{S}$, i.e., $E = S$. For the sake of convenience, we assume that the links in $\mathcal{E}$ are the first E links of set $\mathcal{L}$. Finally, let us define $F = L - E$, $0 \leq F \leq L - 1$.

Let $\boldsymbol{x} = [x_1, \ldots, x_L]^T$ be the $L \times 1$ *vector of flows*, where x_ℓ denotes the flow assigned to the ℓ-th link. An *information transport cost function* $\phi_\ell : \mathbb{R} \mapsto [0, 1]$, taking x_ℓ as input and returning the corresponding transport cost, is associated with every link $\ell \in \mathcal{L}$. Moreover, the first E links are subject to an *information distortion cost function* $\omega_\ell : \mathbb{R} \mapsto [0, 1]$ as well, which takes x_ℓ as input and returns the corresponding distortion cost.

4.1.3 Graph Characterization

Let $\boldsymbol{A}$ be the $N \times L$ incidence matrix of $\mathcal{G}$. For the sake of convenience, we assume that the first S rows of $\boldsymbol{A}$ relate to the nodes in $\mathcal{S}$; the rows in the range $\{S + 1, \ldots, S + R\}$ relate to the nodes in $\mathcal{R}$, and the last row to the sink d. Hence, the general structure of matrix $\boldsymbol{A}$ is

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{I}_S & \boldsymbol{0}_{S \times F} \\ \boldsymbol{\Phi}_{R \times E} & \boldsymbol{X}_{R \times F} \\ \boldsymbol{\Psi}_{1 \times E} & \boldsymbol{\Omega}_{1 \times F} \end{bmatrix} \in \{\pm 1,\ 0\}^{N \times L}, \tag{4.2}$$

where $\boldsymbol{I}_S$ and $\boldsymbol{0}_{S \times F}$ denote the identity matrix of order S and the $S \times F$ null matrix, respectively. The structure of matrices $\boldsymbol{\Phi}$ and $\boldsymbol{X}$, instead, depends on the network topology: the generic element in position (i, j), $S < i \leq S + R$, $\forall j \in \mathcal{L}$, is

$$A_{ij} = \begin{cases} 1 & \text{if edge } j \text{ leaves relay } i\,, \\ -1 & \text{if edge } j \text{ enters relay } i\,, \\ 0 & \text{otherwise}\,. \end{cases} \tag{4.3}$$

Finally, elements in row vectors $\boldsymbol{\Psi}$ and $\boldsymbol{\Omega}$ are such that

$$A_{Nj} = \begin{cases} -1 & \text{if edge } j \text{ enters the sink}\,, \\ 0 & \text{otherwise}\,. \end{cases} \tag{4.4}$$

We define vector $\boldsymbol{b} = [b_1, \ldots, b_N]^T$ as the $N \times 1$ *vector of net flow* through a node. It is $b_k > 0$ for $1 \leq k \leq S$ because sources inject data into the network; $b_k = 0$ for $S + 1 \leq k \leq S + R$ because relays just forward packets, and $b_N = -\sum_{\ell=1}^{E} x_\ell < 0$ because the sink gathers data generated by sources. We remark that the sum of the elements in vector $\boldsymbol{b}$ is greater than or equal to zero, because the aggregate data flow generated by the sources, thanks to the compression process, may be greater than the flow received at the sink; the equality holds in case no compression is performed at the source nodes.

Finally, it is worth noting that the case of star network topology, i.e., no relaying allowed, can be obtained imposing $\mathcal{R} = \emptyset$ and $\mathcal{E} = \mathcal{L}$.

4.1.4 MAC Protocol Design

As the Medium Access Control (MAC) protocol we assume that a Time Division Multiple Access (TDMA) scheme coordinates the concurrent channel access for every transmitter-receiver pair. Although random access may be

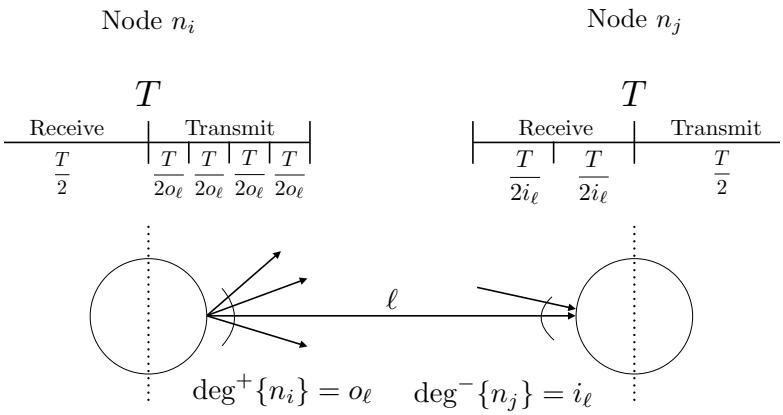


Figure 4.1: Time-division-based access scheme at a generic transmitter-receiver pair. The nodes receive data during the first half of the access frame, of duration T, and transmit for the remaining half. The time share that each node allots to different outgoing or incoming links is assumed to be equal.

preferred to TDMA due to its lack of coordination, we highlight that determinism is becoming increasingly important for commercial deployments, as one can see from recent standardization efforts [167]. A diagram explaining our MAC allocation strategy is provided in Figure 4.1, where edge ℓ is assumed to connect nodes n_i and n_j. The outdegree of n_i is $\deg^+\{n_i\} = o_\ell = 4$, while the indegree of n_j is $\deg^-\{n_j\} = i_\ell = 2$. Time synchronization is assumed for the entire network and the channel access uses *access frames* of duration T. Every node in $\mathcal{R}$ receives data for half of the access frame and forwards packets for the remaining half of it. On the other hand, source nodes remain idle during half of the access frame to gather data and the sink sends the gathered data to the core network during half of the access frame. For the sake of simplicity, an equal slot duration is considered for both incoming transmissions and outgoing transmissions by every device. Therefore, considering a link ℓ characterized by a generic transmitter outdegree and receiver indegree, respectively equal to o_ℓ and i_ℓ, we compute the *transmit interval fraction* for link ℓ, i.e., the fraction of time during which the capacity of the ℓ-th wireless link can be fully exploited, as:

$$\xi_\ell = \frac{T/2}{\max\{o_\ell, i_\ell\}} \times \frac{1}{T} = \frac{1}{2\max\{o_\ell, i_\ell\}} \leq \frac{1}{2}\,. \tag{4.5}$$

Then, the actual capacity of link ℓ is equal to the theoretical channel capacity multiplied by the *utilization* parameter ξ_ℓ. This amounts to a scheduler where equal shares are allotted to incoming and outgoing links and where transmission and reception activities last for $T/2$. In our future work, we plan to extend the channel access model to more general cases, where unequal resources can be allocated to the links according to energy or priority considerations.

4.2 Optimization Problem

In the proposed optimization framework, we aim at minimizing the *aggregate information transport cost* ρ, defined as

$$\rho(x_1, \ldots, x_L) = \sum_{\ell=1}^{L} \phi_\ell(x_\ell)\,, \tag{4.6}$$

where $\phi_\ell(x_\ell)$ returns the transport cost for link ℓ, given that flow x_ℓ is carried over the link. At the same time, we account for an *aggregate information distortion cost* σ, defined as

$$\sigma(x_1, \ldots, x_E) = \sum_{\ell=1}^{E} \omega_\ell(x_\ell)\,, \tag{4.7}$$

where $\omega_\ell(x_\ell)$ is the distortion cost associated with source $\ell = 1, \ldots, E$. These are contrasting objectives: indeed, in order to decrease the transport cost, the network flows need to be reduced and this can only be achieved by compressing the data before injecting it into the network, which however entails a higher distortion. Also, the cost functions in (4.6) and (4.7) represent heterogeneous quantities: while the former can be associated with an energy cost, the latter is a Quality of Service (QoS) metric.

The objective function that we want to minimize is the weighted sum of Equations (4.6) and (4.7), i.e.,

$$f(x_1, \ldots, x_L) = \alpha\rho(x_1, \ldots, x_L) + (1-\alpha)\sigma(x_1, \ldots, x_E)\,, \tag{4.8}$$

where $\alpha \in [0, 1]$ is a parameter that weighs the information transport cost ρ and the information distortion error σ. The following set of constraints on the variables $\boldsymbol{x}$ must be met to ensure the proper operation of the WSN:

- *maximum distortion constraints*, i.e.,

$$D_\ell(x_\ell) \leq \Delta_\ell^{\text{thr}}\,, \qquad \ell = 1, \ldots, E\,; \tag{4.9}$$

- *source flow constraints*, i.e.,

$$\sum_{\ell=1}^{L} A_{k\ell} x_\ell \leq b_k\,, \qquad k = 1, \ldots, S\,; \tag{4.10}$$

- *relay flow conservation constraints*, i.e.,

$$\sum_{\ell=1}^{L} A_{k\ell} x_\ell = 0\,, \qquad k = S+1, \ldots, N-1\,; \tag{4.11}$$

- *destination flow constraint*, i.e.,

$$\sum_{\ell=1}^{L} A_{k\ell} x_\ell + \sum_{\ell=1}^{E} x_\ell = 0\,, \qquad k = N\,; \tag{4.12}$$

- *wireless link flow constraints*, i.e.,

$$0 \leq x_\ell \leq C_\ell\,, \qquad \ell = 1, \ldots, L\,, \tag{4.13}$$

  where C_ℓ is the channel capacity for link ℓ.

We remark that constraint (4.12) can be rewritten as

$$\sum_{\ell=1}^{E}(A_{N\ell}+1)x_\ell + \sum_{\ell=E+1}^{L} x_\ell = \left[\boldsymbol{\Psi} + \mathbf{1}_{1\times E} \quad | \quad \boldsymbol{\Omega}\right] \cdot \boldsymbol{x} = 0\,, \tag{4.14}$$

where $\mathbf{1}_{1\times E}$ denotes the row vector of ones with E entries. We also remark that, since $D_\ell(\cdot)$ is a monotonically decreasing function $\forall \ell \in \mathcal{E}$, the constraints in Equation (4.9) can be converted into *minimum outgoing flow* constraints for source nodes. For a generic source link $\ell = 1, \ldots, E$, the minimum outgoing flow can be determined imposing $D_\ell(x_\ell^{\min}) = \Delta_\ell^{\text{thr}}$, which gives $x_\ell^{\min} = D_\ell^{-1}(\Delta_\ell^{\text{thr}})$. Therefore, the final optimization problem is formulated as follows:

$$\min_{\boldsymbol{x}} \; f(x_1, \ldots, x_L) \tag{4.15a}$$

subject to

$$x_\ell^{\min} \leq x_\ell \leq b_\ell\,, \qquad \ell = 1, \ldots, E\,, \tag{4.15b}$$

$$\sum_{\ell=1}^{L} A_{k\ell} x_\ell = 0\,, \qquad k = S+1, \ldots, N-1\,, \tag{4.15c}$$

$$\left[\boldsymbol{\Psi} + \mathbf{1}_{1\times E} \quad | \quad \boldsymbol{\Omega}\right] \cdot \boldsymbol{x} = 0\,, \tag{4.15d}$$

$$0 \leq x_\ell \leq C_\ell\,, \qquad \ell = E+1, \ldots, L\,. \tag{4.15e}$$

As long as $\phi_\ell(\cdot)$ and $\omega_\ell(\cdot)$ are convex functions $\forall \ell \in \mathcal{L}$ and $\forall \ell \in \mathcal{E}$, respectively, the whole optimization problem is convex. Indeed, the objective function is convex and the constraints are linear functions of the optimization variables $\boldsymbol{x}$. Therefore, the optimal solution can be found through standard techniques.

4.3 Performance Evaluation

In this section, we analytically define the cost functions $\{\phi_\ell(\cdot)\}_{\ell\in\mathcal{L}}$ and $\{\omega_\ell(\cdot)\}_{\ell\in\mathcal{E}}$ and evaluate the performance of the proposed compression and flow allocation framework for two network deployment examples.

4.3.1 Definition of ϕ_ℓ

We assume that the transport cost $\phi_\ell(x_\ell)$ of link $\ell \in \mathcal{L}$ is proportional to the transmission power for this link $P_\ell \in [0, P_{\max}]$. Specifically, P_ℓ is subject to path loss attenuation and noise. The path gain (a smaller-than-one coefficient) is a function of the distance between the transmitter and receiver pair d_ℓ and is expressed as

$$G_{\text{path}}(d_\ell) = K\left(\frac{d_0}{d_\ell}\right)^\eta = \left(\frac{\lambda}{4\pi d_0}\right)^2 \left(\frac{d_0}{d_\ell}\right)^\eta\,, \tag{4.16}$$

where d_0 is a reference distance, λ is the radio wavelength, and η is the path loss exponent. If we consider that the channel is affected by Additive White Gaussian Noise (AWGN) with constant power spectral density N_0, the noise power is obtained multiplying N_0 by the bandwidth B. Thus, the Signal-to-Noise Ratio (SNR) is expressed as a function of transmit power P_ℓ and distance d_ℓ as

$$\mathsf{SNR}(P_\ell, d_\ell) = \frac{G_{\text{path}}(d_\ell) P_\ell}{N_0 B}. \tag{4.17}$$

Note that links are interference-free, due to the adopted TDMA scheduling at the MAC layer. Finally, we define the information flow on the ℓ-th link x_ℓ exploiting the definition of ξ_ℓ in Equation (4.5) and Shannon's formula as follows:

$$x_\ell(P_\ell, d_\ell) = \frac{B \log_2 \left[1 + \mathsf{SNR}(P_\ell, d_\ell)\right]}{2 \max\{o_\ell, i_\ell\}} \text{ [bps]}. \tag{4.18}$$

Note that the channel capacity C_ℓ is defined as the information flow obtained using the maximum allowed transmit power, i.e.,

$$C_\ell = x_\ell(P_{\max}, d_\ell) \geq x_\ell(P_\ell, d_\ell). \tag{4.19}$$

The information transport cost function for the ℓ-th link, is obtained by first expressing P_ℓ as a function of x_ℓ, i.e.,

$$P_\ell(x_\ell) = \left(2^{\frac{x_\ell}{\xi_\ell B}} - 1\right) \frac{N_0 B}{G_{\text{path}}(d_\ell)}, \quad \ell = 1, \ldots, L. \tag{4.20}$$

Then, we define $\phi_\ell(x_\ell)$ normalizing $P_\ell(x_\ell)$ to the maximum transmit power $P_{\max}$, i.e.,

$$\phi_\ell(x_\ell) = \frac{P_\ell(x_\ell)}{P_{\max}} = \frac{1}{P_{\max}} \left(2^{\frac{x_\ell}{\xi_\ell B}} - 1\right) \frac{N_0 B}{G_{\text{path}}(d_\ell)}, \quad \ell = 1, \ldots, L. \tag{4.21}$$

Note that $\phi_\ell(\cdot)$ is a convex function $\forall \ell \in \mathcal{L}$.

4.3.2 Definition of ω_ℓ

For the signal's reconstruction accuracy at the application layer, we consider parameterized rate-distortion curves. Usually, closed-form expressions are available for idealized compression algorithms operating on Gaussian information sources. For practical algorithms these are however unknown and are generally obtained experimentally [161]. We denote the rate-distortion curve for a generic lossy compression algorithm by $D_\ell(x_\ell)$, which returns the distortion associated with link $\ell \in \mathcal{E}$ by transmitting the data over such link at rate x_ℓ. Using the empirical fittings of [161], and defining two parameters $\mu_\ell \in \mathbb{R}^+$ and $\delta_\ell \in \mathbb{R}^+$, we have: $D_\ell(x_\ell) = \mu_\ell[(b_\ell/x_\ell)^{\delta_\ell} - 1]$. The information distortion cost is thus defined as

$$\omega_\ell(x_\ell) = \frac{D_\ell(x_\ell)}{\Delta_\ell^{\text{thr}}} = \frac{\mu_\ell}{\Delta_\ell^{\text{thr}}} \left[\left(\frac{b_\ell}{x_\ell}\right)^{\delta_\ell} - 1\right], \quad \ell = 1, \ldots, E, \tag{4.22}$$

normalizing the distortion function $D_\ell(x_\ell)$ to the maximum distortion value allowed Δ_ℓ^{thr}. Note that $\omega_\ell(\cdot)$ is convex $\forall \ell \in \mathcal{E}$, as well.

Table 4.1: Physical layer parameters

Parameter	Values
η	3
f	2.4 GHz
N_0	-174 dBm/Hz
B	10 kHz
d_0	15 m
$P_{\max}$	20 dBm

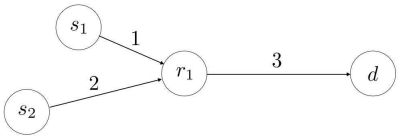


Figure 4.2: Network deployment example #1

4.3.3 Network Setup and Graphical Results

Next, the performance of two network deployments is evaluated using the physical layer parameters listed in Table 4.1.

Table 4.2: Network parameters of example #1

Parameter	Values
d_ℓ, $\ell = 1, 2, 3$	50, 45, 20 m
C_ℓ, $\ell = 1, 2, 3$	62, 63, 144 kbps
μ_ℓ, $\ell = 1, 2$	20, 3.44
δ_ℓ, $\ell = 1, 2$	0.35, 0.63
b_k, $k = 1, 2$	60, 60 kbps
Δ_ℓ^{thr}, $\ell = 1, 2$	5%, 2%

Network example #1. The first network topology is shown in Figure 4.2, with two source nodes ($S = E = 2$), one relay ($R = 1$), and the sink d. $L = 3$ wireless links are established in the network and the incidence matrix of the graph is

$$\boldsymbol{A} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -1 & -1 & 1 \\ 0 & 0 & -1 \end{bmatrix}. \tag{4.23}$$

The other network parameters are reported in Table 4.2.

Figure 4.3 shows the optimal flow values $\boldsymbol{x} = [x_1, x_2, x_3]^T$ versus the weighting parameter α. It can be seen that when α tends to zero the link flows are high: in particular, the rates of source links $\ell = 1, 2$ approach the respective link capacity, i.e., C_1 and C_2. However, as α grows, the information rates decrease. In fact, as shown in Figure 4.4, when $\alpha \to 0$ the main contribution to the objective function $f(\boldsymbol{x})$ is given by the aggregate information distortion cost σ. Conversely, when $\alpha \to 1$, the main contribution to $f(\boldsymbol{x})$ is due to the aggregate information transport cost ρ. Looking at the cost of the single links, from Figure 4.5 we see

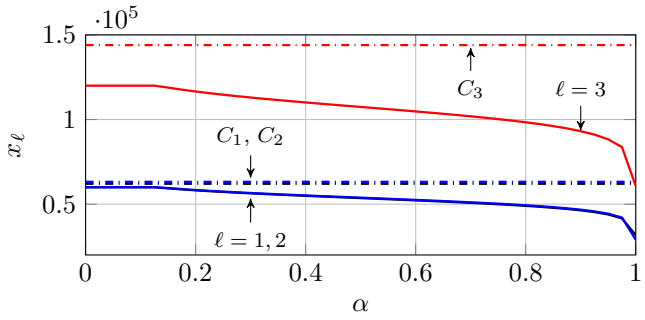


Figure 4.3: x_ℓ vs α for network example #1

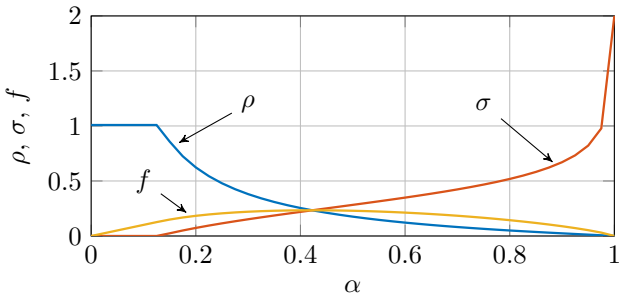


Figure 4.4: ρ, σ, and f vs α for network example #1

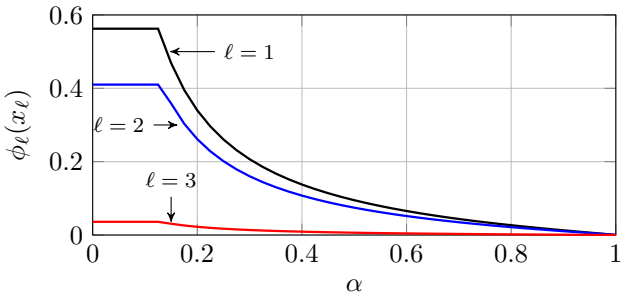


Figure 4.5: $\phi_\ell(x_\ell)$ vs α for network example #1

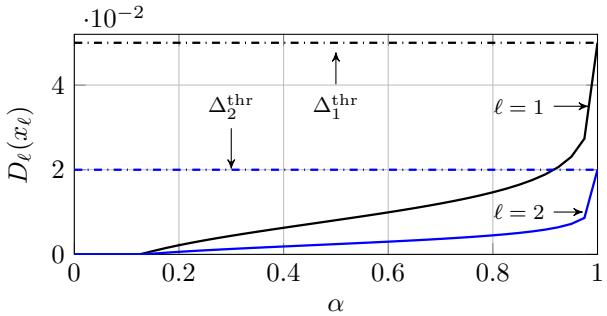


Figure 4.6: $D_\ell(x_\ell)$ vs α for network example #1

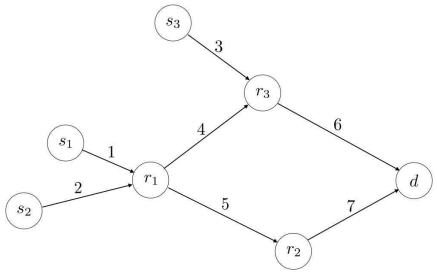


Figure 4.7: Network deployment example #2

Table 4.3: Network parameters of example #2

Parameter	Values
d_ℓ, $\ell = 1, \ldots, 7$	25, 25, 25, 25, 20, 20, 20 m
C_ℓ, $\ell = 1, \ldots, 7$	70, 70, 70, 70, 72, 72, 72 kbps
μ_ℓ, $\ell = 1, 2, 3$	20, 20, 20
δ_ℓ, $\ell = 1, 2, 3$	0.35, 0.35, 0.35
b_k, $k = 1, 2, 3$	65, 65, 65 kbps
Δ_ℓ^{thr}, $\ell = 1, 2, 3$	3%, 4%, 5%

that the transport cost $\phi_\ell(x_\ell)$ decreases for every link $\ell \in \mathcal{L}$ as α approaches one. As a consequence, in Figure 4.6 we can see that when $\alpha \to 1$ the distortion values of source data reach the respective upper bounds Δ_ℓ^{thr}.

From the results of this network deployment, we infer that source nodes have to compress their data if the objective is to reduce the transport cost, until the maximum allowed distortion is reached. The parameter α controls the tradeoff between transport and distortion costs.

Network example #2. A slightly more involved network topology is shown in Figure 4.7, with $S = E = 3$ source nodes, $R = 3$ relays, and the sink d. The incidence matrix is

$$\boldsymbol{A} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ -1 & -1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & 1 \\ 0 & 0 & -1 & -1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & -1 & -1 \end{bmatrix} \tag{4.24}$$

and the network parameters are provided in Table 4.3. Similar considerations hold: as α increases, the transport costs have a higher weight in the optimization, thus the data flows are reduced and the distortion values increase up to the respective upper bounds. However, it is worth observing from Figure 4.8 that links $\ell = 6, 7$, connecting r_3 and r_2 to d, are the bottlenecks of this network deployment. Therefore, in contrast to example #1, where the source links were the bottlenecks, in this case the bottlenecks are away from the sources. Hence, thanks to our approach, the reduction of information flows injected into the WSN by the sources offloads the links under pressure, which may be some hops ahead.

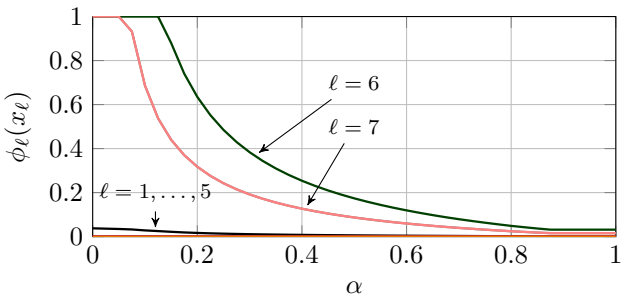


Figure 4.8: $\phi_\ell(x_\ell)$ vs α for network example #2

Finally, a plot of the Pareto frontier between σ and ρ is given in Figure 4.9: as α increases we move from the top-left corner where ρ is high and σ is low to the bottom-right one, where the situation is reversed.

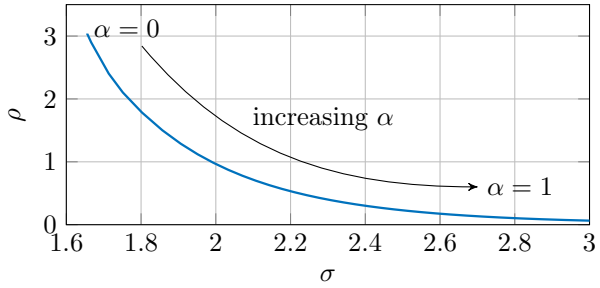


Figure 4.9: Pareto frontier for network example #2

4.4 Conclusions and Ways Forward

In this chapter, we presented a multi-objective optimization framework for the joint optimization of compression at the sources and transport costs in WSNs, and have shown some relevant tradeoffs for two selected network deployments. Our approach is an initial step toward the solution of this class of problems and is amenable to several extensions. First, the channel access scheduling at the nodes can be abstracted by considering a generic orthogonal multiplexing scheme (either in time or frequency) and improved by allowing uneven allocations. Also, the optimal solution can be obtained through distributed algorithms. Finally, additional node qualities can be added, such as amount of residual energy in their batteries, amount of energy harvested, flow priorities or delay constraints.

Summary and Conclusions

In this doctoral thesis, we addressed the support of massive Machine-to-Machine (M2M) traffic in heterogeneous networks and fifth-generation (5G) cellular networks.

Our research contributions can be divided into two parts. In the first part, we dealt with wireless communication standards for the Internet of Things (IoT). In particular, we focused on two enabling technologies for the IoT exploiting long-range wireless links, i.e., the 5G cellular network and Long-Range^{TM} (LoRa), one of the most prominent Low-Power Wide Area Network (LPWAN) technologies. Both 5G systems and LPWANs are based on a star network topology, that is, every end device is connected to a single radio concentrator via a single hop; on the other hand, while 5G operates on licensed frequency bands, LPWANs utilize unlicensed spectrum to communicate.

As for the 5G, we first identified and discussed the issues of the current cellular network, i.e., the Long-Term Evolution (LTE) standard, in supporting massive uplink (UL) traffic coming from IoT terminals. Then, we surveyed the state of the art, to understand how other researchers tackled the so-called *massive access problem* in LTE. Finally, we presented a contention-based radio access protocol for 5G which overcomes the issues of LTE; the core of this work consists in the mathematical model of the various radio access protocols to accommodate IoT traffic on cellular networks. The proposed approach provides a much shorter delivery delay and a massive reduction in the downlink (DL) feedback.

Regarding LoRa, after a brief survey about this technology and its competitors, we evaluated the performance of large-scale LoRa deployments through extensive simulation campaigns, showing that tens of thousands nodes can be served efficiently. Moreover, we studied the negative impact induced by acknowledgement messages sent in DL on a LoRa network.

The second part of the thesis, instead, dealt with fundamental research on the IoT, proposing innovative research approaches about the IoT paradigm at large. In particular, we focused on physical layer security mechanisms for IoT, with the aim of authenticating messages exchanged in the network and verifying the location of Machine-Type Devices (MTDs) by exploiting the channel estimates of a group of trusted anchor nodes. Then, we proposed innovative flow allocation strategies for Wireless Sensor Networks (WSNs), investigating the trade-off between the cost of transmitting data (transport cost) and the cost of compressing them (compression cost) in order to optimize the allocation of flows of data on the wireless links of a WSN. We found that by enabling the compression of raw data at the source nodes, we can offload the wireless links that generate a bottleneck in the transport network.

Bibliography

[1] A. Zanella, N. Bui, A. Castellani, L. Vangelista, and M. Zorzi, "Internet of Things for Smart Cities," *IEEE Internet Things J.*, vol. 1, no. 1, pp. 22–32, Feb. 2014.

[2] L. Atzori, A. Iera, and G. Morabito, "The Internet of Things: A survey," *Comput. Netw.*, vol. 54, no. 15, pp. 2787–2805, Oct. 2010.

[3] P. Bellavista, G. Cardone, A. Corradi, and L. Foschini, "Convergence of MANET and WSN in IoT Urban Scenarios," *IEEE Sensors J.*, vol. 13, no. 10, pp. 3558–3567, Oct. 2013.

[4] NGMN, "5G White Paper," Feb. 2015. [Online]. Available: https://www.ngmn.org/uploads/media/NGMN_5G_White_Paper_V1_0.pdf

[5] C. Pielli, D. Zucchetto, A. Zanella, L. Vangelista, and M. Zorzi, "Platforms and Protocols for the Internet of Things," *EAI Endorsed Trans. IoT*, vol. 15, no. 1, Oct. 2015.

[6] S. Iraji, P. Mogensen, and R. Ratasuk, "Recent Advances in M2M Communications and Internet of Things (IoT)," *Int. J. of Wireless Inf. Netw.*, vol. 24, no. 3, pp. 240–242, Sep. 2017. [Online]. Available: https://doi.org/10.1007/s10776-017-0362-3

[7] J. Gubbi, R. Buyya, S. Marusic, and M. Palaniswami, "Internet of Things (IoT): A Vision, Architectural Elements, and Future Directions," *Future Generation Comput. Syst.*, vol. 29, no. 7, pp. 1645–1660, Sep. 2013. [Online]. Available: http://dx.doi.org/10.1016/j.future.2013.01.010

[8] D. Miorandi, S. Sicari, F. De Pellegrini, and I. Chlamtac, "Internet of Things," *Ad Hoc Netw.*, vol. 10, no. 7, pp. 1497–1516, Sep. 2012. [Online]. Available: http://dx.doi.org/10.1016/j.adhoc.2012.02.016

[9] K. Zheng, T. Lv, Y. Li, and Y. Lu, "The Analysis and Implementation of AllJoyn Based Thin Client Communication System with Heartbeat Function," in *Proc. Int. Conf. Cyberspace Technol. (CCT)*, Beijing, China, Nov. 2014, pp. 1–4.

[10] H. Cha, W. Lee, and J. Jeon, "Standardization Strategy for the Internet of Wearable Things," in *Int. Conf. Inform. and Commun. Technol. Convergence (ICTC)*, Jeju, South Korea, Oct. 2015, pp. 1138–1142.

[11] M. Hernandez and R. Kohno, "UWB Systems for Body Area Networks in IEEE 802.15.6," in *Proc. IEEE Int. Conf. Ultra-Wideband (ICUWB)*, Bologna, Italy, Sep. 2011, pp. 235–239.

[12] S. A. Salehi, M. A. Razzaque, I. Tomeo-Reyes, and N. Hussain, "IEEE 802.15.6 Standard in Wireless Body Area Networks From a Healthcare Point of View," in *Proc. Asia-Pacific Conf. Commun. (APCC)*, Yogyakarta, Indonesia, Aug. 2016, pp. 523–528.

[13] C. Anton-Haro and M. Dohler, *Machine-to-Machine (M2M) Communications: Architecture, Performance and Applications*, 1st ed. Woodhead Publishing Ltd., Jan. 2015.

[14] S.-Y. Lien, K.-C. Chen, and Y. Lin, "Toward Ubiquitous Massive Accesses in 3GPP Machine-to-Machine Communications," *IEEE Commun. Mag.*, vol. 49, no. 4, pp. 66–74, Apr. 2011.

[15] Vodafone Group Plc., "New Study Item on Cellular System Support for Ultra Low Complexity and Low Throughput Internet of Things," 3GPP TSG GERAN#62, Tech. Rep. GP-140421, May 2014.

[16] F. Boccardi, R. W. Heath, A. Lozano, T. L. Marzetta, and P. Popovski, "Five disruptive technology directions for 5G," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 74–80, Feb. 2014.

[17] Cisco, "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2016–2021," Tech. Rep., Feb. 2017. [Online]. Available: http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white_paper_c11-520862.pdf

[18] H. Holma and A. Toskala, *WCDMA for UMTS: HSPA Evolution and LTE*, 4th ed. John Wiley & Sons, Inc., Sep. 2007.

[19] E. Dahlman, S. Parkvall, and J. Skold, *4G: LTE/LTE-Advanced for Mobile Broadband*, 1st ed. Academic Press, Mar. 2011.

[20] M. Z. Shafiq, L. Ji, A. X. Liu, J. Pang, and J. Wang, "A First Look at Cellular Machine-to-Machine Traffic: Large Scale Measurement and Characterization," in *Proc. ACM SIGMETRICS/PERFORMANCE Joint Int. Conf. Meas. and Modeling of Comp. Syst.*, London, UK, Jun. 2012, pp. 65–76.

[21] D. Chornaya, A. Paramonov, and A. Koucheryavy, "Investigation of machine-to-machine traffic generated by mobile terminals," in *Proc. Int. Congr. Ultra Modern Telecommun. and Control Syst. and Workshops (ICUMT)*, St. Petersburg, Russia, Oct. 2014, pp. 210–213.

[22] W. Fischer and K. Meier-Hellstern, "The Markov-Modulated Poisson Process (MMPP) Cookbook," *Perf. Eval.*, vol. 18, no. 2, pp. 149–171, Sep. 1993.

[23] M. Laner, P. Svoboda, N. Nikaein, and M. Rupp, "Traffic Models for Machine Type Communications," in *Proc. Int. Symp. Wireless Commun. Syst. (ISWCS)*, Ilmenau, Germany, Aug. 2013, pp. 1–5.

[24] R. Nelson, *Probability, Stochastic Processes, and Queueing Theory: The Mathematics of Computer Performance Modeling*, 1st ed. Springer-Verlag New York, Inc., Jun. 1995.

[25] N. Nikaein, M. Laner, K. Zhou, P. Svoboda, D. Drajic, M. Popovic, and S. Krco, "Simple Traffic Modeling Framework for Machine Type Communication," in *Proc. Int. Symp. Wireless Commun. Syst. (ISWCS)*, Ilmenau, Germany, Aug. 2013, pp. 1–5.

[26] 3GPP, "Study on RAN Improvements for Machine-Type Communications," Tech. Rep. 37.868 V11.0.0, Sep. 2011.

[27] ——, "Study on provision of low-cost Machine-Type Communications (MTC) User Equipments (UEs) based on LTE," Tech. Rep. 36.888 V12.0.0, Jun. 2013.

[28] ——, "Cellular System Support for Ultra-Low Complexity and Low Throughput Internet of Things (CIoT)," Tech. Rep. 45.820 V13.1.0, Nov. 2015.

[29] IEEE, "IEEE Standard for Air Interface for Broadband Wireless Access Systems – Amendment 1: Enhancements to Support Machine-to-Machine Applications," *IEEE Standard 802.16p-2012*, Oct. 2012. [Online]. Available: http://standards.ieee.org/findstds/standard/802.16p-2012.html

[30] ——, "M2M Traffic Characteristics," Tech. Rep. IEEE C802.16p-11/0062, May 2011. [Online]. Available: http://ieee802.org/16/m2m/

[31] ——, "Machine-to-Machine (M2M) Communications Technical Report," Tech. Rep. IEEE 802.16p-10/0005, Nov. 2010. [Online]. Available: http://ieee802.org/16/m2m/#10_0005

[32] ——, "IEEE 802.16p Machine-to-Machine (M2M) Evaluation Methodology Document (EMD)," Tech. Rep. IEEE 802.16p-11/0014, May 2011. [Online]. Available: http://ieee802.org/16/m2m/#11_0014

[33] M. Amirijoo, P. Frenger, F. Gunnarsson, J. Moe, and K. Zetterberg, "On Self-Optimization of the Random Access Procedure in 3G Long Term Evolution," in *Proc. IFIP/IEEE Int. Symp. Integr. Netw. Manag.*, New York, NY, USA, Jun. 2009, pp. 177–184.

[34] P. Bertrand and J. Jiang, "Random Access," in *LTE – The UMTS Long Term Evolution: From Theory to Practice*, 1st ed., S. Sesia, I. Toufik, and M. Baker, Eds. John Wiley & Sons, Inc., Feb. 2009, ch. 19, pp. 421–457.

[35] 3GPP, "Base Station (BS) Radio Transmission and Reception," *Tech. Spec. 36.104 V13.0.0*, pp. 1–156, Jul. 2015.

[36] G. C. Madueno, C. Stefanovic, and P. Popovski, "Reengineering GSM/GPRS towards a dedicated network for massive smart metering," in *Proc. IEEE Int. Conf. Smart Grid Commun. (SmartGridComm)*, Venice, Italy, Nov. 2014, pp. 338–343.

[37] [Online]. Available: https://www.nsnam.org/

[38] J. Blumenstein, J. C. Ikuno, J. Prokopec, and M. Rupp, "Simulating the Long Term Evolution Uplink Physical Layer," in *Proc. ELMAR-2011*, Zadar, Croatia, Sep. 2011, pp. 141–144.

[39] A. Virdis, G. Stea, and G. Nardini, "SimuLTE - A Modular System-Level Simulator for LTE/LTE-A Networks Based on OMNeT++," in *Proc. Int. Conf. Simulation and Modeling Methodologies, Technol. and Appl. (SIMULTECH)*, Vienna, Austria, Aug. 2014, pp. 59–70.

[40] G. Piro, L. A. Grieco, G. Boggia, F. Capozzi, and P. Camarda, "Simulating LTE Cellular Systems: An Open-Source Framework," *IEEE Trans. Veh. Technol.*, vol. 60, no. 2, pp. 498–513, Feb. 2011.

[41] F. J. López-Martínez, E. del Castillo-Sánchez, E. Martos-Naya, and J. T. Entrambasaguas, "Performance Evaluation of Preamble Detectors for 3GPP-LTE Physical Random Access Channel," *Digit. Signal Process.*, vol. 22, no. 3, pp. 526–534, May 2012.

[42] N. Baldo, M. Miozzo, M. Requena-Esteso, and J. Nin-Guerrero, "An Open Source Product-oriented LTE Network Simulator Based on ns–3," in *Proc. ACM Int. Conf. Modeling, Anal. and Simulation of Wireless and Mobile Syst.*, Miami, Florida, USA, Nov. 2011, pp. 293–298.

[43] 3GPP, "Physical Layer Procedures," *Tech. Spec. 36.213 V12.6.0*, pp. 1–241, Jun. 2015.

[44] ——, "Medium Access Control (MAC) Protocol Specification," *Tech. Spec. 36.321 V12.6.0*, pp. 1–77, Jun. 2015.

[45] R. C. D. Paiva, R. D. Vieira, and M. Saily, "Random Access Capacity Evaluation with Synchronized MTC Users over Wireless Networks," in *Proc. IEEE Veh. Technol. Conf. (VTC Spring)*, Budapest, Hungary, May 2011, pp. 1–5.

[46] G. C. Madueno, C. Stefanovic, and P. Popovski, "How Many Smart Meters Can Be Deployed in a GSM Cell?" in *Proc. IEEE Int. Conf. Commun. (ICC)*, Budapest, Hungary, Jun. 2013, pp. 1263–1268.

[47] P. Jain, P. Hedman, and H. Zisimopoulos, "Machine Type Communications in 3GPP Systems," *IEEE Commun. Mag.*, vol. 50, no. 11, pp. 28–35, Nov. 2012.

[48] Ericsson, "Cellular Networks for Massive IoT," Jan. 2016. [Online]. Available: https://www.ericsson.com/assets/local/publications/white-papers/wp_iot.pdf

[49] 3GPP, "Study on Enhancements to Machine-Type Communications (MTC) and Other Mobile Data Applications," Tech. Rep. 37.869 V12.0.0, Sep. 2013.

[50] R. Ratasuk, B. Vejlgaard, N. Mangalvedhe, and A. Ghosh, "NB-IoT System for M2M Communication," in *Proc. IEEE Wireless Commun. and Netw. Conf. (WCNC)*, Doha, Qatar, Apr. 2016, pp. 1–5.

[51] A. Laya, L. Alonso, and J. Alonso-Zarate, "Is the Random Access Channel of LTE and LTE-A Suitable for M2M Communications? A Survey of Alternatives," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 1, pp. 4–16, First Quarter 2014.

[52] A. Lo, Y. W. Law, M. Jacobsson, and M. Kucharzak, "Enhanced LTE-Advanced Random-Access Mechanism for Massive Machine-to-Machine (M2M) Communications," in *Proc. World Wireless Res. Forum Meeting (WWRF)*, Düsseldorf, Germany, Oct. 2011, pp. 1–7.

[53] K. S. Ko, M. J. Kim, K. Y. Bae, D. K. Sung, J. H. Kim, and J. Y. Ahn, "A Novel Random Access for Fixed-Location Machine-to-Machine Communications in OFDMA Based Systems," *IEEE Commun. Lett.*, vol. 16, no. 9, pp. 1428–1431, Sep. 2012.

[54] T. Taleb and A. Kunz, "Machine Type Communications in 3GPP Networks: Potential, Challenges, and Solutions," *IEEE Commun. Mag.*, vol. 50, no. 3, pp. 178–184, Mar. 2012.

[55] L. M. Bello, P. Mitchell, and D. Grace, "Application of Q-Learning for RACH Access to Support M2M Traffic over a Cellular Network," in *Proc. Eur. Wireless Conf.*, Barcelona, Spain, May 2014, pp. 1–6.

[56] Y. C. Pang, S. L. Chao, G. Y. Lin, and H. Y. Wei, "Network Access for M2M/H2H Hybrid Systems: a Game Theoretic Approach," *IEEE Commun. Lett.*, vol. 18, no. 5, pp. 845–848, Jun. 2014.

[57] C. Y. Tu, C. Y. Ho, and C. Y. Huang, "Energy-Efficient Algorithms and Evaluations for Massive Access Management in Cellular Based Machine to Machine Communications," in *Proc. IEEE Veh. Technol. Conf. (VTC Fall)*, San Francisco, CA, USA, Sep. 2011, pp. 1–5.

[58] C. Y. Ho and C. Y. Huang, "Energy-Saving Massive Access Control and Resource Allocation Schemes for M2M Communications in OFDMA Cellular Networks," *IEEE Wireless Commun. Lett.*, vol. 1, no. 3, pp. 209–212, Jun. 2012.

[59] S. Y. Lien, K. C. Chen, and Y. Lin, "Toward Ubiquitous Massive Accesses in 3GPP Machine-to-Machine Communications," *IEEE Commun. Mag.*, vol. 49, no. 4, pp. 66–74, Apr. 2011.

[60] S. Y. Lien and K. C. Chen, "Massive Access Management for QoS Guarantees in 3GPP Machine-to-Machine Communications," *IEEE Commun. Lett.*, vol. 15, no. 3, pp. 311–313, Mar. 2011.

[61] P. Si, J. Yang, S. Chen, and H. Xi, "Adaptive Massive Access Management for QoS Guarantees in M2M Communications," *IEEE Trans. Veh. Technol.*, vol. 64, no. 7, pp. 3152–3166, Jul. 2015.

[62] S. Zhenqi, Y. Haifeng, C. Xuefen, and L. Hongxia, "Research on Uplink Scheduling Algorithm of Massive M2M and H2H Services in LTE," in *Proc. IET Int. Conf. Inf. Commun. Technol. (IETICT)*, Beijing, China, Apr. 2013, pp. 365–369.

[63] S. Bayat, Y. Li, Z. Han, M. Dohler, and B. Vucetic, "Distributed Massive Wireless Access for Cellular Machine-to-Machine Communication," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Sydney, Australia, Jun. 2014, pp. 2767–2772.

[64] M. Hasan, E. Hossain, and D. Niyato, "Random Access for Machine-to-Machine Communication in LTE-Advanced Networks: Issues and Approaches," *IEEE Commun. Mag.*, vol. 51, no. 6, pp. 86–93, Jun. 2013.

[65] H. S. Dhillon, H. Huang, H. Viswanathan, and R. A. Valenzuela, "Fundamentals of Throughput Maximization With Random Arrivals for M2M Communications," *IEEE Trans. Commun.*, vol. 62, no. 11, pp. 4094–4109, Nov. 2014.

[66] H. Wu, C. Zhu, R. J. La, X. Liu, and Y. Zhang, "FASA: Accelerated S-ALOHA Using Access History for Event-Driven M2M Communications," *IEEE/ACM Trans. Netw.*, vol. 21, no. 6, pp. 1904–1917, Dec. 2013.

[67] C. Stefanovic and P. Popovski, "ALOHA Random Access that Operates as a Rateless Code," *IEEE Trans. Commun.*, vol. 61, no. 11, pp. 4653–4662, Nov. 2013.

[68] C. Stefanovic, M. Momoda, and P. Popovski, "Exploiting Capture Effect in Frameless ALOHA for Massive Wireless Random Access," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Istanbul, Turkey, Apr. 2014, pp. 1762–1767.

[69] X. Wang and J. J. Garcia-Luna-Aceves, "Embracing Interference in Ad Hoc Networks Using Joint Routing and Scheduling with Multiple Packet Reception," *Ad Hoc Netw.*, vol. 7, no. 2, pp. 460–471, Mar. 2009.

[70] A. Zanella and M. Zorzi, "Theoretical Analysis of the Capture Probability in Wireless Systems with Multiple Packet Reception Capabilities," *IEEE Trans. Commun.*, vol. 60, no. 4, pp. 1058–1071, Apr. 2012.

[71] F. Schaich, T. Wild, and R. Ahmed, "Subcarrier Spacing - How to Make Use of This Degree of Freedom," in *Proc. IEEE Veh. Technol. Conf. (VTC Spring)*, Nanjing, China, May 2016, pp. 1–6.

[72] M. Fuhrwerk, J. Peissig, and M. Schellmann, "On the design of an FBMC based AIR interface enabling channel adaptive pulse shaping per sub-band," in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, Nice, France, Aug. 2015, pp. 384–388.

[73] L. Zhang, A. Ijaz, P. Xiao, A. Quddus, and R. Tafazolli, "Subband Filtered Multi-Carrier Systems for Multi-Service Wireless Communications," *IEEE Trans. Wireless Commun.*, vol. 16, no. 3, pp. 1893–1907, Mar. 2017.

[74] S. Saur, A. Weber, and G. Schreiber, "Radio Access Protocols and Preamble Design for Machine Type Communications in 5G," in *Proc. Asilomar Conf. Signals, Syst. and Comput.*, Pacific Grove, CA, USA, Nov. 2015, pp. 8–12.

[75] W. Xu and G. Campbell, "A Near Perfect Stable Random Access Protocol for a Broadcast Channel," in *Proc. IEEE Int. Conf. Commun. (ICC)*, vol. 1, Chicago, IL, USA, Jun. 1992, pp. 370–374.

[76] J. Alonso-Zarate, C. Verikoukis, E. Kartsakli, A. Cateura, and L. Alonso, "A Near-Optimum Cross-Layered Distributed Queuing Protocol for Wireless LAN," *IEEE Wireless Commun.*, vol. 15, no. 1, pp. 48–55, Feb. 2008.

[77] 3GPP, "Multiplexing and Channel Coding," *Tech. Spec. 36.212 V8.8.0*, pp. 1–60, Dec. 2009.

[78] C. Kahn and H. Viswanathan, "Connectionless Access for Mobile Cellular Networks," *IEEE Commun. Mag.*, vol. 53, no. 9, pp. 26–31, Sep. 2015.

[79] H. S. Dhillon, H. Huang, and H. Viswanathan, "Wide-area Wireless Communication Challenges for the Internet of Things," *IEEE Commun. Mag.*, vol. 55, no. 2, pp. 168–174, Feb. 2017.

[80] K. Zhou, N. Nikaein, R. Knopp, and C. Bonnet, "Contention Based Access for Machine-Type Communications over LTE," in *Proc. IEEE Veh. Technol. Conf. (VTC Spring)*, Yokohama, Japan, May 2012, pp. 1–5.

[81] R. P. Jover and I. Murynets, "Connection-Less Communication of IoT Devices Over LTE Mobile Networks," in *Proc. IEEE Int. Conf. Sens. Commun. Netw. (SECON)*, Seattle, WA, USA, Jun. 2015, pp. 247–255.

[82] F. Schaich and T. Wild, "Relaxed Synchronization Support of Universal Filtered Multi-Carrier Including Autonomous Timing Advance," in *Proc. Int. Symp. Wireless Commun. Syst. (ISWCS)*, Barcelona, Spain, Aug. 2014, pp. 203–208.

[83] L. Zhang, P. Xiao, A. Zafar, A. ul Quddus, and R. Tafazolli, "FBMC System: An Insight into Doubly Dispersive Channel Impact," *IEEE Trans. Veh. Technol.*, vol. 66, no. 5, pp. 3942–3956, May 2017.

[84] X. Zhang, M. Jia, L. Chen, J. Ma, and J. Qiu, "Filtered-OFDM – Enabler for Flexible Waveform in the 5th Generation Cellular Networks," in *Proc. IEEE Global Commun. Conf. (Globecom)*, San Diego, CA, USA, Dec. 2015, pp. 1–6.

[85] J. J. Nielsen, D. M. Kim, G. C. Madueno, N. K. Pratas, and P. Popovski, "A Tractable Model of the LTE Access Reservation Procedure for Machine-Type Communications," in *Proc. IEEE Global Commun. Conf. (Globecom)*, San Diego, CA, USA, Dec. 2015, pp. 1–6.

[86] G. Bianchi, "Performance Analysis of the IEEE 802.11 Distributed Coordination Function," *IEEE J. Sel. Areas Commun.*, vol. 18, no. 3, pp. 535–547, Mar. 2000.

[87] A. G. D. Kok and H. G. Tijms, "A Queueing System with Impatient Customers," *J. Appl. Probability*, vol. 22, no. 3, pp. 688–696, Sep. 1985.

[88] N. K. Boots and H. Tijms, "A Multiserver Queueing System with Impatient Customers," *Manag. Sci.*, vol. 45, no. 3, pp. 444–448, Mar. 1999.

[89] J. C. Guey, "The Design and Detection of Signature Sequences in Time-Frequency Selective Channel," in *Proc. Int. Symp. Personal, Indoor and Mobile Radio Commun. (PIMRC)*, Cannes, France, Sep. 2008, pp. 1–5.

[90] S. Zeltyn, "Call Centers With Impatient Customers: Exact Analysis and Many-Server Asymptotics of the M/M/n+G Queue," Ph.D. dissertation, Israel Institute of Technology, Oct. 2004.

[91] Y. H. Nam, B. L. Ng, K. Sayana, Y. Li, J. Zhang, Y. Kim, and J. Lee, "Full-Dimension MIMO (FD-MIMO) for Next Generation Cellular Technology," *IEEE Commun. Mag.*, vol. 51, no. 6, pp. 172–179, Jun. 2013.

[92] E. G. Larsson, O. Edfors, F. Tufvesson, and T. L. Marzetta, "Massive MIMO for Next Generation Wireless Systems," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 186–195, Feb. 2014.

[93] N. Bhushan, J. Li, D. Malladi, R. Gilmore, D. Brenner, A. Damnjanovic, R. T. Sukhavasi, C. Patel, and S. Geirhofer, "Network Densification: the Dominant Theme for Wireless Evolution Into 5G," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 82–89, Feb. 2014.

[94] T. S. Rappaport, S. Sun, R. Mayzus, H. Zhao, Y. Azar, K. Wang, G. N. Wong, J. K. Schulz, M. Samimi, and F. Gutierrez, "Millimeter Wave Mobile Communications for 5G Cellular: It Will Work!" *IEEE Access*, vol. 1, pp. 335–349, 2013.

[95] P. Demestichas, A. Georgakopoulos, D. Karvounas, K. Tsagkaris, V. Stavroulaki, J. Lu, C. Xiong, and J. Yao, "5G on the Horizon: Key Challenges for the Radio-Access Network," *IEEE Veh. Technol. Mag.*, vol. 8, no. 3, pp. 47–53, Sep. 2013.

[96] CEPT, "ERC 70-03 Relating to the Use of Short Range Devices (SRD)," Tech. Rep., May 2017.

[97] G. Gardasevic, S. Mijovic, A. Stajkic, and C. Buratti, "On the Performance of 6LoWPAN Through Experimentation," in *Proc. Int. Wireless Commun. Mobile Computing Conf. (IWCMC)*, Dubrovnik, Croatia, Aug. 2015, pp. 696–701.

[98] U. Raza, P. Kulkarni, and M. Sooriyabandara, "Low Power Wide Area Networks: An Overview," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 2, pp. 855–873, Second Quarter 2017.

[99] M. Weyn, G. Ergeerts, R. Berkvens, B. Wojciechowski, and Y. Tabakov, "DASH7 Alliance Protocol 1.0: Low-Power, Mid-Range Sensor and Actuator Communication," in *Proc. IEEE Conf. Standards for Commun. and Netw. (CSCN)*, Tokyo, Japan, Oct. 2015, pp. 54–59.

[100] M. Centenaro, "DASH7: Come Coniugare Facilmente Comunicazione e Identificazione a Radiofrequenza," Bachelor's thesis (in Italian), University of Padova, Jul. 2012.

[101] ISO, "Information Technology – Radio Frequency Identification for Item Management – Part 7: Parameters for Active Air Interface Communications at 433 MHz," *ISO/IEC 18000-7:2014*, pp. 1–202, Sep. 2014.

[102] IEEE, "IEEE Standard for Local and Metropolitan Area Networks – Part 15.4: Low-Rate Wireless Personal Area Networks (LR-WPANs) – Amendment 5: Physical Layer Specifications for Low Energy, Critical Infrastructure Monitoring Networks," *IEEE Standard 802.15.4k-2013*, pp. 1–149, Aug. 2013.

[103] On-Ramp Wireless Inc., "Light Monitoring System Using a Random Phase Multiple Access System," US Patent 8 477 830, Jul. 2, 2013. [Online]. Available: https://patents.google.com/patent/US8477830B2/en

[104] F. Sforza, "Communication System," US Patent 8 406 275, Mar. 26, 2013. [Online]. Available: https://patents.google.com/patent/US8406275B2/en

[105] A. J. Berni and W. Gregg, "On the Utility of Chirp Modulation for Digital Signaling," *IEEE Trans. Commun.*, vol. 21, no. 6, pp. 748–751, Jun. 1973.

[106] Semtech Corp., "LoRa Modulation Basics," Tech. Rep. AN1200.22, May 2015. [Online]. Available: http://www.semtech.com/images/datasheet/an1200.22.pdf

[107] ETSI, "Electromagnetic Compatibility and Radio Spectrum Matters (ERM); Short Range Devices (SRD); Radio Equipment to Be Used in the 25 MHz to 1000 MHz Frequency Range with Power Levels Ranging up to 500 mW; Part 1: Technical Characteristics and Test Methods," Tech. Rep. EN 300 220-1 V2.4.1, Jan. 2012.

[108] LoRa Alliance, "LoRaWAN Specification," Tech. Rep. V1.0, Jan. 2015.

[109] T. Rebbeck, M. Mackenzie, and N. Afonso, "Low-Powered Wireless Solutions Have the Potential to Increase the M2M Market by Over 3 Billion Connections," Analysis Mason, Tech. Rep., Sep. 2014. [Online]. Available: http://www.analysysmason.com/Research/Content/Reports/Low-powered-wireless-solutions-have-the-potential-to-increase-the-M2M-market-by-over-3-billion-connections/

[110] J. Manyika, M. Chui, P. Bisson, J. Woetzel, R. Dobbs, J. Bughin, and D. Aharon, "The Internet of Things: Mapping the Value Beyond the Hype," McKinsey Global Institute, Tech. Rep., Jun. 2015. [Online]. Available: http://www.mckinsey.com/business-functions/business-technology/our-insights/the-internet-of-things-the-value-of-digitizing-the-physical-world

[111] C. Goursaud and J.-M. Gorce, "Dedicated Networks for IoT: PHY/MAC State of the Art and Challenges," *EAI Endorsed Trans. IoT*, Oct. 2015. [Online]. Available: https://hal.archives-ouvertes.fr/hal-01231221

[112] A. Augustin, J. Yi, T. Clausen, and W. M. Townsley, "A Study of LoRa: Long Range & Low Power Networks for the Internet of Things," *Sensors*, vol. 16, no. 9, Sep. 2016. [Online]. Available: http://www.mdpi.com/1424-8220/16/9/1466

[113] *SX1301 Datasheet*, Semtech Corp., Jun. 2014, V2.01.

[114] J. C. Ikuno, M. Wrulich, and M. Rupp, "System Level Simulation of LTE Networks," in *Proc. IEEE Veh. Technol. Conf. (VTC)*, Taipei, Taiwan, May 2010, pp. 1–5.

[115] 3GPP, "Radio Frequency (RF) System Scenarios," Tech. Rep. 36.942 V13.0.0, Jan. 2016.

[116] T. Petrić, M. Goessens, L. Nuaymi, A. Pelov, and L. Toutain, "Measurements, Performance and Analysis of LoRa FABIAN, a Real-World Implementation of LPWAN," Jun. 2016, working paper or preprint. [Online]. Available: https://hal-institut-mines-telecom.archives-ouvertes.fr/hal-01331966

[117] S. S. Szyszkowicz, H. Yanikomeroglu, and J. S. Thompson, "On the Feasibility of Wireless Shadowing Correlation Models," *IEEE Trans. Veh. Technol.*, vol. 59, no. 9, pp. 4222–4236, Nov. 2010.

[118] R. Fraile, J. F. Monserrat, J. Gozálvez, and N. Cardona, "Mobile radio bidimensional large-scale fading modelling with site-to-site cross-correlation," *Eur. Trans. Telecommun.*, vol. 19, no. 1, pp. 101–106, Feb. 2007.

[119] M. Gudmundson, "Correlation model for shadow fading in mobile radio systems," *Electron. Lett.*, vol. 27, no. 23, pp. 2145–2146, Nov. 1991.

[120] H. Claussen, "Efficient Modelling of Channel Maps With Correlated Shadow Fading in Mobile Radio Systems," in *Proc. IEEE Int. Symp. on Personal, Indoor and Mobile Radio Commun. (PIMRC)*, Berlin, Germany, Sep. 2005, pp. 512–516.

[121] S. Schlegel, N. Korn, and G. Scheuermann, "On the Interpolation of Data with Normally Distributed Uncertainty for Visualization," *IEEE Trans. Vis. Comput. Graphics*, vol. 18, no. 12, pp. 2305–2314, Dec. 2012.

[122] N. Benvenuto and M. Zorzi, *Principles of Communications Networks and Systems*. Wiley, Aug. 2011.

[123] N. Varsier and J. Schwoerer, "Capacity Limits of LoRaWAN Technology for Smart Metering Applications," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Paris, France, May 2017, pp. 1–6.

[124] A.-I. Pop, U. Raza, P. Kulkarni, and M. Sooriyabandara, "Does Bidirectional Traffic Do More Harm Than Good in LoRaWAN Based LPWA Networks?" Apr. 2017, ArXiv Preprint. [Online]. Available: arxiv.org/abs/1704.04174

[125] F. Van den Abeele, J. Haxhibeqiri, I. Moerman, and J. Hoekebe, "Scalability Analysis of Large-Scale LoRaWAN Networks in ns-3," *IEEE Internet Things J.*, vol. 4, no. 6, pp. 2186–2198, Dec. 2017.

[126] *SX1276/77/78/79 Datasheet*, Semtech Corp., Mar. 2015, rev. 4.

[127] N. Sastry and D. Wagner, "Security Considerations for IEEE 802.15.4 Networks," in *Proc. ACM Workshop on Wireless Security (WiSe)*, Philadelphia, PA, USA, Oct. 2004, pp. 32–42.

[128] IEEE, "IEEE Standard for Low-Rate Wireless Networks," *IEEE Std. 802.15.4-2015*, pp. 1–709, Apr. 2016.

[129] NIST, "Recommendation for Block Cipher Modes of Operation: The CCM Mode for Authentication and Confidentiality," Tech. Rep. Special Publication 800-38C, May 2004.

[130] E. Jorswieck, S. Tomasin, and A. Sezgin, "Broadcasting Into the Uncertainty: Authentication and Confidentiality by Physical-Layer Processing," *Proc. IEEE*, vol. 103, no. 10, pp. 1702–1724, Oct. 2015.

[131] G. J. Simmons, "Authentication Theory/Coding Theory," in *Proc. CRYPTO 84 on Advances in Cryptology*, Santa Barbara, California, USA, Aug. 1985, pp. 411–431.

[132] G. Simmons, "A Survey of Information Authentication," *Proc. IEEE*, vol. 76, no. 5, pp. 603–620, May 1988.

[133] U. Maurer, "Authentication Theory and Hypothesis Testing," *IEEE Trans. Inf. Theory*, vol. 46, no. 4, pp. 1350–1356, Jul. 2000.

[134] P. L. Yu, J. S. Baras, and B. M. Sadler, "Physical-Layer Authentication," *IEEE Trans. Inf. Forensics Security*, vol. 3, no. 1, pp. 38–51, Mar. 2008.

[135] P. L. Yu, J. Baras, and B. M. Sadler, "Power Allocation Tradeoffs in Multicarrier Authentication Systems," in *Proc. IEEE Sarnoff Symp.*, Princeton, NJ, USA, Mar. 2009, pp. 1–5.

[136] L. Lai, H. El Gamal, and H. Poor, "Authentication Over Noisy Channels," *IEEE Trans. Inf. Theory*, vol. 55, no. 2, pp. 906–916, Feb. 2009.

[137] P. Baracca, N. Laurenti, and S. Tomasin, "Physical Layer Authentication over an OFDM Fading Wiretap Channel," in *Proc. Int. Conf. Perf. Eval. Methodologies and Tools (ICST)*, Paris, France, 2011, pp. 648–657.

[138] ——, "Physical Layer Authentication over MIMO Fading Wiretap Channels," *IEEE Trans. Wireless Commun.*, vol. 11, no. 7, pp. 2564–2573, Jul. 2012.

[139] A. Ferrante, N. Laurenti, C. Masiero, M. Pavon, and S. Tomasin, "On the Error Region for Channel Estimation-Based Physical Layer Authentication over Rayleigh Fading," *IEEE Trans. Inf. Forensics Security*, vol. 10, no. 5, pp. 941–952, Jan. 2015.

[140] S. Jiang, "Keyless Authentication in a Noisy Model," *IEEE Trans. Inf. Forensics Security*, vol. 9, no. 6, pp. 1024–1033, Jun. 2014.

[141] S. Kay, *Fundamentals of Statistical Signal Processing: Detection Theory.* Prentice Hall, Mar. 1993.

[142] H. V. Khuong and H. Y. Kong, "General Expression for PDF of a Sum of Independent Exponential Random Variables," *IEEE Commun. Lett.*, vol. 10, no. 3, pp. 159–161, Mar. 2006.

[143] Y. Zeng, J. Cao, J. Hong, and L. Xie, "Secure Localization and Location Verification in Wireless Sensor Networks," in *Proc. IEEE Int. Conf. Mobile Adhoc and Sensor Syst.*, Macau, China, Oct. 2009, pp. 864–869.

[144] A. Boukerche, H. A. B. F. Oliveira, E. F. Nakamura, and A. A. F. Loureiro, "Secure Localization Algorithms for Wireless Sensor Networks," *IEEE Commun. Mag.*, vol. 46, no. 4, pp. 96–101, Apr. 2008.

[145] A. T. Parameswaran, M. I. Husain, and S. Upadhyaya, "Is RSSI a Reliable Parameter in Sensor Localization Algorithms – An Experimental Study," in *Proc. IEEE Field Failure Data Analysis Workshop (F2DA09)*, New York, NY, USA, Sep. 2009. [Online]. Available: https://www.cse.buffalo.edu//srds2009/F2DA/f2da09_RSSI_Parameswaran.pdf

[146] R. T. Ioannides, T. Pany, and G. Gibbons, "Known Vulnerabilities of Global Navigation Satellite Systems, Status, and Potential Mitigation Techniques," *Proc. IEEE*, vol. 104, no. 6, pp. 1174–1194, Jun. 2016.

[147] C. Miao, G. Dai, K. Ying, and Q. Chen, "Collaborative Localization and Location Verification in WSNs," *Sensors*, vol. 15, no. 5, pp. 10 631–10 649, May 2015.

[148] J. Yin, Q. Yang, and L. M. Ni, "Learning Adaptive Temporal Radio Maps for Signal-Strength-Based Location Estimation," *IEEE Trans. Mobile Comput.*, vol. 7, no. 7, pp. 869–883, Jul. 2008.

[149] Q. Chen, G. Huang, and S. Song, "WLAN User Location Estimation Based on Receiving Signal Strength Indicator," in *Proc. Int. Conf. Wireless Commun., Netw. and Mobile Comp.*, Beijing, China, Sep. 2009, pp. 1–4.

[150] K. Lin, M. Chen, J. Deng, M. M. Hassan, and G. Fortino, "Enhanced Fingerprinting and Trajectory Prediction for IoT Localization in Smart Buildings," *IEEE Trans. Autom. Sci. Eng.*, vol. 13, no. 3, pp. 1294–1307, Jul. 2016.

[151] D. Liu, M. C. Lee, and D. Wu, "A Node-to-Node Location Verification Method," *IEEE Trans. Ind. Electron.*, vol. 57, no. 5, pp. 1526–1537, May 2010.

[152] D. Al-Abri and J. McNair, "On the Interaction Between Localization and Location Verification for Wireless Sensor Networks," *Comput. Netw.*, vol. 52, no. 14, pp. 2713–2727, Oct. 2008.

[153] M. Fiore, C. E. Casetti, C. F. Chiasserini, and P. Papadimitratos, "Discovery and Verification of Neighbor Positions in Mobile Ad Hoc Networks," *IEEE Trans. Mobile Comput.*, vol. 12, no. 2, pp. 289–303, Feb. 2013.

[154] Y. Wei and Y. Guan, "Lightweight Location Verification Algorithms for Wireless Sensor Networks," *IEEE Trans. Parallel Distrib. Syst.*, vol. 24, no. 5, pp. 938–950, May 2013.

[155] E. A. Quaglia and S. Tomasin, "Geo-Specific Encryption Through Implicitly Authenticated Location for 5G Wireless Systems," in *Proc. IEEE Int. Workshop on Signal Process. Advances in Wireless Commun. (SPAWC)*, Edinburgh, UK, Jul. 2016, pp. 1–6.

[156] S. Yan, I. Nevat, G. W. Peters, and R. Malaney, "Location Verification Systems Under Spatially Correlated Shadowing," *IEEE Trans. Wireless Commun.*, vol. 15, no. 6, pp. 4132–4144, Jun. 2016.

[157] N. A. Pantazis, A. Spiridonos, S. A. Nikolidakis, and D. D. Vergados, "Energy-Efficient Routing Protocols in Wireless Sensor Networks: A Survey," *IEEE Commun. Surveys Tuts.*, vol. 15, no. 2, pp. 551–591, May 2013.

[158] S. Pattem, B. Krishnamachari, and R. Govindan, "The Impact of Spatial Correlation on Routing With Compression in Wireless Sensor Networks," *ACM Trans. Sensor Netw.*, vol. 4, no. 4, pp. 24:1–24:33, Aug. 2008.

[159] E. Fasolo, M. Rossi, J. Widmer, and M. Zorzi, "In-Network Aggregation Techniques for Wireless Sensor Networks: A Survey," *IEEE Wireless Commun.*, vol. 14, no. 2, pp. 70–87, Apr. 2007.

[160] C. M. Sadler and M. Martonosi, "Data Compression Algorithms for Energy-Constrained Devices in Delay Tolerant Networks," in *Proc. ACM SenSys*, Boulder, Colorado, USA, Oct. 2006.

[161] D. Zordan, B. Martinez, I. Vilajosana, and M. Rossi, "On the Performance of Lossy Compression Schemes for Energy Constrained Sensor Networking," *ACM Trans. Sensor Netw.*, vol. 11, no. 1, pp. 15:1–15:34, Aug. 2014.

[162] Y. Yu, B. Krishnamachari, and V. K. Prasanna, "Data Gathering with Tunable Compression in Sensor Networks," *IEEE Trans. Parallel Distrib. Syst.*, vol. 19, no. 2, pp. 276–287, Feb. 2008.

[163] A. B. Sharma, L. Golubchik, R. Govindan, and M. J. Neely, "Dynamic Data Compression in Multi-hop Wireless Networks," *Perf. Eval. Review*, vol. 37, no. 1, pp. 145–156, Jun. 2009.

[164] E. Zeydan, D. Kivanc, C. Comaniciu, and U. Tureli, "Energy-Efficient Routing for Correlated Data in Wireless Sensor Networks," *Ad Hoc Netw.*, vol. 10, no. 6, pp. 962 – 975, Aug. 2012.

[165] J. Berlinska, "Scheduling for Data Gathering Networks with Data Compression," *Eur. J. Oper. Res.*, vol. 246, no. 3, pp. 744–749, May 2015.

[166] M. Zargham, A. Ribeiro, A. Ozdaglar, and A. Jadbabaie, "Accelerated Dual Descent for Network Optimization," in *Proc. American Control Conf.*, San Francisco, CA, USA, Jun. 2011.

[167] T. Watteyne, M. Palattella, and L. Grieco, "Using IEEE 802.15.4e Time-Slotted Channel Hopping (TSCH) in the Internet of Things (IoT): Problem Statement," IETF, Tech. Rep. RFC 7554, May 2015.

List of Publications

[168] M. Centenaro and L. Vangelista, "A Study on M2M Traffic and Its Impact on Cellular Networks," in *Proc. IEEE World Forum on Internet of Things (WF-IoT)*, Milan, Italy, Dec. 2015, pp. 154–159.

[169] M. Polese, M. Centenaro, A. Zanella, and M. Zorzi, "M2M Massive Access in LTE: RACH Performance Evaluation in a Smart City Scenario," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Kuala Lumpur, Malaysia, May 2016, pp. 1–6.

[170] A. Biral, M. Centenaro, A. Zanella, L. Vangelista, and M. Zorzi, "The Challenges of M2M Massive Access in Wireless Cellular Networks," *Digit. Commun. Netw.*, vol. 1, no. 1, pp. 1–19, Feb. 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S235286481500005X

[171] M. Centenaro and L. Vangelista, "Analysis of Small Packet Traffic Support in LTE," in *Proc. Wireless Telecommun. Symp. (WTS)*, Chicago, IL, USA, Apr. 2017, pp. 1–8.

[172] M. Centenaro, L. Vangelista, S. Saur, A. Weber, and V. Braun, "Comparison of Collision-Free and Contention-Based Radio Access Protocols for the Internet of Things," *IEEE Trans. Commun.*, vol. 65, no. 9, pp. 3832–3846, Sep. 2017.

[173] M. Centenaro, L. Vangelista, A. Zanella, and M. Zorzi, "Long-Range Communications in Unlicensed Bands: the Rising Stars in the IoT and Smart City Scenarios," *IEEE Wireless Commun.*, vol. 23, no. 5, pp. 60–67, Oct. 2016.

[174] D. Magrin, M. Centenaro, and L. Vangelista, "Performance Evaluation of LoRa Networks in a Smart City Scenario," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Paris, France, May 2017, pp. 1–7.

[175] M. Centenaro, L. Vangelista, and R. Kohno, "On the Impact of Downlink Feedback on LoRa Performance," in *Proc. IEEE Int. Symp. on Personal, Indoor and Mobile Radio Commun. (PIMRC)*, Montreal, Canada, Oct. 2017, pp. 1–6. [Online]. Available: http://www.dei.unipd.it/~centenar/articoli/impact-dl-lora.pdf

[176] G. Caparra, M. Centenaro, N. Laurenti, S. Tomasin, and L. Vangelista, "Energy-Based Anchor Node Selection for IoT Physical Layer Authentication," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Kuala Lumpur, Malaysia, May 2016, pp. 1–6.

[177] ——, "Wireless Physical Layer Authentication for the Internet of Things," in *Information Theoretic Security and Privacy of Information Systems*, 1st ed., R. Schaefer, H. Boche, A. Khisti, and V. Poor, Eds. Cambridge University Press, Jun. 2017, ch. 14, pp. 390–417.

[178] G. Caparra, M. Centenaro, N. Laurenti, and S. Tomasin, "Optimization of Anchor Nodes' Usage for Location Verification Systems," in *Proc. ICL-GNSS*, Nottingham, UK, Jun. 2017, pp. 1–6. [Online]. Available: http://www.dei.unipd.it/~centenar/articoli/efficient-location-verification.pdf

[179] M. Centenaro, M. Rossi, and M. Zorzi, "Joint Optimization of Lossy Compression and Transport in Wireless Sensor Networks," in *Proc. IEEE Global Commun. Conf. (Globecom)*, Washington, DC, USA, Dec. 2016, pp. 1–6.

[180] M. Centenaro, M. Pesce, D. Munaretto, A. Zanella, and M. Zorzi, "A Comparison Between Opportunistic and Fair Resource Allocation Scheduling for LTE," in *Proc. IEEE Int. Workshop on Comput. Aid. Model. and Des. of Commun. Links and Netw. (CAMAD)*, Athens, Greece, Dec. 2014, pp. 239–243.

[181] S. Saur and M. Centenaro, "Radio Access Protocols with Multi-User Detection for URLLC in 5G," in *Proc. Eur. Wireless Conf.*, Dresden, Germany, May 2017, pp. 1–6.