Doctoral Thesis

NB-IoT M2M Communications in 5G Cellular Networks

Author:
Pilar Andrés Maldonado

Supervisors:
Dr. Juan M. Lopez-Soler
Dr. Pablo Ameigeiras Gutiérrez

A thesis submitted in fulfillment of the requirements
to obtain the International Doctor degree as part of the
Programa de Doctorado en Tecnologías de la Información y las Comunicaciones
in the

Wireless and Multimedia Networking Lab Research Group
Departamento de Teoría de la Señal, Telemática y Comunicaciones

Granada, April 30, 2019

Editor: Universidad de Granada. Tesis Doctorales
Autor: Pilar Andrés Maldonado
ISBN: 978-84-1306-288-4
URI: http://hdl.handle.net/10481/56821

Declaration of authorship

The doctoral candidate Mrs. Pilar Andrés Maldonado, and the thesis supervisors: Dr. Juan Manuel López Soler and Dr. Pablo Ameigeiras Gutiérrez.

Guarantee, by signing this doctoral thesis:

that the research work contained in the present report, entitled ***NB-IoT M2M Communications in 5G Cellular Networks***, has been done by the doctoral candidate under the direction of the thesis supervisors and, as far as our knowledge reaches, in the performance of the work, the rights of other authors to be cited (when their results or publications have been used) have been respected.

Granada, April 30, 2019

Pilar Andrés Maldonado
Ph.D Candidate.

Dr. Juan M. Lopez-Soler
Professor.
University of Granada.

Dr. Pablo Ameigeiras Gutiérrez
Tenured Professor.
University of Granada.

To my parents and sister.

Acknowledgements

First, I would like to thank my supervisors, Prof. Juan M. López Soler and Prof. Pablo Ameigeiras Gutiérrez, for giving me the opportunity to carry out this thesis. I am very glad to work on this interesting topic and learn from them. Their continuous support and guidance were valuable assets for the completion of this thesis. Especially, I would like to thank Pablo for his effort and admirable dedication during this journey.

I am also grateful for the other members of my research group WiMuNet (Wireless and Multimedia Networking Lab) for their collaboration and fellowship. They are a wonderful group of people able to achieve incredible milestones. Special thanks to Prof. Juan J. Ramos Muñoz, José A. Ordóñez Lucena, and Óscar Adamuz Hinojosa for their time, company, and for inspiring me, owing to their unlimited passion in their work. I want also to extend my gratitude to Jonathan Prados Garzón, other inspiring and amazing person to work with. It was a pleasure to have him as a colleague and friend during this long path.

Further, I would like to thank Preben Mogensen for having me during the research stay and giving me the chance to work with him at Aalborg University and Nokia Bell Labs Aalborg. Many thanks go also to Mads Lauridsen for his support, ideas, and guidance during the research stay, and Melisa and Roberto for their invaluable help and friendship. I would also like to extend my gratitude to Germán Madueño, Marek Rohr, and Maxime Remy from Keysight Aalborg for their technical help.

Besides the people already mentioned, I would like to thank those people that in one way or another have contributed to this work. Like the wonderful contributors of ShareTechnote and Netmanias that with their work helped me a lot when I began the study of cellular standards.

Most of all, I am tremendously grateful for the continuous support and love I received from my parents and sister, and the *osho-raging* love and *patience* of Juanjo during the vicissitudes of my peculiar mind in this thesis.

Abstract

Cellular networks are continuously evolving to provide enhanced capabilities and widen the supported use cases beyond the initial focus on mobile broadband. Recently, this evolution has paved the way to support Internet of Things (IoT). The inclusion of IoT in the cellular networks, denoted as Cellular IoT (CIoT), is bringing a larger and more extensive ecosystem of use cases than ever to cellular networks.

The next cellular generation, denoted as 5G, already considers in its design this foreseen connectivity heterogeneity. However, previous generations have to be optimized to cope with these new requirements. To meet this challenge, the Third Generation Partnership Project (3GPP) has standardized three solutions for CIoT: i) Extended Coverage GSM IoT (EC-GSM-IoT), ii) Long Term Evolution Cat-M1 (LTE-M), iii) Narrowband Internet of Things (NB-IoT). These solutions are Low-Power Wide-Area (LPWA) technologies. That means their design goals are extended coverage, low power and low cost devices, and massive connections.

Particularly for LTE-M and NB-IoT, they are standards that can be deployed today to serve LPWA use cases and will become part of the 5G family. NB-IoT defines a new radio access technology based on LTE and is specifically tailored for ultra-low-end IoT applications.

In addition to the emerging IoT use cases, IoT is built upon Machine-Type Communication (MTC), also denoted as Machine to Machine (M2M) communications. The characteristics of MTC traffic greatly differ from the human-generated traffic. For example, both differ in the uplink and downlink traffic loads, temporal distribution of the traffic, traffic profiles (MTC usually follows a periodic or bursty traffic), or mobility.

Within this context, the main objective of this thesis is to study the inclusion of massive MTC (mMTC) into cellular networks. More precisely, the use of

NB-IoT to support mMTC within the cellular networks.

First, the signaling impact due to MTC in the current cellular network (i.e. 4G), is studied. To that end, a new architecture for the main control plane entity of the core is assumed. This new architecture is based on Network Functions Virtualization (NFV) and the studied entity is the Mobility Management Entity (MME). The analytical model is based on queuing theory. In the study, four possible designs are proposed and three traffic classes are considered: mobile broadband, mMTC, and low latency MTC. The evaluation is carried out considering the resources needed for dimensioning, the cost of the system, and the response time of each traffic class assumed. The results show the level of resource sharing and the target design traffic significantly impact the performance of each traffic class and the number of resources needed.

Second, analytical study of the NB-IoT coverage extension performance. To that end, the evaluation includes all available NB-IoT techniques applied to achieve the target of 164 dB Maximum Coupling Loss (MCL). The proposed analytical expressions are based on the Shannon theorem. The analysis includes the limitations due to realistic channel estimation. The results show the performance of the Signal to Noise Ratio (SNR) gain when doubling repetitions is significantly affected when assuming realistic channel estimation compared to ideal channel estimation. Consequently, NB-IoT devices in weak coverage condition will be challenging to reach even considering the novel NB-IoT techniques to extend coverage.

Third, analytical and experimental NB-IoT performance evaluations are developed. The analytical evaluation is based on Markov chains and the experimental evaluation uses a controlled testbed. This testbed consists of commercial NB-IoT devices connected to a base station emulator. The NB-IoT performance evaluation is done in terms of the device's battery lifetime and latency. Using the testbed, the NB-IoT devices are studied empirically and later the proposed analytical model is experimentally validated. To that end, different traffic and coverage scenarios are considered. The validation results show the analytical model performs well compared to the empirical measurements under the same configuration in both cases. The results reach a maximum relative error of the battery lifetime estimation between the model and the measurements of 21% for an assumed Inter-Arrival Time (IAT) of 6 min. This relative error can be further

reduced if larger IATs are considered or some model simplifications assumed are specified in the model. Additionally, the results demonstrate NB-IoT devices can achieve the targets of 10 years of battery lifetime or 10 seconds of uplink transmission latency for a large range of scenarios when the traffic profile has a large IAT, or the configuration of the radio resources do not require an extensive number of repetitions.

Contents

List of Acronyms and Symbols

Acronyms

3GPP	Third Generation Partnership Project
ACK	Acknowledgment
AL	Aggregation Level
AM	Acknowledgment Mode
ARQ	Automatic Repeat Request
AS	Access Stratum
BLE	Bluetooth Low Energy
BLER	Block Error Ratio
BSR	Buffer Status Report
BS	Baseline Scheme
CAPEX	Capital Expenditure
CDF	Cumulative Distribution Function
CDR	Charging Data Record
C-DRX	Connected DRX
CE	Channel Estimation
CFO	Carrier Frequency Offset
CID	Cell-ID
CIoT	Cellular IoT
CS	Circuit Switched
CSS	Common Search Space
CN	Core Network
COTS	Commercial Off-The-Shelf
CP	Control Plane optimization

D2D Device to Device
DCI Downlink Control Information
DL Downlink
DMRS Demodulation Reference Signal
DRX Discontinuous reception
DUT Device Under Test
EC-GSM-IoT Extended Coverage GSM IoT
ECL Coverage Enhancement Level
ECM EPS Connection Management
eDRX extended/enhanced Discontinuous Reception
EDT Early Data Transmission
eMBB enhanced Mobile Broadband
EMM EPS Mobility Management
eNB evolved NodeB
EPC Evolved Packet Core
EPS Evolved Packet System
E-RAB Evolved Radio Access Bearer
E-UTRAN Evolved Universal Terrestrial Radio Access
FE Front End
FFT Fast Fourier Transform
FIFO First Input First Output
GPRS General Packet Radio Service
GSM Global System for Mobile Communications
GTP GPRS Tunneling Protocol
H2H Human to Human
HARQ Hybrid Automatic Repeat Request
HO Handover
HSS Home Subscriber Server
IAT Inter-Arrival Time
IC Integrated Circuit
ICI Inter Carrier Interference
I-DRX Idle DRX
IMT International Mobile Telecommunications
IoT Internet of Things

IP Internet Protocol
ITU International Telecommunication Union
KPI Key Performance Indicator
lMTC low latency MTC
LoRaWAN LoRa Wide Area Networks
LPWA Low-Power Wide-Area
LPWAN Low-Power Wide-Area Network
LS Least-Square
LTE Long Term Evolution
LTE-A LTE Advanced
LTE-M LTE Cat-M1
M2M Machine to Machine
MAC Medium Access Control
MAR Mobile Autonomous Reporting
MCL Maximum Coupling Loss
MCS Modulation and Coding Scheme
MIB Master Information Block
MME Mobility Management Entity
MMSE Minimum Mean Square Error
mMTC massive MTC
MSE Mean Square Error
MTC Machine-Type Communication
MTCD MTC Device
NACK Negative Acknowledgement
NAS Non-Access Stratum
NB-IoT Narrowband Internet of Things
NCCE Narrowband Control Channel Element
NF Network Function
NFV Network Functions Virtualization
NPBCH Narrowband Physical Broadcast CHannel
NPDCCH Narrowband Physical Downlink Control CHannel
NPDSCH Narrowband Physical Downlink Shared CHannel
NPRACH Narrowband Physical Random Access CHannel
NPSS Primary Synchronization Signal

NPUSCH Narrowband Physical Uplink Shared CHannel
NR New Radio
NRS Narrowband Reference Signal
NSA Non-Standalone
NSSS Secondary Synchronization Signal
OFDM Orthogonal Frequency-Division Multiple
OFDMA Orthogonal Frequency-Division Multiple Access
OPEX Operating Expenditure
OS Overdimensioned Scheme
OTDOA Observed Time Difference of Arrival
PA Power Amplifier
PDCP Packet Data Convergence Protocol
PDN Packet Data Network
PDU Protocol Data Unit
PGW Packet Data Network Gateway
POS Point-of-Sale
PRB Physical Resource Block
PSD Power Spectral Density
PSM Power Saving Mode
PTW Paging Time Window
QN Queuing Network
QoS Quality of Service
RA Random Access
RAI Release Assistance Indication
RAN Random Access Network
RAR Random Access Response
RAT Random Access Technology
RE Resource Element
RFID Radio Frequency Identification
RLC Radio Link Control
RRC Radio Resource Control
RRM Radio Resource Management
RSRP Reference Signal Received Power
RU Resource Unit

RV Redundancy Version
S1 S1 Release
SA Standalone
SC-FDMA Single-Carrier Frequency-Division Multiple Access
SC-PTM Single-Cell Point-to-Multipoint
SCR Scheduling Request
SDB State Database
SDR Software Defind Radio
SF Subframe
SGW Serving Gateway
SIB System Information Block
SL Service Logic
SNR Signal to Noise Ratio
SON Self Organizing Network
SR Service Request
SRB Signaling Radio Bearer
SS Traffic Shaper Scheme
TA Tracking Area
TAU Tracking Area Update
TBS Transport Block Size
TM Transparent Mode
ToA Time of Arrival
TS Traffic Separated Scheme
UE User Equipment
UL Uplink
UM Unacknowledged Mode
UMTS Universal Mobile Telecommunications System
UP User Plane optimization
URLLC Ultra-Reliable and Low Latency Communications
USS UE-specific Search Space
VM Virtual Machine
vMME virtualized MME
VNF Virtual Network Function
VNFC Virtual Network Function Component

W Worker

WLAN Wireless Local Area Network

List of Figures

List of Tables

Chapter 1

Introduction and Motivation

Cellular networks have transformed our society in many ways such as our social interactions, our professional networking, the media distribution, the information access, etc. They have evolved from an initial expensive wireless technology a few decades ago, to the everyday commodity of today. Their continuous evolution has led to today's ubiquitous system used by a majority of the world's population [30]. As the adoption of cellular networks expands, new use cases arise. Until now, four cellular networks generations (1G - 4G) have been deployed, and the fifth generation (5G) has been recently introduced. The first and second generations were voice-centric systems. Later, the third and fourth generation also supported data packet services. Each generation enhances the services of the previous generation and entails a significant technological evolution.

Unfortunately, one thing the first four generations have in common was their human-centric communications design. Before 5G, the focus of cellular networks was to provide more services to human users [13]. However, in recent years the development of Internet of Things (IoT) is becoming a groundbreaking revolution.

The IoT concept embodies the vision of everything connected. That is, physical objects such as devices, machines, and vehicles connected to the Internet. This vision encompasses a vast ecosystem of emerging use cases in markets such as industrial machinery, health care, smart cities, etc.

Cisco estimates that 15 billion devices will be connected by 2022 [31]. This value will significantly outnumber the total number of mobile cellular subscriptions (foreseen to be 8.6 billion by 2022 in the Ericsson mobility report [32]).

To cope with the new requirements for machine connectivity to support IoT, previous cellular networks have to be rethought.

1.1 Machine type communications

Machine-Type Communication (MTC), or also known as Machine to Machine (M2M) communications, is a term that describes any data communication between devices to collect data, share information, and perform actions that do not necessarily need human interaction. Traditionally, the use of these devices has been local. That is, without the exposure of the service to be used for more use cases or applications. This has caused an M2M fragmented market due to the multiple vertical industries (e.g. transportation, health care, energy, industry, etc) and the traditional development of domain or vendor-specific closed M2M solutions.

That means, the typical M2M solutions were conceptualized and deployed for a specific process without considering their integration into other applications. These deployments typically use dedicated software applications and require the design of an entire solution stack to support the intended M2M application.

Figure 1.1 shows an ontology of the M2M communications applications grouped by their scope of application. As can be seen, M2M communications foster the development of a vast number of applications. These applications are part of different verticals such as [1]:

- Health care. Applications related to health, monitoring of patients and medicines, identification of patients in hospitals, or collection of medical data. Some of the common traffic characteristics of this type of applications are low mobility, low-medium data rates, high availability and reliability.

- Utilities. Applications designed for the collection, management and use of information to implement event automation. These applications can be used in houses, offices, industrial plants or cities to make an intelligent environment. Some of the common traffic characteristics of this type of applications are low mobility, small data rates, predictable behavior (due to periodic reporting), and delay tolerance.

- Transportation. Applications related to the monitoring of vehicles, roads or goods, or management of the routes. In this vertical, the vehicles (i.e. cars, trains, buses, etc) equipped with sensors become M2M communication entities. Some of the common traffic characteristics of this type of applications are high mobility, high data rates, and multi-link connectivity.

- Security and public safety. Applications related to security, both in private and public locations, or for people or assets. Some of the common traffic characteristics of this type of applications are no mobility, periodic reporting with unpredictable traffic bursts due to the occurrence of alarms, from low-medium data rates to high peaks of traffic to send multimedia information.

- Industry. Applications designed to further optimize the plant, processes, supply chain, inventory and logistics management. The information from the M2M sensors integrates the plant and its underneath processes to enhance the operation within the plant. Some of the common traffic characteristics of this type of applications are no mobility, low-medium data rates, and restrictive delay requirements.

- Retail. Applications designed to provide better customer service, M2M Point-of-Sale (POS), digital signage, or connected vending. Some of the common traffic characteristics of this type of applications are low or no mobility, low data rate, high reliability and availability.

Due to the extensive number of M2M applications and the M2M communications (or MTC) traffic profile, the characteristics of M2M traffic greatly differ from the human-generated traffic. The main differences between Human to Human (H2H) and M2M communications can be summarized as follows [18, 33, 34]:

- Uplink (UL) dominant traffic load: for M2M devices the UL traffic is higher than the Downlink (DL) traffic.

- Low or no mobility: many M2M devices are stationary or have lower mobility than H2H devices.

- Traffic profile: M2M devices generally send periodic traffic (e.g. measures from metering devices) or bursty traffic (e.g. detection of an event in alarm

devices). While the M2M traffic is uniformly generated throughout the day, the H2H traffic is mostly concentrated during daylight and evening.

- Quality of Service (QoS): The requirements of M2M and H2H communications can significantly diverge in terms of latency, reliability, energy consumption, or security.

1.1.1 IoT revolution

In order to overcome the vertical M2M silos, is it required to develop a modern ecosystem that enables interoperation between verticals and applications. The potential interconnection of device and the way we interact with the environment brings a broader context than M2M communications, what is called IoT. M2M is the forerunner of IoT as shown in Figure 1.2. M2M sets the foundations of the connectivity on which IoT is built upon. Then, IoT considers a more horizontal approach where vertical industries converge into an expansive system to enable new applications [35].

IoT promises a technological revolution. The seamless information exchange and networked interaction of the physical and digital world IoT enables will lead

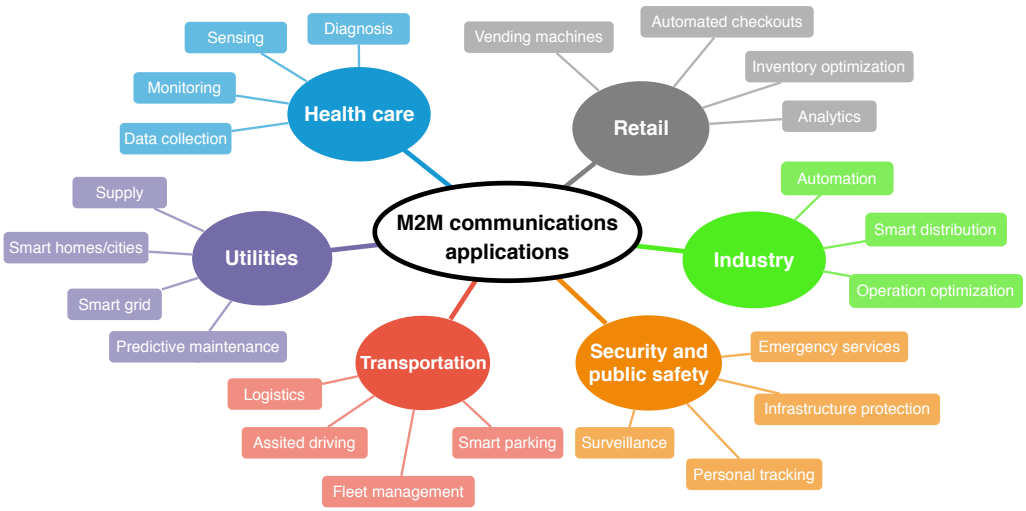


Figure 1.1: M2M communication applications [1, 2].

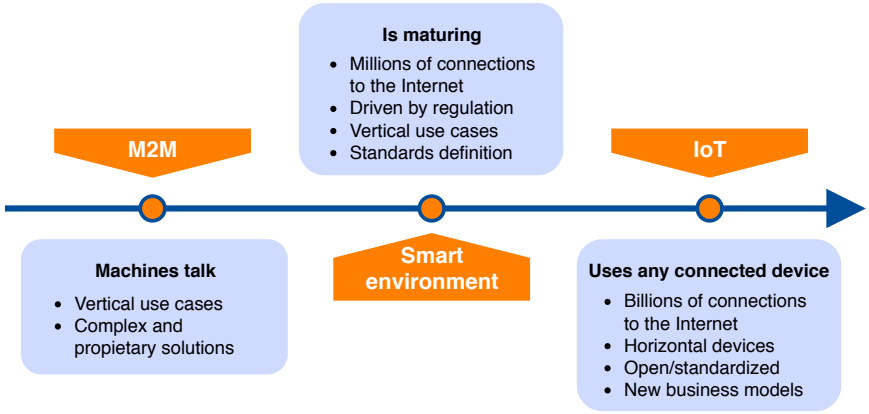


Figure 1.2: M2M to IoT evolution [3].

to the innovation of a large number of relevant application scenarios such as Industry 4.0, autonomous vehicles, geo-fencing, smart agriculture, etc.

There will be a wide range of IoT use cases in the future with their own requirements. The IoT Key Performance Indicators (KPIs) differ regarding the cost, battery lifetime, coverage, throughput, network capacity, security and reliability. Table 1.1 summarizes some generic requirements for different IoT use cases. To classify IoT requirements in a nutshell, we can define two extremes [36]:

- Massive IoT. It is characterized by a massive number of low cost and low energy devices sending small data volumes. Examples of IoT use cases within this type are: smart utilities, logistics, fleet management, smart agriculture, etc.
- Critical IoT. It is characterized by ultra reliable communications with very low latency and high availability. Examples of IoT use cases within this type are: remote health care, industrial control, smart grid, traffic safety, etc.

1.1.2 IoT connectivity landscape

To achieve the development of the IoT vision, a large variety of wireless communication technologies has gradually emerged. The large IoT connectivity landscape

Table 1.1: Generic requirements for typical IoT use cases [3].

IoT use case	Data rate	Mobility	Latency	Battery lifetime
Transportation	100s of kbps UL	10 to 150 km/h	Low (seconds)	3 months
Automation & monitoring	50 - 500 kbps UL	Fixed	High (hours)	10 years
Security & surveillance	0.5 - 8 Mbps UL	Fixed	Real-time UL stream	Connected to electrical grid
Smart Cities	50 - 500 kbps UL	Fixed	Low (seconds)	10 years
Health monitoring	50 -500 kbps UL	< 5 km/h	Low (seconds)	2 years

reflects the diverse communication and application requirements to meet the IoT KPIs. Each communication technology may have a specific application domain or more broader scope and can be summarized in the following list [15]:

Radio Frequency Identification (RFID) is a non-contact identification technology. A passive RFID system is composed of two main components: radio signal transponder (tag) and tag reader. The tag embeds an antenna and an Integrated Circuit (IC)-chip with unique identification code (ID) [37]. The RFID technology is mostly applied to goods logistics. However, recent research is exploring to use the short RFID range for proximity detection and localization within IoT [38].

ZigBee is an open wireless standard. ZigBee is used to create small, low-rate and low-power wireless personal area networks. It is built on the IEEE 802.15.4 physical and Medium Access Control (MAC) standard and adds mesh network and security layers along with an application framework [39]. The major benefits of ZigBee devices are the interoperability assurance and the device's operation as part of a mesh network (i.e. relay their signals using other devices). Thus, the malfunction of one device will not impact the network as a whole.

Bluetooth Low Energy (BLE) is a full protocol stack. Bluetooth was originally aimed as a replacement for wires in mobile devices, and now it has evolved to be used in many different applications. BLE is a low energy version of Bluetooth, still designed for short-range communication. BLE is envisaged as a connectivity solution for short-range communication in the IoT. In addition to the point-to-point and broadcast topologies, in 2017 the mesh networking topology was included in BLE [40].

Wi-Fi is a wireless networking technology for radio Wireless Local Area Network (WLAN) based on the IEEE 802.11 standard. Wi-Fi can deal with high data rates and is widely used for mobile connections. However, it has several limitations for IoT due to its large energy consumption and the limited number of simultaneously associated devices to the access point [15]. To make Wi-Fi suitable for IoT, the Wi-Fi alliance introduced Wi-Fi HaLow as a low-power Wi-Fi solution.

Low-Power Wide-Area (LPWA) unlicensed technologies have recently emerged as a key enabler for IoT. Low-Power Wide-Area Networks (LPWANs) are able to deliver low-power, low-speed and low-cost connectivity with wide-range coverage to IoT applications. They operate in unlicensed spectrum, and there are available many different solutions such as SigFox, LoRaWAN, and Ingenu [41]. Despite the promising characteristics, the use of unlicensed spectrum imposes a significant limitation due to small duty cycles and access mechanisms. Therefore, LPWA technologies are often seen as a complement to existing cellular networks.

Cellular networks have been historically a bad fit for many IoT use cases due to the incurred device energy consumption to communicate with the base station and the cost per-unit basis. However, in several IoT deployments, 2G was selected to communicate with the IoT devices with small data transfers. This is changing as more connectivity options for IoT are available within the cellular networks, such as Cellular LPWA and 5G. Particularly for 5G, its design is expected to cover ultra reliable and low latency communications and massive MTC (mMTC) to provide the basis for the all-connected world of humans and objects. Therefore, 5G is envisioned as a solution for massive IoT and critical IoT.

Cellular LPWA specifically targets the connectivity requirements of mMTC applications (i.e. extensive coverage area, device complexity reduction, and long battery lifetime). To do that, the Third Generation Partnership Project (3GPP) introduced three LPWA technologies: LTE Cat-M1 (LTE-M), Narrowband Internet of Things (NB-IoT), and Extended Coverage GSM IoT (EC-GSM-IoT). The goal is to re-use the legacy cellular networks infrastructure to cope with the massive connectivity of IoT.

Depending on the IoT deployment, one or several of these technologies may provide the set of capabilities required. Table 1.2 summarizes the main KPIs are able to meet each technology previously mentioned extracted from [15]. There is not a solution for all IoT use cases. For example, NB-IoT offers higher data rates and more advanced features (e.g. routing, firmware broadcast, multicast), but LoRaWAN allows the user to manage its own network and the devices have higher battery lifetime. Table 1.3 summarizes the differences in requirements for the IoT connectivity technologies.

Table 1.2: IoT KPIs covered by different wireless connectivity technologies [15].

	Zigbee	**BLE**	**Wi-Fi HaLow**	**LPWA unlicensed**	**LPWA cellular**
Scalability	×	×	✓	×	✓
Reliability	×	✓	✓	×	✓
Low power	✓	✓	✓	✓	✓
Low latency	×	✓	✓	×	✓
Large coverage	×	×	✓	✓	✓
Low module cost	✓	✓	✓	✓	✓
Mobility support	×	×	×	×	✓
Roaming support	×	×	×	×	✓

Table 1.3: Summary of current IoT wireless technologies [16,17].

	Frequency band	**Channel Bandwidth**	**Range**	**Data rate**	**Battery lifetime**	**Topology**	**Organization**
RFID	Low/High/Ultra high frequencies	200 kHz	1 cm - 100 m	1 - 100 kbps	N/A (passive) / 3 - 5 years (active)	P2P	
ZigBee	Sub-1 GHz, 2.4 GHz	2 MHz	10 - 100 m	250 kbps	Months to years	Star/Mesh/Tree	ZigBee Alliance
BLE	2.4 GHz	2 MHz	10 - 100 m	1 Mbps	Months to years	P2P/Star/Mesh	Bluetooth SIG
WiFi	2.4 GHz, 5 GHz Sub-1 GHz (Wi-Fi HaLow)	22 MHz 2 MHz (HaLow)	100 m 1 km (HaLow)	Mbps to Gbps	Days to months	Star	Wi-Fi Alliance
LPWA unlicensed							
LoRaWAN	Sub-1 GHz	125, 250 or 500 kHz	10 - 15 km	50 kbps	10+ years	Star of stars	LoRa Alliance
Sigfox	Sub-1 GHz	100 Hz	10 - 50 km	100 bps	10+ years	Star	Sigfox
Cellular LPWA							
EC-GSM	850 MHz - 950 MHz, 1800 - 1900 MHz (same as GSM)	200 kHz	10 - 15 km	70 - 240 kbps	10+ years	Star	3GPP
LTE-M	450 MHz - 3.5 GHz (same as legacy LTE)	1.08 MHz	10 - 15 km	1 Mbps	10+ years	Star	3GPP
NB-IoT	450 MHz - 3.5 GHz (2G/3G/4G spectrum)	180 kHz	10 - 15 km	250 kbps	10+ years	Star	3GPP

1.1.3 Challenges for MTC deployment in cellular networks

Focusing on cellular networks, the deployment of IoT use cases (i.e. MTC devices) is challenging as cellular networks are mainly designed for human-centric communications. In general, the challenges can be summarized as follows:

- Massive connections: depending on the IoT application, a large number of MTC devices may be connected in a small area, i.e., sharing the same radio resources. This could lead to massive access requests by the MTC devices. Due to the scarcity of radio resources, the radio interface design must enable to maintain the connection of a large number of devices without congesting the network.

- Extremely low power consumption: this feature is crucial for MTC devices battery powered or with limited access to power sources that only wake up infrequently or have infrequent interactions with the network. To support this feature, the network must update several functionalities such as the control procedures or idle and sleep modes.

- Diversity: to address the IoT market, the network should efficiently support diverse requirements from different IoT use cases. This means to be able to support everything from static devices to tracking applications, from delay tolerant communications to extremely low latency, or from infrequent traffic to continuous streaming of data. To support this diversity, the network must optimize the radio and the core to include more features such as enhanced power saving features and mobility management, congestion and overload control, subscription control, etc.

Particularly for the radio interface, Table 1.4 summarizes some MTC features and the associated impact in the radio interface extracted from [18].

1.2 LPWA unlicensed

A few years ago, many non-cellular LPWAN technologies arisen and presented an alternative choice to the existing cellular networks non-optimized for IoT. Examples of these LPWAN technologies are LoRa Wide Area Networks (LoRaWAN),

Table 1.4: MTC features impact in the radio interface [18]

		Standard impacts						
MTC feature	**IoT applications**	**Sleep and idle mode**	**Mobility management**	**Link adaptation**	**Bandwidth request & resource allocation**	**Frame structure**	**Network entry**	**Cooperation**
Massive transmissions	Security Metering Tracking		✓	✓	✓	✓	✓	✓
High reliability	Health Security	✓	✓	✓	✓	✓		✓
Access priority	Health Automation	✓	✓		✓	✓	✓	
Very low power	Tracking Automation	✓	✓	✓	✓	✓	✓	✓
Small data burst	Metering Security	✓						
Low/no mobility	Metering	✓	✓	✓	✓		✓	✓
Monitoring and security	Retail	✓				✓	✓	

SigFox, Ingenu, and Weightless. Within these LPWAN technologies, SigFox and LoRaWAN are the main players.

The SigFox technology was developed in 2010 by the startup Sigfox. Sigfox operates and commercializes its own IoT solution, i.e., it owns from the back-end data and cloud server to the proprietary endpoint software. SigFox can be used to exchange small (12 bytes payload in UL, 8 bytes in DL) and infrequent (140 messages in UL and 4 messages in DL per day) amounts of data. It has bidirectional communication between the device and the base station with a significant link asymmetry. That is, the DL has a different link budget and is very restricted. Therefore, in order to ensure the reception of UL packets, the packet is typically sent multiple times over different frequency channels. For more information, see [42].

LoRaWAN is an open-standard networking layer released by the LoRa Al-

liance in 2015. LoRaWAN defines the MAC layer that enables networking and LoRa is the physical layer. LoRa and LoRaWAN enable inexpensive, long-range connectivity for IoT devices. LoRa uses chirp spread spectrum modulation. This modulation spreads the signal over a wider channel bandwidth. To adapt the data rate, LoRaWAN uses different spreading factors that tune the chirp modulation rates. Thus, lower spreading factors enable higher data rates in a reduced transmission range, whereas higher spreading factors provide longer range at lower data rates. LoRaWAN supports bidirectional communication. The LoRa data rates range from 0.3 kbps to 50 kbps [43]. Additionally, to consider different IoT application requirements (e.g. latency or battery lifetime), LoRaWAN has three classes of end-point devices (i.e. Class A, B and C). However, to avoid network congestion, LoRaWAN defines a duty cycle (commonly set to 1%). For more information, see [44]. Interestingly, recent research efforts such as [45] have studied the integration of LoRaWAN deployments with cellular networks.

1.3 Cellular networks

Despite cellular networks were initially designed for H2H communications, in the last years there has been a significant effort to quickly respond to the emerging IoT market needs. The development of new solution has focused on the accommodation of the IoT KPIs into the cellular networks. Compared to other connectivity technologies previously reviewed, cellular networks are able to offer QoS support, scalability, mobility and roaming support, security, relatively low cost of deployment, and global coverage.

For many years, cellular networks have been evolving. This evolution is commonly known by generations (1G, 2G, 3G, 4G, and 5G). The 3GPP standards development organization covers the radio access, the core transport network, service capabilities, provides the interfaces for non-radio access to the core network, and for interworking with Wi-Fi networks. The work performed by the 3GPP is responsible for the standardization of the different radio access technologies developed during the evolution of the generations: Global System for Mobile Communications (GSM) in 2G, Universal Mobile Telecommunications System (UMTS) in 3G, Long Term Evolution (LTE) in 4G, and New Radio (NR) in 5G. The specification of these radio access technologies has been developed in

stages, known as 3GPP releases. Each release contains a stable set of features. All 3GPP releases are backward compatible, i.e., a device supporting one of the earlier releases can still work on a newer release deployed in the network.

The increasing use of cellular networks has led to the increasing of radio spectrum assigned to them. The International Telecommunication Union (ITU) and regional and national regulators are responsible for the management of the finite amount of radio spectrum available. Within the ITU, the ITU-R is the radio communication sector. While the technical specification of mobile-communication technologies (i.e. UMTS or LTE) is done within the 3GPP, the ITU-R is responsible to turn the technologies into global standards [4]. Additionally, the ITU had helped to drive the development of the last cellular generations by publishing a set of requirements for a mobile communication system, under the name International Mobile Telecommunications (IMT). IMT is a generic term to designate mobile systems worldwide. The following IMT standards are in existence:

- IMT-2000: ITU-R Recommendation M.1457 [46]. It is a worldwide set of requirements for a family of standards for the 3G of mobile communication system. According to the requirements, the data rate in the range 2 Mbps for stationary or walking users, and 348 kbps in a moving vehicle.

- IMT-Advanced: ITU-R Recommendation M.2012 [47]. It is a worldwide set of requirements for a family of standards for the 4G of mobile communication system. According to the requirements, the nominal data rate of 100 Mbps for high mobility and 1 Gbps for low mobility users.

- IMT-2020: The framework and objective for IMT-2020 is outlined in ITU-R Recommendation M.2083 [8]. It is a worldwide set of requirements establishing the requirements and vision for 5G. It identifies three main usage scenarios (i.e. enhanced Mobile Broadband (eMBB), mMTC, and Ultra-Reliable and Low Latency Communications (URLLC)). In the eMBB scenario, the peak data rates are 20 Gbps and 10 Gbps, in the DL and the UL, respectively.

1.3.1 LTE

LTE was first introduced in the 3GPP Release 8. The focus of this technology was mobile broadband services with tough requirements on high data rates, low latency, and high capacity [30]. LTE is the most successful and fastest developing system in the history of cellular communications. LTE has changed the way people use smartphones as it is the enabler of uninterrupted connectivity and new rich multimedia services such as streaming of music, videos and movies at a much faster rate than ever before.

LTE architecture is divided into two components: the access network named as Evolved Universal Terrestrial Radio Access (E-UTRAN) and the core network named as Evolved Packet Core (EPC). Since its introduction, LTE has evolved considerably in the following 3GPP releases. To underline the significant increase in capabilities achieved by the LTE evolution, the 3GPP introduced the names LTE Advanced (LTE-A) and LTE-A Pro for some of the releases. Release 10 includes the extra capabilities that are required for LTE-A. LTE-A is the first standard release that meets the IMT-Advanced requirements for the 4G. Later, Release 13 marks the start of LTE-A Pro.

The evolution of the LTE features ranges from the initial macro-centric deployments with peak data rates of 300 Mbps in a licensed and contiguous bandwidth of 20 MHz, to the support of peak rates of multi-Gbps by means of several improvements in antenna technologies, multisite coordination, carrier aggregation, interoperation of LTE and WLAN networks, etc [30]. Furthermore, the evolution of LTE has also widened the supported use cases beyond the initial focus on mobile broadband, for example, improving support for MTC, group communications, critical communications, and introducing direct Device to Device (D2D) communication. Figure 1.3 shows a summary of LTE features evolution throughout the releases.

1.3.2 5G

Although LTE is still evolving, the road towards the next generation of cellular networks, referred to as 5G, is rapidly coming into the limelight. The overall goal of 5G is to support a technology ecosystem of wireless networks working to provide a seamless communication medium for any kind of device. Unlike

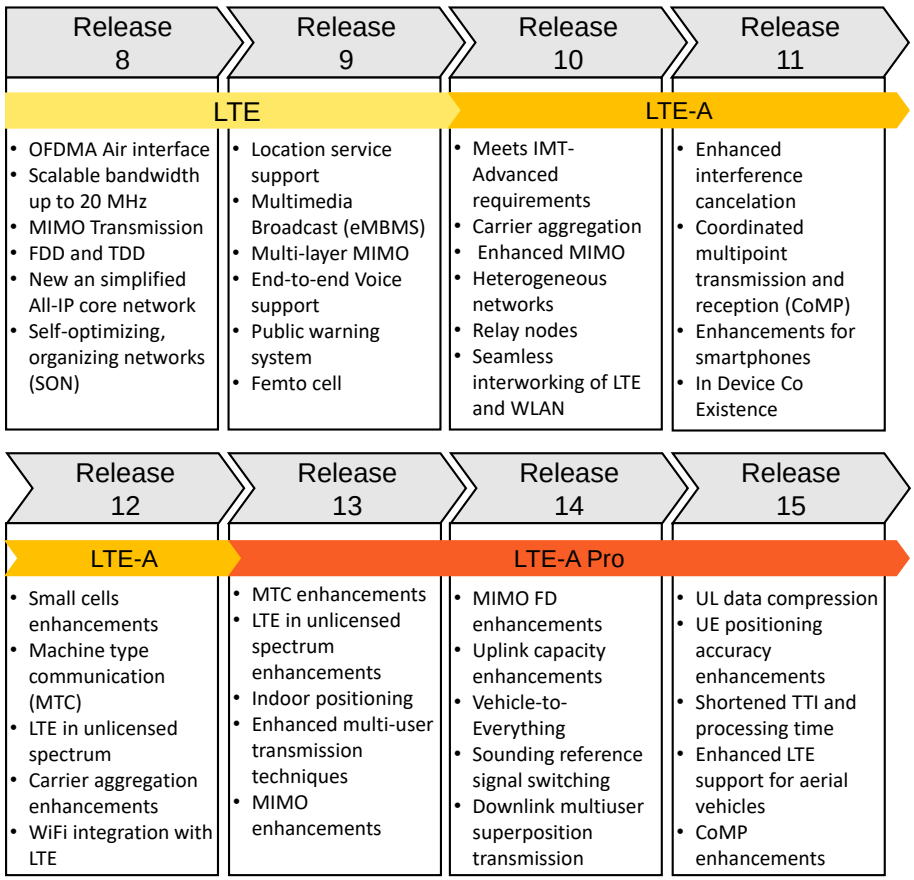


Figure 1.3: LTE evolution in the standard [4–6].

previous generations, 5G brings a shift in the design paradigm from a single-discipline system to a multi-discipline system [48].

For the development of 5G, the ITU has defined a 5G work timetable [8]. The 5G vision and key technical indicators were finished in 2015. Until mid-2017, the ITU specified the 5G technology technical requirements, i.e., the IMT-2020 requirements. In the same year, it begun to collect 5G international standards, i.e., the developing 5G solutions can be submitted as IMT-2020 proposals to ITU. Then, by the end of 2020, the 5G technical specifications will be completed. That means the ITU-approved IMT-2020 specifications will enable a full market deployment of 5G systems. In accordance with this time plan, the 3GPP has

adopted a timetable matching the ITU. In the mid-2017, Release 14 completed the study of a new 5G NR air interface. Release 15 provides the first phase of an NR specification extended with a second phase in Release 16. Figure 1.4 shows a simplified time plan of 5G in ITU-R and 3GPP.

The 3GPP is working in the definition of both a new 5G core network (5GC) as well as the new radio access technology NR. For example, in [49], the 3GPP identifies several scenarios and requirements for next generation access technologies. In [50], the 3GPP describes the service and operational requirements for a 5G system in different scenarios. To provide different options in the 5G deployment, the 3GPP supports two 5G architecture operations [51]:

- Non-Standalone (NSA): This scenario combines NR and LTE radio cells using dual-connectivity and the core network may be either EPC or 5GC. The standardization NSA 5G NR was completed in December 2017.
- Standalone (SA): This scenario only uses one radio access technology (i.e. 5G NR or LTE radio cells) and the core networks are operated alone. The SA 5G NR operation was finished in June 2018.

5G will provide more advanced and enhanced capabilities compared to 4G LTE. To achieve them, there are some key pillars for 5G [48], such as the evolution of existing Random Access Technologies (RATs), hyper-dense small cell deployment, Self Organizing Network (SON) capability, developing millimeter-wave RATs, pursue energy efficiency design approaches, allocation of new spectrum for 5G, virtualization, spectrum sharing, etc.

The 5G principle design is to provide a framework to enable efficient multiplexing of diverse 5G services. Despite the heterogeneity of the foreseen services

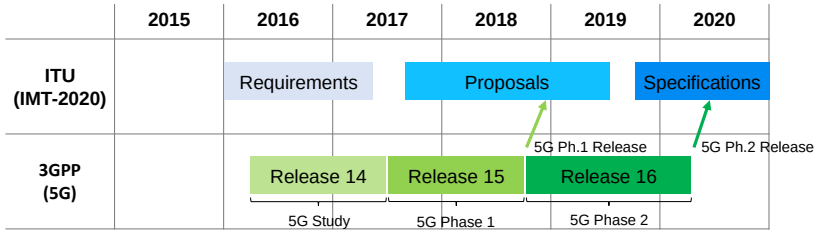


Figure 1.4: Timeline of 5G in ITU-R and 3GPP [7].

and devices, there is a preliminary agreement regarding the three main use cases the 5G must support [8]:

- enhanced Mobile Broadband (eMBB): Data-intensive applications that need lots of bandwidth. The eMBB addresses the human-centric use cases for access to multimedia content, services, and data. The demand within this use case is continuously increasing and new application areas are emerging, setting new requirements. Examples of applications within this use case are 3D video, augmented reality, UHD screens, etc.

- massive MTC (mMTC): Low cost, low energy devices with small data volumes on a mass scale. This use case is pure machine-centric. Due to the possibility of deployment of a large number of devices or their remote locations, the devices are required to be low cost, and have very long battery life. Examples of applications within this use case are smart city, utilities, smart agriculture, etc.

- Ultra-Reliable and Low Latency Communications (URLLC): Latency-sensitive services needing extremely high reliability, availability, and security. This use case is intended to cover both human and machine-centric communication (e.g. critical IoT). Examples of applications within this use case are self-driving cars, industrial automation, remote medical surgery, etc.

Figure 1.5 summarizes the three 5G main use cases and some of the KPIs of IMT-2020 compared to IMT-Advanced.

1.3.3 Cellular IoT

Cellular IoT (CIoT) is defined as a set of technologies under the 3GPP scope to provide IoT connectivity. That means the concept of CIoT encompasses MTC communication via cellular network technologies. Consequently, the new requirements for machine connectivity have emerged transforming the road-map of the standards evolution into the support of the IoT. The common high-level goals CIoT have are: i) decrease device complexity; ii) decrease power consumption; iii) increase coverage.

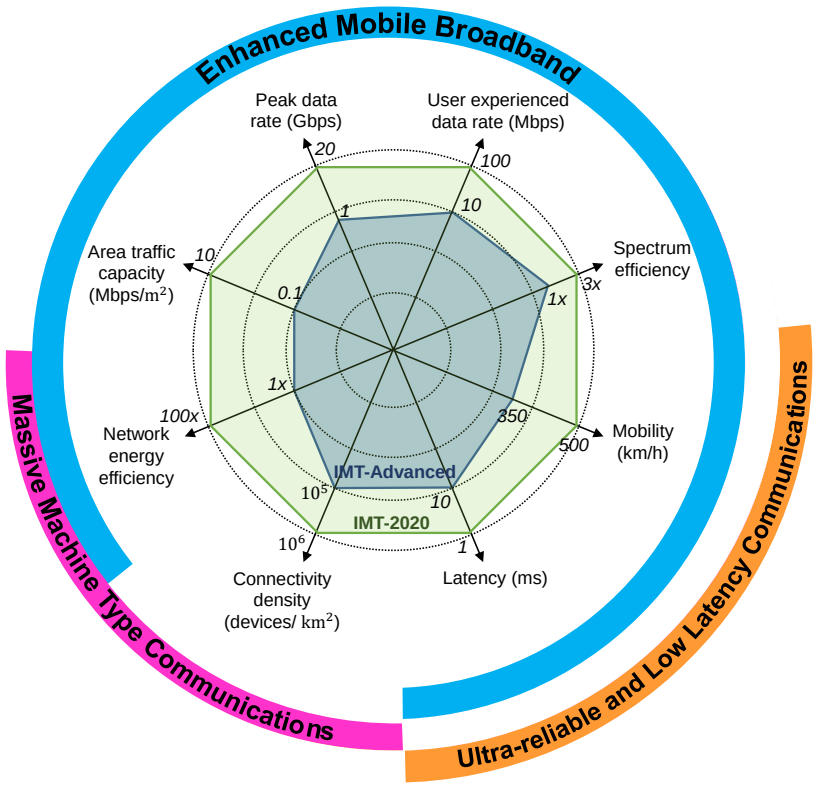


Figure 1.5: 5G usage scenarios and key capabilities of IMT-2020, as compared to IMT-Advanced [8].

In order to optimize the support of IoT by cellular networks, three new CIoT technologies have been introduced within the LPWA segment [13, 52]:

- Extended Coverage GSM IoT (EC-GSM-IoT) is an IoT-optimized evolution of GSM. The main optimizations can be summarized as follows: physical layer adaptation to support extended coverage, streamlining the protocol layer to reduce device complexity, improve higher layers features to increase battery lifetime, and the introduction of a security framework.

- LTE Cat-M1 (LTE-M) is an evolution of LTE optimized for IoT. The main optimizations can be summarized as follows: bandwidth reduction compared to LTE, the specification of a new 20 dBm Power Amplifier (PA)

class to reduce complexity, use of power saving features, and the introduction of small data transmission optimizations.

- Narrowband Internet of Things (NB-IoT) is a new radio access technology based on LTE and optimized for very low-cost and low-power devices. The main optimizations can be summarized as follows: new physical layer design to reduce complexity, the addition of new features to extend battery lifetime and coverage, and higher layer protocols, signaling, and physical-layer processing requirements are significantly simplified to reduce the device power consumption and complexity.

Depending on the targeted IoT service, one solution may be more suitable than the others. The three solutions will operate in spectrum shared with existing LTE or GSM networks. Thus, they do not require the additional deployment of base stations or other hardware. Furthermore, NB-IoT has a specific benefit due to its spectrum flexibility. This is because the narrow system of NB-IoT allows being deployed also in the spectrum that is not used for mobile broadband services today [13]. In locations LTE networks are not so advanced, the extensive coverage of GSM benefits from the use of EC-GSM-IoT. Conversely, LTE-M can achieve significantly higher data rates compared to EC-GSM-IoT and NB-IoT. Table 1.5 summarizes the main characteristics of each solution.

Finally, considering the convergence of 5G and CIoT, the 3GPP has envisioned a backward compatible operation of 5G NR with NB-IoT and LTE-M [13]. Therefore, it is foreseen that NB-IoT and LTE-M will continue to evolve towards 5G future requirements.

Table 1.5: Specification comparison between LTE CAT-1 and the new cellular LPWA solutions (extracted from [17])

Device Category	**LTE Cat-1**	**EC-GSM-IoT**	**LTE Cat-M1 (LTE-M)**	**LTE Cat-NB1 (NB-IoT)**
3GPP Release	8	13	13	13
Technology	LTE	GSM extension	Based on LTE	Clean-state
Peak Data Rate	DL: 10 Mbps UL: 5 Mbps	For DL & UL: 74 kbps (GMSK), 240 kbps (8PSK)	DL: 1 Mbps UL: 1 Mbps	DL: 170 kbps UL: 250 kbps
Duplex Mode	Supports Full duplex FDD/TDD	Supports Half duplex FDD only	Supports Half duplex FDD/TDD	Supports Half duplex FDD only
Bandwidth	20 MHz	0.2 MHz	1.08 MHz (1.4 MHz carrier bandwidth)	180 kHz (200 kHz carrier bandwidth)
Maximum Coupling Loss (MCL)	140.7 dB	164 dB (33 dBm), and 154 dB (23 dBm)	155.7 dB	164 dB
Rx Antenna	Supports double Rx	Supports single Rx	Supports single Rx	Support single Rx
Coverage Support	Complementary to Cat-M1 & NB-IoT	20 dB	+ 15 dB	+ 20 dB
Battery Lifetime	Less than 10 years	Supports within +10 years	Supports within +10 years	Supports within +10 years
Maximum Transmission Power	23 dBm	33 dBm or 23 dBm	20 dBm or 23 dBm	20 dBm or 23 dBm
Spectrum	Supports licensed LTE Bands In-band	Supports licensed GSM bands	Supports licensed LTE Bands In-band	Supports licensed LTE In-band, guard-band & stand-alone

1.4 Objectives of this thesis

The growing use of IoT is revolutionizing the roadmap of cellular networks to be able to enter the IoT market with competitive technologies. To cover the three 5G main use cases defined by the ITU (i.e. enhanced Mobile Broadband (eMBB), massive MTC (mMTC), and Ultra-Reliable and Low Latency Communications (URLLC)), one of the 5G pillars will be to allow multiple access technologies to interwork.

Recent efforts have focused on new cellular standards specifically targeting the connectivity requirements of mMTC applications. These new solutions can provide large-scale connectivity to low power and low-cost devices and are becoming part of the 5G family. One of these solutions, named NB-IoT, is expected to evolve to fulfill the 5G LPWA requirements. Consequently, the NB-IoT network components already operational today will continue to do so and eventually coexist as other 5G NR components.

In this context, the main objective of this thesis is to study the inclusion of mMTC into cellular networks. More precisely, the use of NB-IoT to support

mMTC within the cellular networks. To that end, this thesis addresses the following questions:

1. Focusing on the deployed LTE networks non-optimized for IoT traffic, what are the potential impacts LTE networks will experience trying to support IoT? How could be the expected IoT traffic handled in the core network to satisfy the diverse requirements from different types of devices?

   To answer these questions, we will model and evaluate the performance of different traffic profiles in LTE cellular networks. Within this objective, we will develop and assess theoretical and simulation models to estimate the signaling load of the LTE control plane. This objective is subdivided into the following sub-objectives:

   1.1 Review and analysis of MTC characteristics and LTE signaling procedures.

   1.2 Design and development of analytical and simulation models for the LTE control plane. The analytical model will be based on queuing theory.

2. Considering the new standardized NB-IoT access technology, are the new introduced NB-IoT mechanisms enough to achieve the enhanced coverage target? What are the benefits or drawbacks of each NB-IoT technique?

   To answer these questions, we will analyze the performance of NB-IoT coverage extension techniques under different coverage scenarios and assumptions. This objective is subdivided into the following sub-objectives:

   2.1 Design and development of analytical expressions to estimate the performance of coverage extension techniques in NB-IoT. These expressions are based on the description of the NB-IoT transmission in terms of Signal to Noise Ratio (SNR), bandwidth utilization, and energy per transmitted bit. The analytical expressions will be based on the Shannon theorem.

   2.2 Review of the realistic channel estimation and non-ideal factors to be considered in the low SNR range NB-IoT targets.

2.3 Development of simulators to obtain the relationship between the SNR and the channel estimation error.

2.4 Derivation of analytical expressions to model realistic channel estimation in NB-IoT.

3. In addition to the new NB-IoT access technology, two signaling data procedures were introduced in the standard to optimize data transmissions, are these optimizations efficient for IoT and under what circumstances? What is the expected device performance using NB-IoT and these procedures?

   To answer these questions, we will study the performance of an NB-IoT device in terms of battery lifetime and latency. More precisely, this objective consists of the following sub-objectives:

   3.1 Design and development of an analytical model to estimate the energy consumption of an NB-IoT device and NB-IoT radio resources consumption considering different data transmission procedures. The analytical model will be based on a Markov chain.

   3.2 Empirical evaluation of NB-IoT devices on a controlled testbed. We will set up a testbed consisting of commercial NB-IoT devices connected to a base station emulator to measure the device energy consumption.

   3.3 Analytical NB-IoT model validation. In order to validate our model, we will configure the testbed and the model with the same parameters and compare the results in terms of device's battery lifetime and latency.

1.5 Dissertation road-map

The rest of this thesis is structured as follows:

Chapter 2. Efficient signaling management for MTC. This chapter introduces the concept of Network Functions Virtualization (NFV) and the design of a virtualized LTE Mobility Management Entity (MME) to overcome the foreseen MTC signaling storm. This chapter details the different signaling procedures available in LTE and the new data transmission optimizations standardized to

efficiently support the small data transmissions from MTC devices. The different virtualized MME designs are analyzed in terms of scalability and costs.

Chapter 3. Energy consumption analysis at UE side. This chapter focuses on the estimation of the energy consumption of an NB-IoT device. Firstly, the fundamental background in NB-IoT is given. Next, a detailed system model and the Markov chain based model is described. The proposed model estimates the battery lifetime of the device and the capacity of a NB-IoT carrier under different scenarios.

Chapter 4. Performance evaluation of extended coverage in NB-IoT. This chapter presents analytical expressions to study the coverage extension features available in NB-IoT. Furthermore, this chapter proposes a UL link adaptation algorithm and an analytical NB-IoT evaluation framework. This study is extended to analyze the impact of the new NB-IoT features under different circumstances in terms of latency and battery lifetime.

Chapter 5. Experimental evaluation of NB-IoT. This chapter concludes our analytical NB-IoT analysis by means of two studies: i) experimental analysis of NB-IoT based on a live NB-IoT network and empirical measurements, and ii) analytical model validation. Both studies are done using a testbed composed by commercial NB-IoT devices connected to an NB-IoT base station emulation under different coverage and traffic scenarios.

Chapter 6. Conclusions and Outlook. This chapter draws the main conclusions, and outline the main contributions of this thesis and the future steps.

Appendix A. Resumen. This appendix is a comprehensive summary written in Spanish in order to meet with the requirements imposed by the University of Granada regarding the drafting of the doctoral dissertation.

1.6 List of publications

The following publications have been produced as a result of the work in this thesis:

1. P. Andres-Maldonado, P. Ameigeiras, J. Prados-Garzon, J. J. Ramos-Munoz, and J. M. Lopez-Soler, "MME support for M2M communications

using network function virtualization," in IARIA The Twelfth Advanced International Conference on Telecommunications (AICT 2016), 2016, pp. 106-111.

ISBN: 978-1-61208-473-2

2. P. Andres-Maldonado, P. Ameigeiras, J. Prados-Garzon, J. J. Ramos-Munoz and J. M. Lopez-Soler, "Reduced M2M signaling communications in 3GPP LTE and future 5G cellular networks," 2016 Wireless Days (WD), Toulouse, 2016, pp. 1-3.

   DOI: 10.1109/WD.2016.7461499

3. P. Andres-Maldonado, P. Ameigeiras, J. Prados-Garzon, J. Ramos-Munoz, and J. Lopez-Soler, "Virtualized MME Design for IoT Support in 5G Systems," Sensors, vol. 16, no. 8, p. 1338, Aug. 2016.

   DOI: 10.3390/s16081338

4. J. Prados-Garzon, J. J. Ramos-Munoz, P. Ameigeiras, P. Andres-Maldonado and J. M. Lopez-Soler, "Modeling and Dimensioning of a Virtualized MME for 5G Mobile Networks," in IEEE Transactions on Vehicular Technology, vol. 66, no. 5, pp. 4383-4395, May 2017.

   DOI: 10.1109/TVT.2016.2608942

5. P. Andres-Maldonado, P. Ameigeiras, J. Prados-Garzon, J. J. Ramos-Munoz and J. M. Lopez-Soler, "Optimized LTE data transmission procedures for IoT: Device side energy consumption analysis," 2017 IEEE International Conference on Communications Workshops (ICC Workshops), Paris, 2017, pp. 540-545.

   DOI: 10.1109/ICCW.2017.7962714

6. P. Andres-Maldonado, P. Ameigeiras, J. Prados-Garzon, J. Navarro-Ortiz and J. M. Lopez-Soler, "Narrowband IoT Data Transmission Procedures for Massive Machine-Type Communications," in IEEE Network, vol. 31, no. 6, pp. 8-15, November/December 2017.

   DOI: 10.1109/MNET.2017.1700081

7. P. Andres-Maldonado, P. Ameigeiras, J. Prados-Garzon, J. J. Ramos-Munoz, J. Navarro-Ortiz and J. M. Lopez-Soler, "Analytic Analysis of Narrowband IoT Coverage Enhancement Approaches," 2018 Global Internet of Things Summit (GIoTS), Bilbao, 2018, pp. 1-6.

   DOI: 10.1109/GIOTS.2018.8534539

8. P. Andres-Maldonado, P. Ameigeiras, J. Prados-Garzon, J. Navarro-Ortiz and J. M. Lopez-Soler, "An Analytical Performance Evaluation Framework for NB-IoT," Accepted for publication in the IEEE Internet of Things Journal, 2019.

9. P. Andres-Maldonado, M. Lauridsen, P. Ameigeiras and J. M. Lopez-Soler, "Analytical Modeling and Experimental Validation of NB-IoT Device Energy Consumption," Accepted for publication in the IEEE Internet of Things Journal, 2019.

   DOI: 10.1109/JIOT.2019.2904802

10. P. Andres-Maldonado, M. Lauridsen, P. Ameigeiras and J. M. Lopez-Soler, "Experimental Analysis of NB-IoT Performance Trade-offs," To be submitted, 2019.

Chapter 2

Efficient Signaling Management for MTC

The evolution of cellular networks is commonly known by generations. Each generation generally refers to a change in the fundamental nature of the service, transmission technology, and new frequency bands. Since the first analog generation 1G, to a digital system in 2G, the later support of multimedia in 3G, to the high-speed all-Internet Protocol (IP) switched network in 4G, or the upcoming 5G, all generations have looked ahead to be ready to meet future capacity and performance demands.

5G will continue on the path of 4G, significantly increasing the supported connected devices and speeds, introducing heterogeneous radio interfaces, being much more efficient, and being a platform to provide wireless connectivity for any kind of device or application. That means, 5G will greatly extend the set of use cases supported in cellular networks.

In recent years, the development 4G and the research of 5G have pursued more generic services than the conventional enhanced Mobile Broadband (eMBB) traffic. The growth of Internet of Things (IoT) is revolutionizing the connectivity requirements in all types of networks. Due to the extensive nature of the IoT applications, two new key uses cases are defined [8,13]: massive MTC (mMTC) and Ultra-Reliable and Low Latency Communications (URLLC). mMTC is characterized by low cost/energy devices, small data volumes and a massive number of devices connected. URLLC is characterized by strict requirements, such as ultra

reliability, low latency, and high availability. The distinct requirements depend on the specific application. For instance, the end-to-end latency can reach down to a few milliseconds or even lower [53].

The inclusion of these two generic services in cellular networks entails several challenges in the cellular networks. Specifically for mMTC, the network has to accommodate a huge amount of IoT devices connected with a range of performance and operational requirements significantly different from the usual Human to Human (H2H) communications. Until now, cellular networks were not optimized for such mMTC traffic. This challenge could derive to bottlenecks, congestion, or signaling storms in the network if the resultant traffic is not efficiently handled.

The growing number of diverse MTC devices and H2H devices connected to the network is leading to the faster increase of the control signaling traffic than the data traffic [54]. Additionally, compared to the previous 2G/3G hierarchical multi-protocol architecture, Long Term Evolution (LTE) has a flatter all-IP network architecture owing to the redistribution of network functions within the Random Access Network (RAN) and Core Network (CN). As a consequence of this flatter architecture, the LTE control plane key entity, named Mobility Management Entity (MME), must handle 3 or 4 times more average number of messages per subscriber attached to it than the equivalent core entity of previous generations [55]. This situation leads to a major increase of the processing load on the control plane entities of the network, and specifically on the MME. In order to deal with this increase of signaling and the dynamic traffic patterns generated by different types of users connected, Network Functions Virtualization (NFV) arises as a solution for developing the future cellular networks. NFV offers the possibility of running network functions as software appliances placed on commodity servers. Among other benefits, NFV enables flexible and scalable allocation of resources. Therefore, applying NFV to the control plane can significantly improve the scalability and elasticity of the whole network.

This chapter analyzes and proposes a model for dimensioning the MME virtualization based on NFV considering the three generic services: eMBB, mMTC, and URLLC. To that end, the rest of the chapter is organized as follows. Section 2.1 reviews the basic LTE architecture and its signaling procedures. Section 2.2 introduces the concept of virtualization in cellular networks. Section 2.3 provides some background of queuing networks. Section 2.4 briefly reviews the related

literature. Section 2.5 explains our system model and the main adopted assumptions. Sections 2.6 and 2.7 provide the analysis and evaluation of the virtualized MME (vMME), respectively. Lastly, section 2.8 presents the main conclusions of this chapter.

2.1 LTE overview

The main twist included in LTE compared to previous systems is the definition of an all-IP network topology between the User Equipment (UE) and the Packet Data Network (PDN), compared to the legacy Circuit Switched (CS) domain of previous generations. Since the initial release of LTE, the Third Generation Partnership Project (3GPP) Release 8, it has been continuously enhanced in the following 3GPP Releases. The Release 10 included the standardization of LTE Advanced (LTE-A). LTE-A is the enhanced version of LTE able to fulfill the International Telecommunication Union (ITU) requirements for 4G. LTE-A is designed to be backward compatible with LTE. The requirements of LTE-A are: Downlink (DL) peak data rate of 1000 Mbps, and 500 Mbps in Uplink (UL) [56]. Later, Release 13 included the standardization of LTE-A Pro. LTE-A Pro incorporates many enhanced features already present in LTE-A, among the improvements LTE-A Pro increases the DL data rate up to 3 Gbps and the bandwidth currently available.

In the new LTE architecture, the whole system is known as the Evolved Packet System (EPS). Within the EPS there are two components, the Evolved Universal Terrestrial Radio Access (E-UTRAN), and the Evolved Packet Core (EPC). The EPC is the CN of the LTE system. The E-UTRAN handles the EPC's radio communications with the user, i.e., all RAN functionalities. Figure 2.1 illustrates the reference model of the EPS and the reference points.

2.1.1 Evolved Packet Core

The EPC is an IP based core network with a flat architecture. The main EPC logical nodes are [57]:

Home Subscriber Server (HSS) is a database that contains user-related and subscriber-related information. Subscription data includes credentials

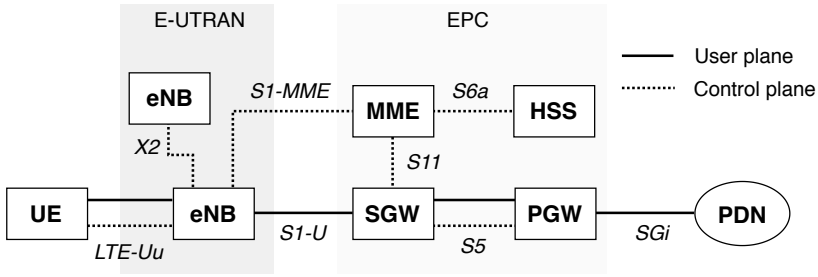


Figure 2.1: LTE network reference model [9].

for authentication and access authorization, and the HSS also provides support functions in mobility management.

Serving Gateway (SGW) deals with the user plane. It is the gateway which terminates the user plane interface towards the E-UTRAN. Among other functions, the SGW buffers DL IP packets destined for UEs not reachable, is the anchor point for inter-3GPP mobility, forwards user plane packets, etc.

Packet Data Network Gateway (PGW) provides connectivity between the LTE network and external IP networks. The PGW assigns the IP address to the UE and routes the user plane packets. Additionally, it provides functions for lawful interception, policy/QoS control, and charging.

Mobility Management Entity (MME) handles all control plane signaling, including mobility and security functions for users attaching over the RAN. The MME is also responsible for the tracking and the paging of user while the user is idle.

All logical node are interconnected by standardized interfaces to provide interoperability between operators.

2.1.2 E-UTRAN

The E-UTRAN consists of a network of evolved NodeBs (eNBs). The E-UTRAN architecture is considered as a distributed architecture as there is no centralized

controller. The E-UTRAN is responsible for all radio functionalities that can be summarized as follows:

- Radio Resource Management (RRM) including all functions related to the management of the radio bearers.
- Header compression to improve the efficiency of the use of the radio interface.
- Security by means of encrypting and integrity protecting the packets sent over the radio interface.
- Connectivity to the EPC consisting in signaling sent to the MME and the bearer path toward the SGW.

The eNBs are interconnected between them by means of an interface called X2 and with the EPC through the S1 interface. There are two categories of protocols between the different entities of the network: i) between the eNBs and the UE, the protocols are known as the Access Stratum (AS) protocols; ii) between the EPC and the UE are referred to as the Non-Access Stratum (NAS) protocols [57, 58]. The high-level signaling messages lie in the NAS and are transported using the AS protocols.

2.1.3 User and control planes

LTE protocol architecture can be separated into control plane and user plane. Each plane uses specific protocols which allow the definition of a different and in-

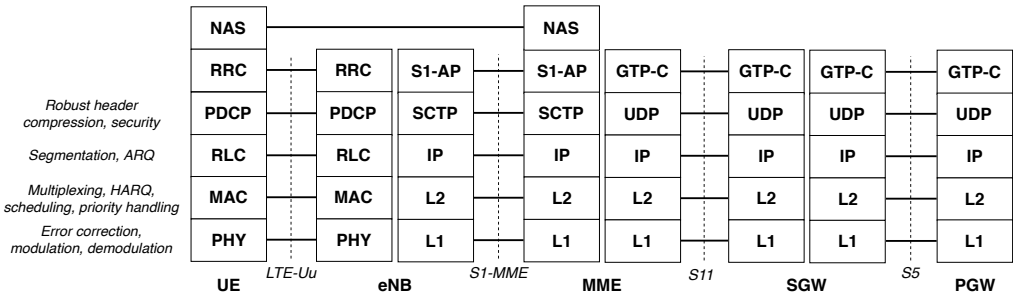


Figure 2.2: LTE control plane protocol stacks.

dependent transport stack and bearers. As an example, Figure 2.2 shows the LTE control plane protocol stack. The functionality of the key LTE stack protocols is summarized as follows [56]:

NAS is responsible of mobility management, bearer management functions, authentication, and security control.

Radio Resource Control (RRC) supports the transfer of the NAS signaling, and the functionality related to the management of the radio resources.

Packet Data Convergence Protocol (PDCP) among others, it is responsible of header compression, ciphering and integrity protection.

Radio Link Control (RLC) ensures the reliable delivery of data streams. It has three modes of operation: i) Transparent Mode (TM), transparent implies that the contents pass through this layer without any alteration; ii) Unacknowledged Mode (UM), it does not require any Acknowledgment (ACK)/Negative Acknowledgement (NACK) from the receiver, but adds operations such as segmentation/concatenation/reassembly, and in-sequence delivery of Protocol Data Units (PDUs); and iii) Acknowledgment Mode (AM), it requires ACK/NACK from the receiver, it also supports segmentation/concatenation/reassembly, in-sequence delivery of PDUs, error correction by Automatic Repeat Request (ARQ), and flow control. Each mode are set up on a channel-by-channel basis [59].

Medium Access Control (MAC) schedules the transmissions that are carried out on the air interface and controls the low-level operation of the physical layer. One control element of this layer is the Buffer Status Report (BSR). The BSR is transmitted by the UE and indicates the eNB about how much data it has available for transmission. There are several types of BSR and triggers for its transmission [22]. One example of trigger is when the periodic BSR Timer expires.

X2AP supports UE mobility and Self Organizing Network (SON) functions.

S1AP handles S1 interface, Evolved Radio Access Bearer (E-RAB), NAS signaling transport, and UE context management.

GPRS Tunneling Protocol (GTP) can be decomposed into separate protocols: GTP-C, GTP-U, and GTP'. GTP-C is responsible of the exchange of information for creation, modification and termination for GTP tunnels. GTP-U is used to forward UE IP packets over different interfaces. GTP' is used for Charging Data Record (CDR) transfer.

Diameter is designed to provide an authentication, authorization, and accounting (AAA) framework.

On the user plane, an IP packet from/to a UE is encapsulated in an EPC protocol and tunneled between the PGW and the eNB. To provide QoS support, LTE introduces the concept of EPS bearer. An EPS bearer is a bi-directional path connecting from a UE to a PGW. Through this path, various types of IP flows are classified by the 5-tuple, i.e., source IP, destination IP, protocol ID, source port, and destination port. When the UE connects to a PDN, the EPC sets up a default bearer. Later, the UE can also have one or more dedicated bearers configured to handle more specific traffic. A dedicated bearer acts as an additional bearer on top of the default bearer.

To transfer RRC and NAS signaling messages over the radio interface (i.e. between the UE and eNB), LTE uses three special radio bearers known as Signaling Radio Bearers (SRBs): SRB0, SRB1, and SRB2. Each SRB has a specific configuration of the radio interface protocols [56]. SRB0 is only used for a few RRC signaling messages. These signaling messages are used to establish the radio connection between the UE and eNB. The messages exchanged in the SRB0 are also used to configure the SRB1. Once the SRB1 is set up, it is used for all subsequent RRC messages and other higher layer messages prior the establishment of the SRB2. While the UE exchanges signaling messages through the SRB1, the SRB2 is configured. The SRB2 is used to transfer all the remaining messages. Both SRB1 and SRB2 always use AM RLC.

2.1.4 Control procedures

The actions the UE can perform with the LTE network depend on its signaling connectivity status. The management of the signaling connectivity can be described by means of three types of states:

- EPS Mobility Management (EMM) is a part of the NAS. EMM includes procedures related to mobility over an E-UTRAN access, authentication, and security.
- EPS Connection Management (ECM) indicates the state of the NAS connection between the UE and MME. This connection consist of an RRC connection over the radio interface and a S1 signaling connection over the S1-MME interface.
- RRC indicates if the UE has an active RRC connection or not with its serving eNB. These states are managed by the RRC protocol.

The states of EMM, ECM and RRC change as the required event proceeds. Figure 2.3 summarizes a few state transitions and the triggering events among EMM, ECM, and RRC in a UE. As seen, several control procedures are involved in the UE transition from one state to another. In this subsection, we focus on those procedures directly related with the session and bearers management, listed as follows [9, 56]:

Attach: It is performed when a UE needs to register with the network. The attach may vary depending on whether it is the first connection to the network or the UE has already been connected before. The attach procedure has four main objectives: i) Register the UE's location with a serving MME;

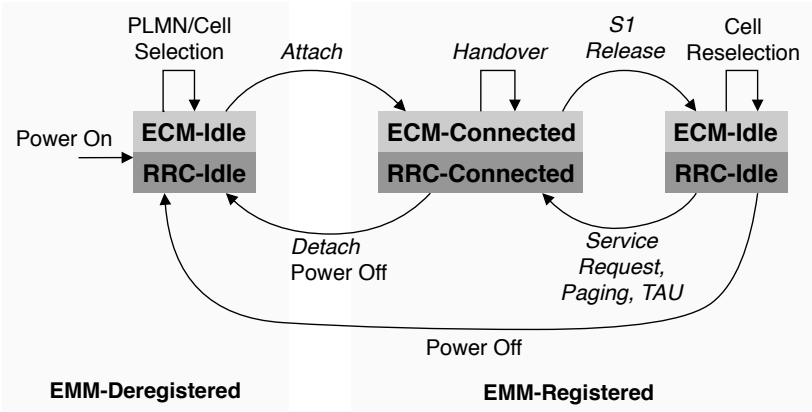


Figure 2.3: EMM, ECM, and RRC state transitions [10].

ii) Configure the signaling radio bearer responsible of the transfer of NAS signaling messages; iii) Provide an IP address to the UE; and iv) Configure the default EPS bearer. This bearer provides the UE with always-on connectivity to a default PDN. If the attach succeeds, the UE stays in EMM-Registered, ECM-Connected and RRC-Connected state and can use the network's service.

S1 Release (S1): This procedure is used to release the SRBs 1 and 2, the UE's data radio bearers and S1 bearers. After the S1 procedure, all UE related context information is deleted in the eNB. The initiation of the S1 procedure can be triggered at the eNB or the MME (e.g. user inactivity, RRC signalling integrity check failure, authentication failure, etc). After the S1 procedure, the UE transits to Idle state, entering EMM-Registered, ECM-Idle and RRC-Idle state.

Paging: It is used by the network to request the establishment of a NAS signaling connection to the UE. In most cases, this paging is triggered when the UE is in Idle state and there is a DL IP packet from the external network to the PGW.

Service Request (SR): It is carried out when a UE in Idle state (EMM-Registered, ECM-Idle, and RRC-Idle state) wants to reuse the LTE service when there is new traffic. This new traffic may be UL traffic generated by the UE, or DL traffic. Therefore, the SR procedure can be triggered by the UE or by paging.

Handover (HO): This procedure changes the serving cell of a UE in RRC-Connected. There are different types of HO depending on the involved nodes: i) Intra-LTE, source and target cells are part of the same LTE network; ii) Inter-LTE, the HO happens towards other LTE nodes; and iii) Inter-RAT, HO between different radio technologies.

Tracking Area Update (TAU): The UE uses the TAU to notify the MME its current location at Tracking Area (TA) (i.e. a group of neighbor eNBs) level. The TAU can be triggered to update the network periodically, or if the UE has moved into a TA in which it was not previously registered.

Detach: This cancels the UE's registration with the network. Either the UE or the network (i.e. MME or HSS) can trigger the detach. Once the UE is detached from the network, it moves to EMM-Deregistered, ECM-Idle and RRC-Idle state.

2.1.5 Efficient signaling procedures for CIoT

The state transitions involve significant signaling. Specifically, moving from RRC-Idle to RRC-Connected using the conventional SR comprises several signaling messages over the radio interface. For the common Machine-Type Communication (MTC) small data transmissions, this can lead to inefficient use of the radio resources. To reduce the signaling generated, two Cellular IoT (CIoT) EPS optimizations for data transfer were introduced in Release 13:

User Plane optimization (UP): This optimization enables the RRC connection to be suspended and resumed by means of two new control procedures: Connection Suspend and Resume. The suspend/resume feature allows to temporarily suspend the RRC connection and store the UE's context in the eNB and the core while the UE is in RRC-Idle state. UP requires an initial RRC connection establishment that configures the radio bearers and the AS security context in the network and UE.

Control Plane optimization (CP): This optimization uses the control plane to forward the UE data packets. To do that, the data packets are sent encapsulated into NAS signaling messages to the MME. Compared to the conventional SR procedure, the UE avoids AS security setup and the user plane bearers establishment required in each data transfer. Additionally, if required/needed, the UE or MME can trigger the establishment of the user plane bearers between the eNB and SGW during data transmissions in CP. This change of functionality implies: i) release of the specific CP user plane bearer between the MME and SGW, called S11-U; ii) user plane bearers establishment; and iii) AS security setup.

Figure 2.4 shows a summarized signaling flow of a mobile originated data transport and its later release of the resources while using the three data transport control procedures previously mentioned (i.e. SR, UP, and CP). As UP and CP

are optimizations of SR, let us see the different steps performed during the SR procedure to understand these optimizations. The first four messages comprise the contention-based Random Access (RA) procedure. The RA procedure allows the eNB to estimate and, if needed, adjust the UE UL transmission timing. The RA begins when the UE transmits a RA preamble. At this point, there could be a risk of contention if two UEs transmit at the same moment using the same preamble. Once the eNB receives the preamble, it replies with a Random Access Response (RAR) message. The RAR identifies the preamble sequence used by the UE and provides a UL scheduling grant. The UE must receive the RAR message before the RAR window size expires. Otherwise, the UE will go back to transmit the preamble. If the RAR is correctly received, using the UL grant within the RAR, the UE starts the RRC connection establishment sending the RRC Connection Request. At this point, the MAC Contention Resolution Timer starts. If the UE does not receive the RRC Connection Setup message before this timer expires, the UE will go back to transmit the preamble. If the RRC Connection Setup message is correctly received, the RA is successfully finished and the UE and eNB continue with the RRC connection establishment.

The subsequent signaling messages include: finish the RRC connection establishment (message 5), UE's authentication at the MME through NAS security level (messages 6 and 7), AS security context establishment between the UE and the eNB (messages 8 and 9), RRC reconfiguration (messages 10 and 11), and data bearers establishment with other core entities (messages 13 and 14). After message 14, UL and DL traffic paths are available allowing delivery of data traffic. If the UE stays inactive a period longer than the defined RRC Inactivity Timer at the network, the eNB will initiate the S1 procedure (messages from 15 to 20).

As previously mentioned, the signaling flow is different in CP and UP. For CP, the data packets are sent encapsulated as NAS signaling messages to the MME (messages 5 and 6). Avoiding the establishment of the user plane bearers reduces significantly the number the required signaling messages. For UP, Figure 2.4 shows the RRC Connection resume/suspend. Due to UP utilizes the usual user plane connectivity, subsequent data packets can be transferred through the data paths.

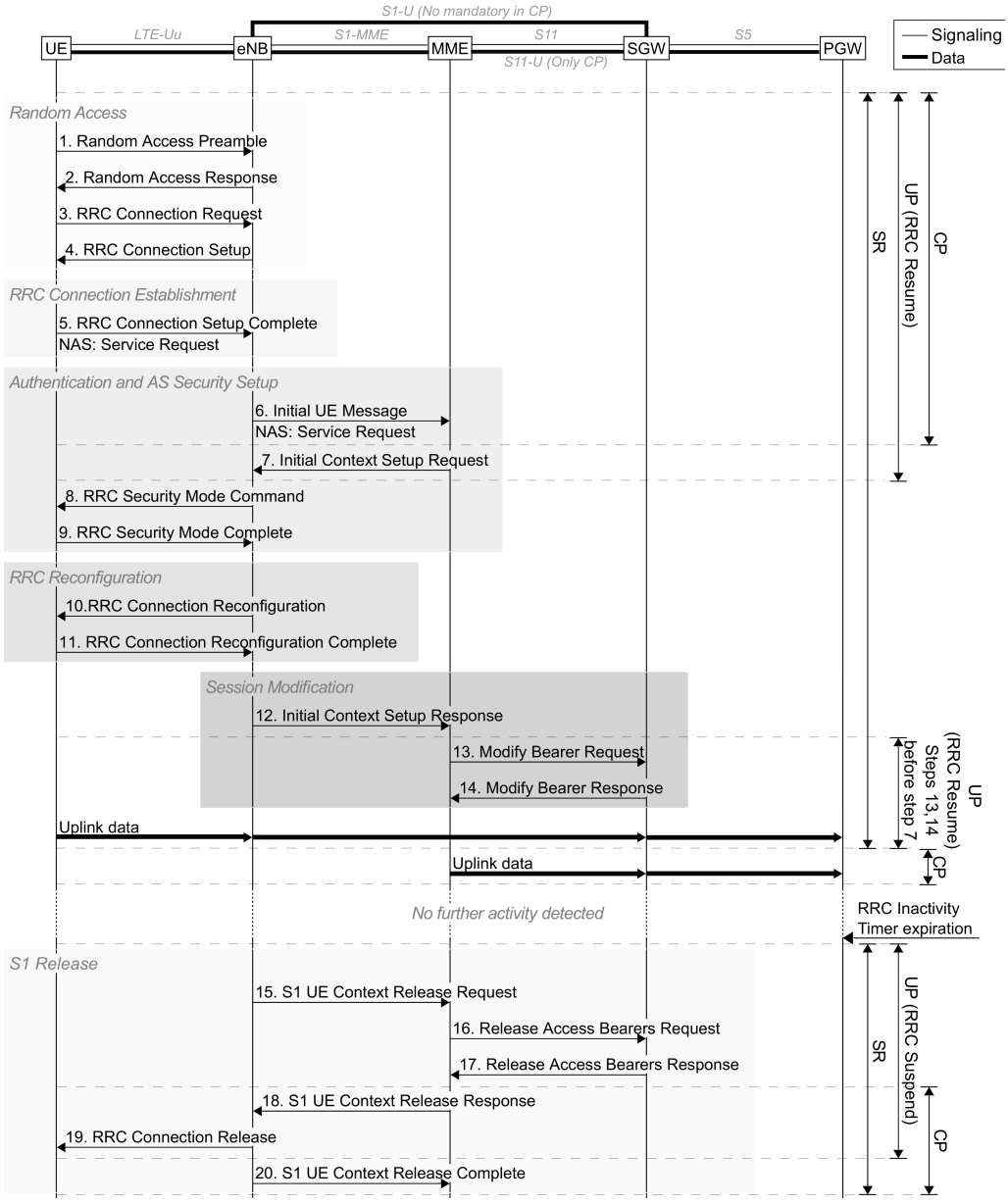


Figure 2.4: Summarized signaling diagram of MO data transport/RRC Connection Resume and S1 Release/RRC Connection Suspend for SR, CP, and UP [9].

2.2 Virtualization

NFV is one of the enabling technologies of 5G systems. Using NFV, network functionalities run as software components, which are called Virtual Network Functions (VNFs), on Commercial Off-The-Shelf (COTS) hardware. This commodity hardware is much less expensive than the very specialized hardware used in 4G networks. Additionally, the VNFs can scale-out/in to adjust its computing and networking capabilities. The design of this dynamic scale-in/out to meet the performance requirements is not a trivial task and involves a cost-performance trade-off. However, with good autoscaling strategies, operators can cope with new challenges, such as the foreseen traffic increase and IoT support while reducing Capital Expenditure (CAPEX) and Operating Expenditure (OPEX).

When using NFV, each Network Function (NF) can run on a Virtual Machine (VM) as a 1:1 mapping model or can be decomposed into smaller components called Virtual Network Function Components (VNFCs) running on multiple VMs as a 1:N mapping model [60]. The infrastructure services can be programmed, instead of re-architecting the infrastructure of the network [61]. Additionally, NFV can facilitate the adaptation of the network entities to the new demand. This is due to each VNFs can be modified to incorporate new features and scale on demand to handle a dynamic load. In this way, operators can create, scale, and deploy network components whenever they are needed, according to the particular traffic conditions.

In future cellular networks, NFV will enable network slicing. It allows to create multiple virtual networks on a shared physical infrastructure. Consequently, future cellular networks can be further abstracted into different network slices constituting an end-to-end logically isolated networks dedicated to different types of services.

Given the substantial benefits of applying NFV into cellular networks, there are also some challenges to implement NFV which need to be addressed such as [62]:

- Portability/Interoperability: Virtual appliances must be able to be loaded and executed in different standardized data-center environments, for whatever vendor or operator.

- Performance Trade-Off: The use of commodity hardware would have a decrease in performance compared to specialized hardware. Then, this challenge focus on the reduction of performance degradation as small as possible by means of modern software technologies.
- Network Stability: It is important to ensure the stability of the network when managing and orchestrating a large number of virtual appliances. This challenge includes when the virtual functions are relocated or during re-configuration events.
- Automation: NFV will only scale if all of the functions can be automated.

2.3 Fundamentals of queuing networks

The goal of Queuing Networks (QNs) is to represent their underlying processes so they can be efficiently managed [63]. Queuing Theory provides tools to answer questions such as the mean waiting time in the queue, the mean system response time, distribution of the number of customers in the queue, etc. The simplest QN model is the single server queue. This model consists of a server that processes customer requests, a queue where customers wait before receiving service, and an arrival of customers to the system. There are an extensive number of QNs system variations from this simple model. QNs can be roughly grouped into four categories:

- Open networks: Customers arrive from outside the system and after they are served, they leave the system.
- Closed networks: A fixed number of customers are in the system and always circulate among the queues.
- Loss networks: Customers arrive from outside the system if there is room in the system and after they are served, they leave the system.
- Mixed networks: A combination of the previous types of QNs.

Within the open QNs there is a special class called Jackson Networks. A Jackson's open queuing network consists of M nodes (or queues) with the following assumptions [64]:

- The service discipline at all queues is First Input First Output (FIFO).
- There is only one class of customers in the network.
- The service time in the queues are independent and identically distributed following an exponential distribution with rate μ_i at node i.
- Upon departure from queue i, the customer chooses the next queue j randomly with the probability $q_{i,j}$ or exits the network with the probability $q_{i,d}$ (i.e. probabilistic routing).
- The network is open to external customer arrival as Poisson processes with rate λ_i at queue i.

Let λ_i be the overall arrival rate to queue i, including both external arrivals and the split output streams from other queues. When the system is in its steady-state, we have the following traffic equation:

$$\lambda_i = \lambda_{0,i} + \sum_{j=1}^{N} p_{j,i}\lambda_j \tag{2.1}$$

where $\lambda_{0,i}$ is the arrival rate of jobs from outside to queue i.

Then, for a open Jackson network of M/M/m queues, Jackson's Theorem [65] states that provided the arrival rate to each queue is such that equilibrium exists, i.e., the condition $\lambda_i < \mu_i m_i$ holds for every $i \in \{1, ..., N\}$, then the steady state probability of the network can be expressed as the product of the state probabilities of the individual queues:

$$\pi(k_1, ..., k_N) = \pi_1(k_1) \cdot \pi_2(k_2) \cdot \pi_N(k_N) \tag{2.2}$$

Therefore, the queues of the network can be considered at independent M/M/1 queues with arrival rate λ_i and service rate μ_i.

2.4 Related works

This section briefly reviews the works proposed in the literature to assess the use of the NFV paradigm to virtualize the EPC entities.

In [66], *Taleb et.al.* discuss the design of a virtualized EPC and its implementation in a cloud computing environment. Additionally, they describe the key elements to offer the EPC "as a Service" (EPCaaS). They present and analyze a number of different implementation options for the virtualized EPC. Focusing on the 1:N mapping option, each entity of the EPC is decomposed into three logical components: Front End (FE), Worker (W), and State Database (SDB). The three-tier design follows the multi-tiered web services paradigm for cloud-based applications. This design has been considered in other works to virtualize the EPC [67–72]. In [67], *Takano et.al.* describe a scaling scheme base on the vitualization of a stateful network entities without interrupting user session continuity. This work exemplifies the proposed scheme applying it to the design of a vMME. Other works differ in their vMME design to allow fully stateless components such as the Ws in [68, 70–72] or the MME Service Logics (SLs) in [69].

Particularly in [70], *Prados et.al.* validate the model by simulation and show the model provides fairly good results for computational resources dimensioning. This work included a MTC traffic model corresponding to a fleet management use case.

Inevitably, the inclusion of MTC in the EPC implies a major increase of the processing load on the MME. However, just a few of the previously mentioned works specifically consider the support for MTC traffic. The analysis presented in this chapter covers this gap focusing on the modeling of the vMME considering MTC use cases.

2.5 System model

Let us assume an access cellular network architecture based on LTE. This network provides service to UEs and MTC Devices (MTCDs). Figure 2.5 depicts the overall system model. The main entities are defined as follows:

UE: Runs the human-users applications. These applications generate or consume network traffic following three traffic profiles: Web browsing, HTTP progressive video, and video calling. We assume the UEs move following a fluid-flow mobility model. This type of user triggers the following signaling procedures: SR, S1, and HO.

MTCD: Sends small data reports infrequently to an IoT server. We assume

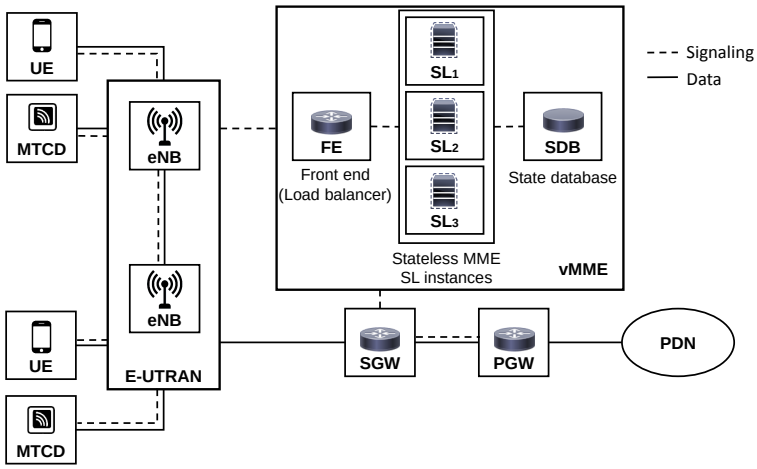


Figure 2.5: Overall system model.

the MTCDs are placed in fixed locations. Following the 3GPP and METIS guidelines, we consider two types of MTCDs: mMTC and low latency MTC (lMTC) devices. mMTC devices run delay-tolerant IoT applications, such as smart utilities applications. lMTC devices run strict low latency IoT applications, such as industrial applications or health care. We also consider a large number of MTCDs connected to the network that will cause traffic peaks due to MTC synchronization or alarm events. In practical deployments, these mMTC peaks may occur corresponding to multiple time periods (e.g. 15 min, 30 min, 1 h) due to the timer-driven nature of most of MTC applications [34]. Consequently, the possible highly synchronization of the MTC devices timers will cause a synchronized communication from a large number of devices to the network in a short period of time.

eNB: Receives signaling messages from the UEs and MTCDs and forward them to the corresponding core entity. The eNB keeps the RRC Inactivity Timer, T_i. Using this timer, the eNB detects the users' inactivity. If T_i expires, the eNB triggers the S1 procedure to release the RRC connection between the user and the eNB.

vMME: Maintains the mobility state of the users and is responsible for the bearers management. We consider the MME is virtualized following the NFV paradigm. Particularly, we adopt the 1:N mapping architectural option. Thus,

following the implementation of [66, 67], the vMME has three tiers: Front End (FE), Service Logic (SL) and State Database (SDB). The FE is responsible for the communication with other entities of the network. This entity also balances the load among the SLs. The SLs implement the processing. Finally, the SDB stores the user session state. This allows a stateless design of the SLs. In terms of scalability, the SL can scale out/in without impacting externally-connected peers and independently from other SLs. However, the scaling of SDBs and FEs is more complicated than the SL as they are stateful components [73].

The general operation of the vMME is as follows: i) at the arrival of a packet, the FE sends the packet to the corresponding SL instance according to its load-balancing scheme; ii) the SL gathers the transaction state of the packet from the SDB; iii) the SL process the packet; iv) after finishing the processing, the SL saves the transaction state into the SDB. This operation enables different messages of the same procedure from the same user can be processed by different SL instances.

As in [73], we assume that all tiers can be replicated within limits. When the processing capacity assigned to the vMME cannot withstand the current load, a new vMME SL instance must be instantiated, and a new processor is added to the processing resources pool. For simplicity, we assume the vMME, the SGW, and the PGW are located in a centralized data center of the network. Additionally, we assume every processor in the data center facility provides the same computational power. The SDB follows a shared-everything architecture [74], which eases the scale-out or scale-in. Furthermore, we assume each tier has enough memory resources to store an unlimited amount of requests that arrive at the tier, while they are waiting to be processed.

2.6 Virtualized MME design for MTC

This section describes the main points relevant to the proposed vMME design. These main points can be summarized as follows:

- Description of the traffic models considered in the analysis and the target arrival rate design. In addition to H2H traffic, in our analysis we include two generic MTC traffic profiles the vMME should support.

- Presentation of four vMME schemes. The goal of these four schemes is to highlight the impact of different dimensioning strategies in the performance of the vMME and the managed traffic.
- Modeling of the vMME based on queuing theory to obtain a set of performance metrics such as response time of the system.

2.6.1 Traffic models

We assume three types of traffic: eMBB, mMTC and lMTC. In the next subsections, we explain each one in detail.

2.6.1.1 enhanced Mobile Broadband traffic model

We adopt the compound data traffic and user behavior model proposed in [69] for eMBB UEs. This traffic model considers three types of applications, namely: i) Web browsing; ii) HTTP progressive video (e.g. YouTube); and iii) Video calling (e.g. Skype service). These applications are redesigned to generate the data rates predicted for future mobile networks in the METIS project [75]. Then, the mean data rates per user used are higher than the current demand. For each type of application, the mean data rate per UE depends on the following:

- Web browsing: the number of web pages visited, the main object size of each web and the number of embedded objects and their sizes.
- HTTP progressive video: the number of downloaded video clips, the size of each video and the video encoding rate.
- Video calling: the constant bit rate of the call.

Consequently, from [69], the mean data rates per user obtained are 233.28 kbps, 5.25 Mbps, and 142.07 kbps for web browsing, HTTP progressive video, and video calling, respectively.

In this study, we only consider those LTE control procedures that are the most frequent [76]: Service Request (SR), S1 Release (S1), and X2 based Handover (HO) procedures. For each of these procedures, the MME process the following messages (see Figure 2.4) [9, 70]:

- Service Request (SR): The analysis consider the UE-triggered SR. During this procedure, the MME receives three messages: Initial UE message (SR_1), Initial Context Setup Response (SR_2), and Modify Bearer Response (SR_3).

- S1 Release (S1): We consider the S1 triggered by user inactivity. During this procedure, the MME receives three messages: S1 UE Context Release Request ($S1_1$), Release Access Bearers Response ($S1_2$), and S1 UE Context Release Complete ($S1_3$).

- X2 based Handover (HO): We assume the MME participates in the X2-based Handover, a sub-type of the Intra-LTE HO. This X2 based HO involves a change of eNB through the X2 interface, but there is no change of SGW or MME. During this procedure, the MME receives two messages: Path Switch Request message (HO_1) and Modify Bearer Response (HO_2).

Let N_{eMBB} be the number of eMBB UEs and $\overline{\lambda}_{SR}$, $\overline{\lambda}_{S1}$ and $\overline{\lambda}_{HO}$ be the mean generation rate per UE for the SR, S1, and HO procedures, respectively. The mean arrival rate of control messages processed by the MME and generated by eMBB UEs can be derived as:

$$\overline{\lambda}_{eMBB} = N_{eMBB} \cdot (3 \cdot \overline{\lambda}_{SR} + 3 \cdot \overline{\lambda}_{S1} + 2 \cdot \overline{\lambda}_{HO}) \tag{2.3}$$

2.6.1.2 Machine-Type Communication traffic model

We consider two classes of MTC devices: massive MTC (mMTC) devices and low latency MTC (lMTC) devices. Let p_h, where $h \in \{mMTC, lMTC\}$, denote the percentage of each type of MTC device, and let r be the ratio of MTC devices per eMBB UEs. Then, the total number of MTC devices of each class, N_h, is given by:

$$N_h = N_{eMBB} \cdot r \cdot p_h \tag{2.4}$$

Both classes follow a two states traffic model: *active* and *alarm*. Figure 2.6 illustrates the traffic model state transitions. The MTC devices generate small data packets following a Poisson process in both states. In the *active* state,

packets are generated infrequently, with a mean rate $\overline{\lambda}_{active}$. When an event happens, a percentage of mMTC or lMTC devices, denoted by a_h, change to the *alarm* state. During the *alarm* state, MTC devices generate packets more frequently, with a mean rate denoted as $\overline{\lambda}_{alarm}$. We assume the event has a fixed time duration t_e, and the probability that in a certain instant of time an MTC device is in the *alarm* state is denoted as p_h^e. After the event, all devices in the *alarm* state return to the *active* state.

Additionally, mMTC events occur at certain instants in time caused by a large number of mMTC devices synchronizing their report transmissions. Some of these events are periodic, whereas others are not. For simplicity, we assume the mMTC synchronization events take place at periodic instants of time. On the other hand, lMTC events are caused by alarm situations, i.e., they are unpredictable. Consequently, we assume lMTC events occur at random instants. Furthermore, when an event occurs, we assume the percentage of MTC devices participating in a synchronization event, a_h, is a discrete random variable. By appropriately choosing $\overline{a}_h$, $\overline{\lambda}_{active}$ and $\overline{\lambda}_{alarm}$, we can model the traffic peaks caused by the alarms.

For simplicity, we assume there is no coordination among mMTC and lMTC events. As the MTC ratio of UL traffic volume is higher than DL, we only consider UL traffic. To transmit a report, the MTCDs triggers the CP procedure. Regarding the CP procedure, we assume the messages 13 and 14 of Figure 2.4 are not considered. The reasoning is there is a connection established between MME and SGW (S11-U connection).

Let us analyze the mean generation rate of the CP procedure per MTC class.

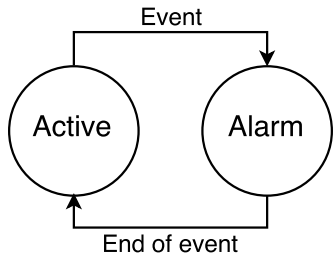


Figure 2.6: MTC device state transition diagram.

As there is one control message per CP procedure processed at the MME, the mean generation rate of the CP procedure equals the mean arrival rate of the processed CP control messages per MTC type. Let us denote this mean arrival rate as $\overline{\lambda}_h^v$. The parameter v defines the time interval used for averaging the arrival rate. Particularly, v can take three possible values $\{active, event, maxevent\}$, which corresponds to the three following different traffic situations (see Figure 2.7):

- *Active*: The interval of time where all MTC devices are in the *active* state, i.e., in the absence of events.

- *Event*: The interval of time where an event takes place, i.e., $N_h \times a_h$ devices are in the *alarm* state. The averaging is carried out over all possible events, each with a given a_h.

- *Maxevent*: The time interval where the largest event takes place, i.e., $N_h \times max(a_h)$ devices are in the *alarm* state.

Then, for each averaging time interval, the mean arrival rate $\overline{\lambda}_h^v$ can be calculated as:

$$\begin{aligned}
&\text{Mean rate in } active && \overline{\lambda}_h^{active} && = N_h \cdot \overline{\lambda}_{active} \\
&\text{Mean rate in } event && \overline{\lambda}_h^{event} && = N_h \cdot \left[(1-\overline{a}_h) \cdot \overline{\lambda}_{active} + \overline{a}_h \cdot \overline{\lambda}_{alarm}\right] \\
&\text{Mean rate in } maxevent && \overline{\lambda}_h^{maxevent} && = N_h \cdot \left[(1-max(a_h)) \cdot \overline{\lambda}_{active} + \right. \\
& && && \left. max(a_h) \cdot \overline{\lambda}_{alarm}\right]
\end{aligned} \tag{2.5}$$

Finally, let $\overline{\lambda}_h$ be the mean arrival rate of CP control messages to the vMME per MTC class. Then, $\overline{\lambda}_h$ follows that:

$$\overline{\lambda}_h = (1-p_h^e) \cdot \overline{\lambda}_h^{active} + p_h^e \cdot \overline{\lambda}_h^{event} \tag{2.6}$$

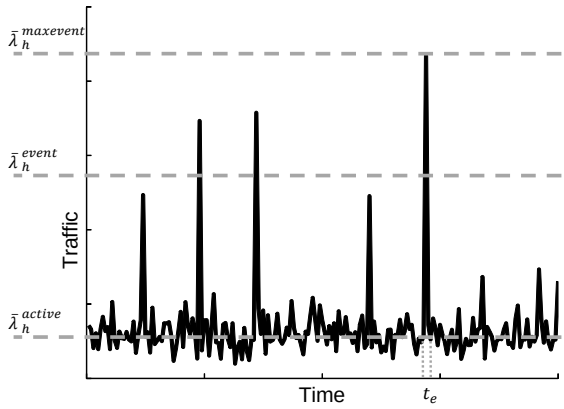


Figure 2.7: MTC traffic model parameters.

2.6.2 vMME design possibilities

2.6.2.1 Target arrival rate design

The vMME dimensioning is done at each tier separately, based on [73]. To dimension the resources needed at each tier of the vMME model in advance, we define a design parameter named *target arrival rate per tier*. The target arrival rate is the maximum arrival rate of signaling messages the vMME has to satisfy a mean response time budget ($\overline{D}_j$) for tier j. That is, the vMME is required to have sufficient processing capacity at tier j to satisfy $\overline{D}_j$ with an arrival rate up to the target.

For eMBB traffic, we set the target arrival rate, λ_{eMBB}^t, equal to the mean arrival rate of eMBB control messages, then $\lambda_{eMBB}^t = \overline{\lambda}_{eMBB}$. For MTC traffic, the target arrival rate λ_h^t, with $h \in \{mMTC, lMTC\}$, depends on whether we choose the vMME to satisfy $\overline{D}_j$ at each tier j during the traffic peaks caused by MTC events. We consider the following three approaches to serve the traffic peaks:

- *Peak*: denoted by λ_h^{peak}. We set the target arrival rate equal to the mean arrival rate during the largest event, $\overline{\lambda}_h^{maxevent}$. With this setting, each tier j of the vMME has to satisfy $\overline{D}_j$ for all traffic peaks, including the largest one. Therefore, it follows that:

$$\lambda_h^t = \lambda_h^{peak} = \overline{\lambda}_h^{maxevent} \tag{2.7}$$

- *Intense smoothing of peaks*: denoted by $\lambda_h^{intsmooth}$. We set the target arrival rate equal to a weighted sum of the mean arrival rates in absence of an event and during an event (see equations (2.5)). Therefore, the target arrival rate can be calculated as:

$$\lambda_h^t = \lambda_h^{intsmooth} = (1 - w_h) \cdot \overline{\lambda}_h^{active} + w_h \cdot \overline{\lambda}_h^{event} \tag{2.8}$$

  where w_h, with $h \in \{mMTC, lMTC\}$, denotes the weighting factor of both MTC arrival rates. We select w_h such that $\lambda_h^{intsmooth}$ is slightly higher than $\overline{\lambda}_h$ (see equation (2.6)). With this setting, each tier j of the vMME is not able to satisfy $\overline{D}_j$ during the traffic peaks. Consequently, part of the signaling messages arriving during a traffic peak will be served after the end of the event.

- *Moderate smoothing of peaks*: denoted by $\lambda_h^{smoothpeak}$. We set the target arrival rate equal to a weighted sum of the mean arrival rates in absence of an event and during the largest event (see equation (2.5)). Therefore, it follows that:

$$\lambda_h^t = \lambda_h^{smoothpeak} = (1 - w_h) \cdot \overline{\lambda}_h^{active} + w_h \cdot \overline{\lambda}_h^{maxevent} \tag{2.9}$$

  With this setting, each vMME tier j is not able to satisfy $\overline{D}_j$ during the traffic peaks. However, $\lambda_h^{smoothpeak} > \lambda_h^{intsmooth}$, thus, the signaling messages arriving during a traffic peak will be served faster than in the previous case.

2.6.2.2 vMME schemes

Within the four vMME presented schemes, two of them are baseline schemes. These two baseline schemes, named the `Baseline Scheme (BS)` and the `Overdimensioned Scheme (OS)`, highlight the impact of the overdimensioned resources due to the inclusion of MTC events. The remain two schemes, named the `Traffic Separated Scheme (TS)` and the `Traffic Shaper Scheme (SS)`, include

new mechanisms to optimize the dimensioning of the vMME, while at the same time they satisfy the *target arrival rate per tier* defined.

Baseline scheme: In this scheme, all resources are shared by all traffic classes, and all tiers (FE, SL, and SDB) have to convey the same target arrival rate. Figure 2.8 shows the design of this scheme.

We set the target arrival rate such that each tier is able to serve the mean arrival rate from eMBB UEs, mMTC devices and lMTC devices. Hence, each tier j is not able to satisfy $\overline{D}_j$ during the MTC traffic peaks. For them, we apply the intense smoothing approach mentioned previously. Consequently, the vMME will be overloaded for a period of time during and after traffic peaks.

Let λ_{FE}, λ_{SL} and λ_{SDB} respectively denote the target arrival rate of control messages at the FE, the SL, and the SDB tiers. For `BS`, these target arrival rates can be calculated as:

$$\lambda_{FE} = \lambda_{SL} = \lambda_{SDB} = \lambda^t_{eMBB} + \lambda^t_{mMTC} + \lambda^t_{lMTC} = \overline{\lambda}_{eMBB} + \lambda^{intsmooth}_{mMTC} + \lambda^{intsmooth}_{lMTC} \tag{2.10}$$

Overdimensioned Scheme: This scheme follows the same design as `BS`, but it is overdimensioned to always satisfy $\overline{D}_j$ at each vMME tier, including during the traffic peaks caused by MTC devices. Therefore, in this case the target arrival rate equals to the sum of the mean arrival rate for eMBB and the mean arrival rate during the largest event of mMTC and lMTC devices. Owing to we assume

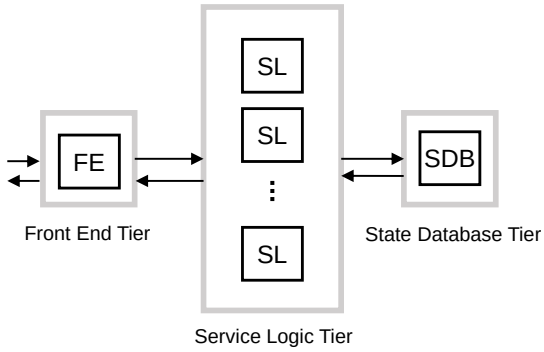


Figure 2.8: `Baseline Scheme` block diagram.

there is no coordination of mMTC and lMTC events, the resources do not need to be designed for simultaneous mMTC and lMTC events. Peaks from mMTC devices are expected to have a bigger impact on the vMME because of the massive number of devices. Then, we focus the arrival rate during mMTC peaks, using the mean arrival rate during the largest mMTC traffic peak, $\overline{\lambda}_{mMTC}^{maxevent}$. Therefore, the target arrival rate can be computed as:

$$\lambda_{FE} = \lambda_{SL} = \lambda_{SDB} = \overline{\lambda}_{eMBB} + \lambda_{mMTC}^{peak} + \lambda_{lMTC}^{intsmooth} \tag{2.11}$$

Traffic Separated Scheme: To mitigate the overdimensioning caused by the mMTC synchronization events, this scheme separates the processing of the traffic classes. To do that, we propose to use dedicated resources and dimension them separately for each traffic class. However, it is challenging to apply such separation at all tiers. Regarding the FE tier, the vMME has to be seen as a single entity by the remaining entities of the EPC. Additionally, the FE has to classify the traffic into the considered traffic classes. For these reasons, we do not consider the option of dividing the FE tier per traffic class. Additionally, the SDB tier does not consider this division per traffic class. The rationale is the SDB is considerably more expensive than the other elements of the design [77], and therefore, such a division per traffic class would yield a costly and, therefore, unattractive design.

Consequently, we apply the separation of dedicated resources per traffic class only at the SL tier (see Figure 2.9). The benefit of this scheme is the target arrival rate of each SL pool can be optimized using the main traffic characteristics of the traffic served. Additionally, the mean response time budget $\overline{D}_j$ at the SL tier can be set differently for each traffic class.

This scheme, unlike previous ones, has a different target arrival rate defined per tier. This is due to the FE tier having to convey all traffic to avoid bottlenecks at this tier. However, SL and SDB tiers can be optimized to reduce overdimensioning. Then, we set the target arrival rate in the FE tier as in the `OS` explained above:

$$\lambda_{FE} = \overline{\lambda}_{eMBB} + \lambda_{mMTC}^{peak} + \lambda_{lMTC}^{intsmooth} \tag{2.12}$$

Due to SLs are dimensioned differently per traffic class, we set their target

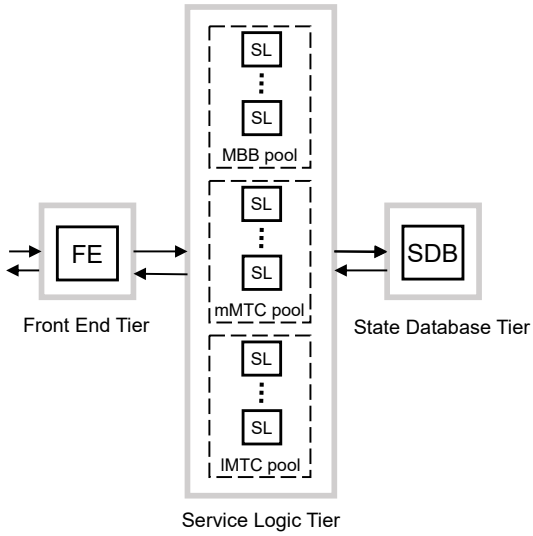


Figure 2.9: `Traffic separated Scheme` block diagram.

arrival rate differently: i) for eMBB traffic, the mean arrival rate; ii) for mMTC traffic, we apply moderate smoothing of traffic peaks, $\lambda_{mMTC}^{smoothpeak}$; and iii) for lMTC traffic, the traffic peak, λ_{lMTC}^{peak}.

Note that the mMTC pool of SL instances is not designed to support mMTC traffic peaks, but this is not expected to be a major issue as mMTC devices are assumed to be delay tolerant. Then, we set the target arrival rate per traffic class as:

$$\begin{aligned} eMBB \quad & \lambda_{SL} = \overline{\lambda}_{eMBB} \\ mMTC \quad & \lambda_{SL} = \lambda_{mMTC}^{smoothpeak} \\ lMTC \quad & \lambda_{SL} = \lambda_{lMTC}^{peak} \end{aligned} \tag{2.13}$$

At the SDB tier, the target arrival rate is a sum of the target arrival rate of each class of SLs, then:

$$\lambda_{SDB} = \overline{\lambda}_{eMBB} + \lambda_{mMTC}^{smoothpeak} + \lambda_{lMTC}^{peak} \tag{2.14}$$

Traffic Shaper Scheme: This scheme adds a traffic shaper after the FE

tier to control the traffic of each class to be processed (see Figure 2.10). This can be considered as a middle approach between the `OS` and the `TS`. In this scheme, all resources are shared among all traffic classes. However, the traffic shaper can smooth traffic peaks and benefit some traffic classes through their shaping criteria. In addition, this scheme avoids the multiplexing loss of processing resources experienced in the `TS` due to the spare capacity of the dedicated workers that cannot be used for other traffics classes.

Each tier has to satisfy $\overline{D}_j$ with a target arrival rate. The target arrival rate used to design the FE tier can be calculated as:

$$\lambda_{FE} = \overline{\lambda}_{eMBB} + \lambda_{mMTC}^{peak} + \lambda_{lMTC}^{intsmooth} \tag{2.15}$$

We assume that the traffic shaper is implemented with a token bucket for each traffic class. The token rates of the buckets equal the target arrival rate of each class: i) for eMBB traffic, the mean arrival rate; ii) for mMTC traffic, we apply moderate smoothing to the traffic peaks, $\lambda_{mMTC}^{smoothpeak}$; and iii) for lMTC traffic, the traffic peak, λ_{lMTC}^{peak}.

The SL and the SDB tiers have the same target arrival rate as they process the same traffic after the shaping of the traffic shaper. Then, the target arrival rate can be calculated as:

$$\lambda_{SL} = \lambda_{SDB} = \overline{\lambda}_{eMBB} + \lambda_{mMTC}^{smoothpeak} + \lambda_{lMTC}^{peak} \tag{2.16}$$

To sum up, Table 2.1 summarizes the target arrival rate criteria of each baseline and proposed scheme.

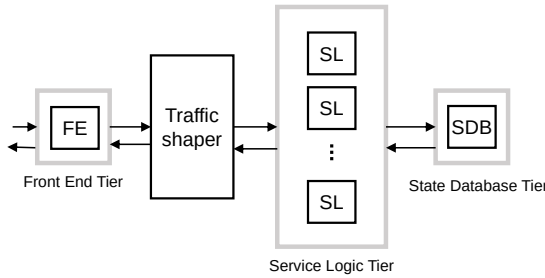


Figure 2.10: `Traffic shaper scheme` block diagram.

Table 2.1: Target arrival rate at each tier for the proposed schemes.

Scheme \ Tier	Front End (FE)	Service Logic (SL)	State Database (SDB)
Baseline Scheme (BS)	$\lambda_{FE} = \overline{\lambda}_{eMBB} + \lambda_{mMTC}^{intsmooth} + \lambda_{lMTC}^{intsmooth}$	$\lambda_{SL} = \lambda_{FE}$	$\lambda_{SDB} = \lambda_{SL}$
Overdimensioned Scheme (OS)	$\lambda_{FE} = \overline{\lambda}_{eMBB} + \lambda_{mMTC}^{peak} + \lambda_{lMTC}^{intsmooth}$	$\lambda_{SL} = \lambda_{FE}$	$\lambda_{SDB} = \lambda_{SL}$
Traffic separated Scheme (TS)	$\lambda_{FE} = \overline{\lambda}_{eMBB} + \lambda_{mMTC}^{peak} + \lambda_{lMTC}^{intsmooth}$	eMBB $\lambda_{SL} = \overline{\lambda}_{eMBB}$ mMTC $\lambda_{SL} = \lambda_{mMTC}^{smoothpeak}$ lMTC $\lambda_{SL} = \lambda_{lMTC}^{peak}$	$\lambda_{SDB} = \overline{\lambda}_{eMBB} + \lambda_{mMTC}^{smoothpeak} + \lambda_{lMTC}^{peak}$
Traffic Shaper Scheme (SS)	$\lambda_{FE} = \overline{\lambda}_{eMBB} + \lambda_{mMTC}^{peak} + \lambda_{lMTC}^{intsmooth}$	$\lambda_{SL} = \overline{\lambda}_{eMBB} + \lambda_{mMTC}^{smoothpeak} + \lambda_{lMTC}^{peak}$	$\lambda_{SDB} = \lambda_{SL}$

2.6.3 vMME modeling

The dimensioning of the schemes previously presented requires a model that provides an estimation of the service response time as a function of the processing resources. To do that, we assume the vMME model proposed in [69], based on the model of a typical cloud processing chain [78]. The notation and main definitions used are summarized in Table 2.2.

This analysis uses Jackson's open queuing network to model the vMME architecture. Figure 2.11 shows the considered vMME queue model. In this model, the SDB and the FE tiers are modeled as M/M/1 queues, and the SL tier is modeled as an M/M/m queue.

Jackson's theorem states that the numbers of messages in the system queues are independent of the other queues, and consequently, the service response time of the complete system is equal to the sum of the service response time of the

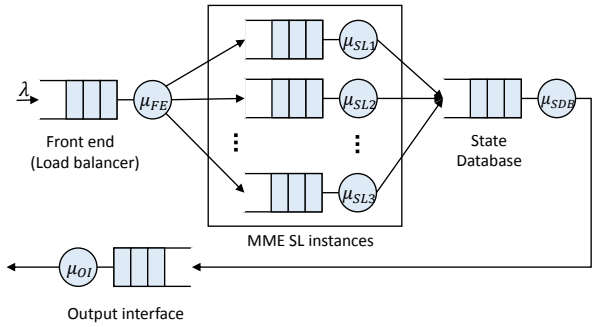


Figure 2.11: vMME queue model.

Table 2.2: vMME modeling primary definitions.

Notation	Description
$\overline{T}_j$	Mean response time in tier j (where $j \in \{FE, SL, SDB\}$)
$\overline{D}_j$	Target mean response time in tier j
N_{eMBB}	Number of eMBB UEs
N_h	Number of MTC devices (where $h \in \{mMTC, lMTC\}$)
$\overline{\lambda}_{SR}$	Service Request mean generation rate per eMBB UE
$\overline{\lambda}_{SRR}$	S1 Release mean generation rate per eMBB UE
$\overline{\lambda}_{HO}$	Handover mean generation rate per eMBB UE
$\overline{\lambda}_{eMBB}$	Mean arrival rate of control messages generated by eMBB UEs
$\overline{\lambda}_{active}$	Mean MTC device packet rate in the *active* state
$\overline{\lambda}_{alarm}$	Mean MTC device packet rate in the *alarm* state
a_h	Percentage of MTC devices that changes to the *alarm* state at an MTC event
p_h	Percentage of each class of MTC devices
p_h^e	Probability of an MTC device is in the *alarm* state in a certain instant of time
r	Ratio of MTC devices per eMBB UE
$\overline{\lambda}_h$	CP control messages mean arrival rate
$\overline{\lambda}_h^v$	CP mean arrival rate during an observation period v (where $v \in \{active, event, maxevent\}$)
w_h	Arrival rate weighting factor for the absence or not of MTC events
λ_h^t	Dimensioning target arrival rate
$\lambda_h^{intsmooth}$	Target arrival rate with intense smoothing of peaks
$\lambda_h^{smoothpeak}$	Target arrival rate with moderate smoothing of peaks
λ_h^{peak}	Target arrival rate of peak traffic
λ_{FE}	Front end target arrival rate
λ_{SL}	Service logic target arrival rate
λ_{SDB}	State database target arrival rate
μ_j	Tier j instance service rate
m_j	Number of instances of tier j

queue in each tier. In the present analysis, we also assume the SDB and the FE can be replicated. When these tiers are replicated, they are modeled as M/M/m queues.

Given a target arrival rate, the aim of our vMME dimensioning is to determine the minimum number of instances required at each tier j to guarantee the mean response time budget $\overline{D}_j$. Due to we consider Jackson's open queuing network, the mean response time of the system $\overline{T}$ can be computed as:

$$\overline{T} = \sum_{j} \overline{T}_j \tag{2.17}$$

where $\overline{T}_j$ is the mean response time at each tier $j \in \{FE, SL, SDB\}$. Assuming each tier is modeled by an M/M/m queue, $\overline{T}_j$ can be derived as:

$$\overline{T}_j = \frac{1}{\mu_j} + \frac{C(m_j, \rho_j)}{m_j \cdot \mu_j - \lambda_j} \tag{2.18}$$

where $\rho_j = \frac{\lambda_j}{\mu_j}$, μ_j is the service rate of one tier instance, λ_j is the target arrival rate considered for dimensioning at tier j (summarized in Table 2.1), m_j is the number of instances of the tier, and $C(m_j, \rho_j)$ is Erlang's C formula. $C(m_j, \rho_j)$ represents the probability that an arriving packet has to wait in the queue of the tier because all of the instances are busy, and it has the following expression:

$$C(m_j, \rho_j) = \frac{\left(\frac{(m_j \cdot \rho_j)^{m_j}}{m_j!}\right) \cdot \left(\frac{1}{1-\rho_j}\right)}{\sum_{k=0}^{m_j-1} \frac{(m_j \cdot \rho_j)^k}{k!} + \left(\frac{(m_j \cdot \rho_j)^{m_j}}{m_j!}\right) \cdot \left(\frac{1}{1-\rho_j}\right)} \tag{2.19}$$

The processing times of the FE, SDB, and output interface are constant. However, the processing time of an SL is different for each control message [69]. Consequently, the mean service time of an SL, $\bar{t}_{SL} = \frac{1}{\mu_{SL}}$, will depend on the frequency of each type of control procedure. For this reason, $\bar{t}_{SL}$ will be different for each scheme considered in this work.

Let t_{SR_i}, t_{S1_i}, t_{HO_i} and t_{CP} denote the processing time of the i-th message of the SR, S1, HO and CP procedures, respectively. The mean service time of an SL for each scheme considered is summarized in Table 2.3.

Note that we perform dimensioning for each tier individually. Then, the dimensioning problem for each tier can be formulated as:

$$m_j = \min\{M_j : \overline{T}_j(\lambda_j, M_j) \leq \overline{D}_j, M_j \in \mathbb{N}\} \tag{2.20}$$

where $\overline{D}_j$ is the target mean response time for each tier. Hence, m_j can be computed with a simple iterative algorithm that increases the number of tier instances until the condition $\overline{T}_j(\lambda_j, M_j) \leq \overline{D}_j$ is met.

Table 2.3: SL mean service time.

Scheme	Mean Service Time
Baseline Scheme (BS)	$\bar{t}_{SL} = \frac{N_{eMBB}\cdot(\bar{\lambda}_{SR}\cdot(t_{SR_1}+t_{SR_2}+t_{SR_3})+\bar{\lambda}_{S1}\cdot(t_{S1_1}+t_{S1_2}+t_{S1_3})+\bar{\lambda}_{HO}\cdot(t_{HO_1}+t_{HO_2}))+(\lambda_{mMTC}^{intsmooth}+\lambda_{lMTC}^{intsmooth})\cdot t_{CP}}{\bar{\lambda}_{eMBB}+\lambda_{mMTC}^{intsmooth}+\lambda_{lMTC}^{intsmooth}}$
Overdimensioned Scheme (OS)	$\bar{t}_{SL} = \frac{N_{eMBB}\cdot(\bar{\lambda}_{SR}\cdot(t_{SR_1}+t_{SR_2}+t_{SR_3})+\bar{\lambda}_{S1}\cdot(t_{S1_1}+t_{S1_2}+t_{S1_3})+\bar{\lambda}_{HO}\cdot(t_{HO_1}+t_{HO_2}))+(\lambda_{mMTC}^{peak}+\lambda_{lMTC}^{intsmooth})\cdot t_{CP}}{\bar{\lambda}_{eMBB}+\lambda_{mMTC}^{peak}+\lambda_{lMTC}^{intsmooth}}$
Traffic separated Scheme (TS)	eMBB $\bar{t}_{SL} = \frac{N_{eMBB}\cdot(\bar{\lambda}_{SR}\cdot(t_{SR_1}+t_{SR_2}+t_{SR_3})+\bar{\lambda}_{S1}\cdot(t_{S1_1}+t_{SR_2}+t_{SR_3})+\bar{\lambda}_{HO}\cdot(t_{HO_1}+t_{HO_2}))}{\bar{\lambda}_{eMBB}}$ mMTC $\bar{t}_{SL} = t_{CP}$ lMTC $\bar{t}_{SL} = t_{CP}$
Traffic Shaper Scheme (SS)	$\bar{t}_{SL} = \frac{N_{eMBB}\cdot(\bar{\lambda}_{SR}\cdot(t_{SR1}+t_{SR2}+t_{SR3})+\bar{\lambda}_{S1}\cdot(t_{SR1}+t_{SR2}+t_{SR3})+\bar{\lambda}_{HO}\cdot(t_{HO1}+t_{HO2}))+(\lambda_{mMTC}^{smoothpeak}+\lambda_{lMTC}^{peak})\cdot t_{CP}}{\bar{\lambda}_{eMBB}+\lambda_{mMTC}^{smoothpeak}+\lambda_{lMTC}^{peak}}$

2.7 vMME performance evaluation

This section presents the simulation setup and results obtained when evaluating the four vMME schemes explained. The goal is to compare the vMME dimensioning, costs, and response time using these four schemes. From this comparison we want to study the impact different dimensioning targets and design schemes can have in the performance of the vMME and the resulting management of the traffic classes considered.

2.7.1 Simulation setup

The evaluation methodology includes three steps, namely:

1. The dimensioning of each tier of the vMME model using the analysis presented in subsection 2.6.3 and the target arrival rate described in Table 2.1 as input. The estimation is done for a range of UEs and a given ratio of MTC devices per UE. Additionally, from the target arrival rate and dimensioning of each scheme we can estimate the associated cost of the scheme used.

2. The generation of signaling traces for each traffic class (eMBB, mMTC, and lMTC). The trace generation considers a specific number of UEs and a given ratio of MTC devices per UE.

3. Finally, the simulation of the vMME queuing model using the signaling traces as input to obtain the vMME response time experienced by a control plane message.

To simulate the vMME queuing model we use MATLAB Simulink framework. The main simulation parameters are summarized in Table 2.4.

To estimate the system running cost, we consider the Amazon EC2 service, with the costs and configuration detailed in Table 2.5. We assume a medium-sized CPU instance *m3.xlarge* with an average of 11.38×10^9 float operations per second [79]. We use the price of the load balancing service provided by Amazon to estimate the cost of the FE tier. Our setup also includes the Amazon Aurora database [80]. The overall cost includes the per-instance cost, the time-based

Table 2.4: Simulation setup configuration.

Simulation Parameters		
Simulation Time		300 s
eMBB UEs		636,000
MTC devices per eMBB UE		10
MTC packet size		200 B
MTC event duration		1 s
FE mean response time budget	$\overline{D}_{FE}$	1 ms
SL mean response time budget	$\overline{D}_{SL}$	1 ms
SDB mean response time budget	$\overline{D}_{SDB}$	1 ms
FE service rate	μ_{FE}	120000 packets/s
SDB service rate	μ_{SDB}	100000 transactions/s
Output interface service rate	μ_{OI}	5000000 packets/s
Unweighted sliding-average smooth		90 ms
w_{mMTC}		0.1
w_{lMTC}		0.0033
Processing times per control message [69]		
Service Request (SR)	$t_{SR_1} = 127.4\ \mu$s	$t_{SR_2} = 94.0\ \mu$s $t_{SR_3} = 93.2\ \mu$s
S1 Release (SR)	$t_{S1_1} = 94.0\ \mu$s	$t_{S1_2} = 94.0\ \mu$s $t_{S1_3} = 93.2\ \mu$s
X2-based Handover (HO)	$t_{HO_1} = 94.0\ \mu$s	$t_{HO_2} = 94.0\ \mu$s
Control Plane optimization (CP)	$t_{CP} = 145.05\ \mu$s	
Traffic Models		
enhance Mobile Broadband (eMBB) [69]	$\overline{\lambda}_{SR}$	0.0045 procedures/s
	$\overline{\lambda}_{S1}$	0.0045 procedures/s
	$\overline{\lambda}_{HO}$	0.0012 procedures/s
MTC devices	$\overline{\lambda}_{active}$	0.0033 packets/s
	$\overline{\lambda}_{alarm}$	0.033 packetss
massive MTC (mMTC)	Percentage of mMTC devices	90 %
	Event period	60 s
	Event magnitude values	$\{10, 30, 50, 8\}$ %
	Event magnitude values probability mass	$\{5, 60, 20, 15\}$ %
Low latency MTC (lMTC)	Percentage of lMTC devices	10 %
	Event period	Unique (at 40 s of the simulation)
	Event magnitude value	33 %

rental fee and the data traffic processed.

2.7.2 Results

This section presents the comparison of the four vMME schemes in terms of: i) dimensioning of the required resources; ii) estimation of the costs based on the model of Amazon EC2; and iii) the evaluation of the response time of the vMME schemes for each traffic class.

2.7.2.1 vMME dimensioning

Following the analytical analysis in subsection 2.6.3, Figure 2.12 shows the resulting dimensioning at each tier of the vMME versus N_{eMBB} for the four vMME schemes. To simplify the comparison in `TS`, we set the same response time budget for eMBB and lMTC. Additionally, for `TS`, the required number of SL instances m_{SL} is the sum of the required number of SL instances for each type of traffic (see Figure 2.12d). Figure 2.13 illustrates the cost per hour for each vMME scheme.

As expected, the `OS` demands the greatest amount of resources, being the most expensive scheme. Conversely, the `BS` is the least expensive one. The `TS` and the `SS` achieve a noticeable reduction in cost in comparison with `OS`. This is mainly due to the isolation between traffic types in the `TS` case and the limitation imposed by the traffic shaper on the mMTC traffic arrival rate in the `SS` case. For both, the dimensioning at the SL and the SDB tiers can be performed

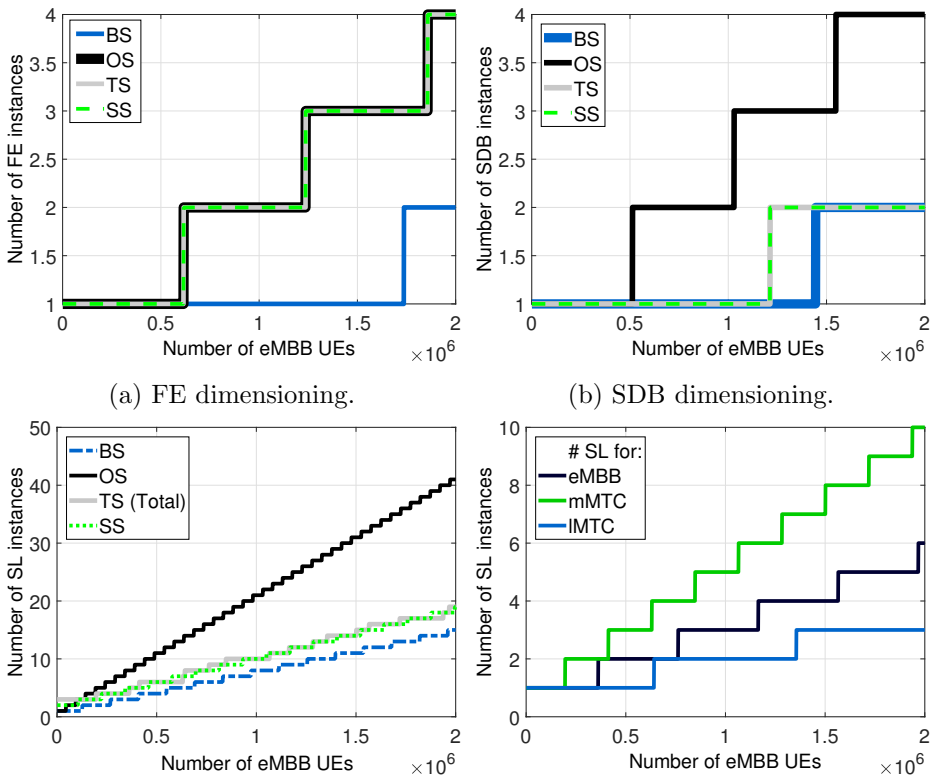
Table 2.5: Cloud service configuration and cost calculation.

Cost	Configuration	Calculation
$C_{ci_{type}}$	*m3.xlarge* instance rental (0.266$/h)	0.266/3600
$C_{ci_{stor}}$	Local storage (10 GB/month) and optimized data access (0.025$/h)	$10 \cdot 0.10 + 0.025/3600$
$C_{ci_{thro}}$	Data sent from the data center ($\lambda(message/s) \cdot 200$ (byte/message))	0.000($)/GB First GB/month 0.090($)/GB Up to 10 TB/month 0.085($)/GB Next 40 TB/month 0.070($)/GB Next 100 TB/month 0.050($)/GB Next 350 TB/month
$C_{db_{type}}$	Aurora $db.r3.8xlarge$ instance (4.64$/h)	4.64/3600
$C_{db_{stor}}$	0.1$ per GB/month, for a total database size of $N_U \cdot 1KB$	$(0.1 \cdot N_U \cdot 1024 \cdot \lambda/10^9)/2{,}628{,}000$
$C_{db_{thro}}$	0.2$ per million transactions/month	$0.2 \cdot \lambda/10^6$
$C_{b_{type}}$	Service fee of 0.025$/month	0.025/2,628,000
$C_{b_{thro}}$	0.008$ per GB serviced, assuming 200 (byte/message)	$\lambda \cdot 0.008 \cdot 200/10^9$

without considering λ_{mMTC}^{peak}, which is around 12.47-times greater than λ_{lMTC}^{peak} in our simulation setup. However, both the `TS` and `SS` cases are designed to satisfy the lMTC delay requirement.

As we assume in the system model each tier has enough memory resources to store the evaluated requests that arrive at the tier, Table 2.6 summarizes a simplified memory consumption estimation for each tier of the vMME to check the amount of memory required in our study. For the estimation, the number of SL instances are obtained from Figure 2.12c. We assume 16,545 packets queued in the system. This number is the worst case of packets queued in the `TS`, obtained from the results of the next subsection.

The different vMME designs show the benefits of shared/dedicated resources



(a) FE dimensioning.

(b) SDB dimensioning.

(c) SL dimensioning.

(d) Detailed SL `TS` dimensioning.

Figure 2.12: Dimensioning at each tier of the vMME model and the four vMME scheme (10 MTC devices per eMBB UE).

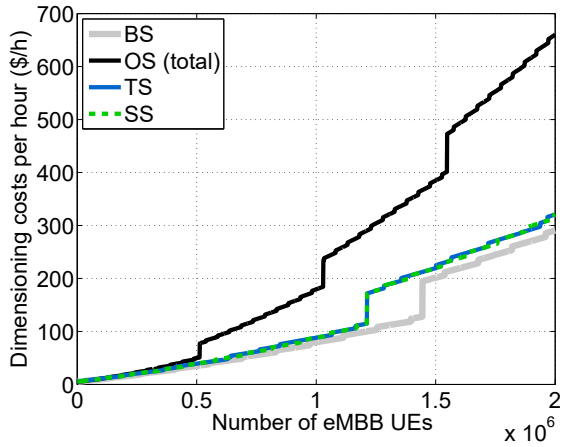


Figure 2.13: Dimensioning costs comparison per evaluated scheme (10 MTC devices per eMBB UE).

definition. The previous figures highlight a correct virtualization design can satisfy the needed requirements while the demanded resources are reduced. However, to be able to do that, the operators need to know the traffic characteristics/requirements of the managed users.

2.7.2.2 vMME delay

In this case, we study the response time of the vMME by means of simulation of the vMME queuing model for each scheme. The simulations return the total response time of the vMME and the response time of each vMME tier for each

Table 2.6: Memory consumption estimation (UE context extracted from [19, 20]).

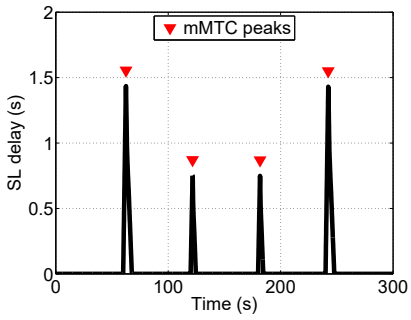
Element	**Memory Consumption**	**Sample Scenario** ($N_{eMBB} = 636,000$ **eMBB UEs** $N_{MTC} = 10 \cdot N_{eMBB}$)
State database	UE context: 264 B/UE	264 B/UE · $(N_{eMBB} + N_{MTC})$ = 1846 MB
Service logic	Operating System ROM: 1000 MB/instance Operating System RAM: 400 MB/instance [81] UE context: 264 B/UE Packet size: 200 B	ROM: 1000 MB · 7 instances = 7000 MB RAM: (400 MB + 264 B/UE) · 7 instances + 16,545 packets · 200 B/packet = 2803 MB

control packet processed. The sample scenario consists of 636,000 eMBB UEs and 300 s of duration. Table 2.7 summarizes the specific dimensioning of the considered sample scenario. The number of UEs is selected such that the processing capacity of the `OS` case and the lMTC SL pool for the `TS` case are about to require an additional SL instance to satisfy $\overline{D}_{SL}$.

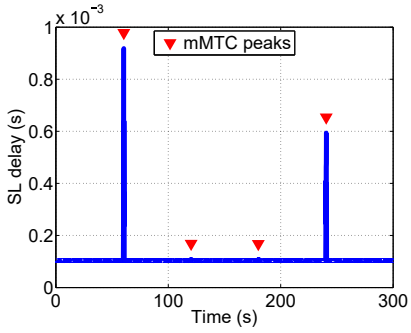
Figure 2.14 shows the response time of a control message at the SL tier for the four vMME schemes. Additionally, Figure 2.15 illustrates the CDF of the overall system response time (i.e. the sum of the delay experienced by a packet at each tier of the vMME). Note Figure 2.14 only includes the response time of the SL tier. Thus, for the `SS` case, the delay experienced by the control messages due to the traffic shaper queues is not included. However, the impact of the waiting time at the traffic shaper queues can be seen on Figure 2.15. The response time results are filtered with a simple 90 ms moving average to smooth the representation of the data.

The results show the `BS` response time is higher than the target mean response time at the SL tier ($\overline{D}_{SL} = 1$ ms) during the mMTC alarm events (Figure 2.14a). This is due to the SL tier is under-dimensioned to support the mMTC traffic peaks. Consequently, in such situations, the mMTC traffic might delay the other traffic types. This effect can be seen in Figure 2.15a. On the contrary, for the `OS` case, the response time at the SL tier always satisfy $\overline{D}_{SL}$ as the system is overdimensioned (Figure 2.14b).

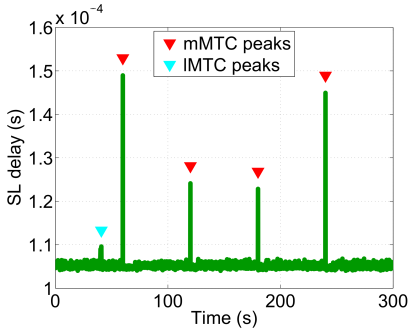
In the `SS` case, the SL tier response time always meets the condition $\overline{T}_{SL} \leq \overline{D}_{SL}$ (Figure 2.14c). That is because the mMTC traffic peaks are limited by the traffic shaper. Moreover, during the lMTC traffic peaks, the system takes advantage of the multiplexing gain. In the `TS` case, the lMTC pool in the SL tier also meets the condition $\overline{T}_{SL} \leq \overline{D}_{SL}$ during the lMTC traffic peak (Figure 2.14f). On the contrary, the mMTC pool exceeds by several orders of magnitude the response time budget during and after the lMTC traffic peaks (Figure 2.14e). This is due to we apply moderate smoothing of the traffic peaks. The eMBB pool also meets the response time budget condition (Figure 2.14d). Note that for the selected number of UEs, only the `OS` case (Figure 2.14b) and the lMTC class in the `TS` case (Figure 2.14f) have a processing load close to the dimensioned capacity.



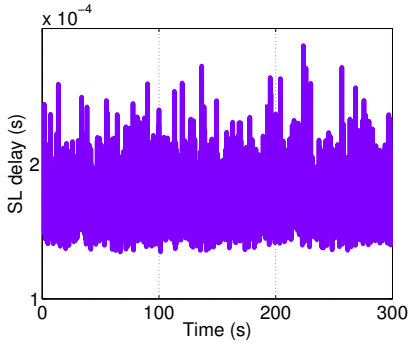
(a) `Baseline Scheme`.



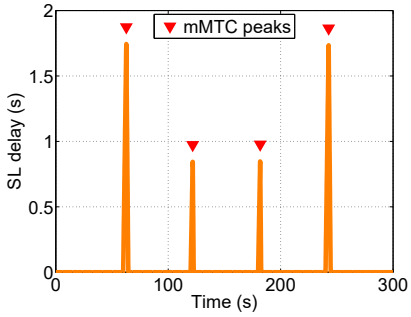
(b) `Overdimensioned Scheme`.



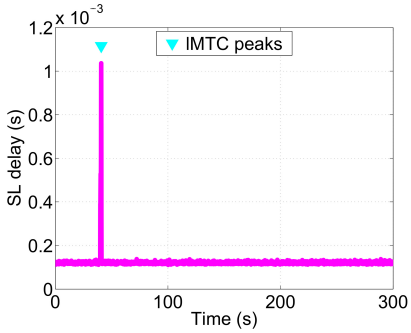
(c) `Traffic Shaper Scheme`.



(d) `Traffic separated Scheme`: eMBB.



(e) `Traffic separated Scheme`: mMTC.

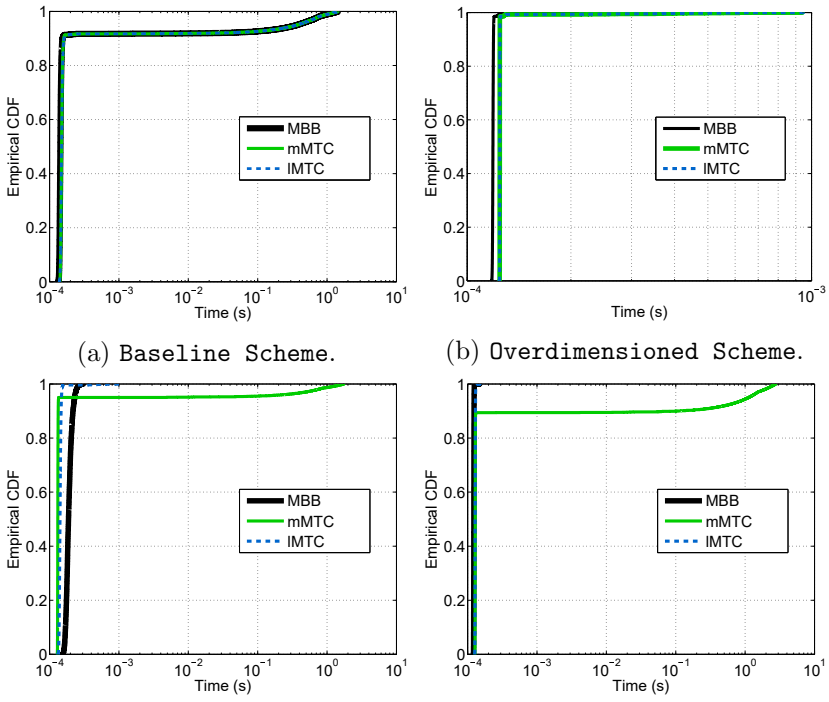


(f) `Traffic separated Scheme`: lMTC.

Figure 2.14: SLs' filtered processing time for each scheme.

Table 2.7: vMME model dimensioning at the simulation point.

Scheme \ Tier	FE instances	SL instances	SDB instances
BS	1	5	1
OS	2	13	2
TS	2	eMBB 2 mMTC 4 lMTC 1	1
SS	2	7	1



(a) `Baseline Scheme`. (b) `Overdimensioned Scheme`.

(c) `Traffic separated Scheme`. (d) `Traffic shaper Scheme`.

Figure 2.15: CDF of the filtered vMME delay for each scheme.

2.8 Conclusions

With a wide range of potential applications, Machine-Type Communication (MTC) devices are rapidly spreading. Cellular networks are being considered as an option to provide connectivity to MTC devices due to their ubiquitous presence, widespread coverage, reliability and support for mobility. However, the widespread provision of MTC entails significant challenges. Among the challenges, this chapter analyzes the scalability of the Mobility Management Entity (MME). The studied MME architecture, named virtualized MME (vMME), is based on NFV paradigm to deploy network services onto virtualized servers. The convergence of cellular networks and NFV improves the scalability and flexibility of the network compared to hardware-based entities. Consequently, the network entities such as the MME can benefit from this paradigm to overcome the foreseen signaling explosion generated by MTC connected devices.

The vMME is decomposed into the three tiers: FE, SL, and SDB. This decomposition has the benefit of higher flexibility, availability, and reduction of the scaling complexity of the VNF. However, this solution also increases the VNF response time, as every packet has to pass through several tiers. In this chapter, we have analyzed and evaluated the scalability and delay of a three-tiered vMME. The evaluation is done considering four vMME design schemes. Two of them are baseline schemes and the remain two are proposed including new mechanisms to optimize the dimensioning of the vMME. The reported comparisons between the schemes include: i) dimensioning of the required resources; ii) estimation of the costs based on the model of Amazon EC2; and iii) the evaluation of the response time of the vMME.

After the conducted simulations, the results show the optimized schemes provide much lower costs than the `Overdimensioned Scheme` while they satisfy the exigent delay requirements of eMBB and low latency MTC. Furthermore, the comparison of the `Traffic Separated Scheme` and the `Traffic Shaper Scheme` shows the multiplexing gain of the latter provides benefits in terms of latency reduction. However, the `Traffic Separated Scheme` can isolate the performance of each traffic class. Regarding the bottlenecks of the proposed schemes, the SDB tier is critical. This is due to the SDB can scale with certain constraints due to its shared-everything architecture. Moreover, the service logic tier design

is determinant, as it considerably affects the performance of each traffic class.

2.8.1 Resulting research contributions

The research contributions resulting from the work done in this chapter are listed below:

- P. Andres-Maldonado, P. Ameigeiras, J. Prados-Garzon, J. J. Ramos-Munoz and J. M. Lopez-Soler, "Reduced M2M signaling communications in 3GPP LTE and future 5G cellular networks," 2016 Wireless Days (WD), Toulouse, 2016, pp. 1-3.

  DOI: 10.1109/WD.2016.7461499

- P. Andres-Maldonado, P. Ameigeiras, J. Prados-Garzon, J. J. Ramos-Munoz, and J. M. Lopez-Soler, "MME support for M2M communications using network function virtualization," in IARIA The Twelfth Advanced International Conference on Telecommunications (AICT 2016), 2016, pp. 106-111.

  ISBN: 978-1-61208-473-2

- P. Andres-Maldonado, P. Ameigeiras, J. Prados-Garzon, J. Ramos-Munoz, and J. Lopez-Soler, "Virtualized MME Design for IoT Support in 5G Systems," Sensors, vol. 16, no. 8, p. 1338, Aug. 2016.

  DOI: 10.3390/s16081338

Chapter 3

Energy Consumption Analysis at UE Side

Internet of Things (IoT) is set to have a major impact on several verticals such as automotive, energy and utilities, financial services, health care, manufacturing, retail, or logistics. Different emerging IoT applications will have different requirements. This heterogeneity ranges from IoT applications with strict requirements on latency to delay tolerant transmissions, from static to high mobility devices, or from a small volume of infrequent data to a continuous high rate of data. Consequently, there is no one-size-fits-all approach to IoT.

For many IoT use cases, Low-Power Wide-Area Networks (LPWANs) will be a good choice [82]. However, a Low-Power Wide-Area (LPWA) wireless IoT radio access network has four conflicting Key Performance Indicators (KPIs): i) Cost; ii) Battery lifetime; iii) Coverage; and iv) Capacity. Traditional cellular networks fall short on meeting all of the four KPIs.

Within this context, the Third Generation Partnership Project (3GPP) introduced a new access technology called Narrowband Internet of Things (NB-IoT) in June 2016. NB-IoT is a set of specifications particularly well fitted to the LPWA segment. NB-IoT is specifically tailored for ultra-low-end IoT applications within the massive MTC (mMTC) use case. It was introduced in Release 13 and its design goals were [83]:

- Maximum latency of 10 seconds on the Uplink (UL).

- Target coverage of 164 dB Maximum Coupling Loss (MCL).
- User Equipment (UE) battery lifetime beyond 10 years, assuming a stored energy capacity of 5 Wh.
- Massive connection density of 1,000,000 devices per square km in an urban environment.

When the work on NB-IoT started, the overall goal was to find a solution competitive in the LPWA segment. Within this goal, an important part of the study was to meet an objective of extending coverage with 20 dB. For NB-IoT, this improvement is in relation to the General Packet Radio Service (GPRS) coverage used as the reference.

To achieve this improvement, two basic solutions are mainly employed: reduced network bandwidth and soft combined retransmissions. However, both solutions implicate a reduction of the data rate that will increase the time on air while transmitting or receiving. Due to NB-IoT is also required to exhibit 10 years of UE battery lifetime are feasible to support in addition to the coverage extension. This situation entails a trade-off between coverage and energy consumption. That is, in order to use more robust radio configurations to be able to operate in weaker radio conditions, the UE will have more time on air, therefore, more energy will be consumed to finish the transfer. For IoT use cases with UEs in extreme coverage and power-limited situations, this trade-off is crucial and depends heavily on how the UE behaves.

As a new 3GPP radio access technology, there is still a lack of NB-IoT analytical modeling [84]. The study proposed in the present chapter for NB-IoT provides an overview and an analysis of the small data transmission optimizations (i.e. Control Plane optimization (CP), and User Plane optimization (UP)) compared to the conventional Service Request (SR) procedure.

Our analysis is based on a Markov chain to model the behavior of the UE while performing these control procedures. We adopt the Markov chain analysis from the work presented in [85]. By using the Markov chain and describing the steps the UE performs in each procedure, we can obtain an estimation of the average energy consumption of the UE. Furthermore, we can estimate the amount of radio resources used by the UE in the different NB-IoT radio channels while performing the control procedures.

We consider four possible communication cases that depend on the direction of the report (i.e. a UL or Downlink (DL) data packet) and if there is a report Acknowledgment (ACK) or not. These four transfer cases allow us to compare the impact different traffic profiles have on the UE performance and the radio resources consumed.

The research carried out aims at answer the following questions:

i) What is the energy reduction achieved by IoT devices using the new small data transmission optimizations.

ii) Evaluate the impact of different Inter-Arrival Times (IATs) (i.e. amount of time that elapses between data packet transfers) resulting from different traffic characteristics of the IoT applications. Additionally, based on the IAT the IoT device may have, we want to study which IoT devices may benefit from using the new small data transmission optimizations.

In order to do that, our analysis follows three main steps:

i) Estimation of the radio resource consumption by the UE while performing a data transmission procedure;

ii) A Markov chain modeling the behavior of a UE and its stationary probabilities; and

iii) The estimates of the average energy consumption required to complete each state of the Markov chain.

The rest of the chapter is organized as follows. Section 3.1 briefly reviews the related literature. Section 3.2 reviews NB-IoT and its features. Section 3.3 provides the fundamental background of Markov chain. Section 3.4 explains the system model and the main assumptions for the evaluation. Section 3.5 presents the proposed analytical model for NB-IoT. Section 3.6 shows the results. Lastly, section 3.7 presents the main conclusions of this chapter.

3.1 Related works

One of the most important requirements for the NB-IoT system is to reduce the energy consumption in the UE compared to conventional cellular networks.

NB-IoT achieves this requirement by means of a set of innovations. Particularly, it includes a new radio interface design, a UE functionality simplification, new signaling reduction optimizations, and power saving features.

Previously to NB-IoT, there is an extensive research on Long Term Evolution (LTE) power consumption. In [86], *Wang et.al.* provide a Markov chain analysis of Discontinuous reception (DRX) impact in terms of power consumption, signal, and delay. This work joins the analysis of DRX and traffic schedulers considering IoT traffic.

Madueño et.al. in [85] evaluate and identify the limitations of the LTE connection establishment. The analytical model, based on a Markov chain, considers the limitation of the different LTE radio channels jointly. From this radio access capacity analysis, we extend the study in [87] to estimate the energy consumption for three different control procedures (i.e. SR, CP, and UP). In that work, we present a Markov chain model to estimate the average energy consumption per packet. The results show CP outperforms the other control procedures in terms of battery lifetime for almost all configurations assumed.

Related to NB-IoT literature, in [88], *Feltrin et.al.* discuss the main sources of latency and present an evaluation of the resource occupation for NB-IoT under different IoT use cases. Their IoT use cases are based on realistic cases considering a set of payload sizes and UE reporting periodicity. For these use cases, the analysis details the resources consumed in each NB-IoT radio channel. They conclude the UL is always more loaded with respect to the DL.

Jörke et.al. in [89] analyze NB-IoT and LTE-M considering three coverage conditions. The analysis is based on a five power consumption state machine and real-life power consumption parameters. The evaluation compares data rate, battery lifetime, latency, and spectral efficiency for both Cellular IoT (CIoT) technologies. Their analysis shows LTE-M outperforms NB-IoT for the assumed normal and robust coverage conditions. For extreme coverage, NB-IoT shows a better performance than LTE-M.

Moreover, [83] presents the 3GPP energy consumption estimates for NB-IoT and other CIoT technologies. This evaluation shows the latency and battery lifetime based on four power consumption levels and traffic profile assumptions. In [90], *Ratasuk et.al.* discuss the design targets of NB-IoT and present a preliminary system design. Their system-level simulation results show the targets can be

achieved under the deployment scenarios evaluated.

Even though these works are a significant advance in the analysis of NB-IoT, there are still unsolved questions concerning the performance of NB-IoT UEs. Additionally, in most cases these works provide final results for specific configurations that hinder the comparison of the results. In this chapter, the proposed analytical NB-IoT evaluation aims to provide a tractable methodology to study and compare different scenarios. The proposed model gives a detailed description of the assumed UE behavior evaluated and can be extended to consider more features of the standard or more complex scenarios.

3.2 NB-IoT overview

NB-IoT focuses on low-cost Machine-Type Communication (MTC) UEs with lower power and higher coverage requirements compared to conventional enhanced Mobile Broadband (eMBB) UEs. NB-IoT meets these demands by means of the use of a small portion of the existing available spectrum, a new radio interface design, and simplified LTE network functions.

3.2.1 Radio design and resource allocation

The new NB-IoT radio interface design is derived from the legacy LTE. The NB-IoT carrier has a 180 kHz bandwidth with support for multi-carrier operation. If the network is operating in multi-carrier mode, the NB-IoT carrier that allows a UE to perform the initial connection setup is referred to as an anchor carrier, and the other carriers are called non-anchor carriers.

In the DL, Orthogonal Frequency-Division Multiple Access (OFDMA) is applied using a 15 kHz subcarrier spacing over 12 subcarriers. This is just the size of one Physical Resource Block (PRB) in LTE standard. In the UL, Single-Carrier Frequency-Division Multiple Access (SC-FDMA) is applied, using either 3.75 kHz or 15 kHz subcarrier spacing [91]. For both the DL and UL, there are 7 OFDMA symbols within a slot. An Subframe (SF) consists of two slots, and one radio frame is made up of 10 SFs. Within this resource grid, one subcarrier in one OFDMA symbol is denoted as Resource Element (RE). An RE carries a complex value with values according to the modulation scheme. NB-IoT supports both single-tone and multi-tone (or subcarrier) operations in UL. Particularly for

multi-tone uplink transmission (12, 6 or 3 tones), only 15 kHz subcarrier spacing is allowed [13, 92].

NB-IoT physical channels and signals are primarily multiplexed in time. Only half-duplex operation is supported. Therefore, the UE is not required to listen to the DL while transmitting in the UL, and vice versa. In NB-IoT, the physical channels defined are:

- Narrowband Physical Broadcast CHannel (NPBCH): master information for system access, i.e., Master Information Block (MIB).
- Narrowband Physical Downlink Control CHannel (NPDCCH): uplink and downlink scheduling information.
- Narrowband Physical Downlink Shared CHannel (NPDSCH): downlink dedicated and common data.
- Narrowband Physical Random Access CHannel (NPRACH): random access.
- Narrowband Physical Uplink Shared CHannel (NPUSCH): uplink data. This channel has two formats. NPUSCH format 1 for UL data transmissions and NPUSCH format 2 for Hybrid Automatic Repeat Request (HARQ) feedback for NPDSCH.

Figure 3.1 shows an example of NB-IoT SFs design. In addition to the physical channels, NB-IoT uses three physical signals:

- Narrowband Reference Signal (NRS): It is used as reference signal strength for the DL and to estimate the DL propagation channel coefficients.
- Primary Synchronization Signal (NPSS) and Secondary Synchronization Signal (NSSS): These signals allow a UE to synchronize to an NB-IoT cell. Both signals are included in specific DL SFs.
- Demodulation Reference Signal (DMRS): It is used at the evolved NodeB (eNB) to allow UL channel estimation. This signal is always multiplexed with NPUSCH data.

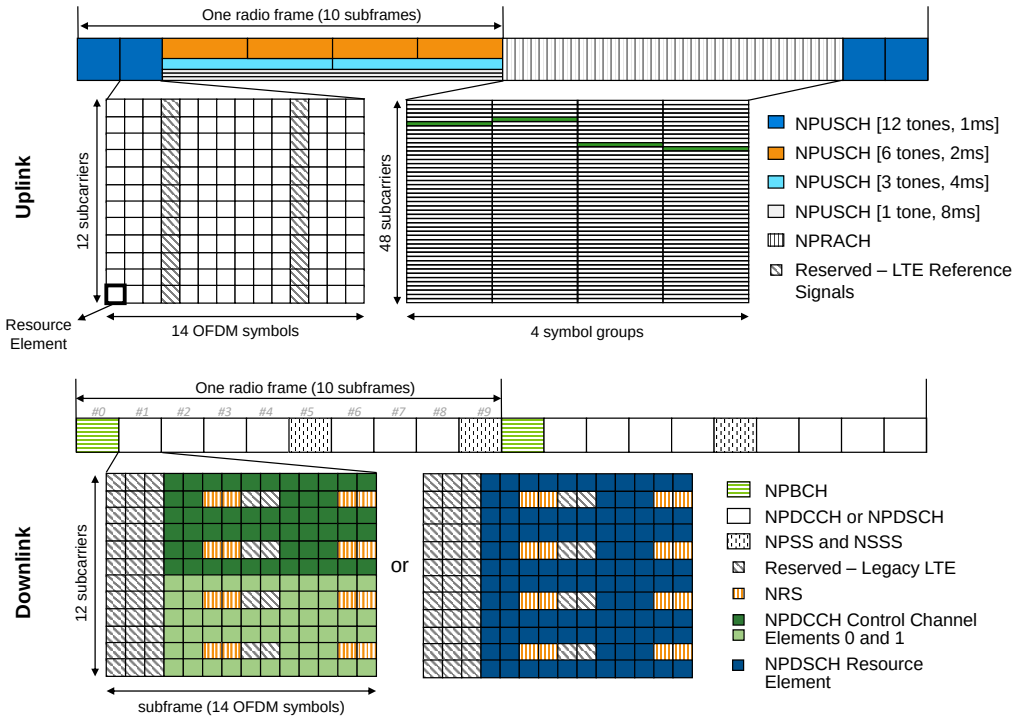


Figure 3.1: NB-IoT in-band physical channels time multiplexing [11].

Furthermore, to provide deployment flexibility there are three NB-IoT operation modes:

- Standalone: utilizing, for example, one or more Global System for Mobile Communications (GSM) carriers.
- Guard-band: utilizing the unused resource blocks within an LTE carrier's guard-band.
- In-band: utilizing resource blocks within an LTE carrier.

As seen in the Figure 3.1, an NPDCCH SF is divided into two Narrowband Control Channel Elements (NCCEs). The number of REs available for the NCCEs and NPDSCH SFs depends on the NB-IoT deployment mode and the number of logical antenna ports (i.e. NRS REs).

3.2.1.1 Coverage extension

The target coverage extension of NB-IoT is achieved by means of reducing the data rate, e.g., lowering the transmission bandwidth, or using repetitions in time.

On the one hand, bandwidth reduction concentrates the limited UE power on a narrower bandwidth. Therefore, it boosts the UL Power Spectral Density (PSD) if the transmission power used is maintained. On the other hand, the successive repetitions can be incrementally soft combined at the receiver before decoding to raise error correction.

In the UL, the transmission can be repeated {1, 2, 4, 8, 16, 32, 64, 128} times, using the same transmission power on each repetition. In the DL, the possible number of repetitions are {1, 2, 4, 8, 16, 32, 64, 128, 192, 256, 384, 512, 768, 1024, 1536, 2048}. The eNB chooses their values based on the signal strength received and reported by the UE.

Specifically for the NPUSCH format 1 (used for UL data transmission), the repetitions have two possible Redundancy Version (RV). An RV specifies a starting point for the extraction of coded bits from the circular buffer with the original input bits and parity bits (for more information see [30]). Then, the exact set of bits extracted to be mapped in the resources assigned for the transmission depends on the RV.

NB-IoT covers an extensive range of radio conditions. To support UEs with different coverage, the network can configure up to 3 Coverage Enhancement Level (ECL). If a network has the 3 ECLs configured, all UEs camping on the NB-IoT cell are separated by their specific ECL depending on their coverage. To group the UEs the network applies two thresholds in terms of Reference Signal Received Power (RSRP). Figure 3.2 shows an example of the distribution of the ECLs in NB-IoT.

The RSRP is a DL reference signal measurement and is defined as the linear average of reference signal power within the considered measurement bandwidth. Therefore, when the UE reads the network broadcast information, it selects an ECL based on its measured RSRP. The ECL selected determines the resources of the NPRACH used to perform the Random Access (RA) procedure (e.g. NPRACH periodicity, number of subcarriers per ECL, number of NPDCCH repetitions during the RA) and other common parameters for all UEs in the cell.

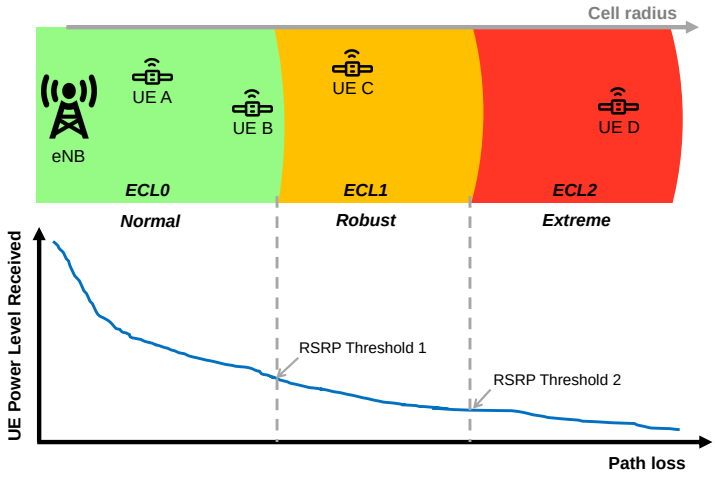


Figure 3.2: ECLs configuration in NB-IoT.

3.2.1.2 Data transmission in NB-IoT

As previously mentioned, the physical channels are primarily multiplexed in time. In the UL, the resources are distributed between the NPRACH and the NPUSCH. The NPRACH resources consist in the assignment of time and frequency resources and occur periodically [92]. These NPRACH resources are provided to each ECL separately. The remain resources not used for NPRACH are available for NPUSCH transmissions. The UE knows the resources reserved to the NPRACH because they are signaled in one of the System Information Blocks (SIBs).

In the DL, the DL physical channels and signals are time-multiplexed. The NPBCH and NPSS are transmitted in SFs 0 and 5 in every frame, respectively, and the NSSS in SF 9 in every two frames [13]. Additionally, the transmission of SIBs will occupy some NPDSCH SFs. For example, SIB1 is transmitted in SF 4 of every other frame in 16 continuous frames. Thus, approximately only 14 out of 20 SFs are available for NPDCCH or NPDSCH. However, from these remain resources, other SFs may be declared as invalid SFs, and thus, the UE will skip monitoring them.

The smallest radio resource allocated to a UE in the DL is the PRB. In the UL, the smallest unit is the Resource Unit (RU). An RU is a new unit for NPUSCH resource allocation and has several configurations. The definition of the

RU depends on the subcarrier spacing and the number of subcarriers allocated in the UL transmission. Table 3.1 shows the possible RU configurations according to the 3GPP specification [29].

Furthermore, the specification [21] provides the allowed configurations of NPUSCH Transport Block Size (TBS) as a function of the number of RUs and Modulation and Coding Scheme (MCS) level. The MCS level describes the modulation order and coding rate that is applied in a channel. Table 3.2 shows the NPUSCH TBS table for multi-tone. Note for single tone configurations the second and third rows shown (i.e. current I_{MCS} 1 and 2 in Table 3.2) are exchanged. The higher MCS value indicates higher data rate and more bits per symbol. For multi-tone configurations, only QPSK modulation is used. For single-tone configurations, the phase rotated $\pi/2$-BPSK or $\pi/4$-QPSK modulations can be used. For the NPDSCH, its TBS table is similar having the same range of MCS levels and number of SFs, except for some TBS values that have a different value.

3.2.2 Scheduling and HARQ operation

When the eNB needs to schedule radio resources to a UE, the eNB signals the information in the NPDCCH through a Downlink Control Information (DCI). A DCI can be transmitted using Aggregation Level (AL) 1 or 2. With AL-1, the DCI is mapped to one NCCE, i.e. two DCIs are multiplexed in one SF. Otherwise, AL-2 is used and the DCI is mapped to both NCCEs, i.e., one SF only carries one DCI. Each DCI includes: i) Time and frequency resource allocation (e.g. number of RU or SFs, number of tones in the UL, scheduling delay, etc); ii) MCS; and iii)

Table 3.1: NB-IoT RU configurations.

NPUSCH format	**Subcarrier spacing (kHz)**	**Number of subcarriers**	**Number of slots**	**RU duration (ms)**
1	3.75	1	16	32
	15	1	16	8
		3	8	4
		6	4	2
		12	2	1
2	3.75	1	4	8
	15	1	4	2

Table 3.2: NPUSCH TBS table for multi-tone [21]

MCS Index (I_{MCS})	**Number of RUs** (N_{RU})							
	1	**2**	**3**	**4**	**5**	**6**	**8**	**10**
0	16	32	56	88	120	152	208	256
1	24	56	88	144	176	208	256	344
2	32	72	144	176	208	256	328	424
3	40	104	176	208	256	328	440	568
4	56	120	208	256	328	408	552	680
5	72	144	224	328	424	504	680	872
6	88	176	256	392	504	600	808	1000
7	104	224	328	472	584	712	1000	1224
8	120	256	392	536	680	808	1096	1384
9	136	296	456	616	776	936	1256	1544
10	144	328	504	680	872	1000	1384	1736
11	176	376	584	776	1000	1192	1608	2024
12	208	440	680	1000	1128	1352	1800	2280
13	224	488	744	1032	1256	1544	2024	2536

Information to support the HARQ operation. There are three formats of DCI: i) N0, used for UL grant; ii) N1, used for DL scheduling; and iii) N2, used for paging.

To provide the UE with an energy efficient mechanism to find its DCIs, the NPDCCH is grouped into search spaces. A search space consist of one or more SFs and there are three types: i) Type-1 Common Search Space (CSS), used for monitoring paging; ii) Type-2 CSS for monitoring RA process; and iii) UE-specific Search Space (USS), used for monitoring DL or UL UE's specific scheduling information. A set of parameters define the NPDCCH periodicity for each search space and the resources occupied [21]:

- R_{max}: Maximum number of repetitions of NPDCCH.
- G: Time offset in a search space period.
- α_{offset}: Offset of the starting SF in a search period.
- T: The search space period calculated as $T = R_{max} \cdot G$ in SF units.

In NB-IoT, the interval between the start of two NPDCCH is referred to as

the PDCCH period (pp), thus $pp = T$. This pp depends on the currently used NPDCCH search space.

As previously mentioned, only half-duplex operation is supported at the UE. This is meant to allow a low-complexity UE implementation, as the standard allows time for the device to switch between transmission and reception modes [13]. Furthermore, there are more scheduling principles to reduce the complexity of the device:

- NB-IoT allows only one HARQ process in both the UL and the DL. Asynchronous, adaptive HARQ process is adopted to support scheduling flexibility.
- There is no simultaneous transmissions of UL and DL HARQ processes.
- Longer UE processing time for both NPDCCH (i.e. DCI) and scheduled data transmission/reception.

Figure 3.3 shows an NB-IoT scheduling example for both a UL transmission and a DL reception. This figure shows the timing relationship operation for the data scheduling and HARQ operation. For example, for the UL transmission, the DCI specifies the resources of the UL scheduling grant. Next, the UE has a time gap to change from reception mode to transmission mode. After the time gap, the UE transmits its data. When the UE completes the NPUSCH transmission, there is another time gap to switch to reception mode and start monitonig the NPDCCH to confirm if the NPUSCH transmission has been received correctly by the eNB.

3.2.3 Power control

NB-IoT supports open loop power control in the UL. In open loop there is no feedback from the eNB and the UE determines the transmit power according to a set of rules. The UE uses its maximum transmit power P_{max} if: i) it is a (re)transmission of the RA response UL grant; ii) enhanced NPRACH power control is not applied (i.e. the UE is in ECLs 1 or 2); or iii) NPUSCH repetitions are greater than 2. Otherwise, the transmit power is determined by [13, 21]:

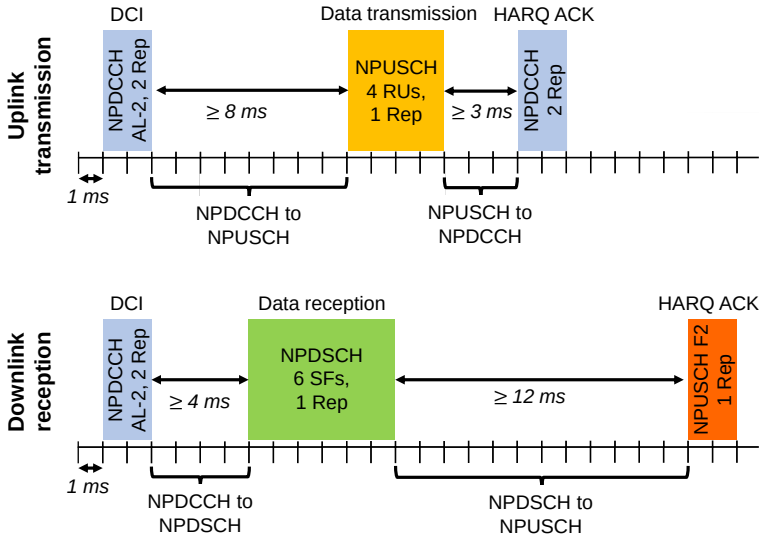


Figure 3.3: Timing relationship operation for a UL transmission and a DL reception [12].

$$P_{NPUSCH} = max\left\{P_{max}, 10log_{10}(M) + P_{target} + \alpha PL\right\} \quad [dBm] \tag{3.1}$$

where P_{target} is the target received power level at the base station [dBm], α is the path loss adjustment factor, PL is the estimated path loss [dB], and M is a parameter that depends on the bandwidth allocation for NPUSCH.

The configuration of the parameters depends on the NPUSCH format, and higher-layer configuration signaling. Particularly, the values of M are $\{12, 6, 3, 1, 1/4\}$ for 12 tones, 6 tones, 3 tones, 1 tone and 15 kHz subcarrier spacing, and 1 tone and 3.75 kHz subcarrier spacing, respectively.

For the NPRACH, the power control is similar as the one used in NPUSCH. The UE uses its maximum transmit power P_{max} if the NPRACH preamble repetitions do not have the lowest repetition level. Otherwise, the transmit power is determined by:

$$P_{NPRACH} = max\left\{P_{max}, P_{target} + \alpha PL\right\} \quad [dBm] \tag{3.2}$$

where P_{target} is the target NPRACH received power level indicated by the higher layers.

3.2.4 Power saving features

NB-IoT reuses LTE power saving mechanisms extending the timers involved to achieve longer battery lifetime. There are two key power saving mechanisms: extended/enhanced Discontinuous Reception (eDRX) and Power Saving Mode (PSM). Both mechanisms enable the UE to enter a power saving state where it is not required to monitor for paging/scheduling information.

DRX defines a cycle where the UE monitors the DL signaling during a short period of time and sleeps the remaining time of the cycle. DRX can be used while the UE has an active Radio Resource Control (RRC) connection with the network (RRC Connected state), named as Connected-mode DRX (C-DRX), or when there is no RRC Connection (RRC Idle state), named as Idle-mode DRX (I-DRX). The transition from the RRC Connected to the RRC Idle state happens at the expiration of the RRC Inactivity Timer. This timer is controlled by the eNB.

A UE in RRC Idle state can be still reachable by the network until the expiration of the Active Timer (T3324). During this period the UE can use either I-DRX or eDRX. I-DRX defines the normal paging cycle the UE uses to monitor the NPDCCH for paging messages while it is in RRC Idle state. If the UE uses I-DRX, it will have continuous I-DRX cycles until the expiration of the Active Timer. Alternatively, eDRX is a mechanism that can extend the sleeping period of the I-DRX. This is because when using eDRX, at each eDRX cycle there is an active phase controlled by a Paging Time Window (PTW) timer where the UE is reachable by means of I-DRX cycles, followed by a sleep phase for the remaining period of the eDRX cycle. In this case, the eDRX cycles will occur until the expiration of the Active Timer.

Later, at the expiration of the Active Timer, the UE moves to PSM mode. The PSM mode disconnects the radio completely and only keeps a basic oscillator running to maintain a time reference. In PSM, the energy consumption is similar to the power-off state. The UE is not reachable, but it is still registered with the network. A UE using PSM remains in deep sleep until a mobile originated

transaction requires initiating a communication with the network. One example is the periodic Tracking Area Update (TAU) procedure (triggered by the expiration of the T3412) or a UL data transmission. Figure 3.4 shows the operation of a UE in RRC Idle state using eDRX and PSM.

3.2.5 Other features

3.2.5.1 Transmission gaps

NB-IoT allows a large set of repetitions to extend coverage. Consequently, the technology also includes transmission gaps. These gaps are breaks during which no transmission and reception happen. In the DL, the transmission gaps are used to avoid blocking DL resources. In the UL, as a different number of subcarriers can be allocated, this enables simultaneous transmission from several UEs. Then, the blocking of UL resources is not the main reason for UL transmission gaps. These UL transmission gaps are used to allow the UE to resynchronize with the network [13]. Figure 3.5 shows an example of the different types of gaps present in NB-IoT.

The UL gap is defined by a periodicity $T^{UL}_{GapPeriod}$ and a gap length T^{UL}_{GapDur}. Hence, if the duration of the UL transmission is greater or equal than $T^{UL}_{GapPeriod}$, the UE applies gaps of T^{UL}_{GapDur} with a periodicity $T^{UL}_{GapPeriod}$ until the transmission is finished. For the DL gap, there are gaps if $R_{max} \geq N^{DL}_{GapThr}$ where N^{DL}_{GapThr} denotes the threshold on the maximum number of repetitions. Like UL, the DL gaps are defined by a periodicity $T^{DL}_{GapPeriod}$ and duration T^{DL}_{GapDur}.

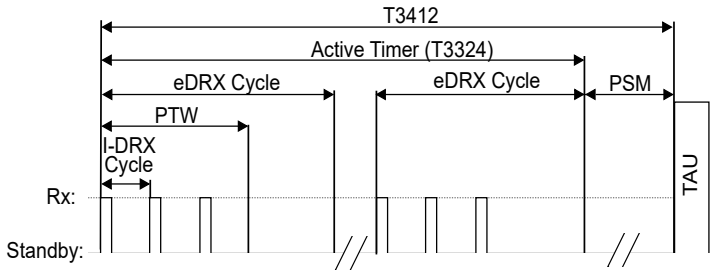


Figure 3.4: Example of the eDRX and PSM behavior.

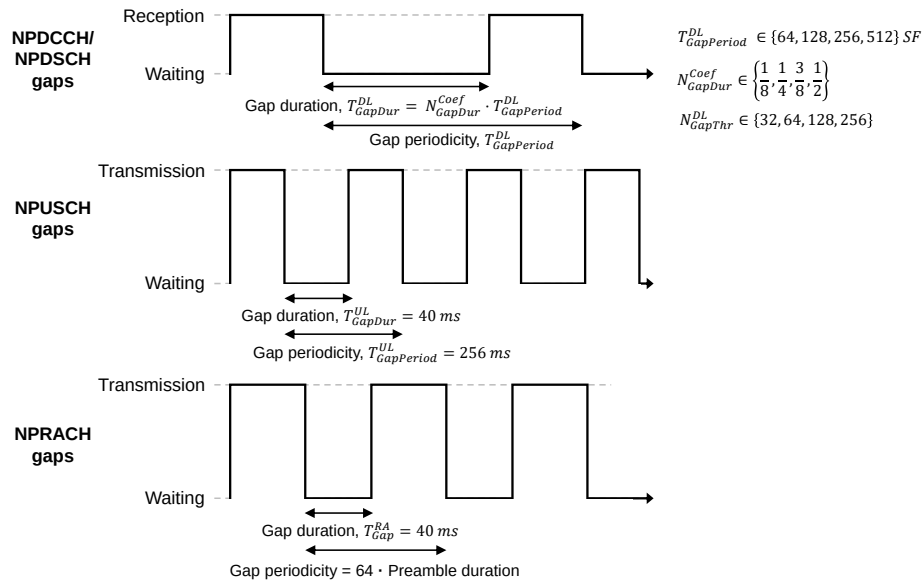


Figure 3.5: Example of NB-IoT gaps.

3.2.5.2 Signaling reduction

The support of CP is mandatory in NB-IoT (see subsection 2.1.5 for a detailed CP description). While the UE is using the CP procedure to communicate with the eNB, the Non-Access Stratum (NAS) signaling message encapsulating the data packet can include a Release Assistance Indication (RAI) information element. This RAI allows the UE to notify the core network if no further UL or DL data transmissions are expected, or only a single DL data transmission subsequent to the current UL data transmission is expected. Therefore, the core can immediately trigger the S1 Release (S1) procedure if there are no user plane bearers established between the eNB and the core [9].

Additionally, in the 3GPP Release 15 a new data transmission mechanism is specified for NB-IoT and LTE-M. This mechanism, named Early Data Transmission (EDT), enables the data transmission during the RA procedure. EDT is specified for both UP and CP procedures. To use EDT, the UE selects a special NPRACH preamble to indicate to the eNB that it wants to send data in Msg3 of the RA procedure. The special preambles and the maximum allowed TBS for EDT are indicated by the eNB in the SIB. EDT enables earlier data transmission,

thus it reduces the total signaling exchange for small data transmissions [93].

3.2.6 NB-IoT enhancements from Release 13

Since the introduction the NB-IoT in Release 13, several enhancements have been introduced and studied to improve the functionalities of NB-IoT. The standardization of Release 14 finished in June 2017 and its main feature enhancements are [94, 95]:

- Support for positioning: inclusion of Observed Time Difference of Arrival (OTDOA) and Cell-ID (CID)/E-CID. OTDOA positioning is based on a UE measuring the Time of Arrival (ToA) on a set of the DL narrowband positioning reference signals transmitted from a set of time-synchronized eNB surrounding the UE, and then the UE reports the reference signal time difference to a positioning server [13]. The E-CID uses serving cell identity and the round-trip time between the UE and the eNB.
- Support for multicast: to do that, the NB-IoT channels are reused to use Single-Cell Point-to-Multipoint (SC-PTM) (earlier standardized in Rel-13 for LTE). This feature is crucial for firmware upgrade to a group of UEs.
- Non-Anchor PRB enhancements: These enhancements allow the support of paging and the RA procedure on NB-IoT non-anchor carriers.
- Higher peak data rates: increasing the maximum TBS and an optional support for 2 HARQ processes.
- New power class with the maximum output power reduced to 14 dBm.
- Mobility and service continuity enhancements.

Additionally, in Release 15, expected to end in June 2019, the NB-IoT evolution continues. The main features enhancements are [93, 94, 96, 97]:

- Further latency and power consumption reduction: inclusion of techniques such as EDT, quick release of RRC connection after the last data transmission, wake-up signal, physical layer scheduling request, etc.

- NPRACH reliability and range enhancements: reducing the false alarm probability for NPRACH detection due to inter-cell interference on NPRACH and introducing additional cyclic prefixes for NPRACH to support cell radius of at least 100 km.
- Narrowband measurement accuracy improvements by using additional existing signals.
- Specified NB-IoT small-cell: support for NB-IoT in microcell, picocell, and femtocell deployments.
- Reduced system acquisition time: improving cell search and system information acquisition performance.
- UE differentiation: consideration of UE-specific information for UE Information Transfer procedure.
- TDD support.
- Access barring enhancement.
- Enhancements to standalone operation mode: support pairing of standalone carrier with in-band/guard-band carrier.
- Support of extended NB-IoT power headroom report range and finer granularity.

3.3 Fundamentals of Markov chains

A Markov chain is a stochastic process, $X = \{X_n, n \in N\}$, that satisfies the Markov property. This property means the future behavior of the system depends only on the current state and not on any of the previous states. The possible values of X_n, denoted as S, are referred to as the state space of the process. The Markov chains are used to calculate the probabilities of the occurrence of the events by viewing them as states transitioning into other states, or the same state as before. Each transition is called a step. If the chain is currently in state S_i, then it moves to state S_j at the next step with a probability denoted by p_{ij}.

The probabilities p_{ij} are called transition probabilities [98]. There are different types of states within the Markov chains:

- Absorbing state: A state is absorbing if it is impossible to leave it (i.e. $p_{ii} = 1$).
- Recurrent state: A state is recurrent if the system will return to it after leaving sometime in the future.
- Transient state: If a state is not recurrent, it is transient.
- Periodic state: A state is periodic if it can only return to itself after a fixed number of transitions greater than 1.
- Aperiodic state: If a state is not periodic, it is aperiodic.

A Markov chain is called an ergodic chain (or irreducible) if it is possible to reach every state from every state. For a Markov chain where its entire state space forms an irreducible recurrent set, the steady-state probability distribution, denoted as π, exists and is independent of the initial state, then:

$$\pi(j) = \lim_{n\to\infty} Pr\left\{X_n = j | X_0 = i\right\} \tag{3.3}$$

The vector π is the solution to the following system:

$$\begin{aligned} \boldsymbol{\pi}\boldsymbol{P} &= \boldsymbol{\pi} \\ \sum_{i\in E} \pi(i) &= 1 \end{aligned} \tag{3.4}$$

where E denotes the finite state space and $\boldsymbol{P}$ the transition matrix between the states of the Markov chain.

3.4 System model

Let us assume a cell with an eNB with an NB-IoT carrier deployed in-band, and N_{UE} UEs camping on it. Each UE transfers/receives one report of size L periodically to/from the eNB. We analyze four transmission cases:

- `Uplink (UL)`: The NB-IoT UE sends a report to the IoT application server.
- `Uplink with an acknowledgment (UL-ACK)`: The same as the `UL` case, but the server replies with a downlink acknowledgment packet as a confirmation.
- `Downlink (DL)`: The NB-IoT UE receives an application layer command from the IoT application server.
- `Downlink with an acknowledgment (DL-ACK)`: The same as the `DL` case, but the UE replies with an uplink report.

To send/receive these periodic reports, the UE can perform three data transmission procedures: SR, UP, and CP (see Figure 2.4). Later, if the eNB detects an inactivity period greater than the defined RRC Inactivity Timer, $T_{inactivity}$, the eNB initiates the S1 procedure to switch the UE to RRC-Idle. To save battery, after a period of discontinuous NPDCCH monitoring, defined by the Active timer T_{active} (see Figure 3.4), the UE moves to PSM.

Particularly for the CP, we assume the UE includes the RAI field to notify if there is no further data UL or DL data transmissions are expected. For the cases `UL`, `UL-ACK`, and `DL-ACK` where the UE can provide this information, the network knows there is no more traffic for the UE and it can enter directly PSM to save more battery. Then, in these three cases for CP procedure, we assume after a short waiting time, the UE enters PSM.

Additionally, for `UL` and `UL-ACK` cases, we assume the UE performs a periodic TAU procedure with a period of five days. For `DL` and `DL-ACK` cases, the periodic TAU is configured with the same frequency as the downlink traffic. This means when the UE exits PSM to perform the periodic TAU procedure, the network uses the communication with the UE to notify there is pending DL traffic (when the TAU Accept message is received at the UE). Thus, the subsequent signaling sequence is the same as for a UE-triggered data transmission procedure.

3.4.1 Resource allocation

In this analysis, we consider three different ECLs. For each ECL we define a specific ECL dependent configuration that fixes how the radio channels are configured

in terms of repetitions, modulation, the MCS, and the number of subcarriers and subcarrier spacing for the UL. Consequently, to select the number of resources (i.e. SFs or RUs) required to send/receive a packet, we extract the row of the MCS corresponding to the current ECL evaluated from the TBS table. For the extracted row, there are a few possible configurations depending on the number of resources used. The selected configuration is the one that needs the least total resources. This is estimated as follows:

$$
\begin{aligned}
N_{seg} &= \left\lceil \frac{L_{packet}}{TBS(MCS_{ECL}, \mathbf{N_{resources}}) - H_{RLCMAC}} \right\rceil \\
N_{total} &= min\left(N_{seg} \cdot \mathbf{N_{resources}}\right)
\end{aligned}
\tag{3.5}
$$

where N_{seg} is the number of resulting segments done to fit in the selected TBS configuration, L_{packet} is the size of the packet in bits, $\mathbf{N_{resources}}$ is a vector containing the possible values of the number of resources, MCS_{ECL} is the MCL lecel corresponding to the current ECL, $TBS(MCS_{ECL}, \mathbf{N_{resources}})$ is the extracted row from the TBS table, H_{RLCMAC} is the size of the Radio Link Control (RLC) and Medium Access Control (MAC) headers, and N_{total} is the total number of resources (i.e. SFs or RUs). Figure 3.6 illustrate an example of the methodology explained previously for a UL packet.

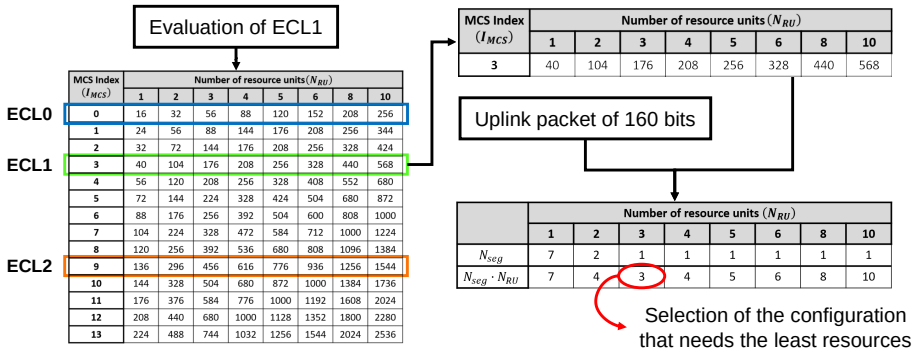


Figure 3.6: Example of the resource estimation for an uplink packet of 160 bits and ECL1.

3.4.2 UE power model

To model the energy consumption of the UE, we assume its behavior can be described as shown in Figure 3.7. The model defines four UE power levels:

- Transmission (P_{TX}): The UE is active transmitting a packet to the network, i.e., the TX branch of the UE is on. To obtain the power used by the UE when transmitting, we use the 3GPP's power control equations of the different UL physical channels [21]. Hence, P_{TX}^{RA} denotes the transmission power for NPRACH, and P_{TX} for the NPUSCH.
- Reception (P_{RX}): The UE is active receiving information from the network, the RX branch of the UE is on.
- Inactive (P_i): The UE is not transmitting or receiving, thus it is inactive. The accurate clock is ON to maintain the synchronization in the air interface.
- Standby (P_s): The UE is in deep sleep low power operation.

3.5 Analytical model for energy consumption

This section presents the main considerations underlining in the energy consumption model. The analysis has three main steps: i) Estimation of the radio resource

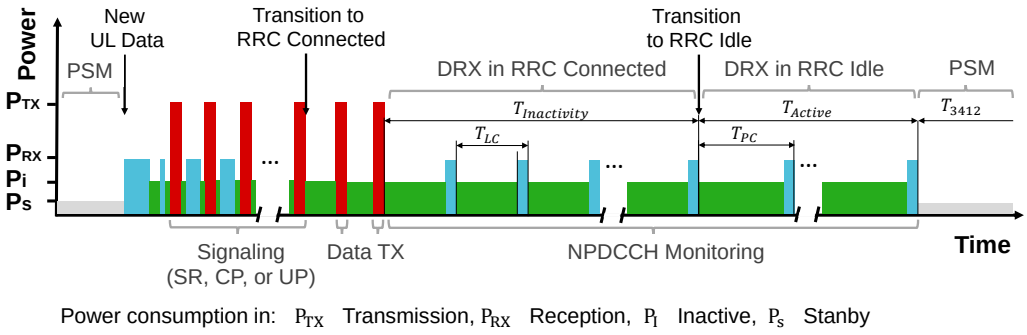


Figure 3.7: Example of power consumption transitions during a UE's connection.

consumption by the UE while performing a data transmission procedure; ii) A Markov chain modeling the behavior of a UE and its stationary probabilities; and iii) The estimates of the average energy consumption required to complete each state of the Markov chain.

3.5.1 Radio resources analysis

In this first step of the evaluation we derive analytical expressions to analyze the capacity of the radio resources in NB-IoT. To do that, we first analyze the amount of load the model will have considering there will be retries from UEs with unsuccessful connections. Next, we derive the probability of connection failure while establishing the RRC connection due to the scarcity of radio resources. These probabilities rely on the assumption each radio channel can be modeled as a separate queue with impatient customers. Then, to obtain these probabilities we need to estimate the arrival rate, the service rate, and the maximum waiting time in the queue. Later, these probabilities will be part of the transition probabilities of our Markov chain to model the UE behavior.

We assume the packet generation process of each UE follows a Poisson model with rate λ_{app} packets per ms. The data rate of the UE is derived from its average IAT in ms, therefore $\lambda_{app} = \frac{1}{IAT}$. For each new data packet, up to m retries of the connection establishment are allowed. If a connection establishment fails, an additional load is added on the RA procedure (i.e. the beginning of the connection establishment). To represent the different loads the model have, we distinguish different rates in packets/s as follows:

- λ_A denotes the mean number of activated preambles in the contention phase per RA opportunity.
- λ_S denotes the mean number of successful preambles activations per RA opportunity (i.e. non-collided preambles). Then $\lambda_S < \lambda_A$ since in case of a preamble collision only one preamble is activated.
- λ_T denotes the total rate (including retransmissions).

For a successful connection establishment, the UE needs a successful RA procedure and there must to be sufficient resources in the radio channels (i.e.

NPRACH, NPUSCH, NPDCCH, and NPDSCH) to exchange the required signaling messages. We start by computing the probabilities of a successful RA. Let d denote the number of available preambles. Then, the probability of no selecting the same preamble as other UE is $1-\frac{1}{d}$. Assuming Poisson arrivals with rate λ_T, the probability of collision at the preamble contention can be derived as:

$$p_c\left(\lambda_T\right)=\sum_{i=1}^{+\infty}\left[1-\left(1-\frac{1}{d}\right)^{i-1}\cdot\mathbb{P}\left(N_T=i,\lambda_T\cdot T_{RAO}\right)\right] \tag{3.6}$$

where $\mathbb{P}\left(N_T=i,\lambda_T\cdot T_{RAO}\right)$ is the probability mass function of the Poisson distribution with arrival rate $\lambda_T\cdot T_{RAO}$, N_T denotes the contending UEs, and T_{RAO} is the interval between RA opportunities in ms. Applying Jensen's inequality to the concave function $1-\left(1-\frac{1}{d}\right)^x$, we obtain a upper bound on the preamble collision probability:

$$p_c\left(\lambda_T\right)=1-\left(1-\frac{1}{d}\right)^{\lambda_T\cdot T_{RAO}-1} \tag{3.7}$$

From these non-collided preambles, we assume the eNB is unable to discern between preambles activated by a single UE or multiple UEs. Therefore, the λ_A and λ_S can be approximated as follows:

$$\begin{aligned}\lambda_A&=\left[1-\mathbb{P}\left(X=0\right)\right]\cdot\frac{d}{T_{RAO}}\\ \lambda_S&=\mathbb{P}\left(X=1\right)\cdot\frac{d}{T_{RAO}}\end{aligned} \tag{3.8}$$

where $\mathbb{P}\left(X=k\right)$ denotes the probability of k successes. For the assumed Poisson distribution, this probability can be approximated as:

$$\mathbb{P}\left(X=k\right)\approx\frac{\left(\lambda_T\frac{T_{RAO}}{d}\right)^k e^{-\left(\lambda_T\frac{T_{RAO}}{d}\right)}}{k!} \tag{3.9}$$

In case of no collision, we assume a probability of preamble detection p_d to take into account the effects present in the radio channels (e.g. path loss, fading, interference, etc). Thus, the preamble detection probability equals $p_d(i)=1-\frac{1}{e^i}$, where i indicates the ith RA attempt [99].

Even when the preamble is correctly detected at the eNB, the establishment

of the connection could fail if there are no enough radio resources. To model this limitation, we use the same methodology as [85] adapted for NB-IoT. Thus, let $p_{fr}(\lambda_T)$ and $p_{fc}(\lambda_T)$ denote the probability of failure due to starvation of resources in the radio channels (i.e. NPUSCH, NPDCCH, and NPDSCH) for two different stages of the communication. These two stages are two possible points of failure when the UE establishes the RRC connection. We consider the failures happen if the Random Access Response (RAR) or the RRC Connection Setup messages are not received. Specifically, $p_{fr}(\lambda_T)$ considers the resources needed until the transmission of the RRC Connection Request message. The probability $p_{fc}(\lambda_T)$ comprises the resources from the reception of the RRC Connection Setup message until all data transmissions/receptions end (see Figure 2.4 for an example of the signaling messages exchanged in `UL` case). Therefore, both $p_{fr}(\lambda_T)$ and $p_{fc}(\lambda_T)$ can be defined as follows:

$$\begin{aligned} p_{fr}(\lambda_T) = 1 - \big(1 - p_{DC_{fr}}(\lambda_{NPDCCH}, \mu_{NPDCCH}, T_{NPDCCH})\big) \\ \cdot \big(1 - p_{DS_{fr}}(\lambda_{NPDSCH}, \mu_{NPDSCH}, T_{NPDSCH})\big) \\ \cdot \big(1 - p_{US_{fr}}(\lambda_{NPUSCH}, \mu_{NPUSCH}, T_{NPUSCH})\big) \\ p_{fc}(\lambda_T) = 1 - \big(1 - p_{DC_{fc}}(\lambda_{NPDCCH}, \mu_{NPDCCH}, T_{NPDCCH})\big) \\ \cdot \big(1 - p_{DS_{fc}}(\lambda_{NPDSCH}, \mu_{NPDSCH}, T_{NPDSCH})\big) \\ \cdot \big(1 - p_{US_{fc}}(\lambda_{NPUSCH}, \mu_{NPUSCH}, T_{NPUSCH})\big) \end{aligned} \tag{3.10}$$

where $p_{X_Y}(\lambda, \mu, T)$ is the loss probability at the radio channel X for the communication stage $Y \in \{fr, fc\}$, λ is the arrival rate at the radio channel, μ the service rate, and T the maximum waiting time in the queue. As explained in [85], p_{X_Y} can be seen as the long-run fraction of customers that are lost in a queuing system with impatient customers. Despite NB-IoT has fixed time slots that leads to use the M/D/1 queue model, the expression to compute the fraction of lost customers $p_{X_Y}(\lambda, \mu, T)$ for the M/D/1 queue does not have a closed form solution. To solve that, the authors of [85] used the equivalent expression for the M/M/1 queue as they found within the parameter ranges used, there is no noticeable difference in the results. Then, using the M/M/1 model, the loss probability can be calculated as follows:

$$p_{X_Y}(\lambda, \mu, T) = \frac{(1-\rho) \cdot \rho \cdot e^{-\mu(1-\rho)\left(T-\frac{1}{\mu}\right)}}{1-\rho^2 \cdot e^{-\mu(1-\rho)\left(T-\frac{1}{\mu}\right)}} \tag{3.11}$$

where $\rho = \frac{\lambda}{\mu}$ is the queue load, λ describes the number of used channel resources (i.e. SFs or RUs) per SF, and μ is the number of available channel resources per SF. The respective values of λ, μ, and T depend on the radio channel.

The assumed timeout value T is a simplification to define a maximum packet waiting time in the queue in ms. This is due to there are several timers involved in NB-IoT during the signaling and data message exchange to be accurately modeled with only one timer. Consequently, using the simplified parameter T, we assume the capacity in NB-IoT is limited by resource scarcity and not by timeouts. For T, the assumed values are $2pp$, $8pp$, and $8pp$ for NPDCCH, NPDSCH, and NPUSCH, respectively. The timeout value of NPDCCH is the minimum RAR window timeout. The timeout value of the other channels have one of the possible values of the MAC Contention Resolution timer. Through a few evaluations, we have found that with greater values of these timeouts, there is no significant difference in the capacity results with the configuration that we use.

To estimate the service rate of a channel, we consider each ECL has a percentage of the total resources available r_{ECL}. For the UL, the available resources for NPUSCH are estimated considering the selected NPUSCH configuration and the remain number of resources within a RAO after subtracting the resources for one preamble transmission. For the DL, the distribution of the radio resources depends on two parameters:

- Ratio of DL resources occupied by DL signals and the NPBCH, r_{DL}: due to the DL SFs carrying NPBCH, NPSS, and NSSS, 5 out of 20 SFs are occupied. This is without considering the possibility of invalid SFs or the DL resources occupied by the SIBs (see subsection 3.2.1.2). Thus, $r_{DL} = 0.25$.

- Ratio of DL resources for data, r_{DLdata}: the remain DL resources have to be distributed between the NPDCCH and NPDSCH. The occurrence of NPDCCH depends on the needed search spaces. Therefore, the distribution of DL resources is variable. To simplify the distribution of DL resources in

the model, we assume a fixed ratio of the SFs from the total dedicated to the NPDCCH. Through an study of the possible range of values, we have selected the value that obtains the best results in terms of capacity. Thus, $r_{DLdata} = 0.6$.

Hence, the service rate of each channel in units of available resources (i.e. SFs or RUs) per SF can be estimated as follows:

$$\begin{aligned} \mu_{NPDCCH} &= r_{ECL} \cdot (1 - r_{DL}) \cdot (1 - r_{DLdata}) \cdot \frac{1}{T_{SF}} \\ \mu_{NPDSCH} &= r_{ECL} \cdot (1 - r_{DL}) \cdot r_{DLdata} \cdot \frac{1}{T_{SF}} \\ \mu_{NPUSCH} &= r_{ECL} \cdot \frac{T_{RAO} - T_{preamble}}{T_{RAO}} \cdot \frac{M_{RU}}{T_{RU}} \end{aligned} \tag{3.12}$$

where M_{RU} and T_{RU} are the number of simultaneous RUs and the duration of a RU, respectively. Both parameters depend on the current NPUSCH configuration used. Note to estimate μ_{NPDCCH} and μ_{NPDSCH}, we are dividing by the duration of one SF T_{SF}, i.e., 1 ms.

For the λ values (in units of number of used channel resources per SF), we have to consider more variables, i.e., the radio channel, the control procedure, and the use case. To simplify the definition of λ for each possibility, Tables 3.3, 3.4 and 3.5 summarize the packet sizes and the amount of radio resources used in each case and channel. Thus, the specific λ of interest can be obtained by adding the elements of one column of the table. The values denoted as N_x represent the number of resources used for the packet x. The total number of resources are estimated as follows N_x = *Number of RUs or SFs* · *Number of segments* · *Number of repetitions*. The number of resources and segments required depends on the packet size (see subsection 3.4.1). For larger packets, more resources will be required, thus, the radio channel will obtain a higher λ.

Table 3.3: Packet sizes and acronyms in each case considered in the analysis. Italic messages are used only in `UL-ACK` or `DL-ACK` cases.

(a) `UL` and `UL-ACK` cases.

	Size (bytes)	Acronym
RAR	7	*rar*
RRC Request	9	*req*
RRC Setup	38	*set*
RRC Setup Comp.	19	*cmp*
RRC Setup Comp. + NAS UL Report	87	*ulCP*
RRC Security Comd.	11	*sec*
RRC Security Comp.	13	*sec_cmp*
RRC Reconf.	61	*rec*
RRC Reconf. Comp.	10	*rec_cmp*
UL Report	59	*ul*
UL Report ACK	SR: 39 UP: 39 CP: 59	SR: *ulAck* UP: *ulAck* CP: *ulAckCP*

(b) `DL` and `DL-ACK` cases.

	Size (bytes)	Acronym
RAR	7	*rar*
RRC Request	9	*req*
RRC Setup	38	*set*
RRC Setup Comp. + TAU Request	99	*tau_req*
RRC Security Comd.	11	*sec*
RRC Security Comp.	13	*sec_cmp*
RRC Reconf.	61	*rec*
RRC Reconf. Comp.	10	*rec_cmp*
RRC DL Info. Transf. + TAU Accept	30	*tau_acp*
DL Report	SR: 59 UP: 59 CP: 79	SR: *dl* UP: *dl* CP: *dlCP*
RAR	7	*rar*
Scheduling Request	9	*schreq*
DL Report ACK	SR: 59 UP: 59 CP: 79	SR: *dlAck* UP: *dlAck* CP: *dlAckCP*

Table 3.4: `UL` and `UL-ACK` cases: Amount of radio resources used for NPDCCH, NPDSCH, and NPUSCH and three data transmission procedures. Italic messages are used only in `UL-ACK` case.

	SR			**UP**			**CP**		
	NPDCCH	NPDSCH	NPUSCH	NPDCCH	NPDSCH	NPUSCH	NPDCCH	NPDSCH	NPUSCH
RAR	$\frac{1-e^{-\lambda_T \cdot T_{RAO}}}{T_{RAO}}$	$\lambda_A N_{rar}$		$\frac{1-e^{-\lambda_T \cdot T_{RAO}}}{T_{RAO}}$	$\lambda_A N_{rar}$		$\frac{1-e^{-\lambda_T \cdot T_{RAO}}}{T_{RAO}}$	$\lambda_A N_{rar}$	
RRC Request			$\lambda_S N_{req}$			$\lambda_S N_{req}$			$\lambda_S N_{req}$
RRC Setup	λ_S	$\lambda_S N_{set}$		λ_S	$\lambda_S N_{set}$		λ_S	$\lambda_S N_{set}$	
RRC Setup Comp.	λ_S		$\lambda_S N_{cmp}$	λ_S		$\lambda_S N_{cmp}$			
RRC Setup Comp. + NAS UL Report							λ_S		$\lambda_S N_{ulCP}$
RRC Security Comd.	λ_S	$\lambda_S N_{sec}$							
RRC Security Comp.	λ_S		$\lambda_S N_{sec\_cmp}$						
RRC Reconf.	λ_S	$\lambda_S N_{rec}$							
RRC Reconf. Comp.	λ_S		$\lambda_S N_{rec\_cmp}$						
UL Report	λ_S		$\lambda_S N_{ul}$	λ_S		$\lambda_S N_{ul}$			
UL Report ACK	λ_S	$\lambda_S N_{ulAck}$		λ_S	$\lambda_S N_{ulAck}$		λ_S		$\lambda_S N_{ulAckCP}$

Table 3.5: `DL` and `DL-ACK` cases: Amount of radio resources used for NPDCCH, NPDSCH, and NPUSCH and three data transmission procedures. Italic messages are used only in `DL-ACK` case.

	SR			**UP**			**CP**		
	NPDCCH	NPDSCH	NPUSCH	NPDCCH	NPDSCH	NPUSCH	NPDCCH	NPDSCH	NPUSCH
RAR	$\frac{1-e^{-\lambda_T \cdot T_{RAO}}}{T_{RAO}}$	$\lambda_A N_{rar}$		$\frac{1-e^{-\lambda_T \cdot T_{RAO}}}{T_{RAO}}$	$\lambda_A N_{rar}$		$\frac{1-e^{-\lambda_T \cdot T_{RAO}}}{T_{RAO}}$	$\lambda_A N_{rar}$	
RRC Request			$\lambda_S N_{req}$			$\lambda_S N_{req}$			$\lambda_S N_{req}$
RRC Setup	λ_S	$\lambda_S N_{set}$		λ_S	$\lambda_S N_{set}$		λ_S	$\lambda_S N_{set}$	
RRC Setup Comp. + TAU Request	λ_S		$\lambda_S N_{tau\_req}$	λ_S		$\lambda_S N_{tau\_req}$	λ_S		$\lambda_S N_{tau\_req}$
RRC Security Comd.	λ_S	$\lambda_S N_{sec}$							
RRC Security Comp.	λ_S		$\lambda_S N_{sec\_cmp}$						
RRC Reconf.	λ_S	$\lambda_S N_{rec}$							
RRC Reconf. Comp.	λ_S		$\lambda_S N_{rec\_cmp}$						
RRC DL Info. Transf. + TAU Accept	λ_S	$\lambda_S N_{tau\_acp}$		λ_S	$\lambda_S N_{tau\_acp}$		λ_S	$\lambda_S N_{tau\_acp}$	
DL Report	λ_S	$\lambda_S N_{dl}$		λ_S	$\lambda_S N_{dl}$		λ_S	$\lambda_S N_{dlCP}$	
RAR	λ_S	$\lambda_S N_{rar}$		λ_S	$\lambda_S N_{rar}$		λ_S	$\lambda_S N_{rar}$	
Scheduling Request			$\lambda_S N_{schreq}$			$\lambda_S N_{schreq}$			$\lambda_S N_{schreq}$
DL Report ACK	λ_S	$\lambda_S N_{dlAck}$		λ_S	$\lambda_S N_{dlAck}$		λ_S		$\lambda_S N_{dlAckCP}$

3.5.2 Markov chain for NB-IoT UE

Let us now describe the behavior of an NB-IoT UE. To model this behavior, we apply a two-dimensional Markov chain. The different states of the Markov chain describe the phases the UE will follow to connect and later communicate with the network. From the defined Markov chain and the transition probabilities between states, we estimate the steady-state probability distribution. That is, the stable probability of a UE in a specific state of the chain. Due to the state space of our Markov chain is an irreducible recurrent set, we know the steady-state probability distribution exists. Later, these probabilities will be useful to obtain the average energy consumption of a UE while performing different control procedures.

Figure 3.8 depicts the proposed Markov chain. Additionally, Figure 3.9 shows an example of the signaling considered at each state of the Markov chain when the SR procedure is considered. The states and transitions between are defined as:

- State Off: This state is the beginning of the Markov chain and models the UE in PSM. The UE changes to the $\{0, 0\}$ state when a new UL report is generated or the periodic TAU timer expires and the UE has to perform the TAU procedure. In the latter case, the network will use the connection establishment for the TAU to notify there is a DL packet pending for the UE.

- States $\{i, 0\}$: They represent the ith start of the RA attempt. These states comprise the synchronization (i.e. decoding NPSS and NSSS and obtaining the core cell information from the MIB and SIBs) and transmission of the RA preamble. To connect to the network, up to m retries of the RA procedure are allowed. If a UE fails the RA procedure, i.e., there is a preamble collision, the UE backs off transiting to the state $\{i + 1, k\}$. If the RA procedure is successful, the UE transits to the $DET\,(i)$ state.

- States $\{i, k\}$: These states represent the kth backoff of the ith RA attempt. For each RA retry, there could be different causes of failure in the establishment of the RRC connection. The backoff time is uniformly chosen in the range $(0, W_c - 1)$, being W_c the maximum backoff window size in ms.

- States $DET\,(i)$: They represent the possibility of the preamble not being detected at the eNB. If the preamble is detected, the UE transits to $REQ\,(i)$, otherwise, the UE backs off to the state $\{i+1, k\}$.

- States $REQ\,(i)$: They represent the possibility of lack of resources needed until the transmission of the RRC Connection Request message. If there are no resources, the UE transits to $RESW\,(i)$, otherwise, the UE transits to $CR\,(i)$.

- States $RESW\,(i)$: As there are no resources available, these states represent the UE waiting period during the RAR window due to the no reception of the RAR message. At the expiration of the RAR window, the UE backs off to the state $\{i+1, k\}$.

- States $CR\,(i)$: They comprise the request for the RRC connection. If there are no enough resources for the exchanges from the reception of the RRC Connection Setup message until the data transmission/reception is completed, the UE transits to the $FAIL\,(i)$ state, otherwise, the UE transits to the *Connect* state.

- States $FAIL\,(i)$: This second possible failure is triggered if the RRC Connection Setup message is not received due to the lack of radio resources at this stage of the communication. Thus, the UE waits until the expiration of the MAC Contention Resolution timer. Next, the UE backs off to the state $\{i+1, k\}$.

- State *Connect*: This state comprises the remaining messages for the RRC connection establishment and the data transmission/reception. The signaling exchanges during this state depend on the data transmission procedure used. After the data transmission, the UE transits to the ACK state if there is a pending application acknowledgement, otherwise, it transits to the *Inactive* state.

- State ACK: This state represents the transmission/reception of the application acknowledgement. After the acknowledgement, the UE transits to the *Inactive* state.

- State *Inactive*: This state represents the period the UE is still reachable by the network before entering PSM, i.e., RRC Inactivity Timer period using Connected DRX (C-DRX), S1 Release (S1) procedure, and the Active Timer period using Idle DRX (I-DRX). At the expiration of the Active Timer, the UE transits to the Off state.

- State *Drop*: This state represents the UE dropping the data packet. After this, the UE transits to the Off state.

The transition between the Markov chain states depend on a set of probabilities. In addition to the probabilities explained in the last subsection 3.5.1, let,

- p_{on} denote the probability of having UL traffic in a SF, expressed as $p_{on} = 1 - e^{-\lambda_{app}}$.

- p_{ack} denote the probability of report acknowledgement. For `UL` and `DL` cases, $p_{ack} = 0$. For `UL-ACK` and `DL-ACK` cases, $p_{ack} = 1$.

Then, the transition probabilities of the Markov chain can be expressed as:

$$P\,(0,0|Off) = p_{on}$$

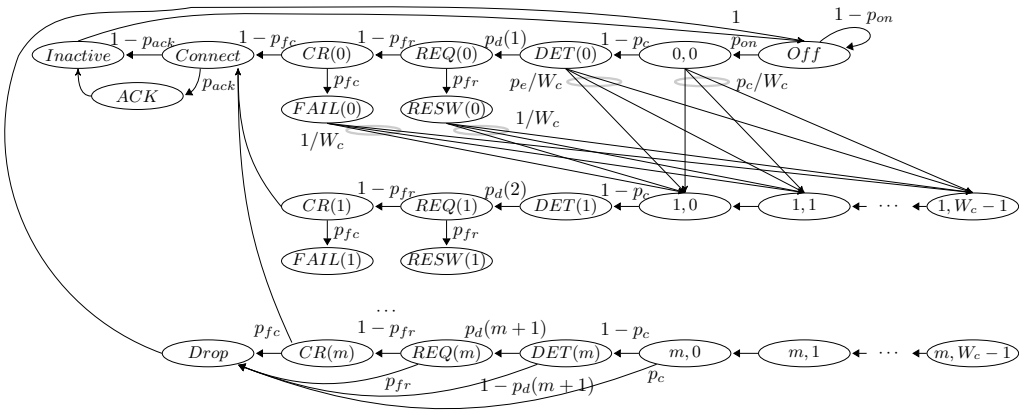


Figure 3.8: NB-IoT UE Markov chain model for m retransmissions.

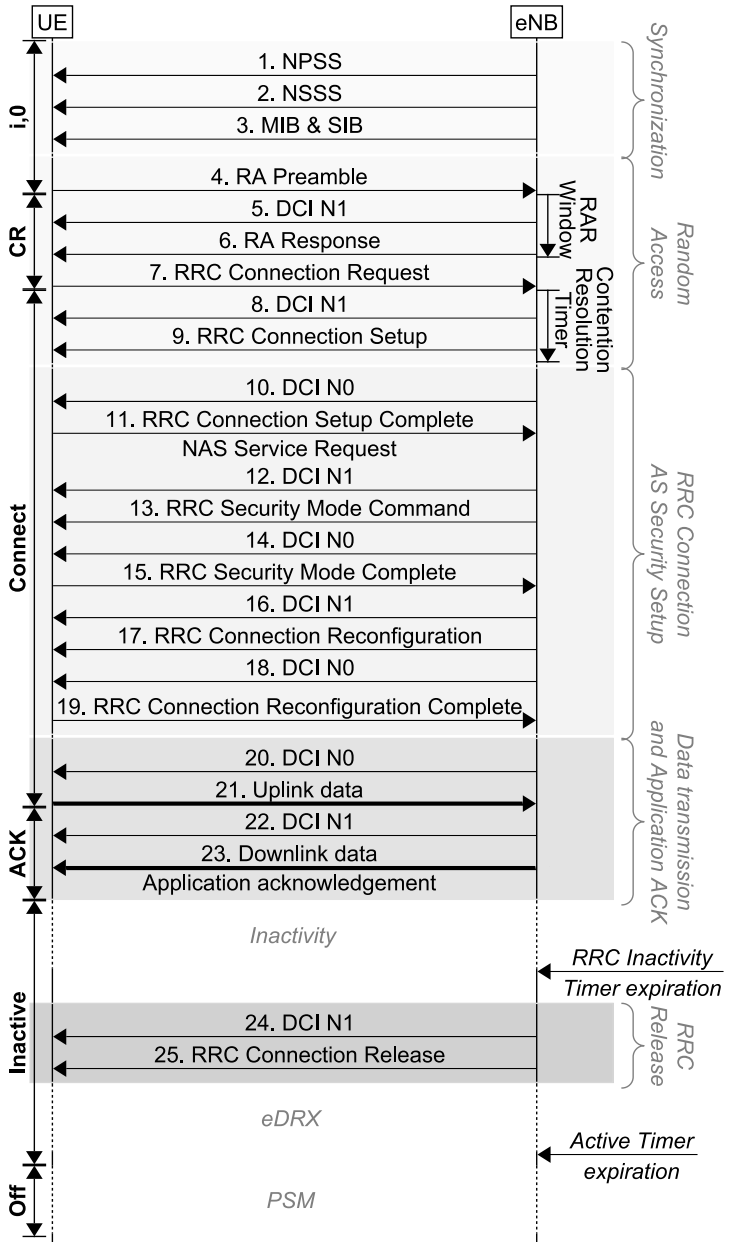


Figure 3.9: `UL-ACK` case signaling flow using the SR procedure [9] and the Markov chain states for a successful connection.

$$P\left(DET\left(i\right)|i,0\right) = 1 - p_c \qquad i \in [0,m]$$

$$P\left(REQ\left(i\right)|DET\left(i\right)\right) = p_d(i+1) \qquad i \in [0,m]$$

$$
\begin{aligned}
P\left(CR\left(i\right)|REQ\left(i\right)\right) &= 1-p_{fr} && i\in[0,m]\\
P\left(Connect|CR\left(i\right)\right) &= 1-p_{fc} && i\in[0,m]\\
P\left(ACK|Connect\right) &= p_{ack}\\
P\left(Inactive|Connect\right) &= 1-p_{ack}\\
P\left(Inactive|ACK\right) &= 1\\
P\left(RESW\left(i\right)|REQ\left(i\right)\right) &= p_{fr} && i\in[0,m]\\
P\left(FAIL\left(i\right)|REQ\left(i\right)\right) &= p_{fc} && i\in[0,m]\\
P\left(i,k|i-1,0\right) &= \frac{p_c}{W_c} && k\in[0,W_c-1],i\in[1,m]\\
P\left(i,k|DET\left(i-1\right)\right) &= \frac{1-p_d(i)}{W_c} && k\in[0,W_c-1],i\in[1,m]\\
P\left(i,k|RESW\left(i-1\right)\right) &= \frac{1}{W_c} && k\in[0,W_c-1],i\in[1,m]\\
P\left(i,k|FAIL\left(i-1\right)\right) &= \frac{1}{W_c} && k\in[0,W_c-1],i\in[1,m]\\
P\left(Drop|m,0\right) &= p_c\\
P\left(Drop|DET\left(m\right)\right) &= 1-p_d(m+1)\\
P\left(Drop|REQ\left(m\right)\right) &= p_{fr}\\
P\left(Drop|CR\left(m\right)\right) &= p_{fc}\\
P\left(Off|Drop\right) &= P\left(Off|Inactive\right)=1
\end{aligned}
\tag{3.13}
$$

Finally, let b_j denote the steady state probability that a device is at the state j. Then, the steady state probabilities of the chain can be derived as:

$$
\begin{aligned}
b_{Off} &= \left(1-p_{on}\right)b_{Off}+b_{Drop}+b_{Inactive}\\
b_{0,0} &= p_{on}b_{Off}\\
b_{DET(i)} &= \left(1-p_c\right)b_{i,0}\\
b_{REQ(i)} &= p_{d(i+1)}b_{DET(i)}=p_{d(i+1)}\left(1-p_c\right)b_{i,0}\\
b_{CR(i)} &= \left(1-p_{fr}\right)b_{REQ(i)}=\left(1-p_{fr}\right)p_{d(i+1)}\left(1-p_c\right)b_{i,0}\\
b_{RESW(i)} &= p_{fr}b_{REQ(i)}=p_{fr}p_{d(i+1)}\left(1-p_c\right)b_{i,0}\\
b_{FAIL(i)} &= p_{fc}b_{CR(i)}=p_{fc}\left(1-p_{fr}\right)p_{d(i+1)}\left(1-p_c\right)b_{i,0}
\end{aligned}
$$

$$
\begin{aligned}
b_{i,0} &= p_c b_{i-1,0} + \left(1 - p_{d(i)}\right) b_{DET(i-1)} + b_{RESW(i-1)} + b_{FAIL(i-1)} \\
&= \left[1 - p_{d(i)} \left(1 - p_c\right) \left(1 - p_{fr}\right) \left(1 - p_{fc}\right)\right] b_{i-1,0} = b_{0,0} \prod_{n=0}^{i} s(n) \\
b_{i,k} &= \frac{W_c - k}{W_c} b_{i,0} \\
b_{Drop} &= p_c b_{m,0} + \left(1 - p_{d(m+1)}\right) b_{DET(m)} + p_{fr} b_{REQ(m)} + p_{fc} b_{CR(m)} \\
&= b_{0,0} \prod_{n=0}^{m+1} s(n) \\
b_{Connect} &= \left(1 - p_{fc}\right) \sum_{i=0}^{m} b_{CR(i)} = b_{0,0} \left(1 - \prod_{n=0}^{m+1} s(n)\right) \\
b_{ACK} &= p_{ack} b_{Connect} = p_{ack} b_{0,0} \left(1 - \prod_{n=0}^{m+1} s(n)\right) \\
b_{Inactive} &= (1 - p_{ack}) b_{Connect} + b_{ACK} = b_{Connect} = b_{0,0} \left(1 - \prod_{n=0}^{m+1} s(n)\right)
\end{aligned} \tag{3.14}
$$

where $s(i)$ is a replacement to reduce the length of the equations and equals $s(i) = \left[1 - p_{d(i)} \left(1 - p_c\right) \left(1 - p_{fr}\right) \left(1 - p_{fc}\right)\right]$.

Next, by imposing the probability normalization condition we have:

$$
\begin{aligned}
1 = & b_{Off} + b_{0,0} + \sum_{i=1}^{m} \sum_{k=0}^{W_c - 1} b_{i,k} + \sum_{i=0}^{m} b_{DET(i)} + \sum_{i=0}^{m} b_{REQ(i)} + \sum_{i=0}^{m-1} b_{RESW(i)} + \\
& \sum_{i=0}^{m} b_{CR(i)} + \sum_{i=0}^{m-1} b_{FAIL(i)} + b_{Drop} + b_{Connect} + b_{ACK} + b_{Inactive}
\end{aligned} \tag{3.15}
$$

Rearranging (3.15), we can compute b_{Off} as:

$$
\begin{aligned}
b_{Off} = & \left(\left(1 + p_{on}\right) \left[1 + \frac{W_c + 1}{2} \sum_{i=1}^{m} \prod_{n=0}^{i} s\left(n\right) + \right. \right. \\
& \left. \left. \left(1 - p_c\right) \left[\sum_{i=0}^{m} p_{d(i+1)} \prod_{n=0}^{i} s(n) \left(1 + \left(1 - p_{fr}\right)\right) + \right. \right. \right.
\end{aligned}
$$

$$\sum_{i=0}^{m-1} p_{d(i+1)} \prod_{n=0}^{i} s(n) \left(p_{fr} + p_{fc}\right) + \sum_{i=0}^{m} \prod_{n=0}^{i} s(n) \Bigg] + \prod_{n=0}^{m+1} s(n) + \left(1 - \prod_{n=0}^{m+1} s(n)\right) \left(2 + p_{ack}\right) \Bigg] \Bigg)^{-1} \tag{3.16}$$

Additionally, we can derive the failure of a connection establishment, i.e., the outage probability from $b_{Connect}$ and b_{Drop} as follows:

$$p_{outage} = \frac{b_{Drop}}{b_{Drop} + b_{Connect}} = \frac{b_{0,0} \prod_{n=0}^{m+1} s(n)}{b_{0,0} \prod_{n=0}^{m+1} s(n) + b_{0,0} \left[1 - \prod_{n=0}^{m+1} s(n)\right]} = \prod_{n=0}^{m+1} s(n) \tag{3.17}$$

Thus, the average number of required connection establishment attempts can be approximated form the number of failures as:

$$N_{attempts}\left(\lambda_T\right) = \sum_{i=0}^{m} \prod_{n=0}^{i} \left(1 - p_{d(n)} \left(1 - p_{fr}\right) \left(1 - p_{fc}\right) \left(1 - p_c\right)\right) = \sum_{i=0}^{m} \prod_{n=0}^{i} s(n) \tag{3.18}$$

Using (3.18), the value of λ_T can be obtained by solving the following iterative equation:

$$\lambda_T = N_{attempts}\left(\lambda_T\right) \cdot N_{UE} \cdot \lambda_{app} \tag{3.19}$$

where N_{UE} is the number of UEs. For the estimation of $N_{attempts}$ and λ_T, we apply the fixed-point method using (3.18) and (3.19) as the starting values until the solution reaches the maximum p_{outage} defined using (3.17).

3.5.3 Energy consumption and delay analysis

Now let us enter the final step of our analysis. In this part we calculate the average energy consumption when performing the SR, the UP, or CP procedures. The energy consumption is based on the average power and duration of each Markov

chain state. That is, we use the states of the Markov chain that define the behavior of the UE to estimate the energy the UE will consume when it enters each state. The aim of the model is to obtain an estimation of the average energy consumption to later derive the UE battery lifetime. To do that, the analysis is divided into four parts. Firstly, we detail the energy consumption while receiving or transmitting packets or signaling in NB-IoT. Secondly, we introduce a few preliminary energy consumption estimations to simplify the analysis. Thirdly, we present estimates of the energy consumption per Markov chain state. Finally, the battery lifetime is estimated. Unless indicated otherwise, the units of the power levels and the timers are in mW and ms, respectively.

Table 3.6 contains the definition of the parameters used in the following analysis.

Table 3.6: Variables and parameters of the model [22, 23].

	Parameter	Value	Description
Energy	$E_j^{\{v,z\}}$	Variable	Average energy consumption in state j, the v procedure (where $v \in \{SR, CP, UP\}$), and the z direction of the report (where $z \in \{ul, dl\}$)(μJ)
	$D_j^{\{v,z\}}$	Variable	Average delay in state j (μJ)
	$E(x)$	Variable	Average energy consumption of the packet x(μJ)
	P_{TX}	Variable	Transmission power consumption (mW)
	P_{max}	545	Maximum transmission power consumption (mW)
	P_{RX}	90	Reception power consumption (mW)
	P_i	3	Inactive power consumption (mW)
	P_s	0.015	Standby power consumption (mW)
Sync.	T_{sync}	Variable	Average initial synchronization time (ms)
	T_{MIB-I}	Variable	MIB waiting time (ms)
	T_{MIB-RX}	Variable	MIB reception time (ms)
RA	T_{RAO}	40	RA periodicity (ms)
	T_{PRE}	5.6	Preamble format 0 duration (ms)
	N_{REP}^{RA}	Variable	Number of preamble repetitions
	T_{RARwdo}	2	RAR window size (*pp*)
	$T_{RARwdo_{start}}$	4, 41	RAR window start (ms). If NPRACH repetitions > 64, $T_{RARwdo_{start}} = 41$, otherwise $T_{RARwdo_{start}} = 4$
	$T_{MAC_{cr}}$	2	MAC contention resolution timer (*pp*)
Gaps	$T_{GapPeriod}^{UL}$	296	UL gap periodicity (ms)
	T_{GapDur}^{UL}	40	UL gap duration (ms)
Scheduling	T_{wDC2US}	8	Start of NPUSCH transmission after the end of its associated DCI (ms)
	T_{wDC2DS}	5	Start of NPDSCH transmission after the end of its associated DCI (ms)
	T_{RU}	Variable	RU duration (ms)
	N_{REP}	Variable	Number of repetitions for NPUSCH or NPDSCH
	$N_{REP_{dci}}$	Variable	Number of DCI repetitions
	N_{RU}	Variable	Number of RUs
	N_{SF}	Variable	Number of SFs
	L_x	Variable	Size of packet x (bits)
C-DRX & I-DRX	T_{DRXi}	1	Period the UE should remain monitoring NPDCCH before starting C-DRX (*pp*)
	$T_{inactivity}$	20	RRC Inactivity timer (s)
	T_{waitCP}	5	Short waiting time before entering PSM when using RAI (*pp*)
	T_{onD}	1	On duration timer (*pp*)
	T_{LC}	2.048	C-DRX Long Cycle (s)
	T_{PC}	2.048	I-DRX Paging Cycle (s)
	T_{active}	14.16	Active Timer duration (s)
	T_{3412}	Variable	Periodic TAU timer
	H_{RLCMAC}	2	RLC/MAC headers size (bytes)

3.5.3.1 Packet energy estimation

Prior to the connection establishment, the UE transmits/receives different types of messages. Each type is analyzed as follows:

DCI allocations: This case happens when receiving a UL grant or a DL assignment. In order to estimate the energy consumption, we first derive the reception time needed for the DCI in ms, $T_{rx}(dci)$. Due to the DL SF duration is 1 ms and we assume each DCI copy requires a whole DL SF, the number of DCI repetitions, $N_{REP_{dci}}$, equals the duration of the DCI, thus $T_{rx}(dci) = N_{REP_{dci}}$. Next, we can estimate the DCI's energy consumption $E_{rx}(dci)$ as:

$$E_{rx}(dci) = P_{RX} \cdot T_{rx}(dci) \tag{3.20}$$

UL packet: The estimated transmission time for packet x is:

$$\begin{aligned} T^{UL}(x) &= N_{REP} \cdot N_{RU} \cdot T_{RU} \cdot N_{seg}(x) \\ N_{seg}(x) &= \left\lceil \frac{L_x}{TBS(MCS, N_{RU}) - H_{RLCMAC}} \right\rceil \end{aligned} \tag{3.21}$$

where N_{REP}, N_{RU}, and $N_{seg}(x)$ are the number of repetitions, RU, and segments, respectively. T_{RU} is the duration in ms of the RU, L_x is the size of the packet x in bits, TBS is the Transport Block Size for the NPUSCH resulting from the selection of MCS and N_{RU}, and H_{RLCMAC} is the size of the RLC/MAC headers. Note that this analysis assumes there is a fixed MCS per ECL, but the N_{RU} can be chosen. Then, the selected combination of MCS and N_{RU} is the one that requires the less number of RUs. Using $T^{UL}(x)$, the total duration of the UL gaps is derived:

$$T^{UL}_{Gap}(x) = \left\lfloor \frac{T^{UL}(x)}{T^{UL}_{GapPeriod} - T^{UL}_{GapDur}} \right\rfloor \cdot T^{UL}_{GapDur} \tag{3.22}$$

Finally, the estimated energy consumption and delay due to the transmission of the packet x is:

$$\begin{aligned} E_{tx}(x) &= P_{TX} \cdot T^{UL}(x) + P_i \cdot T^{UL}_{Gap}(x) \\ T_{tx}(x) &= T^{UL}(x) + T^{UL}_{Gap}(x) \end{aligned} \tag{3.23}$$

The UL packets considered in this analysis along with their sizes are summa-

rized in Table 3.4.

DL packet: The estimation in this case is similar to the UL packet. The reception time needed for the packet y is:

$$\begin{aligned} T^{DL}(y) &= N_{REP} \cdot N_{SF} \cdot N_{seg}(y) \\ N_{seg}(y) &= \left\lceil \frac{L_y}{TBS(MCS, N_{SF}) - H_{RLCMAC}} \right\rceil \end{aligned} \tag{3.24}$$

where N_{SF} is the number of SF, and TBS is the Transport Block Size for the NPDSCH resulting from the selection of MCS and N_{SF}. Due to the DL SF duration is 1 ms, the total number of DL resources for the reception (i.e. $N_{REP} \cdot N_{SF} \cdot N_{seg}(y)$) equals the duration of the reception. Like for the UL estimation, this analysis assumes there is a fixed MCS per ECL, but the N_{SF} can be chosen. Then, the selected combination of MCS and N_{SF} is the one that requires the less number of SFs. Finally, the estimated energy consumption and delay due to the reception of the packet y is:

$$\begin{aligned} E_{rx}(y) &= P_{RX} \cdot T^{DL}(y) \\ T_{rx}(y) &= T^{DL}(y) \end{aligned} \tag{3.25}$$

The DL packets considered in this analysis along with their sizes are summarized in Table 3.5. Note in this analysis we always assume the DL gap threshold is greater than the maximum number of repetitions in the DL, $N^{DL}_{GapThr} > R_{max}$. Thus, there are no DL gaps.

3.5.3.2 Preliminary energy consumption estimations

To ease the understanding of the energy consumption per Markov chain state, we explain in this subsection a few independent energy consumption estimations. In this way, the following section will reuse these estimations as parameters for the analysis. Each independent estimation is done as follows:

C-DRX: While the UE is in RRC Connected state and there is no traffic between the network and the UE, the UE can use C-DRX to discontinuously monitor NPDCCH. Figure 3.10 illustrates an example of C-DRX operation. When using C-DRX, after expiry of DRX inactivity time T_{DRXi}, the UE repeatedly starts a C-DRX cycle. This will happen until the connection is released with a RRC

Release procedure due to the expiration of the RRC Inactivity timer $T_{inactivity}$ configured in the network. A C-DRX cycle, i.e., T_{LC}, involves an active listening period T_{onD}, and an inactive period. Then, the energy consumption due to C-DRX E_{CDRX} can be estimated as follows:

$$\begin{aligned} E_{CDRX} =& P_i \left(\left\lfloor \frac{T_{inactivity} - T_{DRXi}}{T_{LC}} \right\rfloor \cdot (T_{LC} - T_{onD}) \right) + \\ & P_{RX} \left(T_{DRXi} + \left\lfloor \frac{T_{inactivity} - T_{DRXi}}{T_{LC}} \right\rfloor \cdot T_{onD} \right) \end{aligned} \tag{3.26}$$

where $\lfloor A \rfloor$ rounds the element A to the nearest integer less than or equal to A.

S1 procedure: After the expiration of $T_{inactivity}$, the eNB releases the RRC connection sending a RRC Release message to the UE. The estimated energy consumption due to this procedure is:

$$E_{S1Rel} = P_i \left(pp/2 + T_{wDC2DS} \right) + E_{rx}(rel) \tag{3.27}$$

where $pp/2$ denotes the average waiting time for NPDCCH period. In this occasion, this value is used to add a wait until the possible NPDCCH SFs finish. T_{wDC2DS} denotes the wait until the start of the NPDSCH reception after the end of its associated DCI (see subsection 3.2.2 for a detailed description of the timing relationship between actions in NB-IoT), and $E_{rx}(rel)$ is the energy consumption due to the reception of the RRC Release message.

I-DRX: While the UE is in RRC Idle state and until the expiration of the

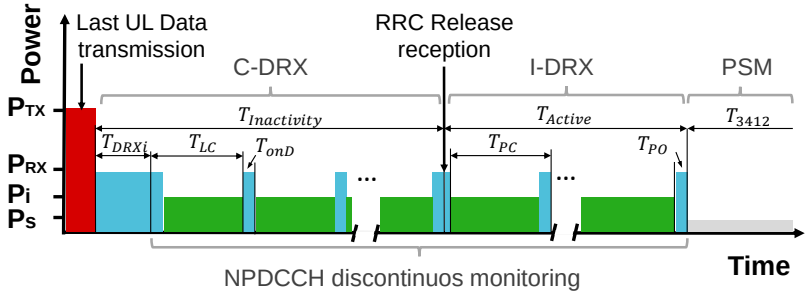


Figure 3.10: Example of power consumption transitions during C-DRX and I-DRX.

Active Timer T_{active}, the UE can use I-DRX. An I-DRX cycle, i.e., T_{PC}, involves an active listening period named paging occasion, and an inactive period. For simplicity, in this analysis we assume C-DRX and I-DRX are configured with the same active and inactive periods, and the difference between both mechanisms is the duration of their application (i.e. RRC Inactivity or Active timers). Consequently, the energy consumption due to I-DRX can be estimated as follows:

$$E_{IDRX} = \left\lfloor \frac{T_{active}}{T_{PC}} \right\rfloor (P_i (T_{PC} - T_{onD}) + P_{RX} T_{onD}) \tag{3.28}$$

TAU procedure: As previously mentioned in the system model, for `UL` and `UL-ACK` cases the UE will perform periodic TAUs with a periodicity of 5 days. This procedure is done while the UE is in PSM. Therefore, we need to estimate the energy consumed due to the periodic TAU to add this consumption to the energy consumption estimation of the UE in PSM. To simplify the correlation between the steps the UE performs in this procedure and the estimation, Figure 3.11 illustrates the signaling flow for the TAU. Then, the energy consumption is estimated as follows:

$$\begin{aligned} E_{TAU} =& P_i (pp/2 + 3T_{wDC2DS} + 2T_{wDC2US} + T_{wDC1} + T_{wDC2} + T_{wDC3}) + \\ & 4E_{rx}(dci) + E_{rx}(rar) + E_{tx}(req) + E_{rx}(set) + E_{rx}(tau\_req) + \\ & E_{rx}(tau\_acpt) + E_{S1Rel} + E_{IDRX} \end{aligned} \tag{3.29}$$

where $pp/2$ denotes the average waiting time for the NPDCCH occurrence as at the beginning there are no steps to be used as a reference to estimate this waiting time. Note that the different elements of the equation are not sorted by their sequential occurrence. T_{wDC2US} is the wait from the reception of the DCI with the UL allocation to the NPUSCH transmission start and $E_{rx}(dci)$ is the energy consumed due to the reception of a DCI.

The parameters T_{wDCX} denote different waiting times for the NPDCCH occurrence (see Figure 3.11). Their values can be estimated as $T_{WDCX} = pp - mod(T_{x_1} + T_{x_2} + ... + T_{x_n}, pp)$, where $x_1, x_2, ..., x_n$ are the considered steps occurred between NPDCCH occasions, T_{x_n} is the duration of the x_n step, and $mod()$ is the modulus after division function. In this analysis, most of the x_n

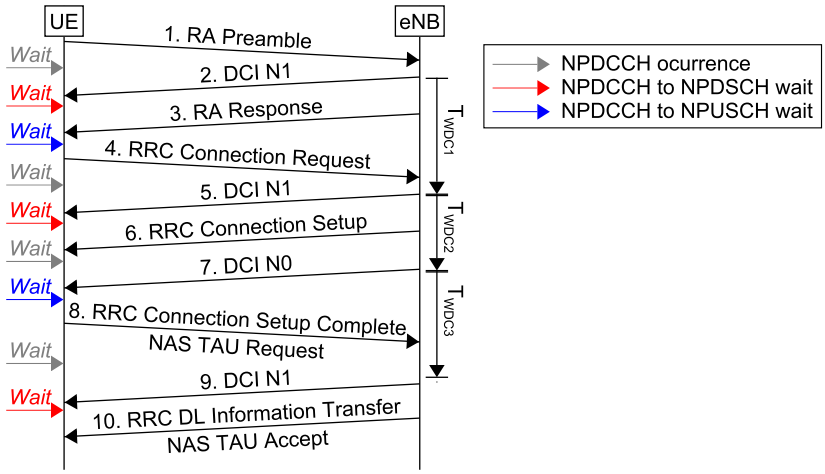


Figure 3.11: Signaling flow of a periodic TAU and the considered NB-IoT waits between actions.

steps between two occurrences of the NPDCCH are: i) DCI's reception time; ii) Wait for the start of NPDSCH/NPUSCH reception/transmission after the end of its associated DCI; and iii) Packet reception/transmission time. Therefore, in (3.29) the different waits can be estimated as follows:

$$\begin{aligned} T_{WDC1} =& pp - mod(T_{rx}(dci) + T_{wDC2DS} + T_{rx}(rar) + T_{wDC2US} + T_{tx}(req), pp) \\ T_{WDC2} =& pp - mod(T_{rx}(dci) + T_{wDC2DS} + T_{rx}(set), pp) \\ T_{WDC3} =& pp - mod(T_{rx}(dci) + T_{wDC2US} + T_{tx}(tau\_req), pp) \end{aligned} \tag{3.30}$$

Additionally, at the end of (3.29), we sum the energy consumed due to the reception of the RRC Release E_{S1Rel} and I-DRX E_{IDRX}. However, the energy consumption due to C-DRX is not included. This is because when the TAU finishes, the Mobility Management Entity (MME) sends a command to the eNB in order to release the connection used. Therefore, there is no waiting for UE's inactivity previously the release of the connection.

3.5.3.3 Energy consumption per Markov chain state

Let us now estimate the energy consumption per Markov chain state. To do that, E_j and D_j denote the average energy consumption and delay of the j state, respectively. As a reminder, unless indicated otherwise, the units of the power levels and the timers are in mW and ms, respectively. Note the delay can be estimated by removing the power components (P) of the equations. To clarify the different steps performed during each state, Figure 3.9 details the signaling flow for the SR procedure. For the following analysis, let $E_j^{\{v,z\}}$ denote the average energy consumption for the j state, the v procedure (where $v \in \{SR, CP, UP\}$), and the z direction of the report (where $z \in \{ul, dl\}$). Therefore we have:

Off **state:** The UE does not have a pending UL or DL packet in current SF. For the `UL` and `UL-ACK` cases, the UE will perform the periodic TAU at the expiration of the periodic TAU timer T_{3412}. Therefore, the energy consumption in Off state includes the energy spent when performing this periodic procedure:

$$E_{Off}^{\{SR,ul\}} = E_{Off}^{\{UP,ul\}} = E_{Off}^{\{CP,ul\}} = \left(1 - \frac{1}{T_{3412}}\right) P_s \cdot T_{SF} + \frac{1}{T_{3412}} E_{TAU} \quad (3.31)$$

where the first addend is the probability of the normal SF multiply by the standby power P_s and the duration of the SF T_{SF} (1 ms), and the second addend is the probability of the occurrence of the periodic TAU multiply by the energy consumed due to the periodic TAU E_{TAU} (this parameter is estimated in subsection 3.5.3.2).

For `DL` cases, the timer T_{3412} is configured greater than the DL report IAT. Therefore, the periodic TAU is not needed:

$$E_{Off}^{\{SR,dl\}} = E_{Off}^{\{UP,dl\}} = E_{Off}^{\{SR,dl\}} = P_s \cdot 1 \quad (3.32)$$

$0,0$ **state:** The UE synchronizes and starts the RA procedure. In this state the equations are the same for the three procedures, and all four cases:

$$\begin{aligned} E_{0,0}^{\{v,z\}} =& P_i \cdot (T_{MIB-I} + T_{RAO}/2) + \\ & P_{RX} \cdot (T_{sync} + T_{MIB-RX}) + P_{TX} \cdot N_{REP}^{RA} \cdot T_{PRE} \end{aligned} \quad (3.33)$$

where $T_{RAO}/2$ denotes the average waiting time for NPRACH resource occur-

rence, T_{sync} is the average required synchronization time, T_{MIB-I} is the waiting time for the occurrence of the MIB, T_{MIB-RX} is the MIB's reading time, N_{REP}^{RA} the number of preamble repetitions, and T_{PRE} the preamble duration.

DET **state:** Detection of the preamble. There is no energy consumed in this state: $E_{DET(i)}^{\{v,z\}} = 0$.

REQ **state:** Checking radio resources availability to start the establishment of the connection. There is no energy consumed in this state: $E_{REQ(i)}^{\{v,z\}} = 0$.

CR **state:** The UE performs a connection request:

$$E_{CR(i)}^{\{v,z\}} = P_i \left(pp/2 + T_{wDC2DS} + T_{wDC2US}\right) + E_{rx}(dci) + E_{rx}(rar) + E_{tx}(req) \tag{3.34}$$

where $pp/2$ denotes the average waiting time for the NPDCCH occurrence as at the beginning there are no steps used as reference to estimate this waiting time, $E_{rx}(rar)$ is the energy consumed due to the reception of the RAR, and $E_{tx}(req)$ is the energy consumed due to the transmission of the RRC Connection Request message.

RESW **state:** The UE waits the RAR window and the RAR window start:

$$E_{RESW(i)}^{\{v,z\}} = P_i \left(T_{RARwdo} + T_{RARwdo_{start}}\right) \tag{3.35}$$

where T_{RARwdo} is the RAR window size and $T_{RARwdo_{start}}$ is the RAR window start.

FAIL **state:** The UE waits until the MAC contention resolution timer, T_{MACcr}, expires:

$$E_{FAIL(i)}^{\{v,z\}} = P_i \cdot T_{MACcr} \tag{3.36}$$

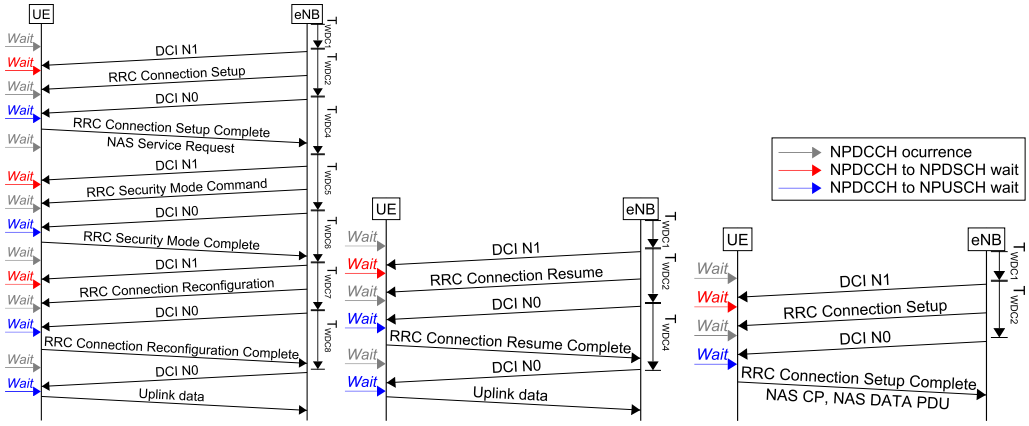
i, k **state:** kth backoff wait of the ith attempt, thus $E_{i,k}^{\{v,z\}} = P_i \cdot T_{SF}$.

$i, 0$ **state:** After a unsuccessful connection attempt and a backoff time, the UE retries the RA procedure. The energy consumption in this state can be estimated as:

$$E_{i,0}^{\{v,z\}} = P_i \cdot T_{RAO}/2 + P_{TX}^{RA} \cdot N_{REP}^{RA} \cdot T_{PRE} \tag{3.37}$$

Connect **state:** After a successful connection, the UE sends its data packet. For the CP procedure setup, the data is transmitted piggybacked in the RRC Connection Setup Complete message. As this state comprises several signaling exchanges, Figure 3.12 illustrates the different signaling messages exchanged per control procedure evaluated for the `UL` case. For the remain cases, their specific signaling messages exchanged are listed in Tables 3.4 and 3.5. The estimates at this state are as follows:

$$
\begin{aligned}
E_{Connect}^{\{SR,ul\}} =& P_i\,(3T_{wDCDS} + 4T_{wDC2US} + T_{wDC1} + T_{wDC2} + T_{wDC4} + T_{wDC5} + \\
& T_{wDC6} + T_{wDC7} + T_{wDC8}) + 7E_{rx}(dci) + E_{rx}(set) + E_{tx}(cmp) + \\
& E_{rx}(sec) + E_{tx}(sec\_cmp) + E_{rx}(rec) + E_{tx}(rec\_cmp) + E_{tx}(ul) \\
T_{wDC4} =& pp - mod(T_{rx}(dci) + T_{wDC2US} + T_{tx}(comp), pp) \\
T_{wDC5} =& pp - mod(T_{rx}(dci) + T_{wDC2DS} + T_{rx}(sec), pp) \\
T_{wDC6} =& pp - mod(T_{rx}(dci) + T_{wDC2US} + T_{tx}(sec\_cmp), pp) \\
T_{wDC7} =& pp - mod(T_{rx}(dci) + T_{wDC2DS} + T_{rx}(rec), pp) \\
T_{wDC8} =& pp - mod(T_{rx}(dci) + T_{wDC2US} + T_{tx}(rec\_cmp), pp)
\end{aligned}
$$



(a) SR signaling flow. (b) UP signaling flow. (c) CP signaling flow.

Figure 3.12: Example of signaling flow and the considered NB-IoT waits in in *Connect* state for `UL` case.

$$
\begin{aligned}
E_{Connect}^{\{SR,dl\}} =& P_i\,(5T_{wDCDS} + 3T_{wDC2US} + T_{wDC1} + T_{wDC2} + T_{wDC9} + T_{wDC5} + \\
& T_{wDC6} + T_{wDC7} + T_{wDC8} + T_{wDC10}) + 8E_{rx}(dci) + E_{rx}(set) + \\
& E_{tx}(tau\_req) + E_{rx}(sec) + E_{tx}(sec\_cmp) + E_{rx}(rec) + \\
& E_{tx}(rec\_cmp) + E_{rx}(tau\_acp) + E_{rx}(dl) \\
T_{wDC9} =& pp - mod(T_{rx}(dci) + T_{wDC2US} + T_{tx}(tau\_req), pp) \\
T_{wDC10} =& pp - mod(T_{rx}(dci) + T_{wDC2DS} + T_{rx}(tau\_acp), pp) \\
E_{Connect}^{\{UP,ul\}} =& P_i\,(T_{wDCDS} + 2T_{wDC2US} + T_{wDC1} + T_{wDC2} + T_{wDC4}) + 3E_{rx}(dci) + \\
& E_{rx}(set) + E_{tx}(cmp) + E_{tx}(ul) \\
E_{Connect}^{\{UP,dl\}} =& P_i\,(3T_{wDCDS} + T_{wDC2US} + T_{wDC1} + T_{wDC2} + T_{wDC9} + T_{wDC10}) + \\
& 4E_{rx}(dci) + E_{rx}(set) + E_{tx}(tau\_req) + E_{rx}(tau\_acp) + E_{rx}(dl) \\
E_{Connect}^{\{CP,ul\}} =& P_i\,(T_{wDCDS} + T_{wDC2US} + T_{wDC1} + T_{wDC2}) + 2E_{rx}(dci) + E_{rx}(set) + \\
& E_{tx}(ulCP) \\
E_{Connect}^{\{CP,dl\}} =& P_i\,(3T_{wDCDS} + T_{wDC2US} + T_{wDC1} + T_{wDC2} + T_{wDC9} + T_{wDC10}) + \\
& 4E_{rx}(dci) + E_{rx}(set) + E_{tx}(tau\_req) + E_{rx}(tau\_acp) + E_{rx}(dlCP)
\end{aligned}
\tag{3.38}
$$

where the different waits between actions the UE experiences are accumulated on each evaluation, thus, we obtain addends such as $X \cdot T_{wDC2DS}$.

ACK **state:** The report acknowledgment in sent/received:

$$
\begin{aligned}
E_{ACK}^{\{SR,ul\}} =& E_{ACK}^{\{UP,ul\}} = P_i\,(T_{wDC11} + T_{wDC2DS}) + E_{rx}(dci) + E_{rx}(ulAck) \\
T_{wDC11} =& pp - mod(T_{rx}(dci) + T_{wDC2US} + T_{tx}(ul), pp) \\
E_{ACK}^{\{SR,dl\}} =& E_{ACK}^{\{UP,dl\}} = P_i\,(T_{RAO}/2 + pp/2 + 2T_{wDC2DS} + 2T_{wDC2US} + \\
& T_{wDC1} + T_{wDC2}) + P_{TX} \cdot N_{REP}^{RA} \cdot T_{PRE} + 3E_{rx}(dci) + E_{rx}(rar) + \\
& E_{tx}(schreq) + E_{rx}(set) + E_{tx}(dlAck) \\
E_{ACK}^{\{CP,ul\}} =& P_i\,(T_{wDC12} + T_{wDC2DS}) + E_{rx}(dci) + E_{rx}(ulAckCP) \\
T_{wDC12} =& pp - mod(T_{rx}(dci) + T_{wDC2US} + T_{tx}(ulCP), pp)
\end{aligned}
$$

$$
\begin{aligned}
E_{ACK}^{\{CP,dl\}} =& P_i \left(T_{RAO}/2 + pp/2 + 2T_{wDC2DS} + 2T_{wDC2US} + \right. \\
& \left. T_{wDC1} + T_{wDC2}\right) + P_{TX} \cdot N_{REP}^{RA} \cdot T_{PRE} + 3E_{rx}(dci) + E_{rx}(rar) + \\
& E_{tx}(schreq) + E_{rx}(set) + E_{tx}(dlAckCP)
\end{aligned}
\tag{3.39}
$$

Note for the `UL-ACK` evaluations, the reception of the application acknowledgement only requires the reception of its DCI and the packet reception. On the contrary, the `DL-ACK` evaluations need the start of the Scheduling Request procedure to request resources to send the acknowledgment to the IoT server. That is, the UE does not have uplink resources allocated to communicate with the network and thus it needs to start the RA procedure to later request resources.

Inactive **state:** The UE stays in this state until the expiration of the Active Timer T_{active}. This state includes C-DRX, the RRC Release, and I-DRX. Particularly for CP, we consider when a UE transmits data in UL, the NAS signaling message encapsulating the data packet includes the RAI field to notify if there is no further data UL or DL data transmissions are expected. When the RAI can be included, we assume the connection is released after a short period T_{waitCP} for S1 processing and transfer delay, and the UE enters PSM straigthforward, i.e., the Active Timer equals $T_{Active} = 0$ s. This situation within CP occurs for the `UL`, `UL-ACK` and `DL-ACK` cases. Therefore, the estimates of the energy consumption are as follows:

$$
\begin{aligned}
E_{Inactive}^{\{SR,ul\}} =& E_{Inactive}^{\{SR,dl\}} = E_{Inactive}^{\{UP,ul\}} = E_{Inactive}^{\{UP,dl\}} = E_{CDRX} + E_{S1Rel} + E_{IDRX} \\
E_{Inactive}^{\{CP,ul\}} =& P_i \cdot T_{waitCP} + E_{S1Rel} \\
E_{Inactive}^{\{CP,dl\}} =& p_{ack} \cdot P_i \cdot T_{waitCP} + (1 - p_{ack}) \left(E_{CDRX} + E_{IDRX}\right) + E_{S1Rel}
\end{aligned}
\tag{3.40}
$$

where E_{CDRX}, E_{S1Rel}, and E_{IDRX} were defined in (3.26), (3.27), and (3.28), respectively.

Drop **state:** If the UE fails all connection attempts, the UE drops the data packet. There is not energy consumption in this state, then, $E_{drop}^{\{v,z\}} = 0$.

3.5.3.4 Battery lifetime estimation

From the prior analysis the energy consumed per day E_{day}^{model} in joules (J) is estimated as:

$$E_{day}^{model} = \left(\left(\sum_j b_j E_j \right) \cdot \frac{D_{day}}{\sum_j b_j D_j} \right) \cdot 1e-6 \tag{3.41}$$

where D_{day} denotes the duration of one day in ms. Finally, the battery lifetime in years Y_{model} can be estimated as:

$$\begin{aligned} E_{day_{Wh}}^{model} &= \frac{E_{day}^{model}}{3600} \\ Y_{model} &= \frac{C_{bat}}{E_{day_{Wh}}^{model} \cdot 365.25} \end{aligned} \tag{3.42}$$

where $E_{day_{Wh}}^{model}$ is the energy consumption per day in watt-hour units, and C_{bat} is the battery capacity.

3.6 NB-IoT UE energy evaluation

This section presents the evaluation setup and the results in terms of energy consumption and capacity obtained with the NB-IoT analytical model proposed in this chapter. The goal of this evaluation is to compare the UE performance under different coverage scenarios, i.e, ECLs. Additionally, we want to study the benefits of the new optimized data transmission procedures (CP and UP) compared to the conventional SR.

3.6.1 Evaluation setup

To evaluate different ECLs, we choose three configurations from the link level evaluations available in the 3GPP references [24] and [25]. From these sources, we select the configurations that obtain similar MCLs values in our link budget for the three radio channels and the different ECLs we consider. That is, these sources provide the required Signal to Noise Ratio (SNR) for specific radio configurations of the radio channels. From these values, we estimate the MCL using a link budget. Table 3.7 summarizes the link budget of the different channels.

Note for NPDCCH we could not find a configuration for ECL 0. However, as the required SNR is similar for NPDCCH and NPDSCH, we will assume the NPDCCH has the same number of repetitions than NPDSCH for ECL 0. Additionally, due to NPDCCH and NPDSCH are the channels with fewer evaluations available in [24] and these evaluations reach a MCL of 161 dB under our link budget assumptions, we assume ECL 2 has 161 dB MCL.

For the evaluation, we assume the IAT ranges from 30 min to 25 h. Additionally, Table 3.8 summarizes the main NB-IoT parameters considered in the evaluation. The configuration of both CSS and USS is equal, the UE can retry up to $m = 7$, the back-off window W_c equals 256 ms, $\alpha = 1$, and $P_{target} = -100dB$. We assume the path loss model of [83] with a distance from the eNB of 1.5 km, 7 km, and 10 km, for ECLs 0, 1, and 2, respectively. Consequently, the obtained uplink transmission power in all ECLs reaches the maximum power P_{max} due to the assumed PL, or the number of repetitions, or because the UE is not in ECL

Table 3.7: Link budget for NPUSCH, NPDSCH, and NPDCCH [24, 25]

	NPUSCH			**NPDSCH**			**NPDCCH**	
ECL	0	1	2	0	1	2	1	2
Subcarrier spacing (kHz)	15	15	3.75	15	15	15	15	15
Modulation	QPSK	QPSK	BPSK	QPSK	QPSK	QPSK	QPSK	QPSK
MCS	9	3	0	4	4	4		
Repetitions	2	16	1	1	32	256	64	512
Number of subcarriers in a burst	12	3	1	12	12	12	12	12
Transmitter								
(1) TX power	23	23	23	35	35	35	35	35
Receiver								
(2) Thermal noise density (dBm/Hz)	-174	-174	-174	-174	-174	-174	-174	-174
(3) Receiver noise figure (dB)	3	3	3	5	5	5	5	5
(4) Interference margin (dB)	0	0	0	0	0	0	0	0
(5) Occupied channel bandwidth (kHz)	180	45	3.75	180	180	180	180	180
(6) Effective noise power = (2) + (3) + (4) + 10log((5)) (dBm)	-118.4	-124.5	-135.3	-116.4	-116.4	-116.4	-116.4	-116.4
(7) Required SINR (dB)	-0.7	-8.1	-1.9	6.9	-4.8	-9.8	-3.9	-10
(8) Receiver sensitivity = (6) + (7) (dBm)	-119.1	-132.6	-137.2	-109.5	-121.2	-126.2	-120.3	-126.4
(9) Rx processing gain	0	0	0	0	0	0	0	0
(10) MCL = (1) -(8) + (9) (dB)	142.1	155.6	160.2	144.5	156.2	161.2	155.3	161.4

0 (see subsection 3.2.3 for power control details).

As previously mentioned, the analysis is done for three data transmission procedures (i.e. SR, UP, and CP) and four communication cases (i.e. `UL`, `UL-ACK`, `DL`, and `DL-ACK`). Following the Mobile Autonomous Reporting (MAR) periodic traffic model from [83], the application payload size of the data report is 20 bytes and the application acknowledgement size is assumed to be 0 bytes. The total size of the data report and acknowledgement considering the IP header (29 bytes) and other protocols overheads specific of the control procedure are shown in Table 3.3. Note that the energy consumption analysis does not include UE processing consumption, nor access barring.

Furthermore, the proposed Markov chain is used to estimate the capacity of the system in the different cases considered. In the capacity analysis, we assume a fixed IAT of 1 h. We examine the capacity gain of UP and CP optimizations compared to the conventional SR. This capacity gain is obtained computing the number of UEs supported N_{UE} using (3.19). More specifically, iterating (3.19) until the outage probability $p_{outage} = 0.1$ is reached using (3.17). Next, we get N_{UE} for each control procedure and derive the gain compared to the SR as:

$$gain = \left(N_{UE}^{optimization} - N_{UE}^{SR}\right) / N_{UE}^{SR} \tag{3.43}$$

where $N_{UE}^{optimization}$ is the number of UEs using UP or CP, and N_{UE}^{SR} the number of UEs supported using SR procedure.

Table 3.8: Evaluation configuration.

		Normal	**Robust**	**Extreme**
	Coverage Enhancement Level (ECL)	0	1	2
	Target MCL (dB)	≃144	≃154	≃161
RA & Sync.	NPRACH repetitions	1	8	32
	$T_{Sync}(ms)$	327	341	597
	$T_{MIB-I}(ms)$	103	103	441
	$T_{MIB-RX}(ms)$	8	8	38
NPDCCH	R_{max}	1	64	512
	G	32	1.5	1.5
	$N_{REP_{dci}}$	1	64	512
	Agregation Level	2		
NPDSCH	MCS	4	4	4
	NPDSCH repetitions	1	32	256
	Modulation	QPSK		
NPUSCH	Number of subcarriers, N_T	12	3	1
	Subcarrier spacing (kHz), SCS	15	15	3.75
	MCS	9	3	0
	NPUSCH repetitions	2	16	1
	Modulation	QPSK	QPSK	BPSK

3.6.2 Results

This subsection presents the results obtained with the proposed NB-IoT model. The evaluation of the energy consumption uses the analysis of the subsection 3.5.3. Additionally, the capacity analysis uses the radio resources analysis of subsection 3.5.1 and the gain formula (3.43).

3.6.2.1 Energy consumption

Figure 3.13 shows the battery lifetime of an NB-IoT UE for the `UL` case. The figure presents the results of the three ECLs, the three data transmission procedures, and different values of IATs. The evaluation includes a baseline consumption due to PSM for reference and the results from [23], which uses a configuration similar to our ECL 0. The evaluation of [23] estimates the energy consumption of a UL transmission after an RA without the specific signaling required to perform the data transmission procedure. These results are included to confirm our model obtains similar battery lifetime as [23] under similar configuration.

The battery lifetime decreases significantly for small IATs. In ECL 0, for

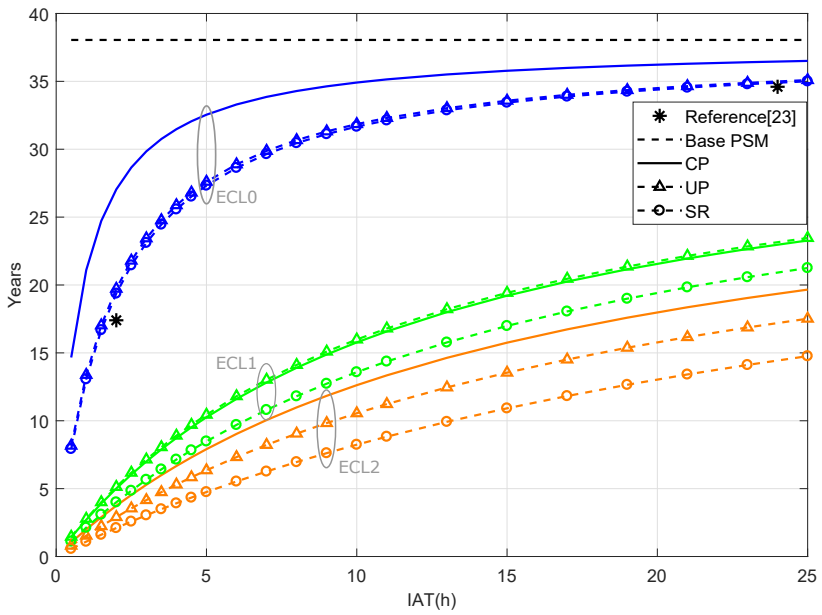


Figure 3.13: NB-IoT UE battery lifetime considering the UL case.

IATs smaller than 5–10 h the consumption caused by the synchronization time in the initial RA procedure and the inactive time spent in DRX makes up a large proportion of the overall consumption (58% in the UL case for UP with an IAT of 1 h). For IATs larger than 5–10 h, the overall energy consumption is dominated by the time spent in PSM (84% in the UL case for UP with an IAT of 1 h). The total period the NB-IoT UE spends performing DRX comprises Inactivity and Active Timers (see Table 3.6). Keeping the UE in DRX decreases the battery lifetime. However, it reduces the probability of reestablishing the connection once it has been released. This may be interesting for traffic comprising bursts of packets not considered in this analysis. Additionally, CP enables the MME to know if there is more pending traffic through the RAI notification. Then, the UE can avoid the need for DRX. For ECL 0, this improves the battery lifetime from 78% when the IAT is 30 min to 4% when the IAT is 24 h, compared to UP. If UP is configured with the same Inactivity and Active Timers as CP, both optimizations provide similar results.

When the UE has poor radio conditions and changes to a worse ECL, there is

a significant reduction of battery lifetime. This is mainly due to the repetitions required and the lower MCS used. The lower the MCS, the higher the number of RUs or SFs to cope with the packet size. For worse ECLs, the energy consumption is dominated by the messages exchanged from the completion of the RA procedure up to the end of the data transmission for almost the entire range of the evaluated IATs (35% and 49% in the `UL` case for UP with an IAT of 10 h for ECLs 1 and 2, respectively). The rest of the energy consumption is due to PSM. However, for very large IATs, the energy consumption in PSM prevails (67% and 42% in the `UL` case for UP with an IAT of 24 h for ECLs 1 and 2, respectively). Particularly for ECL 2, the considered subcarrier spacing of 3.75 kHz partly mitigates the battery lifetime reduction due to the reduced number of NPUSCH repetitions (see Table 3.8).

For the other evaluated cases (i.e. `UL-ACK`, `DL`, and `DL-ACK`), we obtain similar results for all procedures in ECL 0. Except for CP in `DL` case, where the MME cannot receive the RAI notification in the downlink NAS Protocol Data Unit (PDU). Hence, CP optimization uses the DRX mechanism as UP, and the results for both procedures become similar. For ECLs 1 and 2, `UL-ACK` results are similar to the `UL` case. However, for `DL` and `DL-ACK`, there is a significant battery lifetime reduction compared to the `UL` case (up to 30% in `DL` case, and 55% in `DL-ACK`). This is due to downlink traffic being handled by means of periodic TAUs, which implies the transmission/reception of heavy signaling packets.

From the results, we can conclude there is not only one solution able to optimize the energy consumption for all the cases considered. Depending on the UE traffic, some features have to be optimized and others are irrelevant. For example, the energy consumption while the UE is in PSM is crucial for UEs with large IATs. This energy can be reduced if the design of the UE enables an ultra-low power consumption while the UE is in deep sleep. Additionally, the extension of the allowed values in the configuration of the periodic TAU timer can help to save energy consumption if the UE has dominant UL traffic. However, for UEs with small IATs, the signaling reduction is the key issue to optimize.

3.6.2.2 Capacity analysis

Figure 3.14 shows the CP and UP capacity gain relative to the SR procedure for different ECLs and all cases considered in this work, assuming an IAT of 1 h.

From all cases evaluated, `UL` shows the best results, as the CP and UP achieve the greatest reduction in signaling compared to the SR. For example, in `UL` case the capacity gain relative to SR for both procedures reach 162% and 120% in ECL 0. The signaling inefficiency of SR was also shown in [85] where the SR procedure was compared to an assumed lightweight signaling access for an LTE system.

Our results show the radio channel limiting the capacity varies depending on the considered case and ECL. Furthermore, the use of the radio channels is different for each procedure. For ECL 0 and `UL` case, the capacity is mostly limited by the uplink channels' resources. On the contrary, for the `UL` case and the rest of ECLs, the capacity limitation comes from the downlink. This limitation is due to the number of repetitions in both NPDCCH and NPDSCH and the sharing of downlink SFs with downlink signals. Moreover, the CP procedure reduces the required resources at NPDCCH compared to CP and SR. Therefore, for worst ECLs increasingly limited by NPDCCH resources, its gain keeps increasing too.

Regarding uplink and downlink evaluation, downlink gains reach something less than two times the gains of the uplink. This is due to downlink traffic being

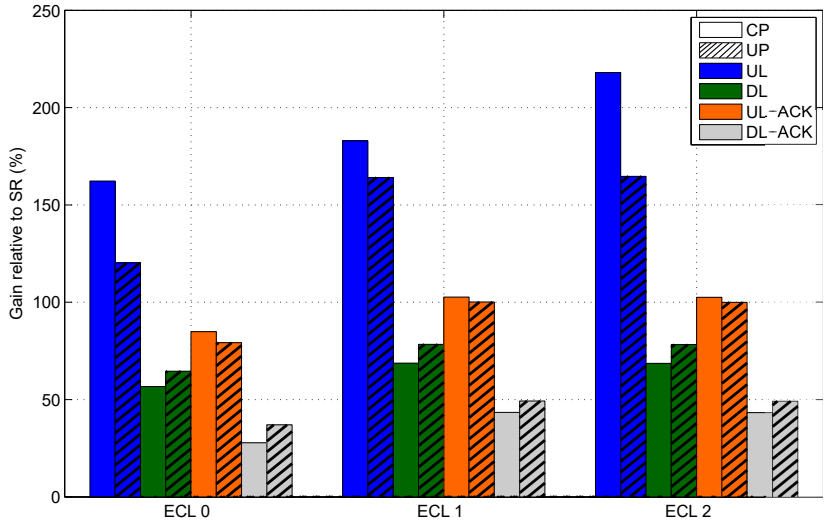


Figure 3.14: CP and UP capacity gain relative to SR in different ECLs and cases.

handled by means of periodic TAUs. This causes additional signaling for the network in both uplink and downlink channels. Particularly, in the `DL` and `DL-ACK` cases, CP achieves slightly worse results than UP. This is because CP increases the load at NPDSCH, compared to UP, and both cases are more intensive in NPDSCH.

Additionally, for larger data payloads than the 20 bytes evaluated, there would be a significant UE battery lifetime reduction if the number of NPUSCH repetitions is high. In terms of capacity, CP and UP would achieve a lower capacity gain relative to SR in cases limited by uplink channels. This is because the signaling overhead could be negligible compared to the data transferred.

3.7 Conclusions

The large-scale deployments of mMTCs involve several challenges on cellular networks. To address these challenges, the addition of NB-IoT provides a set of specifications with the potential to support new IoT use cases. As a new technology, there is still open questions about its possibilities, its performance, or its issues. In this chapter, an NB-IoT analytical model is presented. The goal of this model is to provide a tractable model to analyze the performance of NB-IoT in terms of the UE energy consumption. The analysis is extended to consider the NB-IoT mandatory Control Plane optimization (CP) procedure and two additional data transmission procedures: Service Request (SR) and User Plane optimization (UP). The evaluation of the three procedures is used to compare the conventional data transmission through the SR procedure and the two optimizations UP and CP. The analytical model includes several details of the steps the UE has to perform to complete the data transmission procedures. However, we do not include in the model the UE's processing consumption, nor access barring.

Regarding battery lifetime results, in ECL 0 for IATs larger than 5–10 h the overall energy consumption is dominated by the energy spent in PSM. On the contrary, for small IATs, the consumption caused by the data transmission procedure prevail. For ECLs 1 and 2, the energy consumption is dominated by the messages exchanged after the RA for almost the entire range of evaluated IATs. However, for very long IATs, PSM still dominates the consumption.

Regarding the cell capacity evaluation, the results highlight both optimizations reach considerable capacity gain relative to SR. In the UL case, CP and UP achieve gains of 162% and 120% in ECL 0, respectively. If PSM is used to extend battery lifetime, downlink gains reach something less than two times the gains of the uplink. This is due to the additional signaling generated to perform the TAU to manage downlink traffic.

The comparison of CP and UP optimizations yields similar results, except for some specific configurations. CP achieves up to 78 % of battery lifetime improvement in the UL case and ECL 0 due to its RAI indication. Furthermore, the use of fewer resources at the NPDCCH improves CP's cell capacity gain for the UL case. However, CP is not convenient for long data transmissions, as the network is expected to force the UE to establish the data bearers if a maximum number of messages is exceeded.

3.7.1 Resulting research contributions

The research contributions resulting from the work done in this chapter are listed below:

- P. Andres-Maldonado, P. Ameigeiras, J. Prados-Garzon, J. J. Ramos-Munoz and J. M. Lopez-Soler, "Optimized LTE data transmission procedures for IoT: Device side energy consumption analysis," 2017 IEEE International Conference on Communications Workshops (ICC Workshops), Paris, 2017, pp. 540-545.

  DOI: 10.1109/ICCW.2017.7962714

- P. Andres-Maldonado, P. Ameigeiras, J. Prados-Garzon, J. Navarro-Ortiz and J. M. Lopez-Soler, "Narrowband IoT Data Transmission Procedures for Massive Machine-Type Communications," in IEEE Network, vol. 31, no. 6, pp. 8-15, November/December 2017.

  DOI: 10.1109/MNET.2017.1700081

Chapter 4

Performance Evaluation of Extended Coverage in NB-IoT

In several Internet of Things (IoT) use cases, an IoT device must be reachable even if it is deployed in remote locations or indoor environments. This implicates the need for an extended coverage at the same time the required battery lifetime is satisfied. For most wireless IoT devices, this is an issue, especially for IoT devices with costly access for maintenance.

Theoretically, each doubling of repetitions should provide a 3dB gain in signal energy if we have a perfectly coherent combining of the received information bits. As Narrowband Internet of Things (NB-IoT) targets devices with poor Signal to Noise Ratio (SNR) conditions, the combining performance depends on the quality of the Channel Estimation (CE) at the receiver. This limits the coverage improvement as the performance of the channel estimator is expected to be poor when the signal power is much smaller than the noise power [14].

Additionally, NB-IoT is designed for allowing relaxed oscillator accuracy in the device [13]. However, low-cost oscillators are generally incapable of realizing temperature compensation. Consequently, they can have frequency drifts resulting from self-heating. The possible mismatch of the clock signal generated by the NB-IoT User Equipment (UE) and the evolved NodeB (eNB) introduces clock error. The clock error can affect the performance of the radio link due to the arising of Carrier Frequency Offset (CFO) and sampling errors in the network [100].

This chapter analyzes the coverage extension performance of NB-IoT con-

sidering ideal CE and realistic CE. In our analysis, we assume the UEs have fixed locations. These UEs will experience small CFO that is expected to be approximately constant over several Subframes (SFs) [101]. Therefore, considering the channel estimator is resistant enough for the assumed CFO, we focus on the analysis of the CE performance.

In this chapter, we first derive analytical expressions based on the Shannon theorem to study the coverage extension performance. The analysis includes the limitations due to realistic CE. We adopt the analytical bound derived in [14] to extend the analysis of the performance of NB-IoT considering non-ideal factors. Next, we use this analysis as a part of an evaluation framework that jointly illustrates the coverage extension and the resulting UE battery lifetime and latency. The aim of this analysis is threefold. Firstly, to provide an analytical evaluation framework to analyze the performance of NB-IoT. Secondly, to study the limitations and trade-offs of the repetitions in NB-IoT when considering realistic CE. Thirdly, to estimate the impact of the coverage extension in the final performance of the NB-IoT UE in terms of Uplink (UL) packet transmission latency and battery lifetime.

The rest of the chapter is organized as follows. Section 4.1 provides the fundamental background for the analysis. Section 4.2 briefly reviews the related literature. Section 4.3 explains the system model and the main assumptions for the evaluation. Section 4.4 describes the transmission properties and the non-ideal factors analyzed in this study. Section 4.5 presents the analytical evaluation framework for NB-IoT. Section 4.6 shows the results. Lastly, section 4.7 presents the main conclusions of this chapter.

4.1 Fundamentals of the analysis

4.1.1 NB-IoT pilots

To allow for coherent demodulation of the UL and Downlink (DL) channels, NB-IoT has two reference signals: Demodulation Reference Signal (DMRS) in the UL, and Narrowband Reference Signal (NRS) in the DL. The reference symbols, hereafter called in this chapter as *pilot symbols*, are inserted in the time-frequency resource grid to allow CE. Note that the quality of the estimation is limited by

the number of pilot symbols and the received SNR [102]. Figure 4.1 depicts a resource mapping example for NRS and DMRS.

NRS is included in all SFs that may be used for broadcast or DL transmission. It is mapped to specific REs of the resource grid. The exact REs used for NRS depend on the NB-IoT deployment mode, the cell identity, and the logical antenna port number [29]. The signal generation method for NRS is similar to Long Term Evolution (LTE). The NRS signal uses a pseudorandom QPSK sequence and is generated based on cell identity and port number [13, 29, 103].

DMRS is only multiplexed with the data. Depending on the Narrowband Physical Uplink Shared CHannel (NPUSCH) format, DMRS is included in either one or three Single-Carrier Frequency-Division Multiple Access (SC-FDMA) symbols per slot. Additionally, the specific SC-FDMA symbols for DMRS depend on the subcarrier spacing. Then, from these known transmitted pilot symbols, the receiver can estimate the channel response. The bandwidth and modulation of the DMRS is identical to the associated NPUSCH data. DMRS symbols are constructed from a base sequence of the form $e^{j\phi(n)\pi/4}$ multiplied by a phase factor α [13, 29, 92].

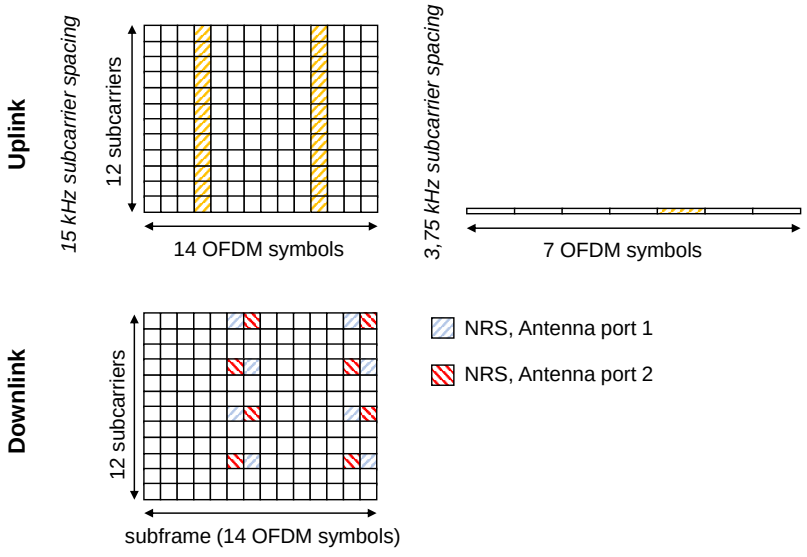


Figure 4.1: Example of REs used for DMRS and NRS [13].

4.1.2 Channel estimators

Channel Estimation (CE) is used to estimate the channel characteristics. This poses a great challenge as the channel is susceptible to several factors that vary in time such as path loss, shadowing, or multipath fading [104]. To be able to recover the distorted received signal, it is necessary to estimate and compensate the channel effect in the receiver.

In the context of NB-IoT, the channel is estimated by using pilot symbols known at the UE and eNB. From these pilot symbols, the channel response of the remain subcarriers without pilots can be estimated using interpolation techniques. Within the CE based on pilot symbols, the Least-Square (LS) and Minimum Mean Square Error (MMSE) techniques are widely used [105, 106].

In a nutshell, the LS CE method finds the channel estimate minimizing the squared differences between the received and estimated signal. This method has very low complexity but tends to amplify noise at frequencies where the transmitted symbol has low energy. The MMSE CE method uses second-order statistics of the channel (channel autocovariance) to minimize the square error. This method has a better performance than LS at the cost of a higher complexity [105, 107, 108].

To improve the CE accuracy, multiple SFs can be jointly used for the CE, known as cross-subframe CE. Using a bigger window in cross-subframe CE requires a higher baseband complexity at the receiver [13]. The performance improvement due to cross-subframe would be constrained by the coherence time of the channel, i.e., the time duration over which the propagation coefficient of the channel remains constant [105]. Consequently, the length of the cross-subframe CE time-domain filter has to be carefully chosen to avoid performance degradation.

4.1.3 Carrier frequency offset

The carrier frequency at the UE and eNB usually differ by a small amount, known as CFO. This frequency offset is mainly caused due to [101]:

- Frequency mismatch between the local oscillators at the transmitter and receiver. The causes, such as temperature or hardware impairments, are expected to be approximately constant over several SFs.

- Doppler shift as a result of the relative motion between the transmitter and receiver. For the massive MTC (mMTC) UEs considered in this analysis (i.e. small stationary battery powered devices located in remote areas), the Doppler shift can be assumed very low (1 Hz) since the UEs are assumed to stay in a fixed location.

The CFO can lead to Inter Carrier Interference (ICI) in the Orthogonal Frequency-Division Multiple (OFDM) system. The ICI implies a subcarrier frequency component can be affected by other subcarrier frequency components. Thus, the OFDM system orthogonality is degraded [109]. The receiver can use the reference signals to obtain a CFO estimation to recalibrate the internal clock. However, this technique depends on the measurement quality. Therefore, there is always a residual frequency offset that degrades the receiver performance [110]. For extreme coverage conditions, the UE may rely on accumulating detection metrics over many Primary Synchronization Signal (NPSS) SFs [13].

4.1.4 Shannon theorem

The Shannon theorem establishes a relationship between bandwidth, power, and capacity in an additive white Gaussian noise channel by the expression [111]:

$$C = BW \cdot \log_2\left(1 + \frac{S}{N}\right) \tag{4.1}$$

where BW is the bandwidth available for the communication, S is the received signal power, C is the channel capacity in bits/s, and N denotes the noise power. From (4.1), we can see the capacity is limited by the available SNR and the bandwidth. Assuming a specific data rate R_b in bits/s. The components of the SNR can be derived as:

$$\begin{aligned} S &= E_b \cdot R_b \\ N &= N_o \cdot BW \end{aligned} \tag{4.2}$$

where E_b is the received energy per information bit and N_o is the constant thermal noise density measured in W/Hz. As the data rate can never exceed the channel capacity, this yields:

$$R_b \leq C = BW \cdot \log_2\left(1 + \frac{E_b \cdot R_b}{N_o \cdot BW}\right) \tag{4.3}$$

by defining the bandwidth utilization $\gamma = R_b/BW$ in bps/Hz, the inequality can be expressed as:

$$\gamma \leq \log_2\left(1 + \gamma \cdot \frac{E_b}{N_o}\right) \tag{4.4}$$

Then, the inequality can be reformulated to obtain a lower bound on the required received energy per information bit, normalized to the noise power density as [112]:

$$\frac{E_b}{N_o} \geq \frac{2^\gamma - 1}{\gamma} \tag{4.5}$$

Additionally, using the Shannon theorem, if we assume an extreme coverage condition where $\frac{S}{N} \ll 1$, using the approximation $\ln(1+x) \approx x$ for $x \ll 1$, the channel capacity under the low SNR range can be estimated as:

$$C = \frac{S}{N_o}\log_2(e) \tag{4.6}$$

From (4.6) we can observe the channel capacity only depends on the ratio of S and N_o in this range of low SNR. This means at extreme radio conditions it is better to reduce the bandwidth allocated to a UE to be more spectrally efficient [13].

4.2 Related works

NB-IoT provides substantially enhanced coverage compared to normal LTE. As a key pillar of the Low-Power Wide-Area (LPWA) technologies, it is very important to understand how much coverage NB-IoT can actually provide and its cost.

Recent works on NB-IoT present an insight on different issues related to the coverage enhancement. In [113], *Lauridsen et.al.* evaluate the coverage and capacity performance of NB-IoT and LTE-M. They simulate the configuration of a real operator deployed base stations located in a rural area, and calibrate the simulation using drive test measurements. Their simulation results show

NB-IoT can provide better coverage performance than LTE-M. However, the cost of providing deep indoor coverage in NB-IoT involves not only a lower number of supported UEs per sector due to overheads, but also 2-6 times higher UE power consumption as compared to LTE-M.

Vejlgaard et.al. in [114] compare the coverage and capacity of SigFox, LoRa, GPRS and NB-IoT. They simulate the link loss between the UEs and the base stations, and compare it with the link budget of each technology. The base stations locations are based on the Telenor's sub 1 GHz cellular network grid in North Jutland, Denmark. The results show all four technologies were able to reach 99% outdoor coverage. However, for indoor coverage, GPRS is unable to provide coverage for 40% of the UEs, while Sigfox, LoRa, and NB-IoT cover more than 95% of the UEs experiencing 20 dB penetration loss.

In [115], *Adhikary et.al.* provide a detailed evaluation of the coverage performance of NB-IoT. Their performance evaluation includes inter-cell interference, a practical channel estimator, and frequency and timing errors. The simulations show NB-IoT achieves a coverage enhancement of up to 20 dB compared with the current LTE system in various deployment scenarios. Additionally, they illustrate the SNR degradation when restricting the channel estimation window at the cost of lower complexity, and prove that NB-IoT provides good co-existence performance with the existing LTE system.

Ali et.al. in [116] propose two DMRS assisted channel estimation algorithms as modifications of the traditional LS and MMSE estimators. They test their proposed algorithm by theoretical analysis and link-level simulations. The authors corroborate by simulations their proposed algorithms provide better estimation precision compared to the traditional LS and MMSE estimators at low SNR conditions.

Additionally, in [117], *Rusek et.al.* propose a sequential MMSE estimator. They consider the presence of random phase noise caused by the fluctuations of oscillators and the residual frequency offset. According to their simulation results, their proposed sequential MMSE estimator improves the channel estimation Mean Square Error (MSE) by 1 dB in the low SNR range, compared to the traditional sequential MMSE estimator that does not thoroughly consider the impact of random phase noises.

In [14], *Beyene et.al.* show the NB-IoT coverage enhancement is limited by

the channel estimation quality. Their results are obtained through simulations and measurement from a Software Defind Radio (SDR) based implementation of NB-IoT testbed. The authors derive an analytical bound for the SNR gain from repetitions considering the channel estimation quality.

Other pioneering work has been conducted in the NB-IoT interference mitigation scheme [118], cell search and initial synchronization [119], physical channel performance evaluation by simulation [94, 120, 121], and link adaptation [122].

As can be seen, recent works in the literature address the coverage performance of NB-IoT channels and provide an insight on this topic. The analysis presented in this chapter is complementary to the existing literature and joins the analysis of the challenges when extending the coverage (i.e. realistic channel estimation and cross-subframe estimation) with the impact these challenges impose at the UE side.

4.3 System model

Let us assume a cell with an eNB with an NB-IoT carrier deployed in-band, and one UE camping on it. We assume a UE in a fixed location that transfers/receives one report of size b periodically to/from the eNB. Therefore, prior UL data, the UE needs to reestablish the Radio Resource Control (RRC) connection between the UE and the eNB. To do that, we assume the UE performs an RRC Resume procedure, i.e., the UE uses the User Plane optimization (UP) optimization. The rationale assuming this optimization instead of the NB-IoT's mandatory Control Plane optimization (CP) is the later comparison of the results with a 3GPP evaluation. Please note that the analysis presented in this chapter focuses on the study of the UL transmission, although the analysis is reciprocal for DL considering SFs instead of Resource Units (RUs), and the specific parameters of DL.

We adopt the following notations. Bold upper case letters stand for matrices. Superscripts $(.)^H$ and $(.)^{-1}$ respectively denote the Hermitian transpose and matrix inverse, whereas the operator $\mathbb{E}\{.\}$ denotes the expectation.

Additionally, we assume a very slowly time-variant channel and low Doppler frequency (1Hz). We only consider channel losses because of path loss, denoted as L, then $MCL = L$. To compensate channel losses, the UE adjusts its transmission

power P_{TX} up to a maximum allowed value P_{max}. Consequently, the SNR, denoted as $\frac{S}{N}$, can be calculated as

$$\frac{S}{N} = \frac{P_{TX}}{L \cdot F \cdot N_o \cdot BW} \tag{4.7}$$

where N_o is the thermal noise density, F is the receiver noise figure, $BW = SCS \cdot N_T$ is the allocated bandwidth, SCS is the subcarrier spacing, and N_T is the number of tones. Figure 4.2 depicts the overall system model. Additionally, Table 4.1 provides the adopted notation.

When the eNB configures UL transmissions, we consider four parameters: number of RUs, Modulation and Coding Scheme (MCS) level, the bandwidth allocated, and the number of repetitions. For DL receptions, we consider three parameters: number of SFs, MCS level, and number of repetitions. When applying repetitions, we assume all repetitions have the same Redundancy Version (RV). For both the DL and UL, the same information is included in each repetition and combined at the receptor using Chase Combining. In this analysis, we consider QPSK modulation is always used. Thus, the combination of the MCS, number of RUs, and allocated bandwidth determine the data rate of the transmission R_b, derived as:

$$R_b = \frac{b + CRC}{N_{RU} \cdot T_{RU}} \tag{4.8}$$

where R_b is measured in bits/s, b is the size of the data packet in bits, CRC is the size in bits of the Cyclic Redundancy Check code, N_{RU} is the number of RUs allocated to the UE, and T_{RU} is the duration in seconds of an RU. Note that the duration of the RU depends on the bandwidth allocated to the UE (see Table 3.1 in previous chapter). As the number of tones decreases, T_{RU} increases. Herein,

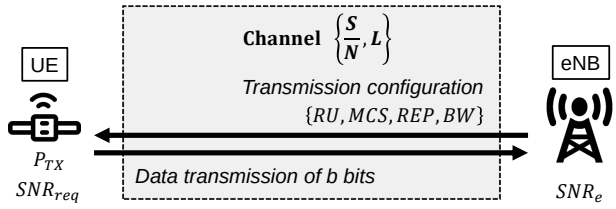


Figure 4.2: System model.

Table 4.1: Main definitions

Notation	Description
L	Path loss
MCL	Maximum Coupling Loss
P_{TX}	Transmission power
P_{max}	Maximum transmission power
F	Noise figure
N_o	Thermal noise density
BW	Bandwidth
SCS	Subcarrier spacing
N_T	Number of tones
R_b	UE's data rate
b	UE's data packet size
CRC	Cyclic Redundancy Check code
N_{RU}	Number of RUs
T_{RU}	Duration of an RU
SNR_{req}	Required SNR at the UE
SNR_e	SNR at the eNB
BW_{eff}	Bandwidth efficiency
SNR_{eff}	SNR required efficiency
N_{REP}	Number of repetitions
σ	Channel estimation error
σ^2	Mean square error
$\mathbf{H}$	Channel response vector
$\hat{\mathbf{H}}$	Estimated channel response vector
γ	Bandwidth utilization
$\hat{E}_b$	Energy per transmitted bit
W_{cs}	Cross-subframe window
η_p	Cross-subframe correction factor

we denote this dependency on the bandwidth as $T^{(BW)}$.

The selected configuration of the transmission parameters determines the $\frac{S}{N}$ at the UE's receiver. We define SNR_{req} as the minimum $\frac{S}{N}$ to successfully decode the UL transmission, then $\frac{S}{N} \geq SNR_{req}$. When applying repetitions or bandwidth reduction, the values of SNR_{req} and $\frac{S}{N}$ can be modified. Specifically

for UL repetitions, the same data is repeatedly transmitted N_{REP} times. The received transmission's copies at the eNB can be combined to raise error correction. The resulting SNR after the coherent combining of the copies is defined as effective SNR, denoted as SNR_e. For ideal CE, the SNR_e can be expressed as:

$$SNR_e = \sum^{N_{REP}} SNR_{req} = N_{REP} \cdot SNR_{req} \tag{4.9}$$

For realistic CE, there is an estimation error, denoted as σ. This CE error will impact the system's performance [123] and limit the gain from repetitions [124]. To model realistic CE, both transmit and receive chains and the propagation channel are shown in Figure 4.3.

Let us consider the received UL pilot signal in frequency domain as (in matrix notation):

$$\mathbf{Y} = \mathbf{XH} + \mathbf{Z} \tag{4.10}$$

where $\mathbf{Y}$ is the received pilot signal vector given as $\mathbf{Y} = [Y[0], Y[1], \cdots, Y[N-1]]^T$. The notation $Y[k]$ indicates the received pilot tone at the kth subcarrier. The diagonal matrix $\mathbf{X}$ is the transmitted pilot signal, with $\mathbb{E}\left\{|X[k]|^2\right\} = 1$, and $\mathbf{Z}$ denotes the Gaussian noise vector given as $\mathbf{Z} = [Z[0], Z[1], \cdots, Z[N-1]]^T$, with $\mathbb{E}\{Z[k]\} = 0$ and $Var\{Z[k]\} = \left(\frac{S}{N}\right)^{-1}$. The variance is obtained from the L assumed in the channel. Furthermore, $\mathbf{H}$ is the frequency response of the channel vector given as $\mathbf{H} = [H[0], H[1], \cdots, H[N-1]]^T$.

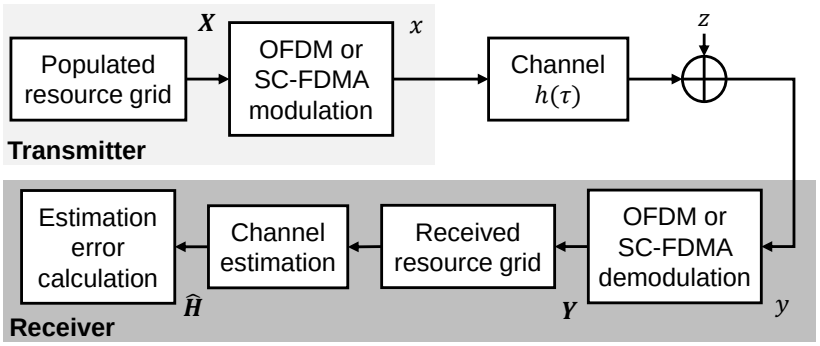


Figure 4.3: Realistic CE block diagram.

We consider a fading multipath channel model of M distinct paths [125]. The channel impulse response during the OFDM symbol in the time domain can be expressed as

$$h(\tau) = \sum_{i}^{M} a_i \delta(\tau_i) \tag{4.11}$$

where a_i and $\delta(\tau_i)$ respectively represent the path gain and the propagation delay of the ith path.

In the channel model, the total power is normalized to one. $\mathbf{H}$ is the Fast Fourier Transform (FFT) transformation of $h(\tau)$, *i.e.*, $\mathbf{H} = FFT_N[h(\tau)]$. For realistic CE, the channel estimate, denoted as $\hat{\mathbf{H}}$, can be modeled as:

$$\hat{\mathbf{H}} = \mathbf{H} + \mathbf{H_e} \tag{4.12}$$

where $\mathbf{H_e}$ is the estimation error which has a variance σ.

We assume a LS channel estimator. Therefore, from the known reference UL pilot symbols, we estimate the channel response as $\hat{\mathbf{H}} = \mathbf{X}^{-1}\mathbf{Y}$ [105].

From the channel estimation, the MSE of the channel estimator σ^2 can be expressed as:

$$\sigma^2 = \mathbb{E}\left\{\left(\mathbf{H} - \hat{\mathbf{H}}\right)^H \left(\mathbf{H} - \hat{\mathbf{H}}\right)\right\} \tag{4.13}$$

Note that for the DL, the channel coefficients of the subcarriers without pilot symbols have to be obtained by means of linear interpolation. Then, from the computed MSE, we obtain the CE error as $\sigma = \sqrt{\sigma^2}$.

Finally, we adopt the analytical bound for the SNR gain from signal repetition defined in [14]. This analytical bound provides the approximated SNR_e from the $\frac{S}{N}$, the CE error σ, and the number of repetitions N_{REP} [14]:

$$SNR_e = \frac{N_{REP} \cdot \left(\sigma + \frac{S}{N}\right)}{\left(\sigma + 1 + \frac{\sigma}{\frac{S}{N}}\right) \cdot \left(1 + \frac{\sigma}{2 \cdot \frac{S}{N}}\right)} \tag{4.14}$$

Figure 4.4 shows an example of the obtained SNR_e using (4.14) for two different values of the number of repetitions and estimation error. We can observe SNR_e is impacted by the $\frac{S}{N}$ and σ. Moreover, as $\frac{S}{N}$ decreases, the quality of the

channel estimation decreases thereby increasing the estimation error σ.

If $\frac{S}{N} = SNR_{req}$, from equations (4.9) and (4.14) we can observe the relationship between SNR_{req} and SNR_e when considering repetitions for ideal and realistic CE. Unlike the analytical bound for the SNR gain when doubling repetitions of (4.14), we consider there is a direct improvement on the SNR_{req} when bandwidth reduction technique is applied for both ideal and realistic CE, and the transmission power is maintained. This improvement is due to uplink Power Spectral Density (PSD) boosting as the bandwidth reduction technique concentrates a given power on a narrower bandwidth [126]. This PSD boost can be calculated using (4.7).

4.4 Analytical transmission analysis

The presented analysis is based on the Shannon theorem. We consider three UL transmission configuration approaches: i) RU number modification; ii) bandwidth reduction; and iii) repetitions. Figure 4.5 shows an example of the possible UE transmission configuration to enhance its coverage with these approaches. These three approaches are represented in the analysis as $(.)^{(RU,BW,REP)}$, respectively. Thus, from the Shannon theorem, we obtain the minimum bounds of three transmission properties:

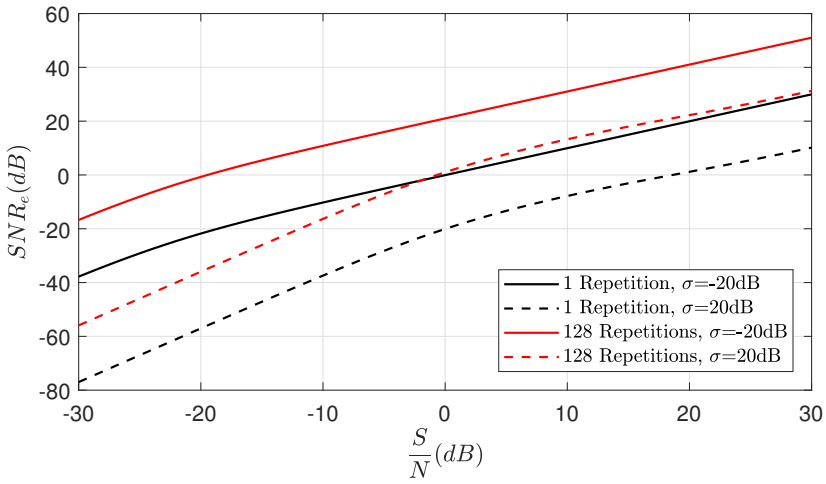


Figure 4.4: SNR after coherent combining [14].

- The required SNR, denoted as SNR_{req}.
- The energy per transmitted bit $\hat{E}_b$.
- The bandwidth utilization γ.

To satisfy the analysis, rearranging (4.7) and assuming $\frac{S}{N} = SNR_{req}$, the transmission power is computed as:

$$P_{TX} = SNR_{req}^{(RU,BW,REP)} \cdot L \cdot F \cdot N_o \cdot BW \tag{4.15}$$

However, equation (4.15) does not include the constraints of the 3GPP's open loop UL power control mechanism for NB-IoT defined in [21] (see subsection 3.2.3). Therefore, for higher values of P_{TX}, the resulting values of the parameters analyzed will exceed the values obtained through Shannon theorem.

Based on the Shannon theorem and assuming ideal CE, the SNR at the eNB, denoted as SNR_e, can be derived as:

$$SNR_e = 2^{R_b/BW} - 1 = 2^{\frac{b+CRC}{BW \cdot N_{RU} \cdot T_{RU}^{(BW)}}} - 1 \tag{4.16}$$

By substituting (4.9) into (4.16), the required SNR for a UL transmission at the UE can be expressed as:

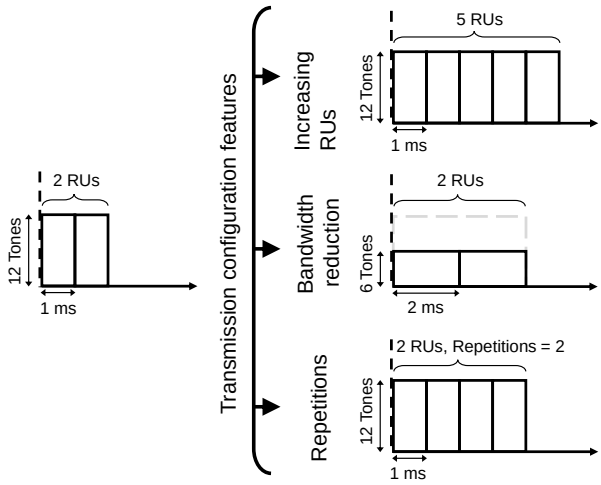


Figure 4.5: Example of NPUSCH transmission approaches in NB-IoT.

$$SNR_{req}^{(RU,BW,REP)} = \frac{2^{R_b^{(RU,BW,1)}/BW} - 1}{N_{REP}} = \frac{2^{\frac{b+CRC}{BW \cdot N_{RU} \cdot T_{RU}^{(BW)}}} - 1}{N_{REP}} \tag{4.17}$$

where $(.)^{(RU,BW,1)}$ denotes the number of repetitions is equal to its minimum $N_{REP} = 1$. From (4.17), ideally, doubling repetitions can bring about 3dB gain. From the data rate of the UE, we can obtain the bandwidth utilization γ of the transmission, given by:

$$\gamma^{(RU,BW,REP)} = \frac{R_b^{(RU,BW,1)}}{N_{REP} \cdot BW} = \frac{b + CRC}{N_{REP} \cdot BW \cdot N_{RU} \cdot T_{RU}^{(BW)}} \tag{4.18}$$

Furthermore, let us now estimate the energy per bit to noise power spectral density ratio $\frac{E_b}{N_o}$. As previously explained in subsection 4.1.4, the received energy per information bit is derived as $E_b = S/R_b$, and the noise power can be expressed as $N_0 = N/BW$. Thus, the $\frac{E_b}{N_o}$ can be computed as $\frac{E_b}{N_o} = \frac{S}{N}/\gamma$. Let $\frac{E_b}{N_o}$ be the lower bound of the received energy per bit to noise power spectral density ratio, and $\hat{E_b}$ be the energy per transmitted bit, then:

$$\begin{aligned} \hat{E_b}^{(RU,BW,REP)} &= \frac{E_b}{N_o}^{(RU,BW,REP)} \cdot L \cdot F \cdot N_o \\ &= \frac{2^{\frac{b+CRC}{BW \cdot N_{RU} \cdot T_{RU}^{(BW)}}} - 1}{\frac{b+CRC}{BW \cdot N_{RU} \cdot T_{RU}^{(BW)}}} \cdot L \cdot F \cdot N_o \end{aligned} \tag{4.19}$$

From the previous analysis we can derive some observations:

- $\hat{E_b}^{(RU,BW,REP)}$ in (4.19) is no longer a function of the number of repetitions. However, the utilization of repetitions reduces the SNR_{req} at the expense of the reduction of γ.
- As the number of RUs is greater, the transmission properties analyzed (i.e. SNR_{req}, γ, and $\hat{E_b}$) decrease.
- The increase of the RU's duration $T_{RU}^{(BW)}$ is the same as the reduction in the bandwidth for all multi-tone configurations. Therefore, while the UE

maintains a multi-tone configuration, the analyzed transmission properties preserve their values. This holds true if the transmission power is reduced in accordance with the bandwidth. However, when moving from multi-tone to single-tone configuration, the increase of $T_{RU}^{(BW)}$ and the reduction of the bandwidth is unequal. Thereby, single-tone configurations present higher SNR_{req}, γ, and $\hat{E}_b$ than multi-tone configurations. If the transmission power is maintained, this approach concentrates the limited power on a narrower bandwidth. This enhances the received SNR, and thus the coverage can be extended.

Let us now consider realistic CE. The next subsections detail the non-ideal factors considered in the analysis.

4.4.1 Shannon capacity fitting

The Shannon bound is the theoretical maximum data rate of the channel for a given SNR and bandwidth. Although LTE is near the Shannon bound, NB-IoT performance is to be analyzed and compared to the Shannon bound. In order to capture this factor accurately, we use a modified Shannon capacity formula. This approximation (4.20) was originally proposed in [127] for LTE. Therefore, the modified Shannon formula is as follows:

$$C = BW \cdot BW_{eff} \cdot \log_2\left(1 + \frac{SNR_e}{SNR_{eff}}\right) \tag{4.20}$$

where C is the capacity of the channel measured in bits/s, BW_{eff} is the bandwidth efficiency of the used technology (i.e. NB-IoT), and SNR_{eff} is the efficiency of the SNR in NB-IoT.

Once all the parameters of the modified Shannon bound are estimated, the SNR_e can be calculated knowing the bit rate and the bandwidth of the transmission or reception. Considering the UE's data rate $R_b = C$ and rearranging (4.20), the SNR_e can be expressed as:

$$SNR_e = \left(2^{\frac{R_b}{BW \cdot BW_{eff}}} - 1\right) \cdot SNR_{eff} \tag{4.21}$$

4.4.2 Channel estimation

In order to include the effect of the presence of CE errors in realistic CE, we adopt the analytical bound for the SNR gain from signal repetition defined in [14] and shown in Equation (4.14). Then, by substituting (4.21) into (4.14), we obtain the approximation of the SNR_{req} for realistic CE:

$$\left(2^{\frac{R_b}{BW \cdot BW_{eff}}} - 1\right) \cdot SNR_{eff} = \frac{N_{REP} \cdot (\sigma + SNR_{req})}{\left(\sigma + 1 + \frac{\sigma}{SNR_{req}}\right) \cdot \left(1 + \frac{\sigma}{2 \cdot SNR_{req}}\right)} \tag{4.22}$$

In this case there is not a simple solution when solving equation (4.22) for SNR_{req}. Then, we obtain SNR_{req} through an iterative method. Note that σ depends on the SNR_{req}. This is due to the quality of the CE depends on the amplitude of the received pilot symbols, and therefore, the SNR_{req}.

At this point of the analysis, we need to know the dependency of σ and SNR_{req} to be able to use (4.22). To do that, we develop two simulators, one for Orthogonal Frequency-Division Multiple Access (OFDMA) for the DL channels and another for SC-FDMA for the UL channels. The goal of these simulators is to emulate the transmission and reception chains of both systems to estimate the MSE the CE has under different conditions. Later, from the MSE we can obtain σ. Thereby, the main parts of interest of both simulators are: i) Generation of the NB-IoT pilots for the UL and the DL as described in [29], ii) Specific processing of the signal as required for each system, and iii) Obtain pilot estimates using LS estimation and the associated MSE of the estimation. Figure 4.6 illustrates a simplified diagram of the main functional blocks the simulators include to estimate the MSE.

After conducted simulations (see subsection 4.6.1 for details for the simulations configuration), we found the dependency between the σ and the SNR_{req} for the assumed LS CE can be approximated in dB as a linear dependency, given by:

$$\sigma_{dB} = c_1 \cdot SNR_{req,dB} + c_2 \tag{4.23}$$

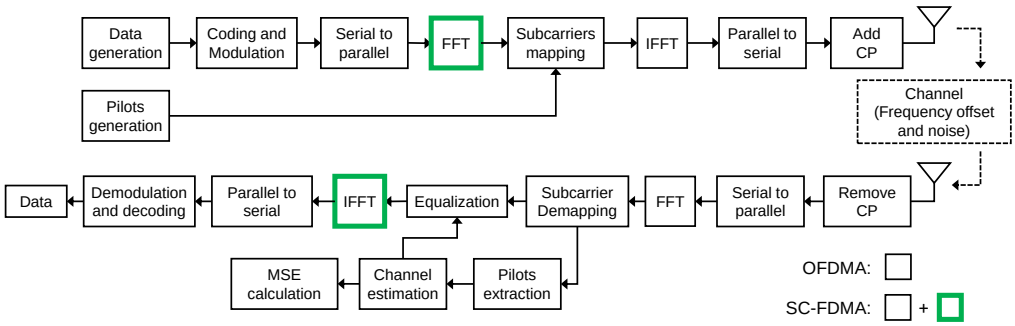


Figure 4.6: Structure of the OFDMA and SC-FDMA simulators.

where $SNR_{req,dB}$ and σ_{dB} are measured in dB, and c_1 and c_2 are constants that depend on the cross-subframe window used in the CE and the modulation technique. Figure 4.7 shows an example of the linear dependency between $SNR_{req,dB}$ and σ_{dB} using our SC-FDMA simulator and LS estimation. This linear behavior can be seen also in other works of the literature such as [128, 129].

4.4.3 Cross-subframe

To minimize the effects of noise on the realistic CE, we consider the CE performance can be improved using multiple consecutive SFs for the estimation, also known as cross-subframe CE. Under the hypothesis of a slowly time-variant

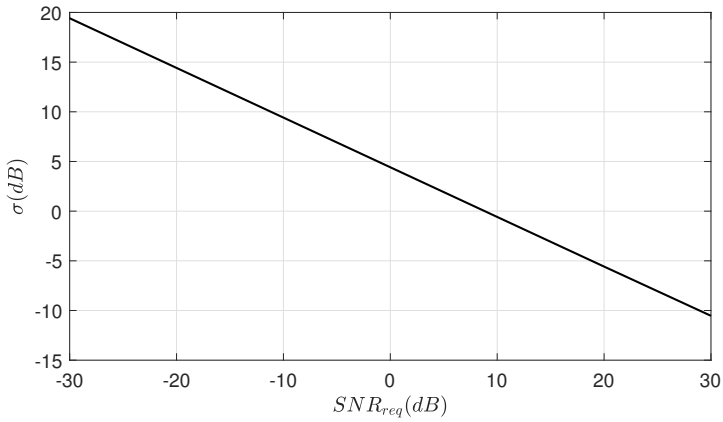


Figure 4.7: σ as a function of the SNR_{req} using the SC-FDMA simulator and LS estimation.

channel, the utilization of cross-subframe CE can produce a substantial noise reduction.

Figure 4.8 depicts a simplified block diagram of the steps performed for uplink cross-subframe CE considering a cross-subframe window $W_{cs} = 2$. The W_{cs} value denotes the number of resources to be considered in the cross-subframe CE. For the UL, these resources are the number of RUs. For the DL, the number of SFs. Following the steps described by Figure 4.8, we obtain the CE error from the MSE for a specific W_{cs} configuration. To do that, the previous mentioned simulators for OFDMA and SC-FDMA perform the following steps:

1. Start a channel realization out of K realizations. For each realization, both simulators will follow the actions described in Figure 4.6.

2. Extract the pilot symbols from their known location within the received matrix of symbols. These received pilots are used to obtain the LS estimation of the channel in these positions, i.e., pilot estimates. Note in the DL the pilots are located in specific time and frequency resources. Therefore, in this case we also need to use interpolation to estimate the entire SF.

3. The pilot estimates are averaged within the cross-subframe window considered to obtain the final CE. From the CE, the associated MSE of the estimation is computed.

4. Finally, if the Kth realization finished, all obtained MSEs are averaged.

When we consider cross-subframe technique in our evaluation, the configuration of the transmission/reception determines the value of W_{cs}. For example, in the UL, W_{cs} is derived as:

$$W_{cs} = min\left(W_{cs}^{max}, 2^{\lfloor \log_2(N_{RU} \cdot N_{REP} \cdot \eta_p) \rfloor}\right) \tag{4.24}$$

where W_{cs}^{max} is the maximum cross-subframe window considered, N_{RU} the number of RUs, N_{REP} the number of repetitions, and η_p is a correction factor. This correction factor is used to include single-tone configurations have less DMRS symbols than multi-tone configurations.

The simulation of different values of W_{cs} provides a set of c_1 and c_2 values of equation (4.23), as can be seen later in subsection 4.6.1. As an example, Figure 4.9

shows the benefit of the cross-subframe technique reducing the resulting CE error σ as the considered window increases. For the largest cross-subframe window of 16 ms, the obtained σ decreases 6 dB compared to single frame estimation. Finally, the value of W_{cs} used in the transmission/reception determines the equation (4.23) to be used in (4.22) when we estimate the SNR_{req}.

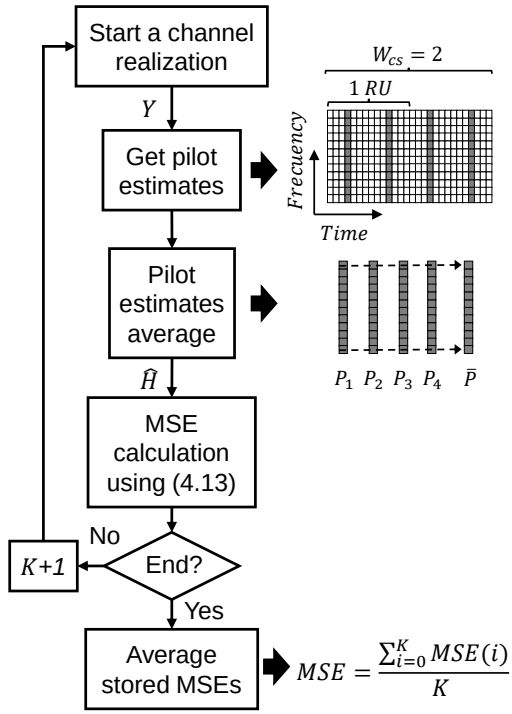


Figure 4.8: Block diagram of a UL CE using cross-subframe with a window of 2, where P_n represents the nth DMRS vector and $\bar{P}$ the time averaged vector.

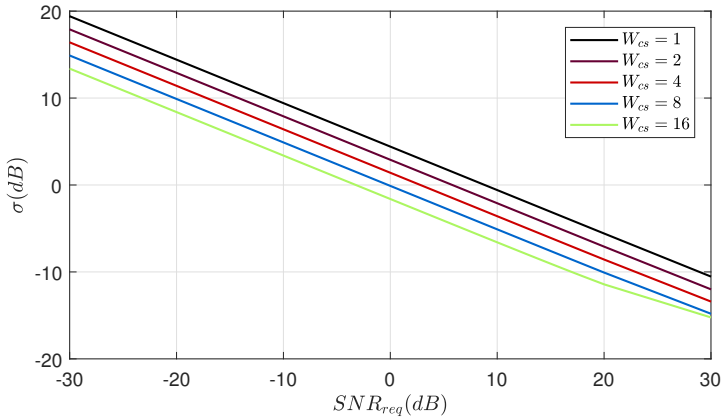


Figure 4.9: CE error σ as a function of the SNR_{req} using the SC-FDMA simulator, LS estimation, and different cross-subframe windows W_{cs}.

4.4.4 Uplink link adaptation

Link adaptation is essential to adjust the transmission to the channel conditions. However, the existing link adaptation mechanisms proposed for LTE do not consider the new dimensions NB-IoT adds [84, 122]. Consequently, as part of the analysis of this chapter we propose a UL link adaptation. We consider the UL link adaptation can be performed in three dimensions: i) MCS or RUs, ii) bandwidth allocated, and iii) repetitions.

To complete the link adaptation, we assume three phases depicted in Figure 4.10. Firstly, from the 3GPP's Transport Block Size (TBS) table for NB-IoT [21], we calculate the R_b corresponding to each combination of MCS and number of RUs allowed. Secondly, from the analysis of section 4.4, we estimate the SNR_{req} for each position of the TBS table. This is done considering the bandwidth reduction and repetitions approaches too. Later, the algorithm searches the optimal configuration using as a criterion the minimization of the transmission time. The rationale for this criterion is that in several use cases (specifically for the coverage extension scenarios we study in this chapter), the resulting transmission power from the uplink power control will quickly use the maximum allowed power (e.g. when the UE may need more than 2 repetitions or the UE is not in the best coverage level). Thus, it is important to reduce the UE time on air to decrease the power consumption.

The algorithm takes as inputs the needed SNR SNR_{in} (i.e. the maximum allowed SNR_{req} at the UE for the UL transmission that has to be satisfied), and the size of the packet b. The following Algorithm 1 shows the pseudo code of our proposed link adaptation. As a result of the three considered dimensions in the UL link adaptation, the algorithm relies on different flags to search more than one possible solution. That means once one candidate point is found, depending on the resultant parameters' configuration of the point, the algorithm could search more candidate points in the next number of repetitions, or radio configuration, or using less RUs. For example, if the candidate point is found in the last NB-IoT radio configuration (i.e. single-tone and 3.75 kHz subcarrier spacing), the next number of repetitions allowed is checked considering up to the half of the number of RUs of the previous candidate point to obtain a lower time on air than the first candidate found. Then, from the found solutions, the optimal solution is selected. Note that we give priority to bandwidth reduction as it preserves the bandwidth utilization.

Despite this algorithm minimizes the transmission time, the energy consumption is more critical in NB-IoT. Using the estimated $\hat{E}_b$, we could optimize the energy consumption. However, both criteria will obtain very similar link adaptation results. From the analysis of section 4.4, we can estimate the packet energy consumption as $\hat{E}_b \cdot (b + CRC)$. Moreover, the $\hat{E}_b$ and the transmission power P_{TX} are related as $P_{TX} = \hat{E}_b \cdot R_b$. If we estimate the $\hat{E}_b$ using (4.19) and later we recalculate the $\hat{E}_b$ values that exceed the maximum transmission power P_{max} considering this limit. For $MCL \geq 110dB$ and both criteria (i.e. to minimize the energy consumption through $\hat{E}_b$ or transmission time), we obtain the same energy consumption using the presented link adaptation algorithm. This is because both criteria (i.e. transmission time or $\hat{E}_b$) have the same optimal positions in the TBS table.

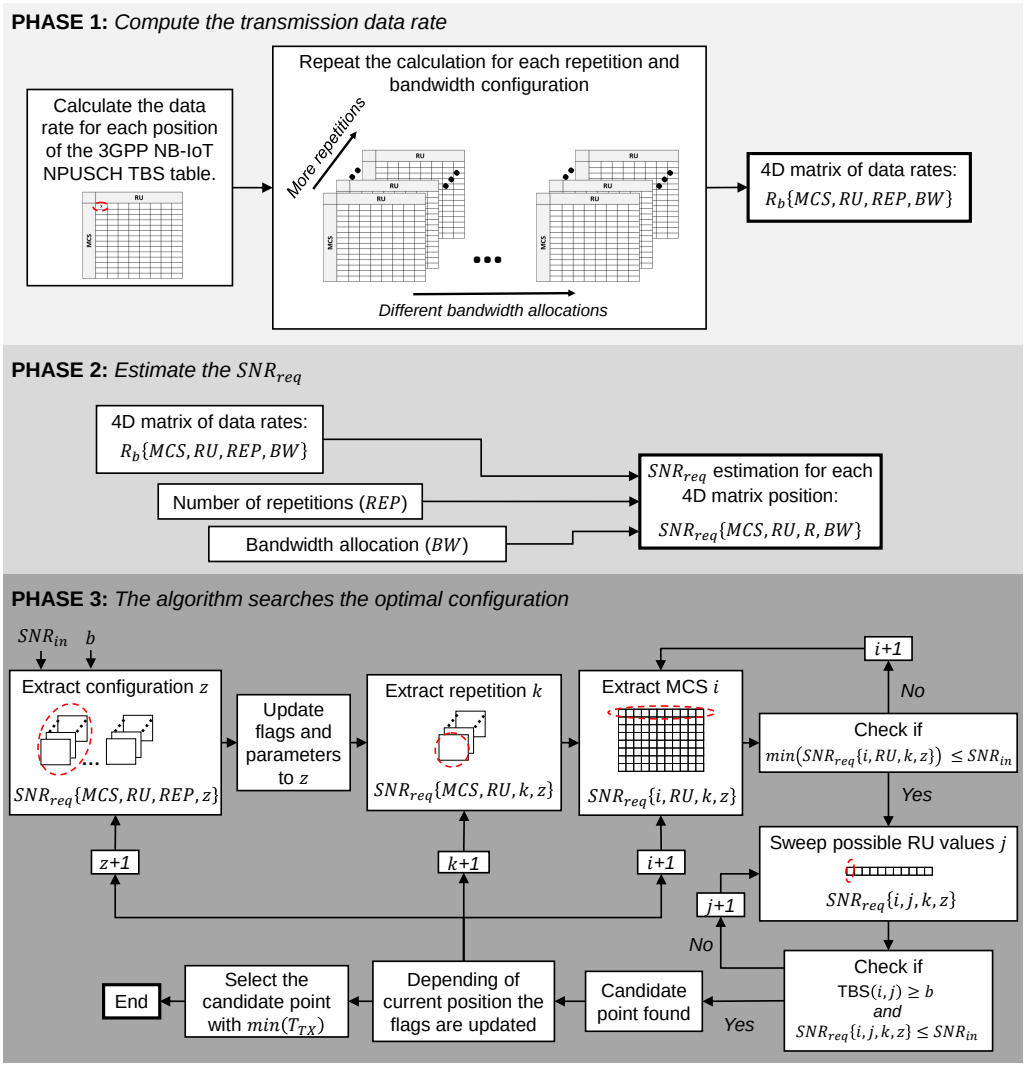


Figure 4.10: UL link adaptation phases.

Algorithm 1 Proposed UL Link Adaptation Algorithm for NB-IoT

Input: SNR_{in} and b

Output: $Best\_Point\_Found\,(C, I_{MCS}, I_{RU}, REP_{idx})$

Initialization
$N_{confs} \leftarrow 5$ // *Bandwidth configurations allowed*
$C \leftarrow 1$ // *Current bandwidth configuration to evaluate*
$REP_{max} \leftarrow$ Maximum number of repetitions allowed
$RU_{max} \leftarrow 8$ // *Maximum number of RUs*
$SNR^{C}_{req}(I_{MCS}, I_{RU}, REP) \leftarrow SNR_{req}$ of the configuration C corresponding to MCS level I_{MCS}, I_{RU} RUs and REP repetitions
$TBS(I_{MCS}, I_{RU}) \leftarrow$ Number of bits corresponding to I_{MCS} combined with I_{RU} in NB-IoT TBS table
while $C \leq N_{confs}$ & *not finished* **do**
  Set REP_{max} according to C
  if $length\,(Points\_Found) > 1$ **then**
    Reinitialization of some parameters and flags
  end if
  while $REP_{idx} \leq REP_{max}$ & *not finished* **do**
    while $I_{MCS} \leq MCS_{max}$ & *not finished* **do**
      if $min(SNR^{C}_{req}(I_{MCS}, :, REP_{idx})) \leq SNR_{in}$ **then**
        while $I_{RU} \leq RU_{max}$ & *not finished* **do**
          if $TBS(I_{MCS}, I_{RU}) \geq b$ & $SNR^{C}_{req}(I_{MCS}, I_{RU}, REP_{idx}) \leq SNR_{in}$ **then**
            $Points\_Found \leftarrow (C, I_{MCS}, I_{RU}, REP_{idx})$
            Depending on C, some parameters are reconfigured to search more points before increasing C or exit the algorithm
          end if
        end while
      end if
    end while
  end while
end while
$Best\_Point\_Found \leftarrow get\,(min\_transmission\_time\,(Points\_Found))$

4.5 Analytical evaluation framework for NB-IoT

In order to jointly illustrate the coverage extension and the resulting UE battery lifetime and latency, we propose an analytical evaluation framework for NB-IoT. This framework summarizes the steps required to consider different CE scenarios and check the performance impact of these scenarios in the UE. Therefore, the proposed evaluation framework involves three sequential steps, namely:

1. SNR_{req} estimation from the analysis presented in section 4.4. This estimation includes the proposed NB-IoT's Shannon fit, ideal or realistic CE, and the cross-subframe CE technique when applied.

2. Utilization of the outcome of step 1 as an input to configure the link adaptation of the signaling packet transmissions/receptions required prior to the UL data transmission. For the UL, the link adaptation is done using the algorithm explained previously. For the DL, the link adaptation algorithm searches the optimal configuration sweeping and comparing all possible configurations.

3. Estimation of the NB-IoT performance from the NB-IoT energy consumption model explained in Chapter 3. This model uses the output of the link adaptation of step 2 to configure the energy consumption and delay of each packet transmitted/received during the control procedure assumed. In this evaluation, the NPUSCH transmission power follows the 3GPP power control for NB-IoT defined in [21] (see subsection 3.2.3).

To clarify the evaluation framework process, Figure 4.11 depicts steps 1 and 2. This figure shows the inputs and outputs of each stage of the evaluation. The evaluation framework begins at step 1 with two parallels estimations. The first estimation returns the combination $\{c_1, c_2\}$ that depends on the cross-subframe window size W_{cs} and the modulation technique. The second estimation returns an SNR_e four-dimensional table. The four dimensions come from: i) the number of resources (RUs or SFs); ii) the MCS; iii) the number of repetitions; and iv) the bandwidth allocation. Next, using both estimations we obtain the SNR_{req} four-dimensional table. Note that for the DL there is only one possible bandwidth allocation, thus, the SNR_e and SNR_{req} tables have only three dimensions.

In step 2, the radio resources required for all packets to be transmitted or received are configured. To that end, this step takes four inputs: i) the SNR_{req} four-dimensional table; ii) the required Maximum Coupling Loss (MCL); iii) the packet size; and iv) the direction of the packet (UL or DL). After the selection of resources for all needed packets, the evaluation framework uses the analysis of Chapter 3 to estimate the battery lifetime of the UE and the latency to finish the control procedure. These specific packet configurations will modify the packet energy estimation (i.e. $T_{tx}(x)$, $T_{rx}(x)$, $E_{tx}(x)$, $E_{rx}(x)$, etc) seen in Chapter 3.

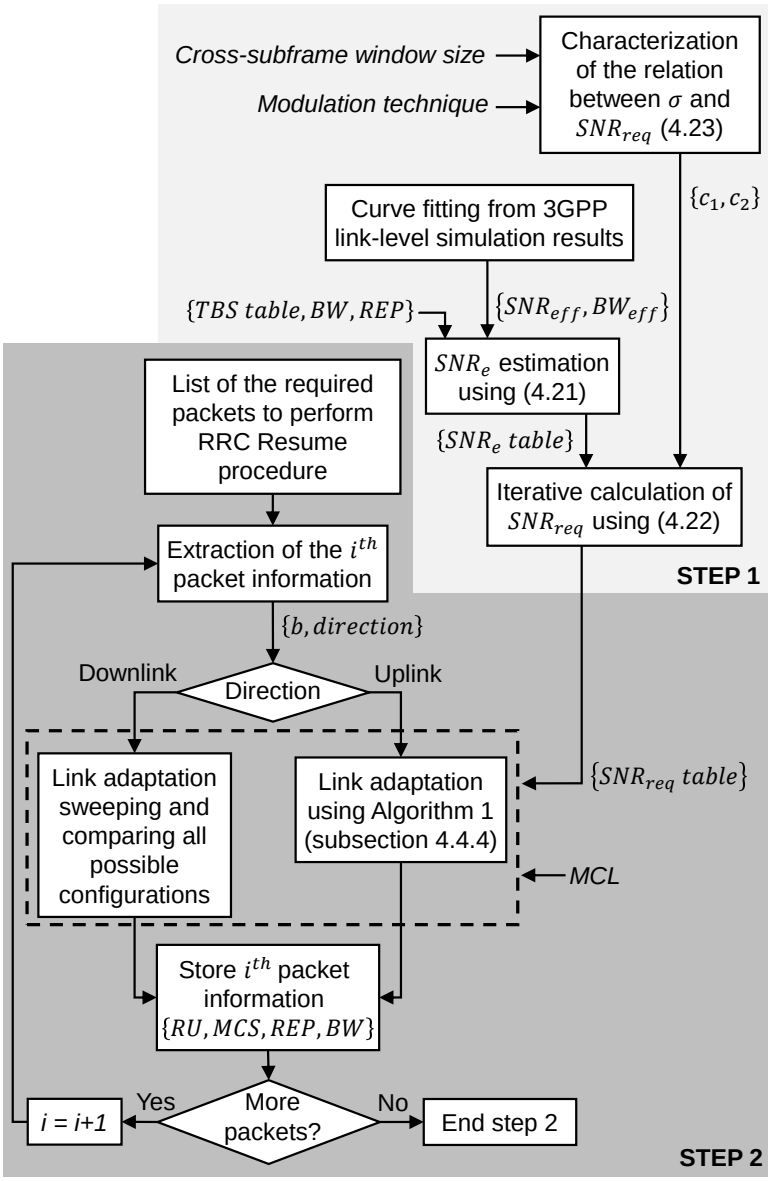


Figure 4.11: Block diagram of steps 1 and 2 of our evaluation framework.

4.6 NB-IoT extended coverage evaluation

This section presents the obtained simulation results and discusses the NB-IoT performance considering three different scenarios, labeled as:

- `ideal`: ideal CE, that is $\hat{\mathbf{H}} = \mathbf{H}$.
- `wcs`: realistic CE and cross-subframe CE utilization. In this scenario, larger transmissions could benefit from the use of larger cross-subframe windows.
- `nocs`: realistic CE without cross-subframe technique.

The NB-IoT performance is measured in terms of the latency needed to send the UL data packet (i.e. finish the UP procedure and send the data packet) and UE's battery lifetime.

4.6.1 Evaluation setup

For the first step of the evaluation framework, the correction values for Shannon formula are derived through curve fitting to the 3GPP's link-level simulation results for NPUSCH [130] and Narrowband Physical Downlink Shared CHannel (NPDSCH) [131]. Table 4.2 summarizes the correction values obtained. In this case, Narrowband Physical Downlink Control CHannel (NPDCCH) performance, i.e., the required number of repetitions to achieve the needed SNR_{req}, is obtained from the results of [94]. For simplicity, the configuration of both Common Search Space (CSS) and UE-specific Search Space (USS) is equal as in Chapter 3, and R_{max} equals the number of repetitions at the NPDCCH.

In order to estimate the SNR_{req} when applying repetitions, step 1 needs the CE error σ used in (4.22). To do that, we need the components c_1 and c_2 of (4.23) that define the dependency between σ_{dB} and $SNR_{req,dB}$. Both components are obtained through simulation of σ for NPUSCH and NPDSCH pilot symbols. The simulation is done with MATLAB modifying the code used in [105, 132]. The simulation of the pilots in the UL and DL follows the 3GPP specification for NB-IoT [29]. The configuration of OFDMA and SC-FDMA is done according to the Third Generation Partnership Project (3GPP) LTE 1.25 MHz carrier bandwidth. Then, the components c_1 and c_2 are obtained for five different cross-subframe windows. For each configuration, 5000 channel realizations are simulated. Table 4.3 summarizes the resulting values.

To determine W_{cs}, we set $\eta_p = 1$ for multi-tone configurations, $\eta_p = 0.6667$ for single-tone configurations, and W_{cs}^{max} equals 16. For the power control, we

assume $\alpha = 1$ and $P_{target} = -100dB$. The CRC length equals 24 bits. Finally, Table 4.4 includes both steps 2 and 3 parameters.

For the second and last steps of the evaluation framework, we evaluate the performance of the UE for a specific MCL under different scenarios. To do that, we use the link budget of each radio channel and the assumed MCL to obtain the $SNR_{req,dB}$ we will have in the analysis as a requirement for the transmission or the reception of packets. The link budget done in this chapter is equal to the one presented in Chapter 3, more specifically in Table 3.7. However, in this chapter, we follow the reverse way.

Table 4.2: Shannon correction parameters values.

Parameter	**NPUSCH Multi-tone**	**NPUSCH Single-tone**	**NPDSCH**
BW_{eff}	0.35	0.35	0.58
SNR_{eff}	1	0.60	1.90

Table 4.3: Values of the parameters of (4.23) for different number of cross-subframe windows for LS channel estimator.

W_{cs}	**SC-FDMA**		**OFDMA**	
	c_1	c_2	c_1	c_2
1	-0.4896	4.4971	-0.4998	14.5262
2	-0.4844	3.0252	-0.4995	13.0035
4	-0.4780	1.5869	-0.4990	11.5017
8	-0.4725	0.1239	-0.4992	9.9952
16	-0.4475	-1.1335	-0.4969	8.5077

Table 4.4: Parameters for NB-IoT performance estimation [21, 23, 26–28].

Energy consumption configuration			
Variable	Value		
Deep sleep power consumption	0.015 mW		
Inactive power consumption	3 mW		
Reception power consumption	80 mW		
Maximum transmission power consumption	500 mW (included 60 mW for other analog and baseband circuitry)		
Power amplifier efficiency	45%		
Battery capacity	5 Wh		
UE's transmit power for NPRACH	$P_{target} = -100dBm$, $\alpha = 1$		
Connected DRX cycle	256 ms		
On duration DRX timer	1 NPDCCH period		
Idle DRX cycle	5120 ms		
Active timer	20 s		
Preamble detection probability	Preamble: $1 - e^{-i}$, where i indicates the ith preamble transmission. Other packets: 1		
Physiscal layer			
Propagation condition	Typical Urban (TU) 20 paths, 1Hz		
Carrier frequency offset	20 Hz		
UE — eNB noise figure	5 dB — 3 dB		
UE — eNB power class	23 dBm — 35 dBm		
Protocol overhead			
Higher layer procedure	RRC Resume		
PDCP/RLC/MAC overheads (bytes)	5 / 2 / 2		
Packet sizes on top of PDCP (bytes)	Latency estimation: UL report 85 Battery lifetime estimation: UL report 50, DL ACK 65		
NB-IoT design			
Deployment	In-band		
Modulation	QPSK		
NPDCCH design	Format: 0, Aggregation level: 2 Periodicity: $T_{NPDCCH} = R_{MAX} \cdot G$, where $G = 1.5$		
Start of NPUSCH transmission after the end of its associated DCI	8 ms		
Start of NPDSCH transmission after the end of its associated DCI	5 ms		
RSRP thresholds	MCL = {144, 154} dB		
ECL configuration for Random Access	ECL 0	ECL 1	ECL 2
NPRACH/NPDCCH repetitions	1	8	32
Random Access Opportunity	40 ms	240 ms	640 ms
Synchronization time	327 ms	341 ms	597 ms
MIB acquisition time	111 ms	111 ms	483 ms

4.6.2 Results

This subsection presents the results obtained with the proposed NB-IoT evaluation framework.

4.6.2.1 SNR derivation and realistic channel estimation

In the first step of the evaluation framework, we estimate the SNR_{req} in different configurations of number of RUs, MCS, BW, and repetitions. Figures 4.12, 4.13, 4.14, and 4.15 are examples of the analysis of this first step.

Figure 4.12 presents the comparison of the spectral efficiency as a function of the MCL for our NB-IoT Shannon fit and the baseline Shannon bound. As it can be seen, NB-IoT presents a great gap between its performance and the Shannon bound. This is due to implementation issues and the repetition coding scheme used in NB-IoT [124]. For large MCLs ($MCL > 150$ dB), there is a performance deficit up to 5 dB approximately of the NB-IoT fits compared to Shannon.

Figure 4.13 shows the analyzed transmission properties mentioned in Section 4.4. In this analytical evaluation, the transmission power follows expression (4.15) without the constraints of the 3GPP, specifically we assume $BW = 180$ kHz, and $L = 100$ dB. The comparison of two different RU allocations highlights the benefit

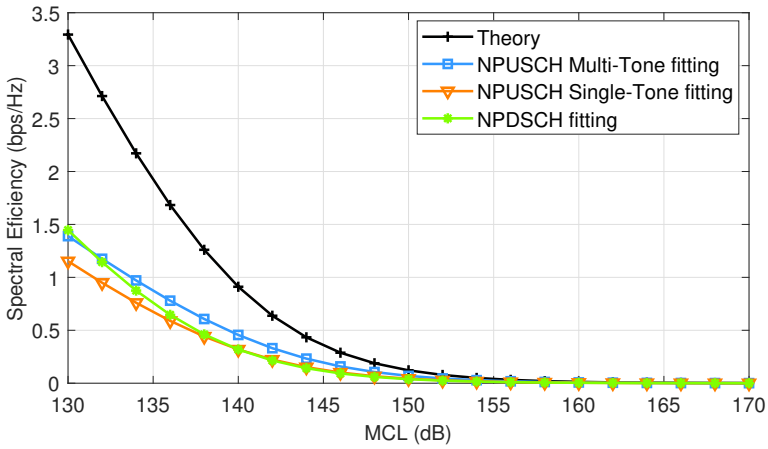


Figure 4.12: Spectral efficiency as a function of the MCL considering the curves from Shannon bound, and the proposed Shannon fit for NB-IoT.

of cross-subframe technique in `wcs` compared to `nocs`. For a given TBS size, a greater number of RU achieves a lower $\hat{E}_b$. Although not shown in the figure, if the transmission power is maintained, increasing the number of repetitions and reducing the allocated bandwidth will reduce the SNR_{req}. However, when applying repetitions, this SNR_{req} reduction is at the expense of reducing the bandwidth utilization γ. Therefore, whenever possible, it is better to first reduce the bandwidth allocated.

Figure 4.14 illustrates the impact of the CE error when doubling repetitions to extend coverage in terms of SNR_{req}. As the SNR_{req} is lower, the CE error increases. Therefore, in this range of SNR_{req} the gain when doubling repetitions is limited in `wcs` and `nocs` scenarios compared to `ideal` scenario. For a number of repetitions greater than 16, the gain when doubling the repetitions applied in `wcs` and `nocs` scenarios is less than 1.5 dB. Furthermore, we can see `wcs` improves the results of `nocs` scenario due to the benefits of cross-subframe CE.

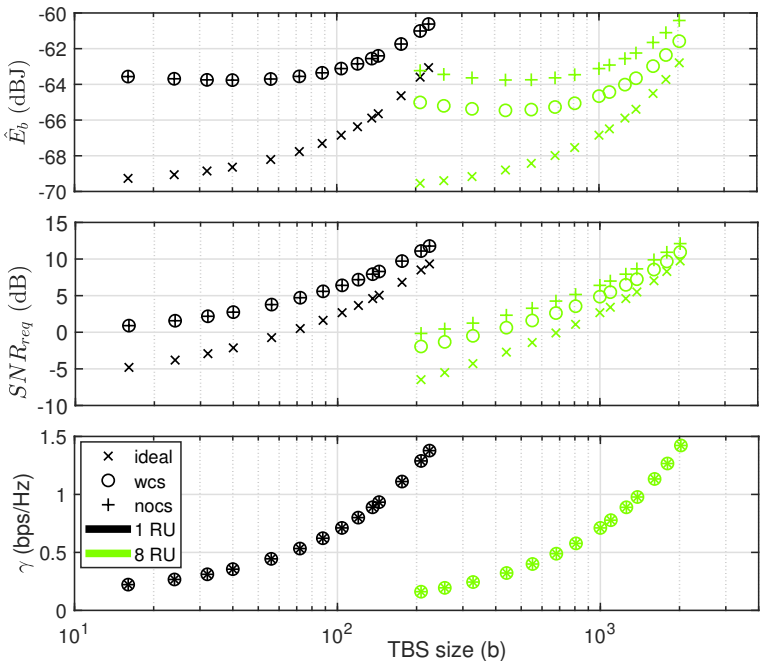


Figure 4.13: Transmission properties comparison as a function of the TBS size for different number of RUs and the three scenarios.

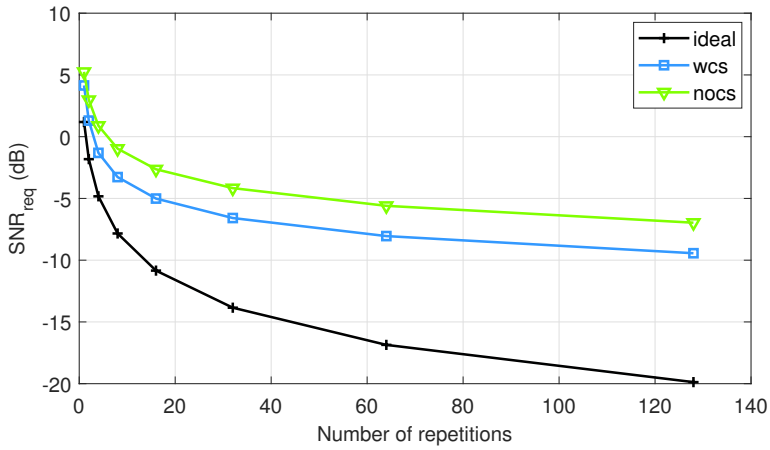
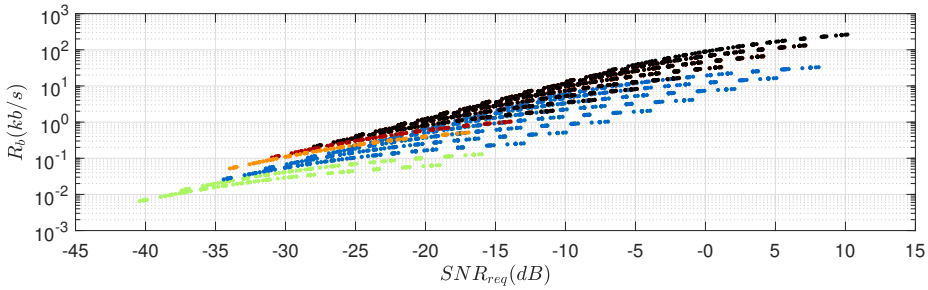


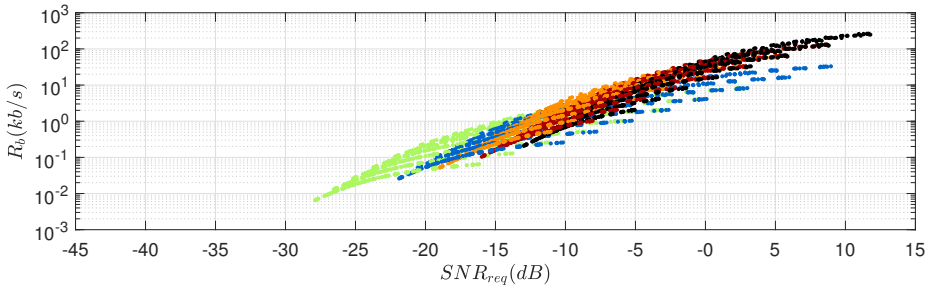
Figure 4.14: Example of degradation of the SNR gain in `wcs` and `nocs` scenarios compared with `ideal` scenario when a higher number of repetitions is used for the UL and a TBS of 504 bits with 5 RUs.

Focusing on the resulting SNR_{req}, Figure 4.15 shows the SNR_{req} as a function of the data rate R_b for the three scenarios considered. This figure represents the 4D table of SNR_{req} we obtain in our analysis, as each point of the figure is a configuration of the NPUSCH TBS with a specific number of repetitions and bandwidth allocated. Note we assume there is an ideal gain in the SNR_{req} when we apply bandwidth reduction and the transmission power is maintained. That means, if we reduce the bandwidth to the half, we obtain 3 dB of gain. This assumption can be noticed in the figures as each configuration of the bandwidth is shifted a little bit in the x-axis of the figure. On the contrary, the SNR_{req} gain due to repetitions in realistic CE depends on the CE error. This effect can be seen when comparing the `ideal` scenario 4.15a with realistic CE scenarios (i.e. 4.15b and 4.15c) as `ideal` scenario reaches a larger range of SNR_{req}.

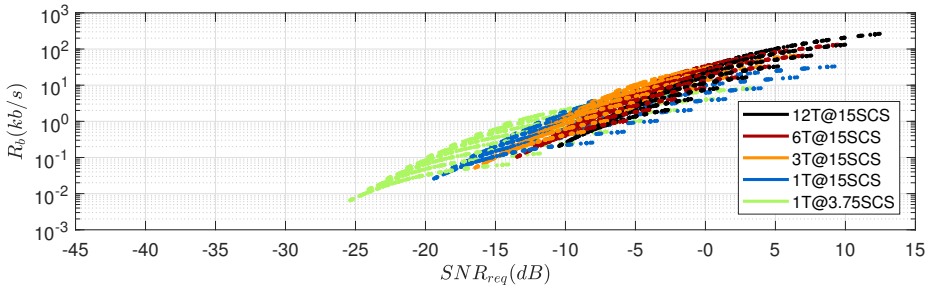
From the previous figures, we can observe the considerable impact realistic CE has. The poor performance of the channel estimator in the NB-IoT targeted low SNR range can significantly impact the coverage extension. In some cases, the coverage extension limitation can be due to is not possible to successfully decode the packets at the receptors, thus, the communication is not possible. However, in other cases, even if the UE is reachable and the communication



(a) `ideal` scenario.



(b) `wcs` scenario.



(c) `nocs` scenario.

Figure 4.15: SNR_{req} as a function of the datarate R_b considering repetitions and bandwidth reduction.

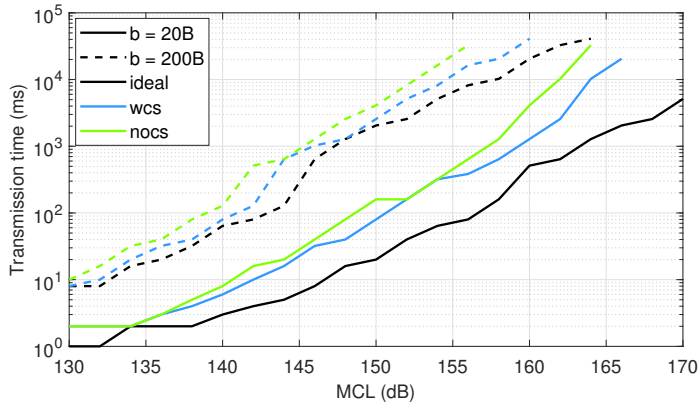
can be established using more robust configurations, the coverage extension can be not used. If the resulting radio resources configuration requires the use of extensive resources from the network, the coverage extension can be restricted by the operator.

4.6.2.2 Uplink link adaptation

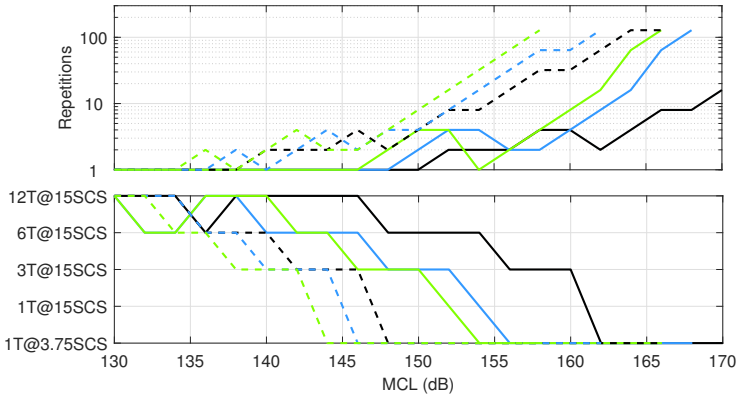
Regarding the second step of our evaluation framework, Figure 4.16 presents an example of UL link adaptation sweeping the MCL for two packet sizes (20 bytes and 200 bytes). Note in this example of UL link adaptation application, we assume there is no segmentation of these two packets. Therefore, the configured TBS has to be big enough to accommodate the assumed packet size.

Figure 4.16a illustrates the impact in terms of transmission time of the ideal CE (`ideal` scenario), compared to realistic CE (`wcs` and `nocs` scenarios). Again, we can see the utilization of the cross-subframe technique improves the results of realistic CE evaluations. For example, for the considered small packet size of 20B, `wcs` scenario increments its transmission time an average of 250% compared to `ideal` scenario. However, for `nocs` scenario, this increment reaches an average of 496%. For larger packet sizes, such as the 200 bytes shown, this increment is less significant. It reaches average values of 68% and 170% in `wcs` and `nocs` scenarios, respectively.

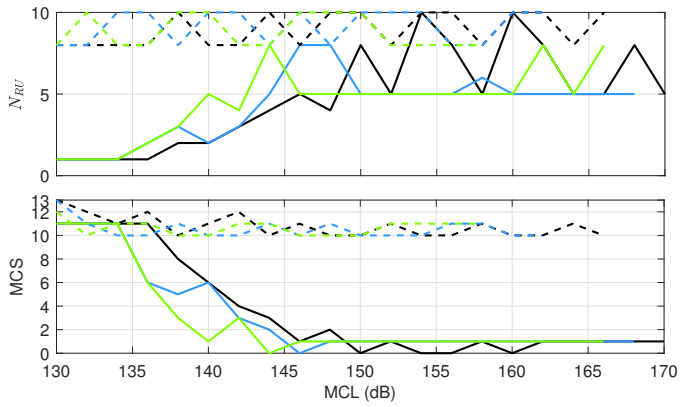
Figure 4.16b shows the configuration of the number of repetitions and the bandwidth allocated. Owing to bandwidth reduction preserves the bandwidth utilization when applied, the algorithm gives priority to this approach. Therefore, an extensive number of repetitions is not applied until single-tone configurations are used. Additionally, Figure 4.16c presents the detailed configuration of the packet in terms of number of RUs N_{RU} and the MCS level. Both parameters are coupled by the NPUSCH TBS table (see subsection 3.2.1.2). As no segmentation is used in this example of link adaptation, the packet size of 200 bytes is limited to TBS configurations with high number of RUs and MCS level. This is because these positions have bigger sizes of TBS. For the packet size of 20B, the algorithm has more positions of the TBS available. In this case, we can see as the MCL worsen, the algorithm selects configurations more redundant. These more redundant configurations involve lower data rate by means of increasing the number of RUs and decreasing the MCS level.



(a) Transmission time.



(b) Repetitions and bandwidth allocation.



(c) Number of RUs and MCS configured.

Figure 4.16: Example of UL link adaptation in terms of a) transmission time; b) Repetitions and bandwidth; c) Number of RUs and MCS as a function of the MCL considering two packet sizes (20 bytes and 200 bytes) and the three scenarios.

4.6.2.3 Coverage extension impact on the UE

Finally, for the last step of our evaluation framework, we need to apply the analysis performed in the previous steps, i.e., the estimation of the SNR_{req} and the link adaptation. With both, we can estimate in this step the battery lifetime and latency the UE will have under different radio conditions.

Figures 4.17 and 4.18 represent the performance of an NB-IoT UE for the three scenarios (i.e. `ideal`, `wcs`, and `nocs` presented in the beginning of Section 4.6). The configuration of the parameters for the estimations of both figures is similar to the parameters of the evaluations done in [26] and [27], such as the size of the UL reports or the DRX configuration. In these results the packets can be segmented to be able to find solutions in the link adaptation if the packets are too big to be sent in one piece under the coverage conditions evaluated.

Figure 4.17 shows the latency estimation when a UE performs an RRC Resume procedure to transfer a UL report. The latency is calculated adding the time components for synchronizing, setting up the connection, and transmitting the UL report of 85 bytes. The latency has two abrupt steps when there is a change of Coverage Enhancement Level (ECL) due to the increase of the time dedicated to synchronization and repetitions in the NPRACH.

As expected, `ideal` scenario obtains the best results and can be evaluated for greater MCLs than `wcs` and `nocs` scenarios. Additionally, the included results from [26] and `ideal` scenario are similar. For both, the support of a latency of at least 10 seconds is achieved up to the MCL of 164 dB. Nevertheless, `wcs` and `nocs` attain worse results than `ideal` scenario. This is owing to the degradation of the SNR gain when doubling repetitions in realistic CE. Therefore, for higher MCLs that rely on repetitions to extend coverage, this degradation exacerbates the difference between realistic CE (`wcs` and `nocs` scenarios) and ideal CE (`ideal` scenario).

Figure 4.18 shows the battery lifetime estimation sweeping the MCL and considering two different Inter-Arrival Times (IATs). This figure also includes the results [27]. In this estimation, we consider the UE transmits a UL report of 50 bytes and waits for the reception of a DL application Acknowledgment (ACK) of 65 bytes. After the DL ACK and before the UE enters Power Saving Mode (PSM), the UE monitors the NPDCCH until the expiration of the Active Timer. The

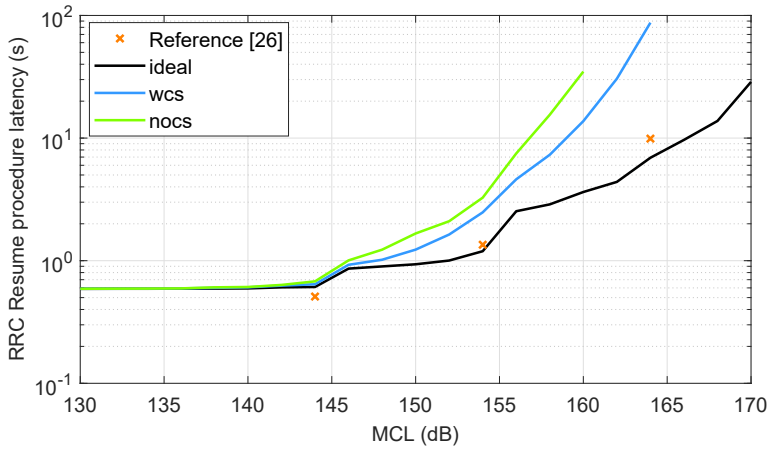


Figure 4.17: RRC Resume procedure latency versus MCL.

`ideal` scenario provides similar results to the source [27]. When we consider realistic CE, the battery lifetime results are more pessimistic. As seen before, this detriment in the battery lifetime is greater as the MCL increases. Particularly for UEs with small IATs, this effect is noted before. For example, when $MCL = 154$ dB, `wcs` presents a battery lifetime reduction of a 50% and 18% compared to `ideal` for IATs of 2 h and 24 h, respectively. However, when $MCL = 164$ dB, both IATs obtain similar values, reaching a battery lifetime reduction of a approximately 90% in `wcs` compared to `ideal` scenario.

In Figures 4.17 and 4.18, we can observe again the great impact realistic CE has but from the UE side. There will be IoT applications with restrictive requirements in terms of latency or expected battery lifetime. The UE side point of view is interesting as allow us to analyze if under the foreseen UE circumstances when it is deployed (e.g. coverage or traffic profile), it will be feasible to use NB-IoT to transfer data packets.

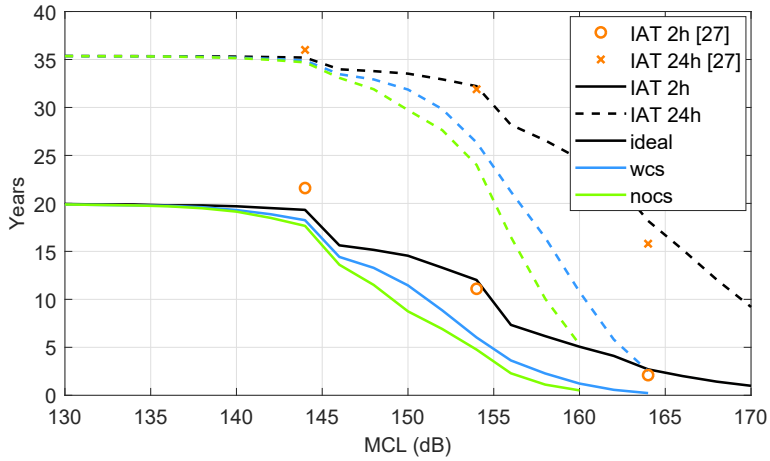


Figure 4.18: Battery lifetime estimation versus MCL of an NB-IoT UE. The figure includes two different IATs and the three scenarios.

4.7 Conclusions

The promising extensive coverage enhancement attainable in NB-IoT can connect IoT UEs that are placed in remote areas or environments with deep penetration loss. This is achieved by means of PSD boosting and repetitions in time. However, both approaches entail a reduction of the data rate, thus, an increase in the energy consumption needed to finish a transfer. This trade-off is even more challenging under the foreseen NB-IoT UEs weak coverage condition where the gain due to repetitions can be significantly limited by the channel estimator performance.

Until now several works have focused on the coverage performance of NB-IoT. However, there are several IoT use cases where the trade-off between coverage and energy consumption is essential to satisfy the IoT application requirements. In this chapter, an analytical evaluation framework for NB-IoT is presented. The goal of this evaluation framework is to provide a tool to analyze the performance of NB-IoT considering the coverage of the UE as well as the resulting battery lifetime. The analytical framework is divided into three issues: i) SNR estimation; ii) link adaptation; and iii) NB-IoT energy and delay estimation. The previous Chapter 3 focuses on the energy and delay estimation. The analysis presented in this chapter delves into the SNR estimation and link adaptation. Specifically,

in this chapter we have worked in the following items to complete the analytical evaluation framework:

- Derivation of analytical expressions based on the Shannon theorem to describe a few transmission properties (i.e. required SNR, energy per transmitted bit, bandwidth utilization).
- Derivation of analytical expressions to describe realistic CE and the use of cross-subframe.
- Development of OFDMA and SC-FDMA simulators to obtain the relationship between the CE error and the required SNR under different conditions.
- Proposal of a UL link adaptation algorithm that jointly considers the new dimensions NB-IoT adds to the link adaptation (i.e. MCS or RUs, bandwidth allocated, and repetitions).

In the evaluation, we consider three different scenarios: i) ideal CE; ii) realistic CE with cross-subframe; and iii) realistic CE without cross-subframe. The NB-IoT UE performance is evaluated in terms of UL packet transmission latency and battery lifetime.

Regarding the UL link adaptation, when the MCL worsen, the algorithm selects more redundant MCS and RUs combinations. Next, bandwidth reduction is applied as it also keeps the bandwidth utilization. However, both techniques can only cover a limited range of coverage extension. Therefore, if the UE has poor coverage, repetitions become essential to reach greater coverage extension. However, in this situation the gain due to repetitions can be limited if the channel estimator performance is poor.

The conducted evaluations show the performance of the SNR gain when doubling repetitions is significantly affected when assuming realistic CE compared to ideal CE. As the MCL increases, this degradation increases due to larger CE errors. When realistic CE is considered, the use of cross-subframe improves its results. Specifically, regarding the UE battery lifetime, for an $MCL = 154$ dB, realistic CE with cross-subframe shows a battery lifetime reduction of a 50% and 18% compared to ideal CE for IATs of 2 h and 24 h, respectively. However, for higher MCLs such as 164 dB, both IATs reach a battery lifetime reduction of approximately 90% in realistic CE with cross-subframe compared to ideal CE.

4.7.1 Resulting research contributions

The research contributions resulting from the work done in this chapter are listed below:

- P. Andres-Maldonado, P. Ameigeiras, J. Prados-Garzon, J. J. Ramos-Munoz, J. Navarro-Ortiz and J. M. Lopez-Soler, "Analytic Analysis of Narrowband IoT Coverage Enhancement Approaches," 2018 Global Internet of Things Summit (GIoTS), Bilbao, 2018, pp. 1-6.

  DOI: 10.1109/GIOTS.2018.8534539

- P. Andres-Maldonado, P. Ameigeiras, J. Prados-Garzon, J. Navarro-Ortiz and J. M. Lopez-Soler, "An Analytical Performance Evaluation Framework for NB-IoT," Accepted for publication in the IEEE Internet of Things Journal, 2019.

Chapter 5

Experimental Evaluation of NB-IoT

As seen in previous chapters, the four Narrowband Internet of Things (NB-IoT) targets (i.e. maximum 10 seconds in the Uplink (UL), target coverage of 164 dB Maximum Coupling Loss (MCL), battery lifetime beyond 10 years, and support of massive connections) are strongly related and depend heavily on the device behavior and the NB-IoT network deployment and configuration. Depending on the Internet of Things (IoT) use case, one or more of these goals will be more important than the others. For example, smart metering devices tend to send or receive data periodically [34]. For battery powered smart meters located in remote areas, coverage and battery lifetime are critical due to their costly access for maintenance. However, for alarms or event detectors, whose traffic is tightly coupled with human activities, the priority is the latency and coverage.

The IoT ecosystem diversity and the specific requirements of each use case make it important to have a clear view of the trade-offs between the NB-IoT targets. For example, providing the coverage extension while maintaining low energy consumption comprises a great conflict as the coverage extension is mainly achieved by trading off data rate, e.g., lowering the transmission bandwidth, or using repetitions in time.

The NB-IoT networks are eventually ready to roll-out in practical deployments. However, the new design of NB-IoT causes several works proposed for Long Term Evolution (LTE) to be unfit for the study of NB-IoT. Thereby, there

is still unawareness if NB-IoT will be able to cope with the IoT Key Performance Indicators (KPIs) due to the vast scenarios and configurations to consider.

Within this context, this chapter tackles the study of the NB-IoT performance, both analytically and experimentally as well. We study the NB-IoT performance considering its conflicting trade-offs between targets. Our aim is to provide a general vision of the User Equipment (UE) performance possibilities. To do that, the specific goals of this chapter are the following:

i) Experimentally validate our energy consumption NB-IoT model proposed in Chapter 3. The validation is done comparing the analytical results from the model with empirical measurements obtained in a controlled testbed. As the model in Chapter 3 is large, due to it takes many NB-IoT features into account, in this chapter we will only validate the UE energy consumption (and latency) estimations. That is, we will assume no capacity constraints in the radio channels. In such case, the connection failure and collision parts of the Chapter 3 Markov chain (see Section 3.5.2) are not required because the corresponding transition probabilities are zero. The conducted tests in this goal focus on examining if our model correctly describes the steps the UE will perform to send a UL data packet.

ii) Experimentally study of the NB-IoT targets. With the vast configurations possibilities of NB-IoT, is a challenge to know under what radio and traffic circumstances the NB-IoT coverage and latency targets can be met and at what battery lifetime-cost for the UE. In order to be as close as possible to future practical NB-IoT deployments, we will use the controlled testbed of the previous goal and the study will be done considering only empirical measurements. To do that, we will observe the configuration of a live NB-IoT network and we will examine the performance of commercial NB-IoT UEs under several configurations. The proposed tests in this goal focus on highlight the impact different factors will have on the UE performance.

As a summary of the parameters studied in this chapter, Table 5.1 illustrates the key parameters that define the configuration of the communication between the UE and evolved NodeB (eNB) and where these parameters are signaled.

The rest of the chapter is organized as follows. Section 5.1 briefly reviews the related literature. Section 5.2 is dedicated to the experimental study of the NB-IoT targets. Section 5.3 describes and validates the analytical model for NB-IoT. Lastly, section 5.4 presents the main conclusions of this chapter.

5.1 Related works

For the further evolution and use of NB-IoT is important to understand the possibilities and trade-offs NB-IoT has. In the current literature, several works [11, 90, 94, 115, 120, 121, 133] and 3GPP evaluations [130, 131, 134, 135] present NB-IoT performance evaluation results. These sources provide final results for specific simulated configurations that are a useful start point to understand the possibilities of NB-IoT.

However, to be able to predict probable limitations or other challenges, there is also necessary to provide an analysis procedure. Within this topic, the authors of [136] provide a mathematical NB-IoT network model to estimate the throughput and success probability. In [137], *Azari et.al.* propose an analytical model based on queuing theory for the channel scheduling in NB-IoT. They investigate the latency-energy trade-off resulting from the resource sharing between channels and the Coverage Enhancement Levels (ECLs). In [138], the authors propose an analytical UL performance estimation. They show the variability of the network performance due to different payload sizes, Narrowband Physical Uplink Shared

Table 5.1: Summary of key parameters configured in the radio interface.

Parameter	Definition	Range of values	Signaled in
Repetitions	Number of repetitions in the time domain. Each channel can have a different number of repetitions	UL $\in [1, 2, 4, 8, 16, 32, 64, 128]$ DL $\in [1, 2, 4, 8, 16, 32, 64, 128, 192, 256, 384, 512, 768, 1024, 1536, 2048]$	NPRACH repetitions: SIB2 NPUSCH or NPDSCH repetitions: DCI NPDCCH repetitions for USS (R_{max}): RRC Connection Setup NPDCCH repetitions for CSS Paging or RA (R_{max}): SIB2
DCI subframe repetitions	Number of repetitions of the DCI in the time domain	The range of values depends on the NPDCCH repetitions (R_{max})	DCI repetitions for USS: DCI DCI repetitions for CSS: SIB2
Start subframe (G)	Starting subframe configuration for anNPDCCH search space	G $\in [1.5, 2, 4, 8, 16, 32, 48, 64]$	G for CSS: SIB2 G for USS: RRC Connection Setup
MCS	Modulation and Coding Scheme	$[0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12]$	DCI
Radio resources	Number of Subframes (SFs) for DL and Resource Units (RUs) for UL	$[1, 2, 3, 4, 5, 6, 8, 10]$	DCI

CHannel (NPUSCH) repetitions, and the Narrowband Physical Random Access CHannel (NPRACH) periodicity. Other NB-IoT analytical/simulation studies have been conducted in the analysis of the energy consumption due to scheduling request procedure [139], paging procedure [140], and cell data rate [141].

Additionally, recent works have developed tools to model NB-IoT. The authors of [142] present their implementation of NB-IoT using the high-performance open-source software radio LTE physical layer library called srsLTE [143]. In [144], *Throha et.al.* develop a Downlink (DL) simulator for NB-IoT network using Matlab. The simulator tests the DL physical channels and signals. Specifically, they evaluate the relationship between Block Error Ratio (BLER), Signal to Noise Ratio (SNR), and repetitions. In [145], *Martiradonna et.al.* propose an open source simulator for NB-IoT based on the LTE-Sim tool [146]. Their preliminary performance evaluation shows the system goodput and the Cumulative Distribution Function (CDF) of the end-to-end packet delay.

Focusing in experimental evaluations like the one presented in this chapter, the authors of [147] analyze the energy consumption, reliability, and delay to send a datagram under different signal quality scenarios. In [148], *Yeoh et.al.* discuss optimizations for the firmware design to prolong the battery lifetime. In [149], *Lauridsen et.al.* present empirical power consumption measurements of two NB-IoT UEs. They also estimate the resulting battery lifetime when considering different power saving features.

In the present chapter, we combine the analytical and experimental analysis of NB-IoT. Our proposed analytical model provides three main benefits compared to previous literature in NB-IoT: i) inclusion of the new NB-IoT features (e.g. resource allocation, coverage extension techniques, power saving features, etc); ii) validation of the model through empirical measurements; iii) detailed analytical methodology that can be extended to assess more complex communication scenarios between the UE and the eNB (i.e. more control procedures can be included in the analysis).

5.2 Experimental performance evaluation

In this first part of the chapter, we present an empirically-based study of the factors that impact NB-IoT UE energy consumption and delay. We carry out

this empirical study in two steps. In the first step, we observe the configuration of a live NB-IoT network. To do that, we connect a commercial NB-IoT UE to a NB-IoT cell and we extract the radio parameters configured from the UE debug logs. The aim of this first step is to examine a realistic evolution of the radio resources configuration considering different UE radio conditions. Later, we use similar radio configuration trends in the defined test cases.

In the second step, we measure the UE energy consumption under different radio configurations using a controlled testbed. The testbed is comprised by a base station emulator connected to commercial NB-IoT UEs. We evaluate four test cases with the goal of identifying the factors that are power-hungry and under what circumstances.

Despite this experimental study of NB-IoT was presented as the second goal in the introduction of this chapter, we present this empirical study first instead of the validation of the analytical model to identify the items the model of the Chapter 3 does not include. These identified items will be listed at the end of this section.

5.2.1 Measurements methodology

To empirically evaluate the NB-IoT performance trade-offs, Figure 5.1 shows the experimental setup used to perform the previously mentioned two steps. The setup employed to analyze the live NB-IoT configuration has the label "STEP 1" and the testbed used to evaluate the UE performance has the label "STEP 2".

Specifically, under "STEP 1" we emulate different radio conditions of a UE connected to the NB-IoT TDC network deployed in the area of Aalborg (Denmark) by placing the UE in a shielded box, which is connected to an antenna through an RF attenuator. The RF attenuator reduces the level of the signal to emulate different coverage levels. In this first set of measurements, we recorded 22 traces by sweeping the RF attenuation, obtaining 1 trace per x dB attenuation level. This x dB attenuation ranges from 0 dB up to 41 dB. For each trace, the UE performs the following actions: attach to the network, send 3 UDP packets (payloads 20, 50, and 200 bytes) using the Control Plane optimization (CP) procedure, detach from the network. Note each trace is a debug log from the UE. Therefore, we can extract information related to the received Master In-

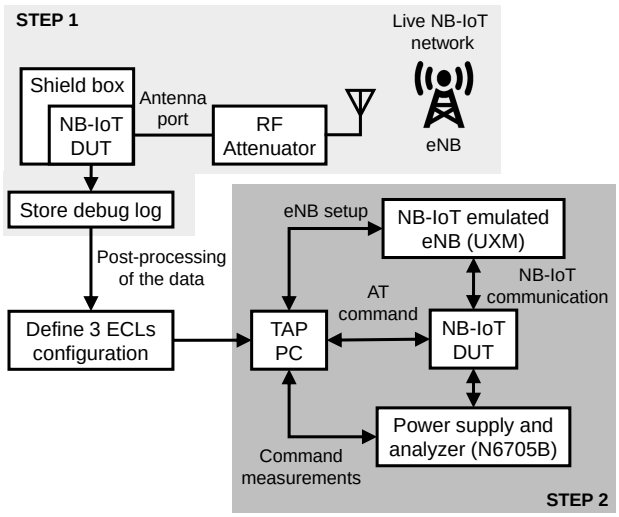


Figure 5.1: Experimental setup.

formation Block (MIB), System Information Blocks (SIBs), Downlink Control Information (DCI) allocations, Radio Resource Control (RRC) messages, Non-Access Stratum (NAS) messages, etc.

The observed configurations are then examined to design the test cases to implement in our testbed and the baseline configuration of three ECLs. Then, we enter Figure 5.1 "STEP 2". The testbed comprises a Device Under Test (DUT) connected to a base station emulator and a power analyzer. The base station emulator and the power analyzer are the Keysight E7515A UXM Wireless Test Set (UXM) and Keysight N6705B DC Power Analyzer, respectively. The UXM supports NB-IoT's Release 13, and this is also the release assumed in this chapter. The N6705B acts as a power supply and measures the voltage and current draw by the DUT. The whole testbed is configured using Keysight's Test Automation Platform (TAP). The TAP provides the following utilities: i) communication with the DUT through AT commands; ii) an unified interface to configure the UXM and N6705B; iii) synchronization of the protocol logs and measurements with $\leq$ 1ms accuracy. The used DUTs are three commercial NB-IoT UEs running firmware from August 2018. The differences between these DUTs are the manufacturer and that only one of the DUTs supports multi-tone configurations at the time the measurements were performed.

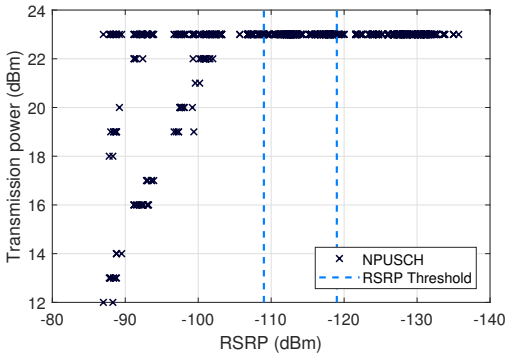
5.2.2 Live NB-IoT coverage extension management

Figure 5.2 presents the measured configuration of the transmission power 5.2a, Modulation and Coding Scheme (MCS) index 5.2b, number of repetitions for the data channels 5.2c, and DCI repetitions 5.2d. The common Figure 5.2 x-axis illustrates the Reference Signal Received Power (RSRP). This RSRP is the last filtered RSRP the UE reports in its debug log previously the message where we extract the parameters configuration. Each point of the figures illustrates one configuration of the parameters represented from a specific type of debug log message. For example, during one trace, the UE and eNB may exchange 14 DCIs N0 (i.e. UL grant). Thus, from this trace we obtain 14 samples of configuration of the NPUSCH MCS, NPUSCH repetitions, and DCI repetitions.

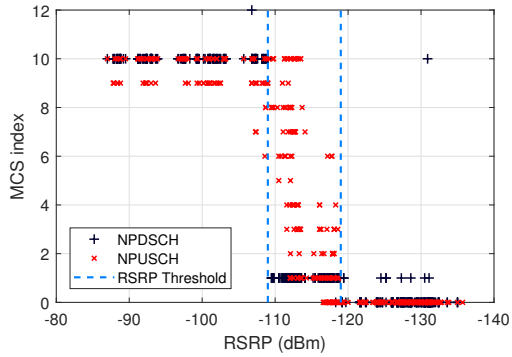
Additionally, the vertical lines represent the ECL RSRP thresholds configured by the network. The RSRP thresholds are included to illustrate the coverage range each ECL cover, although the ECL have no relation with the parameters shown in the Y-axis. Within an ECL, the configuration of the Random Access (RA) procedure and the Common Search Space (CSS) parameters are fixed. Apart from that, the MCS and user-specific repetitions are selected by the network considering the current UE coverage.

Figure 5.2a shows the resulting NPUSCH transmission power as a function of the RSRP. While the UE is in ECL 0 (i.e. from -80 dBm to -109 dBm RSRP), the UE applies the UL power control (see subsection 3.2.3) to configure its transmission power. Within the ECL 0, we can appreciate several samples reach the maximum transmission power (i.e. 23 dBm) and other lower values for the same RSRP. This is due to the two NPUSCH formats (i.e. UL transmissions and Hybrid Automatic Repeat Request (HARQ) feedback) have a different parameters configuration when applying the UL power control. When the UE belongs to a worse ECL than 0, it directly applies the maximum transmission power (i.e. 23 dBm).

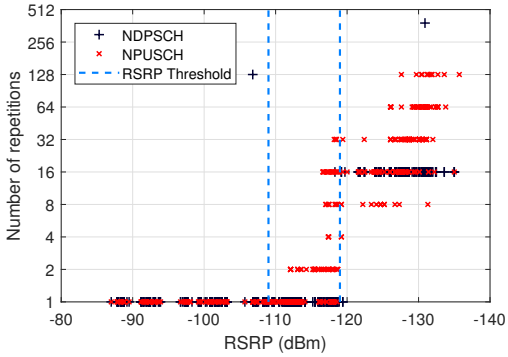
From Figures 5.2b and 5.2c, we can identify that when the UE has good coverage, the eNB sets up a high MCS index (i.e. a higher data rate). However, as the coverage gets worse, the eNB reduces the MCS index and employs repetitions to ensure the successful decoding of the packets at the receiver. This trend in the adaptation of the radio resources configuration is the same trend the UL link



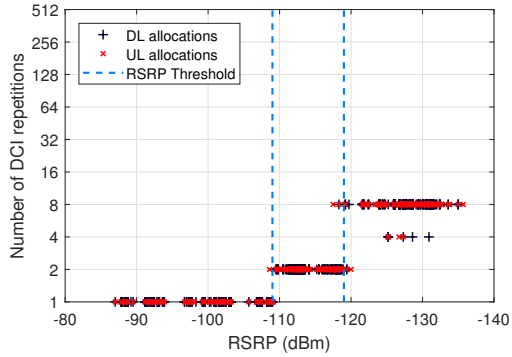
(a) NPUSCH transmission power used as a function of the RSRP.



(b) MCS index configured as a function of the RSRP.



(c) Number of repetitions configured as a function of the RSRP.



(d) Number of DCI repetitions configured as a function of the RSRP.

Figure 5.2: Observed live NB-IoT's network configuration as a function of the RSRP.

adaptation algorithm we proposed in Chapter 4 follows. Figure 5.2c shows the DL configuration is more constant than the UL configuration, both for MCS and repetitions. Additionally, Figure 5.2d illustrates the DCI repetitions configuration has a similar static response than DL configuration. The increase of DCI repetitions matches the change of ECL.

An additional observation during the measurements is, that depending on the network configuration and the UE traffic profile, the period between the last data transfer until the release of the RRC connection can be an issue for the energy saving. The configuration of this period involves a trade-off between DL latency and power consumption due to: i) the use of power saving mechanisms such as

DRX allows the UE to sleep for a certain period and thus, the UE is not reachable by the network; and ii) the RRC Inactivity timer configuration may conflict with the UE traffic and force frequent RRC connection reestablishment. From the traces, we observe the eNB has set the RRC Inactivity timer to 20 s. Because the UE is not configured to use C-DRX, while connected, it will continuously monitor the Narrowband Physical Downlink Control CHannel (NPDCCH) during the 20 s of inactivity in the RRC Connected state. For UEs with UL dominant traffic, the 20 s of continuous reception entail a significant waste of energy.

5.2.3 NB-IoT performance emulation

Let us now analyze the NB-IoT trade-offs using the previously presented testbed. The goal is to study under what radio and traffic circumstances the NB-IoT coverage and latency targets can be meet and at what battery lifetime-cost for the UE. To do that, this section is divided in four parts: i) testbed setup, ii) battery lifetime estimation, iii) detailed energy consumption decomposition, and iv) analysis of the results.

5.2.3.1 Testbed setup

The testbed is used to evaluate the impact of different parameters on the UE energy consumption and delay under diverse coverage and traffic scenarios. To compare the performance of UEs belonging to different ECL, we define three specific ECL dependent configurations. In the testbed, we assume the ECL thresholds seen in the live NB-IoT cell. Then, the testbed RSRP is fixed to force the DUT to connect to the ECL under evaluation. Unlike the live NB-IoT, the configuration of the data channels, Narrowband Physical Downlink Shared CHannel (NPDSCH) and NPUSCH, is fixed in the UXM. This means that different packet sizes would be transferred using the same configuration and relying on segmentation if it is needed.

Table 5.2 summarizes the configuration selected for each ECL. As can be seen, the configuration of the ECLs 1 and 2 is similar for the parameters associated with the RA procedure. This is done to ensure the successful establishment of the RRC connection during the measurements without the possibility of retries that will complicate the comparison of the measurement results. The measurement

results shown are obtained through one empirical measurement at the testbed. Note the latency obtained from the measurements is the period from the UE starts its synchronization to the network, until the RRC connection is released (from packet 1 to 30 in Figure 5.3).

To characterize the impact of different factors in the UE battery lifetime, the

Table 5.2: UXM ECLs configuration [21, 29].

	ECL	**0**	**1**	**2**
Power control	RSRP (dBm/15kHz)	-105	-115	-119
	Reference signal (dBm)	27		
	Estimated path loss (dBm)	132	142	146
	α	0.7		
	$P0_{nominal}$ (dBm)	-67		
	Preamble initial power (dBm)	-104		
NPRACH	Periodicity (ms)	640		
	DCI repetitions for RA	8	16	16
	NPRACH repetitions	2	8	16
	RAR Window Size (pp)	5	5	7
NPDCCH	R_{max} for USS	8	16	32
	R_{max} for CSS RA	8	16	16
	R_{max} for CSS Paging	32		
	G for USS and CSS RA	2		
	DCI repetitions for Paging	32		
	DCI repetitions for USS	1	2	8
	ACK/NACK repetitions	2	4	16
NPDSCH	MCS	10	1	0
	Number of SFs	1	4	6
	NPDSCH repetitions	1	1	16
NPUSCH	Number of subcarriers, N_T	1		
	Subcarrier spacing (kHz)	15		
	MCS	10	5	0
	Number of RUs, N_{RU}	1	2	6
	NPUSCH repetitions	1	8	32
	Modulation	QPSK	QPSK	BPSK
Idle configuration	RRC Inactivity Timer (s)	1		
	Active Timer (s)	120		
	T3412 (h)	11		
	Paging cycle (Radio Frames)	256		
	PTW (s)	20.48		
	eDRX cycle (s)	81.92		

DUT power consumption is evaluated using four test cases:

- `SIZE`: This test focuses on the data payload size. It is performed for the three ECLs.

- `IMCS`: This test analyzes the MCS index for the NPUSCH allocations. To do the comparison fair, the configuration of all evaluated points results in the same Transport Block Size (TBS) of 256 bits by means of adjusting the MCS index together with the number of RUs. Based on the results of the Subsection 5.2.2, only ECL 1 changes the MCS index. Therefore, this test is performed only for ECL 1.

- `ULREP`: This test focuses on the repetitions in NPUSCH. Based on the results of Section 5.2.2, only ECLs 1 and 2 apply repetitions, thus this test includes only ECLs 1 and 2.

- `SCS`: Considering the five possible bandwidth allocations in NB-IoT. This test compares the performance of these configurations. If the UE has not reached its maximum transmission power, following the power control mechanism, ideally, the decrease of bandwidth allocation will increase the duration of the RU and decrease the transmission power equally. To analyze this effect, `SCS` test only evaluates ECL 0 where the UE can use a smaller transmission power. Additionally, due to only one of the considered DUTs supports multi-tone configurations, this test is performed only for the DUT C.

Table 5.3: Test cases with the UXM settings.

Test case	**Sweeping parameter**	**Other settings**
`SIZE`	$Payload = \{20, 50, 200, 500\}$ B	$ECL = \{0, 1, 2\}$
`IMCS`	$MCS^{UL} = \{0, 1, 2, 3, 4, 6, 8\}$	$N_{RU} = \{10, 6, 8, 5, 4, 3, 2\}$ $ECL = 1$
`ULREP`	$N_{REP}^{npusch} = \{1, 2, 4, 8, 16, 32, 64\}$	$ECL = \{1, 2\}$
`SCS`	$SCS = 3.75$ kHz, $N_T = 1$ $SCS = 15$ kHz, $N_T = \{1, 3, 6, 12\}$	$P0_{nominal} = -84dBm$ $ECL = 0$

Table 5.3 summarizes the specific UXM settings for each test case. The parameters not included in Table 5.3 are fixed to the configuration of the specific ECL the UE belongs. For all test cases, the N6705B sampling time is 1 ms, and the current range is set to auto. This enables to automatically change the measurement range during the measurements. The DUT and the UXM are wired (see Figure 5.1) and we do not add noise in the emulation of the connection between both.

Additionally, Table 5.4 lists the measured average power consumption levels for the three devices evaluated. We can see devices A and B have similar average power consumption values. On the contrary, device C have a higher power consumption on transmission and standby modes compared to the other devices. This makes the device C less suitable than the others. However, this could be improved in later firmware updates of the device.

5.2.3.2 Battery lifetime estimation

From the measured energy consumption, we estimate the UE battery life using a methodology similar to [149]. Assuming a smart utility sensor, the applied traffic profile follows a periodic UL reporting with a predefined Inter-Arrival Time (IAT). Prior to the periodic reporting, the UE needs to reestablish the RRC connection and it thus performs the CP procedure.

Let us define a battery lifetime estimation with four phases for modeling the periodic traffic pattern:

P1: UE exits Power Saving Mode (PSM), establishes the RRC connection, and sends the data using the CP procedure.

P2: The UE continuously monitors the NPDCCH until RRC connection is released.

Table 5.4: Measured average power consumption.

	Device A	**Device B**	**Device C**
Transmit at 23 dBm	765 mW	731 mW	1030 mW
Receive	242 mW	215 mW	168 mW
Sleep	29.1 mW	17.8 mW	17.7 mW
Standby	11.13 μW	14.14 μW	24.3 μW

P3: The UE uses extended/enhanced Discontinuous Reception (eDRX) until the Active Timer expires.

P4: The UE sleeps using PSM until the next transmission period begins.

Therefore, the energy consumption when sending one UL report, E_{report}, can be derived as:

$$\begin{aligned} E_{report} &= E_{conn} + E_{rel} + E_{idle} + P_{standby} \cdot T_{sleep} \\ T_{sleep} &= IAT - T_{conn} - T_{rel} - T_{idle} \end{aligned} \tag{5.1}$$

where E_{conn}, E_{rel}, and E_{idle} denote the energy consumed in joules within the phases P1, P2, and P3, respectively. $P_{standby}$ is the average power consumption in PSM, and T_{conn}, T_{rel}, T_{idle}, T_{sleep} the duration in seconds of the phases P1, P2, P3, and P4, respectively. Finally, we can estimate the energy consumed per day E_{day} and the battery lifetime in years Y as follows:

$$\begin{aligned} E_{day} &= \frac{D_{day}}{IAT} \cdot E_{report} \\ Y &= \frac{C_{bat}}{(E_{day}/3600) \cdot 365.25} \end{aligned} \tag{5.2}$$

where D_{day} denotes the duration of one day in seconds, and C_{bat} is the battery capacity in watt-hour. In the testbed, we consider the periodic UL reports are UDP packets with 50 B of payload and $C_{bat} = 5$ Wh [83].

To provide a fair comparison of the battery lifetime estimation for different measurements of the same ECL and different ECLs, we take one measurement of the DUT during the phase P3 per ECL as a reference for the rest of the measurements of the same ECL. In this analysis, the configuration of the I-DRX and eDRX is always the same. Therefore, taking the measurement of a phase P3 per ECL avoids measurements with a different number of Paging Time Window (PTW) that will hinder the comparison of the energy consumption due to signaling.

5.2.3.3 Detailed energy consumption decomposition

In addition to the four phases detailed previously, we divide the UE's steps between the transmission of two periodic reports in six different segments. Figure

5.3 shows the steps considered at each segment. This division allows us to specify the parts that consume more energy in the results. Thus, the different segments are defined as:

- SYNC: UE time and frequency synchronization with a cell, i.e., the decoding of Primary Synchronization Signal (NPSS) and Secondary Synchronization Signal (NSSS), and the later MIB and SIBs.
- RA: RA procedure, i.e., from the transmission of the preamble up to the HARQ Acknowledgment (ACK) of the RRC Connection Setup.
- CONN: The periodic report is transmitted piggybacked as NAS signaling. This segment includes the reception of the NAS Service Accept and its later HARQ ACK.
- REL: This segment includes the UE inactive period while the UE is still in RRC Connected, the reception of the RRC Connection Release, and its later Radio Link Control (RLC) ACK.
- EDRX: UE discontinuously monitoring NPDCCH using eDRX until the expiration of the Active Timer.
- PSM: UE sleeps using PSM until the next periodic report is generated.

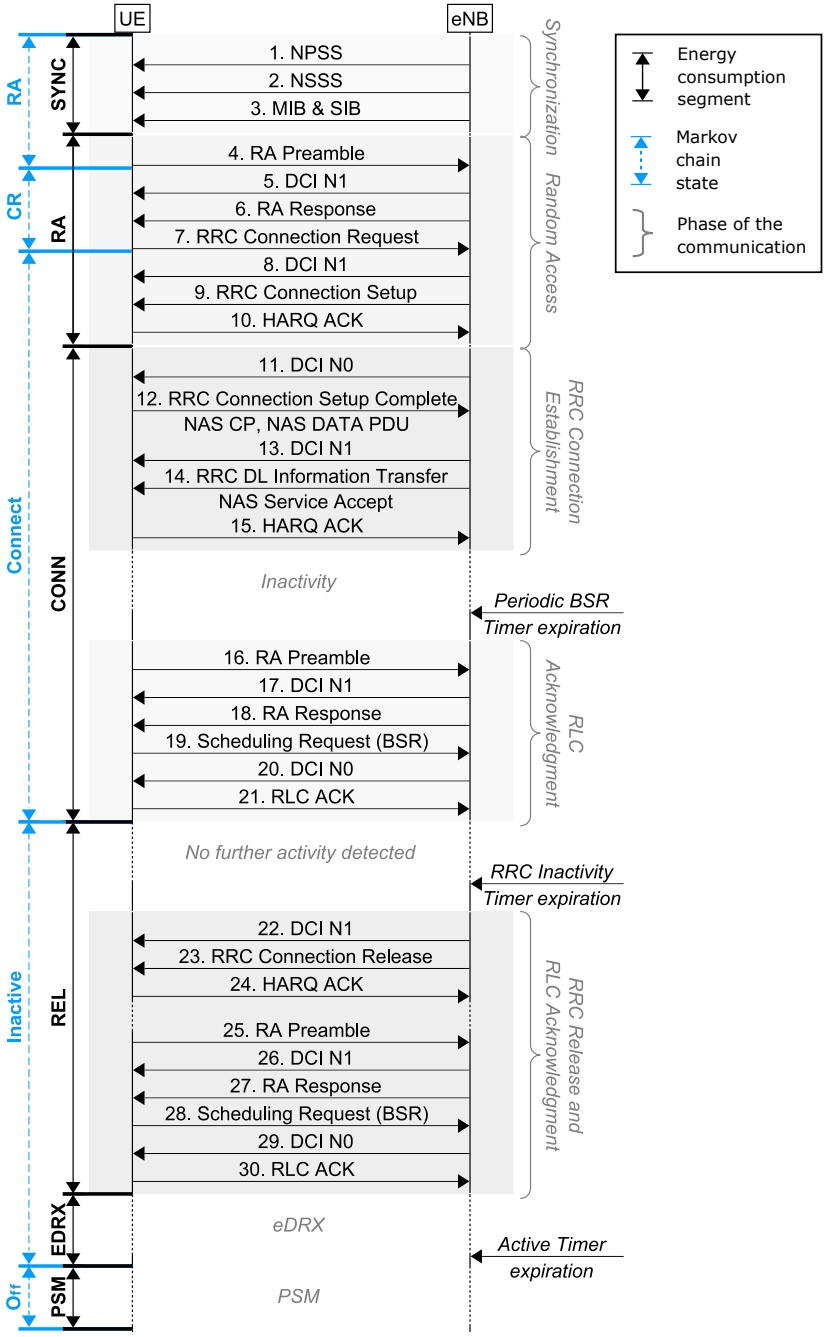


Figure 5.3: Signaling flow of a mobile Originated data transport in CP [9], the energy segments considered, and the Markov chain states.

5.2.3.4 Results

This subsection presents the results obtained with the testbed using three different DUTs. Figure 5.4 shows the `IMCS` test case battery lifetime evolution of the three DUTs for an IAT of 24 h as a function of the MCS index assuming the ECL 1. All configurations result in the same TBS. Hence, as the MCS index increases, the number of RUs is reduced accordingly (see NPUSCH TBS Table 3.2). The measurement results show that the battery lifetime increases with the MCS. The reason is higher MCS indexes have higher data rate, thus, less time on air for the reception or the transmission. This prolongs the sleep time, i.e., battery lifetime, and use of less radio resources. However, these high MCS indexes only can be used while the UE has good coverage as they are less robust. Note there is a reduction of the battery lifetime in MCS 2. It is due to this configuration has a higher number of RUs (8 RUs) than MCS 1 (6 RUs).

Figure 5.5 presents the `ULREP` test case battery lifetime evolution of the three DUTs for an IAT of 24 h as a function of the number of NPUSCH repetitions. The NPUSCH repetitions have significant impact on the battery lifetime. In this figure, selected evaluations with the same number of NPUSCH repetitions for the ECLs 1 and 2 are provided. These values exemplify the effect on the battery lifetime due to RA and the DL channels, as these configurations change for each ECL. The same trade-off as in Figure 5.4 is observed, when using a more robust

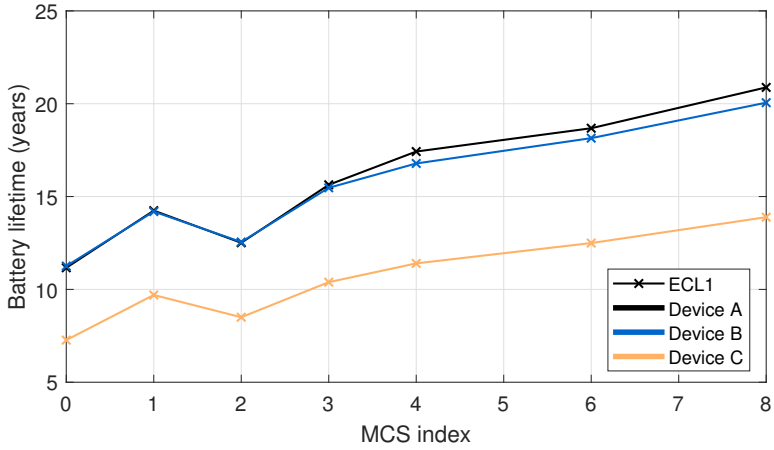


Figure 5.4: Battery lifetime evolution when increasing the MCS index for `IMCS` test case and assuming an IAT of 24 h.

configuration (i.e. extend coverage), more radio resources are used and the energy consumption increases.

In both Figures 5.4 and 5.5, devices A and B have a similar performance. However, device C obtains worse results due to its high power consumption in standby. For smaller IATs the obtained results show the same trend.

Despite not shown in a figure, in `SCS` test case for device C and IAT 24 h, all bandwidth allocation configurations where $SCS = 15$ kHz achieve similar results with a battery lifetime around 19 years. For these configurations, the increase of the duration of the RU can be compensated by the associated reduction of the transmission power. This holds true if the power control mechanism can be still applied and the UE is not using its maximum transmission power. However, when reducing the SCS to 3.75 kHz, the battery lifetime decreases approximately 14%. This is due to two reasons: i) the increase of the duration of the RU and the decrease of the power consumption are unequal in this case; and ii) the increase of the duration of the NPUSCH format 2 transmissions (i.e. HARQ ACKs) that have a different power control configuration than usual NPUSCH format 1 transmissions [21].

Focusing on the NB-IoT targets, the battery lifetime target of 10 years is satisfied for the majority of the configurations evaluated, when the UE does

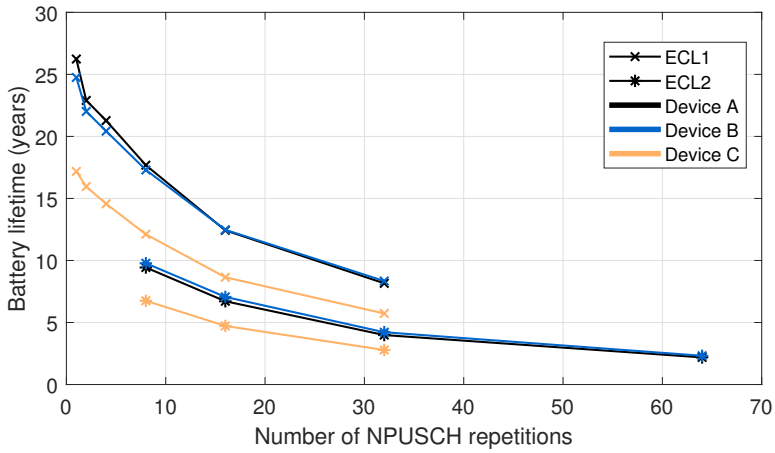


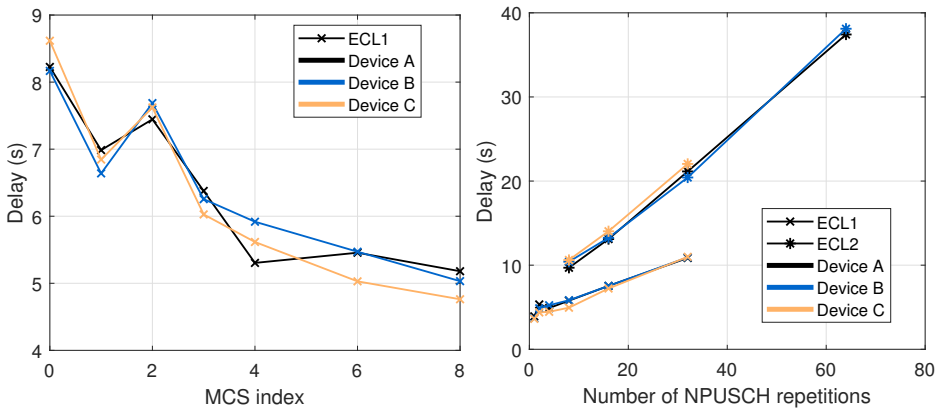
Figure 5.5: Battery lifetime evolution when increasing the number of NPUSCH repetitions for `ULREP` test case and assuming an IAT of 24 h.

not belong to the worst ECL (i.e. ECL 2) or there is no need for extensive use of repetitions (> 16). In terms of the maximum latency of 10 s, the same configurations that do not achieve the previous target also exceed the 10 s, as can be seen in Figure 5.6 for `IMCS` and `ULREP` test cases.

Figure 5.7 represents the energy consumed by the UE during the six defined segments in the Subsection 5.2.3.3. This analysis is done for device A, three IATs, and the three ECLs. From the figure, we can derive some observations:

- The energy consumption during the CONN segment becomes dominant, when the UE belongs to a worse ECL. For ECL 2, the CONN segment requires approximately 90% of the energy consumption. Thereby, for UEs belonging to ECL 2, signaling optimizations such as Early Data Transmission (EDT) are essential. This s due to the EDT mechanism enables data transmission during the RA procedure.

- For UEs with large IATs, the UE will stay more time in PSM. When the IAT equals 24 h, the energy consumed due to PSM is 57% and 34% for ECLs 0 and 1, respectively. Additionally, for ECLs 0 and 1 with IATs smaller than 24 h, the EDRX segment consumes approximately 17%. This value can be reduced if the Active Timer and the eDRX configuration are selected



(a) Latency evolution when increasing the MCS index for `IMCS` and assuming an IAT of 24 h.

(b) Latency evolution when increasing the number of NPUSCH repetitions for `ULREP` and assuming an IAT of 24 h.

Figure 5.6: Latency evolution for `IMCS` (5.6a) and `ULREP` (5.6b) test cases.

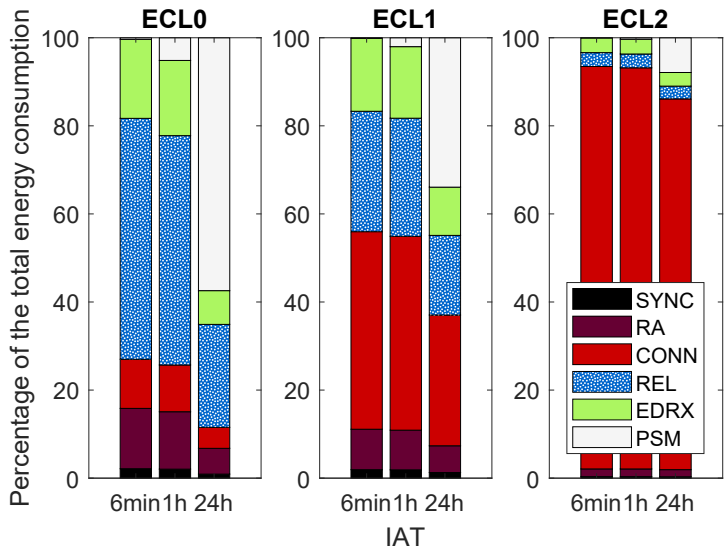


Figure 5.7: Decomposition of the energy consumption components when `SIZE` test case and 50B of payload for device A, three different IATs, and the three ECLs.

accordingly to the UE traffic profile. For example, if the UEs only have UL traffic, there is no need to keep the UE to be reachable by the network. Thereby, in these cases a shorter Active Timer or PTW values can help to reduce active periods before entering PSM. In other cases, UEs may have infrequent DL traffic together with their UL traffic can stay reachable by the network while saving a lot of energy if a long eDRX cycle is combined with a small PTW. Long eDRX cycles allow long sleep periods with a few monitoring periods during the small PTW.

- The ECLs 0 and 1 have a significant consumption in the REL segment. This is mainly due to the RRC Inactivity Timer being fixed to 1 s and the C-DRX is not used. Therefore, the device is continuously monitoring the NPDCCH. To optimize the consumption during the REL segment, features such as C-DRX and Release Assistance Indication (RAI) can be applied. The RAI is an improvement for the CP procedure that allows the UE to notify the core network if there are no more packets pending to immediately release the radio connection if possible [9]. Like in the previous bullet point, here we have another trade-off between the UE reachability (that will impact

the DL latency), and the energy consumption. The configuration of the RRC Inactivity Timer is done trying to avoid connection reestablishments originated by the UE or the network with pending traffic. However, for UEs with infrequent traffic, this period is a waste of energy as no more traffic is expected in this short period. In these cases, C-DRX can save a significant amount of energy if its cycles are configured long enough to allow the device to sleep as much as possible until the connection is released. On the contrary, for UEs with frequent traffic or burst traffic and including DL responses, short C-DRX cycles (or greater active periods during the C-DRX cycles) are more suitable to maintain the DL latency.

- Particularly for ECL 0, the RA segment consumes more than the CONN segment. This is due to the RA procedure being configured more redundant as it is crucial to start the communication with the eNB. This procedure is inevitable, but the combined consumption of RA and CONN segments could be improved by using the EDT feature if it is suitable to the UE traffic profile.

Additionally, Table 5.5 summarizes the delay experienced by the UE during the six segments. Note the PSM segment is shaded in gray as the duration of this segment is assumed equal to the evaluated IAT, see Equation (5.1). In this table, we can observe ECL 0 and 1 successfully meet the 10 s target, but ECL 2 exceeds the limit as the CONN segment alone already consumes 17s. Furthermore, the segments that depend on NPRACH resources (i.e. SYNC, RA, REL) do not return this predictable increment of delay when the UE belongs to a worse ECL. This is due to the fluctuation of the wait for the occurrence of NPRACH resources in the next RA opportunity that can range from a few ms to the NPRACH period value in the measurements.

Finally, Figure 5.8 presents the battery lifetime as a function of the IAT considering four different UL UDP packet sizes for `SIZE` test case. The figure shows the results of device A, although the results are similar for the other devices studied in this work. Achieving 10 years battery lifetime with ECL 2 is a major challenge for any of the tested packet sizes, due to the many radio resource repetitions, unless packets are only sent a few times per week. For ECL 1 sending up to 1 packet every 11 h is feasible for small packets, while ECL 0 supports up

Table 5.5: Decomposition of the delay components considering a UDP report of 50B of payload for device A and the three ECLs.

	Delay per segment (s)					
	SYNC	**RA**	**CONN**	**REL**	**EDRX**	**PSM**
ECL 0	0.20	0.87	0.25	2.72	121.31	
ECL 1	0.69	0.91	1.41	2.81	129.15	
ECL 2	0.66	0.61	17.03	2.3	121.4	

to 500 bytes every 6 h with the 10 years battery lifetime target.

5.2.3.5 DUT behavior examination

Before we start the validation of the NB-IoT analytical model proposed in Chapter 3, we use the previous empirical results to analyze if the model correctly describes the behavior of the UE. Figure 5.9 shows an example of the measured power consumption while the DUT performs the CP procedure using the testbed. With this figure and other measurements, we can identify the different steps the DUT performs while communicating with the UXM. From this analysis, we find the following item that were not considered in the model of Chapter 3:

- Reception of the NAS Service Accept within a RRC DL Information Trans-

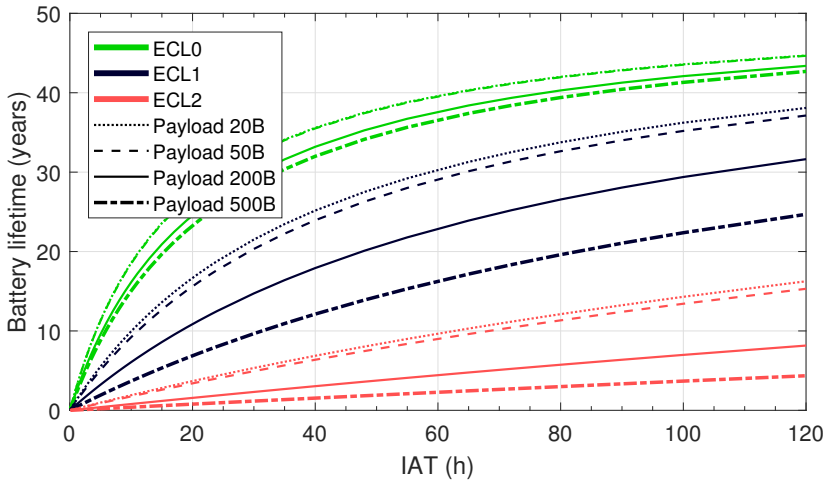


Figure 5.8: `SIZE` test case battery lifetime as a function of the IAT for device A and four UDP data payloads.

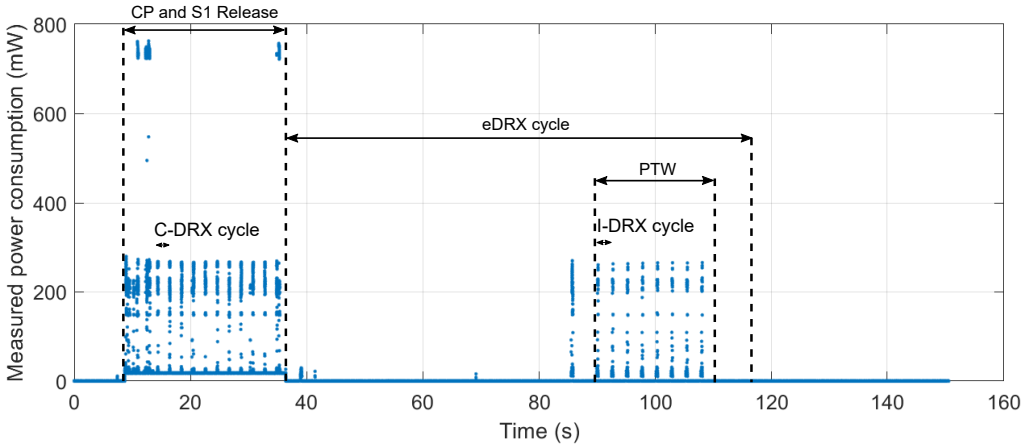
fer message at the end of the CP procedure (message 14 of Figure 5.3).

- HARQ ACK mechanism to confirm the DL receptions.
- Periodic Buffer Status Report (BSR) mechanism.
- Scheduling request to send ACKs at the RLC layer, e.g., if the last RRC DL Information Transfer if not confirmed when the periodic BSR timer expires and to confirm the reception of the RRC Release message.
- DCIs reception between packet segments.
- Short synchronization before paging occasions due to the DUT enters a standby mode and has to resynchronize with the network before the paging occasion occurs. Figure 5.10b illustrates an example of this short synchronization.
- Use of eDRX while the device is in RRC Idle.
- Unexpected noticeable short waits. For example: i) between messages 10 and 11 of Figure 5.3 the DUT has a wait before the reception of the UL grant; ii) after the confirmation of the RRC Release (i.e. message 30 of Figure 5.3), the DUT does not enter RRC Idle state immediately. Figure 5.10a illustrates an example of this wait before using eDRX.

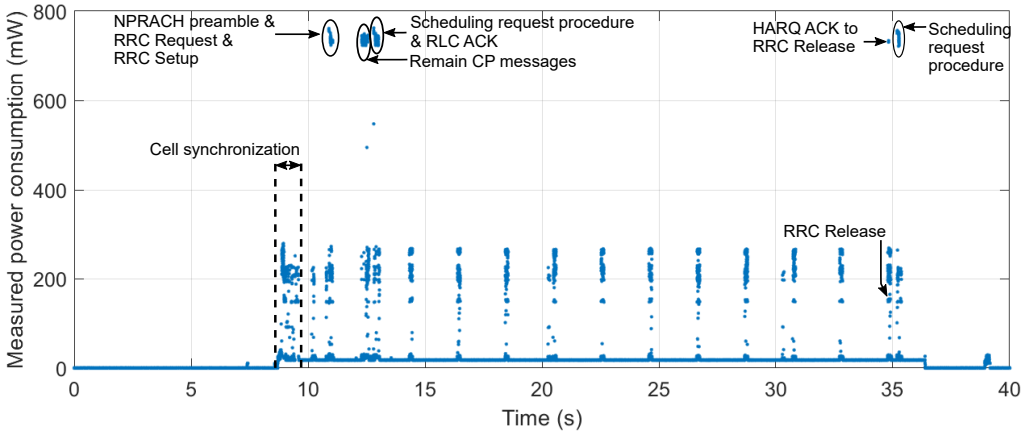
5.3 NB-IoT model validation

Let us now enter the second part of the chapter. In this part, we present and validate an NB-IoT energy consumption model. This model is based on the one presented in Chapter 3 with the following novelties:

- Inclusion of several details in the model that increase its accuracy, such as: HARQ ACKs, eDRX, periodic BSR, and DCI reception between segments.
- The model only focuses on the CP procedure as this procedure is the one used by the DUTs in the testbed. To ease the analysis, we do not consider the Tracking Area Update (TAU) procedure in the model. Therefore, we assume the UE always exits PSM due to the generation of a new UL report.



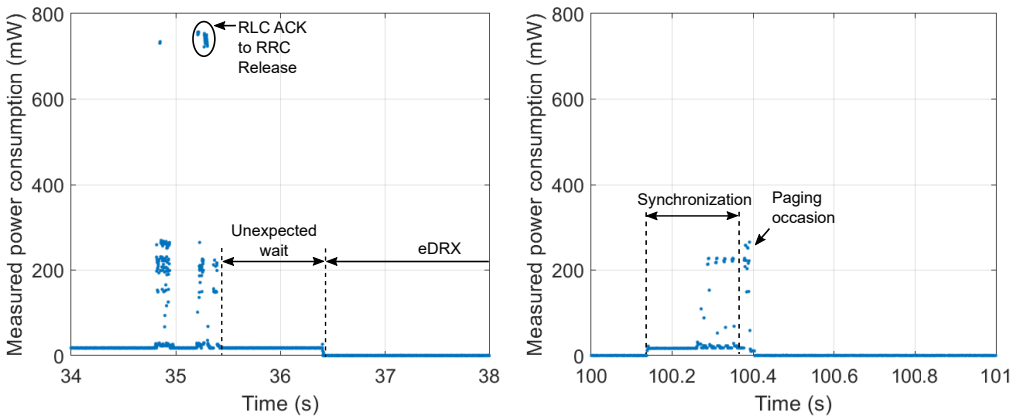
(a) Example of the CP and S1 Release procedures, and eDRX monitoring while the DUT is in RRC Idle.



(b) Zoom of the above example while DUT is performing the CP and S1 Release procedures.

Figure 5.9: DUT measured power consumption example in our testbed.

Compared to the large model of Chapter 3, the analytical model validated in this chapter only includes the Markov chain states involved in a successful connection. That means we do not consider RRC connection establishment failures or capacity constraints. This is due to the later validation of the model does not consider failures in the connection establishment. Therefore, only the UE energy consumption (and latency) estimates are validated.



(a) After RRC Release ACK, the DUT does not enter RRC Idle state immediately.

(b) DUT short synchronization before paging occasion.

Figure 5.10: Example of a few details of the DUTs behavior in our testbed.

To begin with the analysis, let us assume a cell with an eNB with an NB-IoT carrier deployed in-band, and one UE camping on it. The UE transfers UL reports of size L periodically to the eNB. These reports are destined to an IoT server. We assume there is no ACK from the IoT server following the UL report. Additionally, we assume the configuration of CSS and UE-specific Search Space (USS) is equal. To send these periodic UL reports, the UE performs the CP procedure.

The assumed behavior of the UE is as follows (see Figure 5.3). Prior to the periodic UL data transmission, the UE needs time and frequency synchronization with a cell and it thus decodes NPSS and NSSS. Next, the UE gets the core cell information from the MIB and SIBs. At this point, the UE starts the RA procedure to begin the communication with the network. After the successful contention resolution of the RA, the UE and eNB reestablish the RRC connection, and the UE switches to RRC Connected state. Next, as part of the CP procedure, the UE can receive resource allocations through DCI, and send/receive data from the network over NPUSCH/NPDSCH. The CP procedure uses RRC DL Information Transfer packets to forward packets to the UE. While the UE is communicating with the eNB, if the last RRC DL Information Transfer packet has not been acknowledged at RLC layer, the UE requests resources to

send the confirmation when the periodic BSR timer expires, starting a Scheduling Request (SCR) procedure. This is because the RRC DL Information Transfer packet is sent in RLC Acknowledgment Mode (AM). Later, if the eNB detects an inactivity period greater than the defined RRC Inactivity Timer, the eNB initiates the RRC Release procedure to switch the UE to RRC Idle. To save battery, after a period of discontinuous NPDCCH monitoring, the UE moves to PSM.

To organize our explanation, the analytical model validation is divided in five different parts: i) main considerations of the energy consumption estimation in the model, ii) reduced Markov chain description, iii) energy and delay expressions derivation, iv) experimental setup description, and v) analysis of the results.

5.3.1 Detailed energy consumption estimation

The key of the analytical model is define the duration of the transmissions/receptions that will significantly impact the energy consumption. To achieve that, this section presents the main considerations underlining in the model.

5.3.1.1 Power saving features

As detailed in subsection 3.2.4, to prolong battery lifetime, an NB-IoT UE can use eDRX and PSM. Both techniques enable the UE to enter a power saving state where it is not required to monitor for paging/scheduling information.

For simplicity, the model assumes Connected DRX (C-DRX) is only used after the UE ends its communication with the eNB (i.e. after packet 21 in Figure 5.3). Therefore, the number of C-DRX cycles can be approximated as $N_{cycles}^{CDRX} = \left[\frac{T_{inactivity} - T_{DRXi}}{T_{LC}}\right]$, where [] denotes the nearest integer function (a detailed C-DRX description can be found in subsection 3.5.3.2).

Later, when the UE is in RRC Idle state, the Active Timer (T3324) T_{active} controls the period the UE is reachable by the network. During this period, there are a number of eDRX cycles. The number of eDRX can be estimated as $N_{cycles}^{eDRX} = \left[\frac{T_{active}}{T_{eDRX}}\right]$, where T_{eDRX} is the duration of the eDRX. Each eDRX cycle has an active phase controlled by the PTW timer T_{PTW}, and a sleep phase for the remaining period. Withing the PTW, there are several Idle DRX (I-DRX) cycles (or paging cycles) that can be estimated as $N_{cycles}^{IDRX} = \left[\frac{T_{PTW}}{T_{PC}}\right]$, where T_{PC}

is the duration of the I-DRX cycles. After the expiration of the T_{active}, the UE enters PSM.

5.3.1.2 Power analysis

To model the UE energy consumption, we assume its behavior can be described as shown in Figure 5.11. The model uses the four power levels defined in the subsection 3.4.2 (i.e. transmission P_{TX}, reception P_{RX}, inactive P_i, and standby P_s), adding a the new level UL gap (P_{ULgap}) seen in the empirical measurements. This power level occurs when the UE is active and waiting for the end of the UL transmission gap.

The studied NB-IoT UEs will enter standby mode whenever possible (i.e. when the system is quiescent). Then, the P_s power level can be seen in PSM as well as during I-DRX inactive periods.

5.3.1.3 Synchronization

In order for a UE to connect to the network, it must synchronize with the serving cell. The model considers two different types of synchronization in the analysis:

- Initial synchronization: After the UE exits PSM, it needs time and frequency synchronization with the cell.
- Short synchronization before paging: The UE's standby periods while performing I-DRX cause the UE has to wake up shortly before the paging occasion to do a short synchronization.

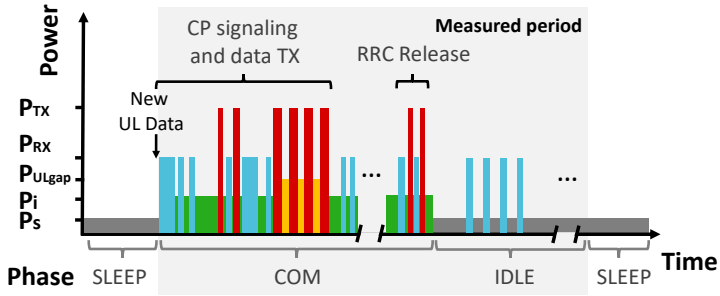


Figure 5.11: Example of the considered power levels for the model and the phases in the measurement setup.

The composition of both synchronization processes will depend on the current coverage of the UE and the cell configuration. To ease their inclusion in the model, we consider a simplified definition. The duration of the initial synchronization depends on three parameters: i) Average required synchronization time T_{sync}; ii) Waiting time for the occurrence of the MIB T_{MIB-I}; and iii) MIB's reading time T_{MIB-RX}. The value of T_{sync} is based on the performance summary found in [150], while the values of T_{MIB-I} and T_{MIB-RX} from [23]. For the short synchronization before paging, its definition is based on empirical measurements performed in this work. Then, to estimate the energy consumed, from the measurements we obtain an average power $P_{IDRX_{sync}}$ and duration $T_{IDRX_{sync}}$.

5.3.1.4 NPDCCH scheduling

The periodic occurrence of NPDCCH in this analysis is the same as the one used in Chapter 3. As a reminder, the waiting time until the next NPDCCH is derived as:

$$T_{WDC}(x_1, x_2, ..., x_n) = pp - mod(T_{x_1} + T_{x_2} + ... + T_{x_n}, pp) \tag{5.3}$$

where $x_1, x_2, ..., x_n$ are the considered steps occurred between NPDCCH occasions, T_{x_n} is the duration of the x_n step, pp is the PDCCH period derived as $R_{max} \cdot G$ (see subsection 3.2.2), and $mod()$ is the modulus after division function. In this analysis, most of the x_n steps between two occurrences of the NPDCCH are: i) DCI's reception time; ii) Wait for the start of NPDSCH/NPUSCH reception/transmission after the end of its associated DCI; and iii) Packet reception/-transmission time.

5.3.2 Reduced Markov chain

To model the UE's behavior we use a Markov chain. Figure 5.12 depicts the proposed Markov chain used to model the UE's behavior. The proposed model does not consider RA failures and access barring. Figure 5.3 shows the steps considered at each state of the Markov chain. The states and transitions between are defined as:

- State Off: This state models the situation in which the UE has no new UL

report to transmit. In this state, the UE is using PSM. The UE changes to the *RA* state when a new UL report is generated.

- State *RA*: The *RA* state represents the synchronization and transmission of the RA preamble. This transmission triggers the transition to the *CR* state.

- State *CR*: This state comprises the request for the RRC connection. After the reception of the Random Access Response (RAR) and later transmission of the RRC Request messages, the UE transfers to the *Connect* state.

- State *Connect*: This state models the establishment of the RRC connection, and the end of the CP procedure (including the RLC AM ACK of the last RRC DL Information Transfer). After the completion of the CP procedure, the UE transfers to the *ACK* state if there is a pending DL response from the IoT server, otherwise, it transfers to the *Inactive* state.

- State *ACK*: This state represents the reception of the DL response from the IoT server. After this reception, the UE transfers to the *Inactive* state.

- State *Inactive*: This state models the period the UE is still reachable by the network before entering PSM, i.e., RRC Inactivity Timer period using C-DRX, reception of the RRC Release, the transmission of its RLC AM ACK, and the Active Timer period using I-DRX. At the expiration of the Active Timer, the UE transfers to the Off state.

We assume the traffic is Poisson distributed with rate λ_{app} packets per ms. The UE's data rate is derived from its average IAT in ms, therefore $\lambda_{app} = 1/IAT$. Let p_{on} denote the probability of having UL traffic in a ms, expressed

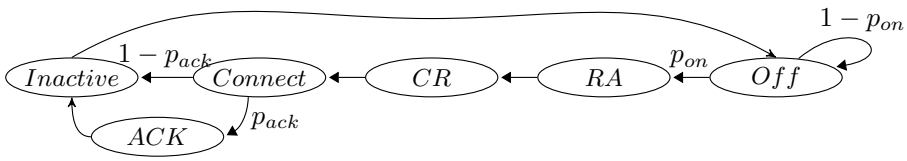


Figure 5.12: Markov Chain model for an NB-IoT's UE.

as $p_{on} = 1 - e^{-\lambda_{app}}$. As we assume there is no DL response from the IoT server, $p_{ack} = 0$.

Denote b_j as the steady state probability that a UE is at j state. As the state space of our Markov chain is an irreducible recurrent set, we know the steady-state probability distribution exists. Then, the stationary probability for each state can be derived as:

$$\begin{aligned} b_{RA} &= p_{on} \cdot b_{off} \\ b_{CR} &= b_{RA} = p_{on} \cdot b_{off} \\ b_{Connect} &= b_{CR} = p_{on} \cdot b_{off} \\ b_{ACK} &= p_{ack} \cdot b_{Connect} = p_{ack} \cdot b_{RA} \\ b_{Inactive} &= b_{Connect} = p_{on} \cdot b_{off} \end{aligned} \tag{5.4}$$

By imposing the probability normalization condition, we obtain b_{off} as:

$$b_{off} = \left(1 + p_{on}\left(4 + p_{ack}\right)\right)^{-1} \tag{5.5}$$

5.3.3 Energy consumption and delay analysis

Now let us calculate the average energy consumption when performing the CP procedure. The energy consumption is based on the average power and duration of each Markov chain state. The analysis presented is similar to the analysis presented in Chapter 3, specifically in subsection 3.5.3, highlighting in this chapter the novelties of the analysis. Table 5.6 contains the definition of the parameters used in the following analysis. Note in the analysis the power unit is the mW and the delay is ms. Therefore, the resulting energy consumption is in μJ.

The analysis is divided into three parts. Firstly, we detail the energy consumption while receiving or transmitting packets or signaling in NB-IoT. Secondly, we present estimates of the energy consumption per Markov chain state. Finally, the battery lifetime is estimated.

5.3.3.1 Packet energy estimation

The estimation of the packet energy is similar to the one presented in subsection 3.5.3.1. Therefore, in this subsection we only describe the novelties included in

Table 5.6: Variables and parameters of the model.

	Parameter	Value	Description
Energy	E_j	Variable	Average energy consumption in state j (μJ)
	$E(x)$	Variable	Packet x average energy consumption (μJ)
	P_l	Variable	Average power consumption in mW for level l (where $l \in \{TX, RX, ULgap, i, s\}$)
Synchronization	$P_{IDRX_{sync}}$	A: 65.6 B: 34.5	Average power consumption while performing short synchronizations in I-DRX (mW) for Device A and B
	$T_{IDRX_{sync}}$	250	Average duration of the short synchronizations during I-DRX (ms)
	T_{sync}	547.5	Average initial synchronization time (ms)
	T_{MIB-I}	103	MIB waiting time (ms)
	T_{MIB-RX}	8	MIB reception time (ms)
RA	$T_{RAPeriod}$	640	RA periodicity (ms)
	T_{PRE}	5.6	Preamble format 0 duration (ms)
	N^{RA}_{REP}	Variable	Number of preamble repetitions
Gaps	T^{RA}_{Gap}	0, 40	NPRACH gap duration (ms). If NPRACH repetitions > 64, $T^{RA}_{Gap} = 40$, otherwise $T^{RA}_{Gap} = 0$
	$T^{s}_{GapPeriod}$	296, 128	Gap periodicity for UL and DL (ms)
	T^{s}_{GapDur}	40, 32	Gap duration for UL and DL (ms)
	$DLGapThr$	64	DL gap threshold
Scheduling	T_{wDC2US}	8	NPUSCH transmission start after the end of its associated DCI (ms)
	T_{wDC2DS}	5	NPDSCH transmission start after the end of its associated DCI (ms)
	T_{ACK-k0}	13	Delay for the ACK of a DL packet (ms)
	$T_{PeriodBSR}$	8	Buffer Status Report (BSR) Timer (pp)
C-DRX & I-DRX	T_{DRXi}	2	Period the UE should remain monitoring NPDCCH before starting C-DRX (pp)
	$T_{inactivity}$	20	RRC Inactivity timer (s)
	T_{onD}	Variable	On duration timer during a C-DRX cycle
	Npp_{onD}	8	Number of consecutive NPDCCH periods to monitor at the start of C-DRX (pp)
	T_{LC}	2.048	C-DRX Long Cycle (s)
	T_{wIDRX}	1.1	Wait before entering I-DRX after send the RRC Release ACK (s)
	T_{PC}	2.56	I-DRX Paging Cycle (s)
	T_{PTW}	20.48	PTW cycle duration (s)
	T_{eDRX}	81.92	eDRX cycle duration (s)
	T_{active}	120	Active Timer duration (s)
	H_{RLCMAC}	4	RLC/MAC headers size (B)

the model. Table 5.7 summarizes the UL and DL messages exchanged and its respective sizes in bytes.

DCI allocations: we first derive the reception time needed for the DCI.

Table 5.7: UL and DL messages exchanged in the analysis.

Message	Acronym	Direction	Size (bytes)
RAR	*rar*	DL	32
RRC Connection Request	*req*	UL	9
RRC Connection Setup	*set*	DL	10
RRC Connection Setup Complete + NAS Data PDU (50 bytes UDP payload)	*setCmp*	UL	108
RRC DL Information Transfer + NAS Service Accept	*accept*	DL	15
Scheduling Request	*scr*	UL	9
RLC ACK	rlc_{Ack}	UL	2
RRC Connection Release	*rel*	DL	2

Compared to the computation at Chapter 3, in this version we consider the reception of the DCI may be extended. This is owing to other channels and signals are present in the DL SFs and the possible occurrence of DL gaps. Due to the broadcast information present in the NB-IoT frame, approximately only 14 out of 20 SFs are available for control and data transmissions. This is a rough assumption based on the broadcast information comes with a low period, as explained in subsection 3.2.1.2. Considering this limitation, the DCI reception time $T^{DL}(dci)$ in ms is calculated as follows:

$$T^{DL}(dci) = \left\lceil N_{REP_{dci}} \cdot \left(\frac{20}{14} - 1 \right) \right\rceil + N_{REP_{dci}} \tag{5.6}$$

where $N_{REP_{dci}}$ is the number of DCI repetitions. Due to the DL SF duration is 1 ms and we assume each DCI copy requires a whole DL SF, the number of DCI repetitions equals the duration of the DCI. If $R_{max} > N^{DL}_{GapThr}$ there will be DL gaps in the reception. The total duration of the gaps $T^{DL}_{Gap}(dci)$ is derived as:

$$T^{DL}_{Gap}(dci) = \left\lfloor \frac{N_{REP_{dci}}}{T^{DL}_{GapPeriod} - T^{DL}_{GapDur}} \right\rfloor \cdot T^{DL}_{GapDur} \tag{5.7}$$

Finally, we can estimate the DCI's energy consumption, $E_{rx}(dci)$, and delay, $T_{rx}(dci)$, as:

$$\begin{aligned} E_{rx}(dci) =& P_{RX} \cdot T^{DL}(dci) + P_i \cdot T^{DL}_{Gap}(dci) \\ T_{rx}(dci) =& T^{DL}(dci) + T^{DL}_{Gap}(dci) \end{aligned} \tag{5.8}$$

UL packet: The estimated transmission time for packet x, $T^{UL}(x)$, the total duration of the UL gaps, $T^{UL}_{Gap}(x)$, and the number of segments $N_{seg}(x)$, are calculated as (3.21) and (3.22). The difference of this estimation compared to the computation at Chapter 3 is the inclusion of the reception of DCIs between segments. If the number of segments is greater than 0, $N_{seg}(x) > 0$, we estimate the energy consumed due to the reception of DCIs between the packet segments as:

$$\begin{aligned} E_{seg}(x) = (N_{seg}(x) - 1) \cdot (P_{RX} \cdot T_{rx}(dci) + \\ P_i \cdot (T_{WDC}(x_1, x_2, ..., x_n) + T_{wDC2US})) \end{aligned} \tag{5.9}$$

where T_{WDC} is the waiting time until the next NPDCCH occurrence and T_{wDC2US} is the wait from the reception of the DCI with the UL allocation to the NPUSCH transmission start. The steps of T_{WDC} (i.e. $x_1, x_2, ..., x_n$) depend on the last transmission/reception, as explained in subsection 5.3.1.4. In this case, the last transmission is a segment of the packet. Finally, the estimated energy consumption and delay due to the transmission of the packet x is:

$$\begin{aligned} E_{tx}(x) =& P_{TX} \cdot T^{UL}(x) + P_{ULgap} \cdot T^{UL}_{Gap}(x) + E_{seg}(x) \\ T_{tx}(x) =& T^{UL}(x) + T^{UL}_{Gap}(x) + (N_{seg}(x) - 1) \cdot (T_{rx}(dci) + \\ & T_{WDC}(x_1, x_2, ..., x_n) + T_{wDC2US}) \end{aligned} \tag{5.10}$$

Note that the RRC Connection Request and Scheduling Request messages are scheduled with the UL grant contained in the RAR message. Thus, the estimation of the energy consumption of these packets is similar to the others except for the fixed allocation of resources that forces the following configuration: $N_{RU} = 4$, $N_{REP} = 1$, and $TBS = 88$ bits.

DL packet: Compared to its analogous analysis from Chapter 3, in this version we include the limitation of the resource sharing between channels and signals in the DL and the possible occurrence of DL gaps. The reception time needed for the packet y is:

$$
\begin{aligned}
T^{DL}(y) &= \left\lceil N_{REP} \cdot N_{SF} \cdot N_{seg}(y) \cdot \left(\frac{20}{14} - 1\right) \right\rceil + N_{REP} \cdot N_{SF} \cdot N_{seg}(y) \\
N_{seg}(y) &= \left\lceil \frac{L_y}{TBS(MCS, N_{SF}) - H_{RLCMAC}} \right\rceil
\end{aligned}
\tag{5.11}
$$

where L_y is the packet size, N_{SF} is the number of SFs allocated to the DL packet, and TBS is the Transport Block Size for the NPDSCH resulting from the selection of MCS and N_{SF}. Due to the DL SF duration is 1 ms, the total number of DL resources for the reception (i.e. $N_{REP} \cdot N_{SF} \cdot N_{seg}(y)$) equals the duration of the reception. If $R_{max} > N^{DL}_{GapThr}$ there will be DL gaps in the reception. Additionally, if $N_{seg}(y) > 0$, we need to include the reception of the DCIs between segments. Therefore, both effects are estimated as:

$$
\begin{aligned}
T^{DL}_{Gap}(y) &= \left\lfloor \frac{N_{REP} \cdot N_{SF} \cdot N_{seg}(y)}{T^{DL}_{GapPeriod} - T^{DL}_{GapDur}} \right\rfloor \cdot T^{DL}_{GapDur} \\
E_{seg}(y) &= (N_{seg}(y) - 1) \cdot \big(P_{RX} \cdot T^{DL}(dci) + \\
&\qquad P_i \cdot (T_{WDC}(y_1, y_2, ..., y_n) + T_{wDC2DS})\big)
\end{aligned}
\tag{5.12}
$$

where the steps of T_{WDC} (i.e. $y_1, y_2, ..., y_n$) are the duration of the last packet segment received, and T_{wDC2DS} is the wait from the reception of the DCI with the DL allocation to the NPDSCH reception start. Finally, the estimated energy consumption and delay due to the reception of the packet y is:

$$
\begin{aligned}
E_{rx}(y) &= P_{RX} \cdot T_{rx}(y) + P_i \cdot T^{DL}_{Gap}(y) + E_{seg}(y) \\
T_{rx}(y) &= T_{rx}(y) + T^{DL}_{Gap}(y) + (N_{seg}(y) - 1) \cdot \big(T^{DL}(dci) + \\
&\qquad T_{WDC}(y_1, y_2, ..., y_n)\big)
\end{aligned}
\tag{5.13}
$$

UL HARQ ACKs: This is a special case as ACKs are sent using NPUSCH format 2. While using this format, the RU is always composed of one subcarrier with a length of 4 slots. Therefore, the energy consumption, $E_{tx}(HARQ_{ack})$, and delay, $T_{tx}(HARQ_{ack})$, due to the transmission of a $HARQ_{ack}$ can be derived as:

$$
\begin{aligned}
T^{ULF2}(HARQ_{ack}) =& N_{REP} \cdot T_{RU} \\
T^{UL}_{Gap}(HARQ_{ack}) =& \left\lfloor \frac{T^{ULF2}(HARQ_{ack})}{T^{UL}_{GapPeriod} - T^{UL}_{GapDur}} \right\rfloor \cdot T^{UL}_{GapDur} \\
E_{tx}(HARQ_{ack}) =& P_{TX} \cdot T^{ULF2}(HARQ_{ack}) + P_{ULgap} \cdot T^{UL}_{Gap}(HARQ_{ack}) \\
T_{tx}(HARQ_{ack}) =& T^{ULF2}(HARQ_{ack}) + T^{UL}_{Gap}(HARQ_{ack})
\end{aligned}
\tag{5.14}
$$

5.3.3.2 Energy consumption per Markov chain state

Let E_j and D_j be the average energy consumption and delay of the j state, respectively. The following equations describe the energy consumption. Figure 5.3 can be used to correlate the packets exchanged during each state of the Markov chain. Note the delay can be estimated by removing the power components (P) of the equations. Then, E_j can be estimated as:

Off **state:** The UE does not transmit UL packet in current SF $E_{off} = P_s \cdot 1$.

RA **state:** The UE synchronizes and starts the RA procedure (i.e. exchanges from 1 to 4 in Figure 5.3):

$$
\begin{aligned}
E_{RA} =& P_i \cdot \left(T_{MIB-I} + T_{RAPeriod}/2 + T^{RA}_{Gap}\right) + \\
& P_{RX} \cdot (T_{sync} + T_{MIB-RX}) + P^{RA}_{TX} \cdot N^{RA}_{REP} \cdot T_{PRE}
\end{aligned}
\tag{5.15}
$$

where $T_{RAPeriod}/2$ denotes the average waiting time for NPRACH resource occurrence.

CR **state:** The UE performs a connection request (i.e. exchanges from 5 to 7 in Figure 5.3):

$$
E_{CR} = P_i \cdot (pp/2 + T_{wDC2DS} + T_{wDC2US}) + E_{rx}(dci) + E_{rx}(rar) + E_{tx}(req)
\tag{5.16}
$$

where $pp/2$ denotes the average waiting time for the NPDCCH occurrence as at the beginning there are no steps used as reference to estimate this waiting time.

$Connect$ **state:** After a successful connection, the UE sends its data packet. For the CP procedure setup, the data is transmitted piggybacked in the RRC

Connection Setup Complete message (i.e. exchanges from 8 to 21 in Figure 5.3):

$$\begin{aligned}
E_{Connect} =& P_i\,(2T_{wDC2DS} + T_{wDC2US} + 2T_{ACK-k0} + T_{WDC1} + T_{WDC2} \\
&+T_{WDC3}) + 3E_{rx}(dci) + E_{rx}(set) + 2E_{tx}(HARQ_{ack}) + \\
&E_{tx}(cmp) + E_{rx}(accept) + (1 - p_{ack})\,E_{schCmp} \\
T_{WDC1} =& pp - mod\,(T_{rx}(dci) + T_{wDC2DS} + T_{rx}(rar) + T_{wDC2US} + T_{tx}(req), pp) \\
T_{WDC2} =& pp - mod\,(T_{rx}(dci) + T_{wDC2DS} + T_{rx}(set) + T_{ACK-k0} + \\
&T_{tx}(HARQ_{ack}), pp) \\
T_{WDC3} =& pp - mod\,(T_{rx}(dci) + T_{wDC2US} + T_{tx}(cmp), pp)
\end{aligned} \tag{5.17}$$

where E_{schCmp} is the energy consumed to perform an Scheduling Request when requesting resources to send an RLC ACK. Then, E_{schCmp} can be estimated as:

$$\begin{aligned}
E_{schCmp} =& P_i \cdot (T_{PeriodBSR} + T_{RAPeriod}/2 + pp/2 + T_{wDC2DS} + 2T_{wDC2US} + \\
&T^{RA}_{GAP} + T_{WDC4}) + 2E_{rx}(dci) + E_{rx}(rar) + P^{RA}_{TX} \cdot N^{RA}_{REP} \cdot T_{PRE} + \\
&E_{tx}(scr) + E_{tx}(rlc_{Ack}) \\
T_{WDC4} =& pp - mod\,(T_{rx}(dci) + T_{wDC2DS} + T_{rx}(rar) + T_{wDC2US} + T_{tx}(scr), pp)
\end{aligned} \tag{5.18}$$

ACK **state:** The UE receives the IoT server's DL response:

$$\begin{aligned}
E_{ACK} =& P_i\,(T_{WDC5} + T_{wDC2DS} + T_{ACK-k0}) + E_{rx}(dci) + E_{rx}(DLack) + \\
&E_{tx}(HARQ_{ack}) + p_{ack} \cdot E_{schCmp} \\
T_{WDC5} =& pp - mod\,(T_{rx}(dci) + T_{wDC2DS} + T_{rx}(accept) + T_{ACK-k0} + \\
&T_{tx}(HARQ_{ack}), pp)
\end{aligned} \tag{5.19}$$

Inactive **state:** The UE stays in this state until the expiration of the Active Timer. This state includes C-DRX, the RRC Release (i.e. exchanges from 22 to 28 in Figure 5.3), and I-DRX:

$$
\begin{aligned}
E_{Inactive} =& E_{CDRX} + P_i \cdot (pp/2 + T_{wDC2DS} + T_{ACK-k0} + T_{wIDRX}) + E_{rx}(rel) + \\
& E_{tx}(HARQ_{ack}) + E_{schRel} + E_{IDRX} \\
E_{CDRX} =& P_i \left(N_{cycles}^{CDRX} \cdot (T_{LC} - T_{onD})\right) + P_{RX} \cdot \left(T_{DRXi} + N_{cycles}^{CDRX} \cdot T_{onD}\right) \\
E_{IDRX} =& N_{cycles}^{eDRX} \left(P_s \cdot (T_{eDRX} - T_{PTW}) + N_{cycles}^{IDRX} \left(P_s \left(T_{PC} - (N_{REP_{dci}} + \right.\right.\right. \\
& \left.\left.\left. T_{IDRX_{sync}})\right) + P_{RX} \cdot N_{REP_{dci}} + P_{IDRX_{sync}} \cdot T_{IDRX_{sync}}\right)\right) + \\
& P_s \cdot \left(T_{active} - N_{cycles}^{eDRX} \cdot T_{eDRX}\right)
\end{aligned}
\tag{5.20}
$$

where E_{schRel} is the energy consumed when requesting resources using the Scheduling Request procedure after the RRC Release. Unlike E_{schCmp}, E_{schRel} does not include the complete Scheduling Request procedure, only up to the request of resources (i.e. exchanges from 25 to 28 in Figure 5.3). This definition of E_{schRel} is included in the model to emulate the behavior seen in the experimental measurements with the evaluated DUTs. Thus, E_{schRel} is estimated as:

$$
\begin{aligned}
E_{schRel} =& P_i \cdot (T_{PeriodBSR} + T_{RAPeriod}/2 + pp/2 + T_{wDC2DS} + T_{wDC2US} + \\
& T_{GAP}^{RA}) + E_{rx}(dci) + E_{rx}(rar) + P_{TX}^{RA} \cdot N_{REP}^{RA} \cdot T_{PRE} + E_{rx}(scr)
\end{aligned}
\tag{5.21}
$$

Additionally, T_{onD} specifies the number of consecutive NPDCCH SFs at the beginning of a C-DRX cycle to monitor. This timer is given in units of pp. However, as the duration of the C-DRX cycle could be smaller than the duration of the T_{onD} due to a large value of pp, the duration of T_{onD} is estimated as:

$$
T_{onD} = min\left(\left[T_{LC}/pp\right], Npp_{onD}\right) \cdot R_{max}^{USS} \tag{5.22}
$$

where Npp_{onD} is the number of pp defined at the onDuration Timer, and R_{max}^{USS} is the maximum number of repetitions for NPDCCH for USS.

5.3.3.3 Battery lifetime estimation

For the validation of the analytical model, we have two estimations of the battery lifetime: i) from the model and ii) from the measurements. On the one hand, the estimation based on the analytical model starts computing the energy consumed

per day E_{day}^{model} in joules (J) as:

$$E_{day}^{model} = \left(\left(\sum_j b_j E_j \right) \cdot \frac{D_{day}}{\sum_j b_j D_j} \right) \cdot 1e-6 \tag{5.23}$$

where D_{day} denotes the duration of one day in ms, b_j is the steady state probability of the j Markov chain state, and E_j and D_j are the energy consumption (μJ) and delay (ms) of the j Markov chain state. Finally, the battery lifetime of the model in years Y_{model} can be estimated as:

$$\begin{aligned} E_{day_{Wh}}^{model} &= \frac{E_{day}^{model}}{3600} \\ Y_{model} &= \frac{C_{bat}}{E_{day_{Wh}}^{model} \cdot 365.25} \end{aligned} \tag{5.24}$$

where $E_{day_{Wh}}^{model}$ is the energy consumption per day in the model in watt-hour units, and C_{bat} is the battery capacity.

On the other hand, the battery lifetime estimation based on measurements, the DUT power consumption is measured from the beginning of the CP to the start of PSM. Figure 5.11 illustrates the mentioned measured period. To compare the results from the measurements with the model, we define three phases during the measurements:

- *COM*: UE wakes up, sends its data using CP, monitors NPDCCH while applying C-DRX, and releases the RRC connection after the reception of the RRC Release packet.

- *IDLE*: UE stays in I-DRX until T_{active} expiration.

- *SLEEP*: UE sleeps using PSM until the next transmission.

From the measured energy consumption when sending one UL report, we can estimate the energy consumed per day considering a specific IAT. For simplicity, when comparing with the model, we assume the duration of the *SLEEP* $T_{SLEEP}^{test} = IAT$. Then, the average energy consumed per day E_{day}^{test} and the battery lifetime in years Y_{test} can be estimated as follows:

$$
\begin{aligned}
N_{reports_{day}} &= \frac{D_{day}}{T^{test}_{COM} + T^{test}_{IDLE} + T^{test}_{SLEEP}} \\
E^{test}_{day_{Wh}} &= \frac{N_{reports_{day}} \cdot \left(E^{test}_{COM} + E^{test}_{IDLE} + E^{test}_{SLEEP}\right)}{3600} \\
Y^{test} &= \frac{C_{bat}}{E^{test}_{day_{Wh}} \cdot 365.25}
\end{aligned}
\tag{5.25}
$$

where $N_{reports_{day}}$ denotes the number of UL reports sent in one day, T^{test}_i and E^{test}_i are the duration and energy consumed in the ith phase, respectively. As mentioned before, the *SLEEP* phase is not measured in the experimental setup. Then, the energy consumed in this phase is estimated as $E^{test}_{SLEEP} = P_s \cdot T^{test}_{SLEEP}$.

5.3.4 Experimental setup

To validate the NB-IoT model, we use the testbed shown in Figure 5.1 under the label "STEP2". This testbed measures the UE's energy consumption while sending a UL report using CP procedure. We consider the periodic UL reports are UDP packets with 50B of payload and the battery capacity of the UE is $C_{bat} = 5Wh$ [83].

The validation of the model is done based on three test cases. These test cases address different main parts of the proposed model:

- `G`: This test focuses on the evaluation of the G parameter. The value of G together with R_{max} defines pp (see subsection 3.2.2). Then, the general scheduling process.

- `REP`: This test is a simplification of the use of repetitions as all the parameters related to repetitions are set equal. The goal is to examine the energy consumption impact due to an increase of repetitions to extend coverage.

- `SCS`: Considering the two subcarrier spacing (SCS) allowed in NB-IoT (15 and 3.75 kHz). This test compares the performance of both single subcarrier configurations. If the UE has not reached its maximum transmission power, following the power control mechanism, ideally, the decrease of SCS will increase the duration of the RU and decrease the transmission power equally.

Each point of the experimental evaluation showed is obtained through one empirical measurement in the testbed (that is, only one realization). We were not able to include more realizations in the validation due to testbed availability constraints. Table 5.8 shows the baseline configuration of the radio interface between the DUT and the UXM. Table 5.9 summarizes the specific UXM settings for each test case considered in this validation. The configuration of the parameters chosen forces the maximum transmission power p_{max}, except for the `SCS` test case. For `SCS`, the power control is configured to use p_{max} when $SCS = 15$ kHz, and reduce the power as obtained from the power control mechanism when reducing the SCS to 3.75 kHz [21]. Table 5.10 lists the measured average power consumption levels for the two DUTs evaluated.

Table 5.8: Baseline configuration of the main parameters.

	Parameter	**Value**
Power control	p_{max}	23 dBm
	RSRP	-110 dBm/15kHz
	Reference Signal indicator	27 dBm
	α	1
	$P0_{nominal}$	-67 dBm
	$P0_{UEspecific}$	0 dBm
	Preamble initial power	-90 dBm
NPRACH	Periodicity	640 ms
	N_{REP}^{RA}	2
	RAR Window Size	$5pp$
NPDCCH	R_{max} for USS and CSS	8
	G for USS and CSS	2
	$N_{REP_{dci}}$ for CSS	8
	$N_{REP_{dci}}$ for USS	1
NPDSCH	MCS	3
	N_{SF}	10
	N_{REP}	1
NPUSCH	MCS	3
	N_{RU}	10
	N_{REP}	1
	Number of subcarriers	1
	Subcarrier spacing (SCS)	15 kHz

Table 5.9: Test cases with UXM settings.

Test case	Sweeping parameter	Other settings
`G`	$G = \{1.5, 2, 4, 8, 16, 32, 48, 64\}$	
`REP`	$R_{max} = \{1, 2, 4, 8, 16, 32, 64\}$ $N_{REP_{dci}} = N_{REP} = R_{max}$	$G = \{8, 4, 2, 2, 2, 2, 2\}$ CSS and USS are set equal
`SCS`	$SCS = \{3.75, 15\}$ kHz	$P0_{nominal} = -68$ dBm $\alpha = 0.7$ $N_{REP} = 2$

Table 5.10: Measured average power consumption.

	Device A	**Device B**
P_{TX}† at 23 dBm	765 mW	731 mW
P_{TX}‡ at 17 dBm	503 mW	311 mW
P_{RX}	242 mW	215 mW
P_{ULgap}	160.4 mW	128.4 mW
P_i	29.1 mW	17.8 mW
P_s	11.13 μ W	14.14 μ W

† This TX power is used in `G`, `REP` test cases, and when $SCS = 15$ kHz in `SCS`.

‡ This TX power is used when $SCS = 3.75$ kHz in `SCS`

5.3.5 Results

This subsection contains the validation results of our proposed analytical NB-IoT model using two different DUTs. The validation is done in terms of battery lifetime and latency to perform the CP. For both results, we compare the values obtained with:

- The estimation from the measurements obtained with the testbed configured as detailed in subsection 5.3.4, labelled in the figures as "Measurements".
- The analytical model presented in subsection 5.3.1, labelled in the following figures as "Model". For the three test cases studied, the analytical model is configured with the same parameter values indicated in subsection 5.3.4.

Figures 5.13 and 5.14 show the battery lifetime obtained when using the analytical model and the measurements for the test cases `G` and `REP`, respectively.

Figure 5.13 shows the increase of the parameter G does not have a noticeable impact on the battery lifetime. Although the increase of G delays the scheduling of resources, the UE stays inactive while waiting and therefore the energy consumption increase is small. As expected, for larger IATs, the lifetime increases. Figure 5.14 shows the significant impact the number of repetitions has on the battery lifetime. Considering the target battery lifetime of 10 years in NB-IoT, the correct use of repetitions to extend coverage, together with the knowledge of the traffic profile of the UE are essential to achieving it.

Additionally, for the `SCS` test case and $SCS = 15$ kHz, both DUTs achieve a similar battery lifetime of average 18 years for an IAT of 24 h. However, when reducing the SCS to 3.75 kHz, the battery lifetime decreases an average of 20% as the power consumption decrease and T_{RU} increase are unequal, and the increase of the duration of NPUSCH format 2 transmissions. Despite the 3.75 kHz SCS obtains worse results, it is an interesting configuration for deep indoor IoT scenarios where a large number of UEs are concentrated in a small area and most of them experience a significant penetration loss. The reason is this configuration provides more robust communication with the eNB and enables more simultaneous connections.

To ease the comparison of the analytical model and the measurements, Figure 5.15 illustrates the relative error resulting from both estimations of the battery

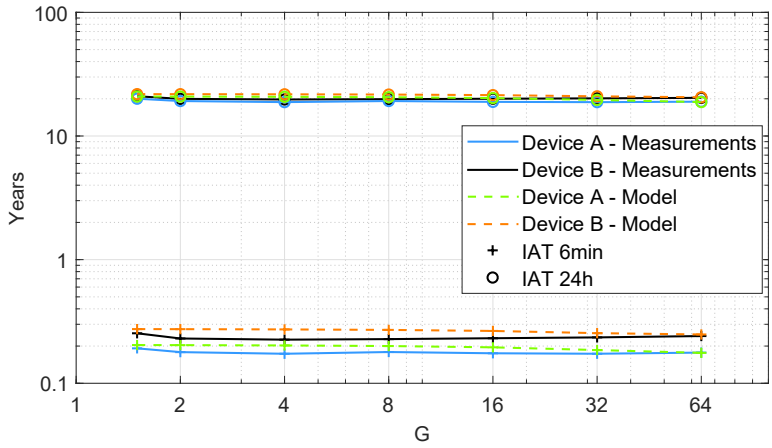


Figure 5.13: Battery lifetime estimation as a function of G and two different IATs for `G` test case.

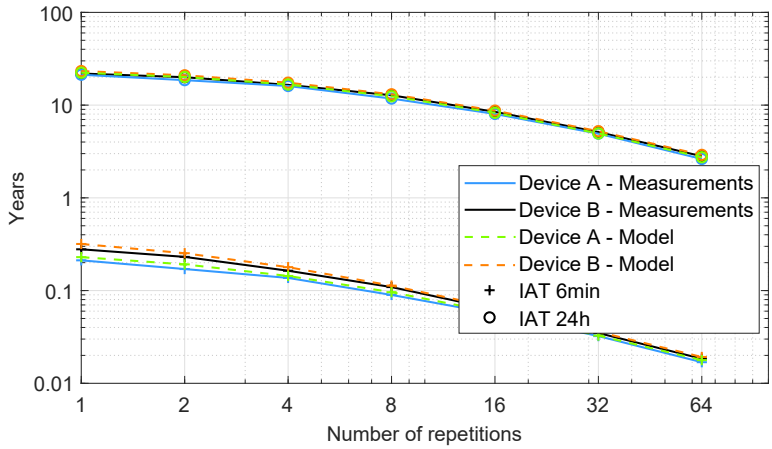


Figure 5.14: Battery lifetime estimation as a function of the number of repetitions and two different IATs for `REP` test case.

lifetime in years for `G` and `REP` test cases, considering the two DUTs and an IAT of 6 min. The maximum relative error obtained between the model and the measurements is 21%. We use the IAT of 6 min as a pessimistic scenario for periodic reporting. For example, the smallest IAT considered in [83] for periodic mobile autonomous reporting is 30 min. The error decreases as the value of the parameter G or the number of repetitions increases. This is because the energy consumed while performing CP increases. Therefore, the model estimation improves as the procedure becomes more important than other assumed simplifications such as the synchronizations. Particularly, the main factors in the relative error are:

- The simplification of the synchronizations (i.e. the initial synchronization and the short synchronizations before paging) modeling: Both power-hungry and robust processes have been modeled with an average duration and power consumption. However, both synchronizations entail several steps and their performance depend on channel quality and the NB-IoT deployment [88, 115].

- The assumed statistical average prior to the transmission of a preamble: In the system model considered, the preamble transmission happens in three different signaling exchanges. We always assume the wait for NPRACH resources to send the preamble is half the NPRACH periodicity (i.e. its

statistical average). However, this wait can range from a few ms to the NPRACH period value in the measurements as we only consider one empirical realization per measurement.

For larger IATs than 6 min, the relative error is smaller. For example, the maximum relative error is 11% when the IAT is 24 h. In this case, the reduction is due to the larger PSM duration that is easily modeled with its average power consumption. Additionally, for `SCS` test case, the resulting average relative error assuming an IAT of 6 min is 6% and 12% for devices A and B, respectively.

Finally, Figure 5.16 shows the latency to finish the CP procedure for the `REP` test case. This figure compares the measured latency of both DUTs and the value obtained with the analytical model. As expected, as the number of repetitions increases, the latency is higher. This increment is less notable in other test cases. For example, the maximum latency reached is 10.35 s for $G = 64$ in `G` test case, and 4.81s for $SCS = 3.75$ kHz in `SCS` test case. Note that the difference between the model and the measurements is greater when estimating the CP's latency than the battery lifetime. This is due to the model does not consider retransmissions the UE could experience, the simplification of the synchronization, and some waits of the UEs seen in the measurements but not included in the model.

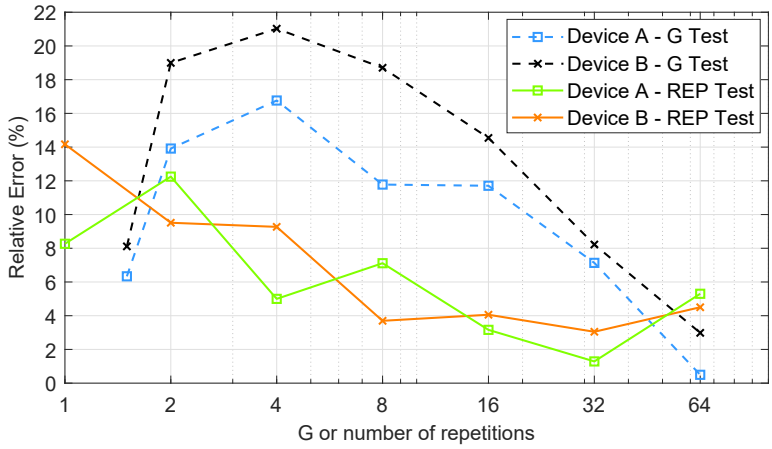


Figure 5.15: Relative error of the battery lifetime estimation in years between the analytical model and the measurements assuming an IAT of 6 min.

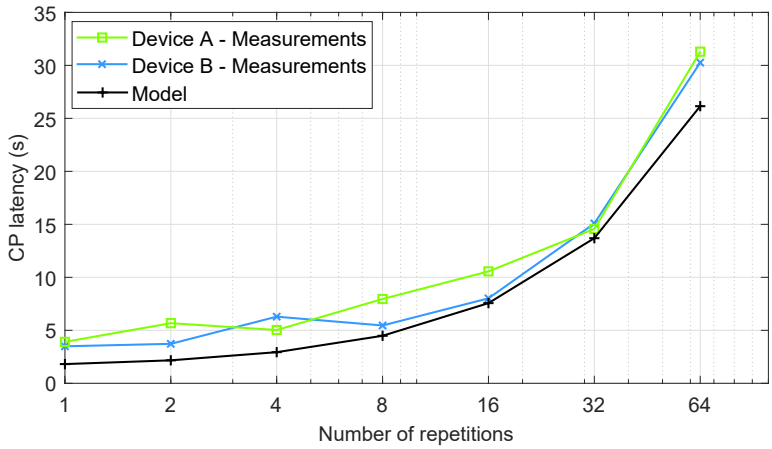


Figure 5.16: Comparison of the latency to finish CP procedure measured in both DUTs and obtained with the analytical model for `REP` test case.

5.4 Conclusions

NB-IoT provides a large set of parameter configurations that defines the communication between the UE and eNB to ensure coverage and battery lifetime with reasonable latency and network resource efficiency. Firstly, all UEs are grouped in ECLs that sets up its RA and CSS configuration. Later, each UE has a user-specific configuration depending on its coverage. These possible configurations of the radio interface result in a significant variability of the UE performance. The aim of the analysis of this chapter is to understand the trade-offs between the NB-IoT targets. To do that, this chapter is divided into two parts.

In the first part, we focus on observe empirical NB-IoT live configurations and use this knowledge to study the performance of a UE under different scenarios in our controlled testbed. The evaluation considers three DUTs and four test cases. The test cases focus on analyze the impact of different factors, namely: i) `SIZE`: the data packet size; ii) `IMCS`: resource allocation for the NPUSCH; iii) `ULREP`: NPUSCH repetitions; and iv) `SCS`: bandwidth allocation. In terms of energy consumption, there are different causes that will take a significant portion of the energy depending on the current ECL or traffic profile. For example, for UEs in good coverage, the configuration of the release of the RRC connection is important. Thus, in this case, solutions such as RAI, or C-DRX can greatly save

energy. On the contrary, for UEs in bad coverage, the energy consumption due to the transfer of packets is the dominant issue. In this case, signaling reduction and optimizations like EDT are essential.

As evident from the empirical results, the 3 ECLs configurations have a great impact on the UE performance. The targets of battery lifetime and latency can be satisfied when the UE is not in extreme coverage conditions or there is no need for extensive use of repetitions (> 16). For example, a UE periodically sending UDP packets of 50 bytes, the 10 years target is satisfied if the packets have an IAT larger than 4 h, 11 h, and 68 h for the ECLs 0, 1, and 2, respectively.

In the second part, we propose and validate an analytical model for NB-IoT. This model is based on the one presented in Chapter 3, improving the estimation as this new version includes more details. The goal is to provide a tractable energy consumption and delay model for NB-IoT. This model not only relates the performance metrics of battery lifetime and latency with the input parameters, but it also provides an analytical methodology that can be easily extended to include more control procedures.

The proposed analytical model estimates the average energy consumption and delay of a UE sending periodic UL reports using CP procedure. The validation is done based on the same testbed as the first part of this chapter, but adapting the test cases to different main parts of the model. These test cases address main parts of the analytical model: i) `G`: the scheduling process of NB-IoT; ii) `REP`: the lengthening of transmissions/receptions to extend coverage; iii) `SCS`: performance of the single subcarrier configurations. In this validation two different DUTs are used. The results show the analytical model performs well, obtaining a maximum relative error of the battery lifetime estimation in years between the model and the measurements of 21% assuming an IAT of 6 min. The main factors in the relative error are the simplification of the synchronizations and the assumed statistical average prior to the transmission of a preamble. As espected, for larger IATs than 6 min, the relative error is smaller (e.g. 11% when the IAT is 24 h) as the PSM duration is larger.

5.4.1 Resulting research contributions

The research contributions resulting from the work done in this chapter are listed below:

- P. Andres-Maldonado, M. Lauridsen, P. Ameigeiras and J. M. Lopez-Soler, "Analytical Modeling and Experimental Validation of NB-IoT Device Energy Consumption," Accepted for publication in the IEEE Internet of Things Journal, 2019.

  DOI: 10.1109/JIOT.2019.2904802

- P. Andres-Maldonado, M. Lauridsen, P. Ameigeiras and J. M. Lopez-Soler, "Experimental Analysis of NB-IoT Performance Trade-offs," To be submitted, 2019.

Chapter 6

Conclusions and Outlook

The proliferation of the Internet of Things (IoT) ecosystem is bringing tremendous challenges in meeting the requirements for connectivity, mainly with regards to long battery life, low device costs, low deployment costs, extended coverage and support for a massive number of devices. Within this context, Low-Power Wide-Area (LPWA) technologies will likely play an important role in the IoT. The concept of Cellular IoT (CIoT), that is, IoT via cellular network technologies, will vastly require to review the cellular connectivity roadmap against the IoT requirements.

As an enabler of LPWA CIoT connections, the Third Generation Partnership Project (3GPP) introduced Narrowband Internet of Things (NB-IoT). NB-IoT is applicable to a large range of IoT application scenarios and will be part of 5G to support 5G LPWA use cases in the foreseeable future. However, NB-IoT is still in its infancy, needing analytical modeling investigation, system performance optimization, and experimental characterization.

In this thesis, we have studied the feasibility to manage IoT in the current (4G) and future (5G) generations of cellular networks. More specifically, we have first addressed the modeling of the signaling load due to Machine-Type Communication (MTC) in the future softwarized cellular networks. Later, we have delved into the performance of a IoT sensor using NB-IoT. We have proposed an analytical model for NB-IoT devices and an evaluation framework to study the performance of NB-IoT. Moreover, we have empirically validated the performance of the NB-IoT design targets.

The rest of the chapter is organized as follows. Section 6.1 draws the main conclusions extracted from the work carried out in this thesis. Section 6.2 lists the main contributions of this thesis. Finally, Section 6.3 discusses directions for future work related to the topics covered in this thesis.

6.1 Main conclusions

The most relevant conclusions extracted from the work developed in this thesis are the following:

Ch1 The development of current and future cellular networks have to consider three generic services with vastly heterogeneous requirements: enhanced Mobile Broadband (eMBB), massive MTC (mMTC) and Ultra-Reliable and Low Latency Communications (URLLC). The introduction of MTC in cellular networks entails several challenges due to the new MTC traffic characteristics compared to traditional human communications.

The addition of new enhancements in Long Term Evolution (LTE) together with the standardization of new LPWA technologies (i.e. Extended Coverage GSM IoT (EC-GSM-IoT), LTE Cat-M1 (LTE-M), and NB-IoT) are meant to help cellular networks to become suitable for the IoT applications. Furthermore, it is envisaged that LTE-M and NB-IoT will continue to coexist alongside other 5G components to fulfill LPWA requirements.

Ch2 Network Functions Virtualization (NFV) paradigm is seen as an enabler to provide a new architecture for cellular networks to cope with the expected increase in traffic. Besides the eMBB traffic is considered as a direct extension of the 4G broadband service, the addition of mMTC and URLLC services will need to cope with a vast service heterogeneity. NFV provides the scalability and flexibility necessary for these new services foreseen.

Within the LTE network, the Mobility Management Entity (MME) must handle several signaling packets per subscriber attached to it. The key role the MME has together with the inclusion of new types of traffic to handle may arise signaling storms and network congestion. This chapter focuses on the MME as one of the most impacted LTE entities due to growing IoT development within cellular networks.

To overcome the above-mentioned traffic challenges, we have proposed a three-tiered virtualized Mobility Management Entity (vMME) following the NFV paradigm. The three tiers are: Front End (FE), Service Logic (SL), and State Database (SDB). We have evaluated the dimensioning of four vMME designs and the performance of the virtualized MME (vMME) in terms of required resources, costs based on the model of Amazon EC2, and response time for each traffic class considered.

The results show a careful design of the SL tier target traffic greatly impacts the resources required to allocate and the performance of each traffic class. If the mMTC traffic is managed together with the other traffic classes, the required resources to satisfy the target delay of URLLC increases significantly. To avoid this, it is necessary to isolate the resources for traffic classes with conflicting requirements. Furthermore, considering the scalability constraints of the SDB due to its shared-everything architecture, this tier may become a bottleneck in all the proposed schemes.

Ch3 The mMTC use case is about massive access by a large number of devices. A typical mMTC device is a low complexity and battery constrained (low-energy) device active intermittently that sends occasional and small data transmissions. Given these characteristics, many of mMTC devices will rely on Low-Power Wide-Area Network (LPWAN) for connectivity.

The introduction of NB-IoT in the 3GPP standards is intended as a solution for the LPWA segment in the cellular networks. NB-IoT is designed for MTC and has four key NB-IoT design targets: maximum latency of 10 seconds in the Uplink (UL), target coverage of 164 dB Maximum Coupling Loss (MCL), battery lifetime beyond 10 years, and support of massive connections.

To analyze the benefits and limitations of NB-IoT for mMTC, we have proposed an analytical model to study the energy consumption of an NB-IoT User Equipment (UE). The analytical model is based on the steady states probabilities of a Markov chain. We have also compared the capacity gain from the conventional control procedure for data transmission named Service Request (SR), and two new optimized data transmission procedures called User Plane optimization (UP) and Control Plane optimization (CP).

From the battery lifetime viewpoint, the results show when the UE has poor radio conditions and changes to a worse Coverage Enhancement Level (ECL), there is a significant reduction of battery lifetime. Additionally, the Inter-Arrival Time (IAT) has a considerable impact on the battery lifetime. As the IAT increases, the UE can spend longer periods in Power Saving Mode (PSM) and the battery lifetime increases. When comparing the data transmission procedures (i.e. SR, UP, and CP), CP obtains the best results due to the signaling reduction and the use of Release Assistance Indication (RAI) notification to release the radio connection as soon as possible. For example, for ECL 0, this improves the battery lifetime from 78% when the IAT is 30 min to 4% when the IAT is 24 h, compared to UP. There are several factors that make more challenging satisfying the battery lifetime target. In our results, the target of 10 years is satisfied for almost the whole IAT range when the UE is in ECL 0. However, for ECLs 1 and 2, this target is fulfilled when the UE has an IAT larger than 6h and 13 h, respectively.

The capacity results show the UP and CP procedures achieve a great reduction in signaling compared to SR. For example, when the UE only sends a UL report, CP and UP achieve gains of 162% and 120% in ECL 0, respectively. Despite the benefits of CP, this procedure is not convenient for long data transmissions, as the network is expected to force the UE to establish the data bearers if a maximum number of messages is exceeded. Therefore, for long data transmissions UP would be more advisable. The results also show the radio channel limiting the capacity varies depending on the considered traffic case and coverage level. For the best ECL the capacity is mostly limited by the uplink channels resources. On the contrary, for the rest of ECLs, the capacity limitation comes from the downlink channels. Regarding the uplink and the downlink evaluation, downlink gains reach something less than two times the gains of the uplink.

Ch4 An important part when proposing NB-IoT was to meet an objective of extending coverage with 20 dB compared to General Packet Radio Service (GPRS). This is because for mMTC devices, energy efficiency, and good coverage are the most important Key Performance Indicators (KPIs).

However, both KPIs are coupled in conflicting ways, when using a more robust configuration to extend coverage, more radio resources are used and the energy consumption increases. Furthermore, the NB-IoT extended coverage results in a very low operating Signal to Noise Ratio (SNR). In the low SNR range, the performance of the channel estimation mechanisms may not be sufficient to ensure adequate channel estimation accuracy. Thus, the performance of NB-IoT physical channels can be limited.

We have studied the performance of NB-IoT in extended coverage when non-ideal factors are considered. We compare the NB-IoT performance under three Channel Estimation (CE) scenarios: ideal CE, realistic CE with cross-subframe, and realistic CE without cross-subframe. We propose a complete analytical evaluation framework for NB-IoT that jointly obtains the coverage extension and the resulting UE battery lifetime and latency. The framework depends on three elements: SNR estimation based on the Shannon theorem, a proposed UL link adaptation, and the previously proposed energy consumption model.

The results show NB-IoT presents a great gap between its performance and the Shannon bound. This is due to implementation issues and the repetition coding scheme used in NB-IoT. Focusing on the UL link adaptation, for a UE in good coverage, the modification of Modulation and Coding Scheme (MCS) and Resource Units (RUs) or bandwidth reduction techniques keep the bandwidth utilization and can cover a limited range of coverage extension. In the case the UE has poor coverage, bandwidth reduction and repetitions become essential to reach greater coverage extension. However, in this situation the gain due to repetitions can be limited.

The SNR gain when doubling repetitions is limited in realistic CE scenarios compared to the ideal 3dB gain. For example, for a number of repetitions greater than 16, the gain when doubling the repetitions applied in realistic CE scenarios is less than 1.5 dB. This poor performance in realistic CE is worse as the MCL increases due to larger CE error present in the estimation. To reduce the CE errors, the use of cross-subframe CE improves the performance of the channel estimator as it averages several Subframes (SFs). From the simulations, for the largest cross-subframe window of 16 ms, the

obtained CE error decreases 6 dB compared to single frame estimation.

As a result of the poor CE at low SNR, the obtained UE performance results when considering realistic CE in terms of latency and battery lifetime significantly worsen compared to ideal CE. Regarding the UE battery lifetime, for an $MCL = 154$ dB, realistic CE with cross-subframe shows a battery lifetime reduction of a 50% and 18% compared to ideal CE for IATs of 2 h and 24 h, respectively. For higher MCLs, such as 164 dB, that rely on repetitions to extend coverage, this degradation of the UE performance worsen reaching a 90% of battery lifetime reduction in realistic CE with cross-subframe compared to ideal CE.

Ch5 NB-IoT has a potentially wide range of IoT applications such as agriculture, health care, security, fleet management, smart cities, smart utilities, etc. Each one may have its own KPIs priorities. However, the knowledge of the NB-IoT performance boundaries and trade-offs is a general concern before be able to further develop and deploy NB-IoT.

We have tackled the study of NB-IoT performance analytically and experimentally. We have observed the configuration of a live NB-IoT network. With this knowledge, we have examined the performance of an NB-IoT UE under different scenarios in our controlled testbed. The testbed is composed by an evolved NodeB (eNB) emulator, a Device Under Test (DUT), and a power analyzer. The experimental evaluation considers three DUTs and four test cases. Each test case focuses on analyzing the impact of different factors in the DUT performance in terms of energy consumption and latency. Additionally, we have validated our analytical NB-IoT model. The validation uses the same tesbed under other three defined test cases to compare the results obtained with the measurements and the model. The proposed tests cases in the validation focus on examining if the model correctly describes the steps the UE will perform to send a UL data packet.

The results show there are different causes that will take a significant portion of the energy consumption depending on the current ECL or the UE traffic profile. As evident from the empirical results and saw in Chapter 3, the 3 ECLs configurations have a great impact on the UE performance. Focusing on the NB-IoT targets, the battery lifetime target of 10 years is

satisfied for the majority of the configurations evaluated, when the UE does not belong to the worst ECL (i.e. ECL 2) or there is no need for extensive use of repetitions (> 16). In terms of the maximum latency of 10 s, the same configurations that do not achieve the previous target also exceed the 10 s. For example, a UE periodically sending UDP packets of 50 bytes, the 10 years target is satisfied if the packets have an IAT larger than 4 h, 11 h, and 68 h for the ECLs 0, 1, and 2, respectively.

Additionally, the validation results conclude the analytical model performs well, obtaining a reasonable maximum relative error of the battery lifetime estimation compared to the empirical measurements. For example, the maximum relative error is 21% and 11% for an IAT of 6 min and 24h, respectively. The relative error decreases as the energy consumed while performing CP increases (e.g. due to repetitions) as the procedure becomes more important than other assumed simplifications such as the synchronizations.

6.2 Research contributions

The research contributions resulting from this thesis are listed below:

- A proposal of a reduced signaling procedure for small data transmissions for LTE, named as Random Access-based Small IP packet Transmission (RASIPT) [151]. Note this publication is related to this thesis but its contribution is not included within the contents of the thesis.
- Performance evaluation of a three-tiered vMME [72, 152]. The evaluation is done in terms of scalability, cost, and response time.
- An analytical model to estimate the average energy consumption per data packet of a LTE UE [87]. In this analysis, the conventional SR procedure is compared with the optimizations UP and CP.
- An analytical model to estimate the energy consumption of an NB-IoT UE and NB-IoT radio resources consumption [153–156]. The model provides a tractable tool to estimate the number of UEs supported in a NB-IoT carrier, and the battery lifetime and latency of an NB-IoT UE under different traffic

and coverage assumptions. Additionally, a simplified version of this model without the consideration of connection failures have been validated by means of empirical measurements using DUTs.

- Analytical expressions have been derived to describe the NB-IoT transmission in terms of SNR, bandwidth utilization, and energy per transmitted bit [154, 155]. These expressions consider ideal or realistic CE.
- A proposal of a UL link adaptation algorithm [154]. This algorithm considers the new features of NB-IoT (i.e. repetitions and bandwidth allocation).
- An integral analytical NB-IoT evaluation framework [155]. This framework summarizes the whole analytical process to evaluate NB-IoT and is composed of three previously proposed elements: the analytical expressions for the NB-IoT transmission, the UL link adaptation algorithm, and the NB-IoT energy consumption model. The framework provides a way to evaluate the performance of NB-IoT in terms of coverage, battery lifetime, latency, and capacity under several scenarios.
- The experimental characterization of a live NB-IoT network and trade-offs between the NB-IoT targets [157].

6.3 Future work

The heterogeneous emerging IoT use cases together with the increasing traffic to handle pose a great challenge in cellular networks. MTC, or Machine to Machine (M2M) communications, are expected to have an important role in current and future cellular networks. Consequently, cellular networks must continuously evolve and cope with new requirements. NB-IoT is a good example of the evolution of cellular networks as it reuses LTE to provide support for IoT. However, as a new access technology and the several new features it has, NB-IoT still has unresolved research issues.

Based on the work carried out in this thesis, several open issues and improvements lie ahead.

- Extension of the queuing model proposed in Chapter 2 to consider more entities that handle control plane messages and more control procedures un-

der realistic scenarios. The idea is to predict the possible bottlenecks of the network with the new data transmission over Non-Access Stratum (NAS) signaling (i.e. CP optimization) and evaluate the mechanisms standardized to control the use of CP.

- Extension of the NB-IoT model proposed in Chapter 3 (and later more detailed for CP in Chapter 5) to include:
  - Inclusion of new features such as access barring, new traffic profiles, and the change of ECL as the UE retries several times the Random Access (RA) procedure.
  - A mechanism to automate the methodology of analysis. From the analysis done, the model can be extended to more complex signaling scenarios to estimate the UE performance. For these complex scenarios, a more agile mechanism would be require to automate the generation of the estimation from the description of the signaling and the radio interface parameters.
- Development of a model to analyze the performance of the new extended/enhanced Discontinuous Reception (eDRX) and PSM. NB-IoT provides a large set of configurations where most of the parameters depend on their values and a few of them on the periodicity of the Narrowband Physical Downlink Control CHannel (NPDCCH). The combination of DRX mechanisms (eDRX, C-DRX, and I-DRX) and PSM enables the UE to sleep for long periods and have relaxed monitoring of radio channels. However, this is at the expense of reducing the periods the device is reachable by the network, thus, increasing the Downlink (DL) latency. Therefore, the idea is to jointly analyze the power saving factor and the wake-up latency when applying these techniques.
- Extension of the analysis of the extended coverage provided in Chapter 4 considering more advanced channel estimators, interference, and the evaluation of the system perspective (i.e. study the overall performance assuming a number of UEs distributed within one cell). UEs in very weak radio conditions could be restricted on the use of coverage enhancements. If the cell is not suitable for the UE to camp on, the UE is required to try to find

another suitable cell. A deeper analysis of the extreme radio conditions an NB-IoT cell may provide a useful way to predict if the desired NB-IoT deployment is feasible.

- Experimental measurement of the battery lifetime considering real batteries and derivation of analytical expressions including the non-ideal characteristics of the batteries, such as self-discharge and different temperatures. As the battery lifetime in NB-IoT is expected to last several years, the non-ideal characteristics of the batteries in practical deployment could have a significant impact on the duration of the UE once deployed.

- Study and research of alternative antenna schemes that can improve the SNR without increasing the complexity of the UE. This research may help to improve the UEs performance under challenging radio conditions (i.e. high MCL) instead of the use of an extensive number of repetitions.

Appendices

Appendix A

Resumen

El presente apéndice incluye un amplio resumen en castellano de la memoria de tesis con el objetivo de cumplir con la normativa de la Escuela de Posgrado de la Universidad de Granada referente a la redacción de tesis doctorales cuando éstas son escritas en inglés.

A.1 Introducción y motivación

Las redes celulares han transformado nuestra sociedad en muchas facetas como por ejemplo en la forma que interactuamos, nuestro trabajo, la distribución de contenido multimedia, acceso a la información, etc. Han evolucionado desde la tecnología inalámbrica costosa de hace unas décadas, a una comodidad diaria en la actualidad. Su continua evolución ha encaminado al sistema ubicuo usado en la actualidad por una gran mayoría de la población mundial. Conforme la adopción de las redes celulares se expande, nuevos casos de uso aparecen.

Hasta ahora, las cuatro generaciones de redes celulares desplegadas (1G - 4G) se han centrado en proporcionar servicios a las comunicaciones entre personas. Sin embargo, en los últimos años el desarrollo y crecimiento del Internet de las Cosas (IoT) se ha convertido en una revolución que está redirigiendo la hoja de ruta de las redes celulares. Esto se puede ver en las últimas agregaciones de funcionalidad del Proyecto de Asociación para la Tercera Generación (3GPP) para mejorar los estándares de las redes celulares actuales, y en el diseño y objetivos de la próxima generación celular (5G).

El concepto de IoT encarna la visión de todo conectado, es decir, dispositivos, máquinas y vehículos conectados a Internet. Esta visión engloba un vasto exosistema de casos de uso emergentes en mercados como maquinaria industrial, salud, ciudades inteligentes, etc. Cisco estima que 15 mil millones de dispositivos estarán conectados en 2022. Para poder hacer frente a los nuevos requisitos de conectividad de las comunicaciones entre máquinas para soportar IoT, las redes celulares deben ser repensadas.

En las siguientes subsecciones se ahondará en las comunicaciones tipo máquina, esenciales para el desarrollo de IoT, en las posibilidades de conectividad disponibles en IoT y sus desafíos, y se presentarán las nuevas soluciones estandarizadas en las redes celulares para soportar eficientemente IoT.

A.1.1 Comunicaciones tipo máquina e IoT

Las Comunicaciones Tipo Máquina (MTC), también conocidas como comunicaciones Máquina a Máquina (M2M), es un término que describe cualquier comunicación de datos entre dispositivos para recopilar datos, compartir información, y realizar acciones que no necesitan necesariamente de interacción humana. Tradicionalmente, el uso de estos dispositivos ha sido local, es decir, sin la exposición del servicio para ser usado por más casos de uso o aplicaciones. Esto ha generado un mercado de M2M fragmentado debido a las diferentes industrias verticales (por ejemplo, transporte, salud, industria, utilidades, etc.) y al desarrollo de soluciones M2M cerradas específicas para un dominio o un vendedor.

Para poder sobrepasar estos silos M2M, se necesita desarrollar un ecosistema moderno que permita la interoperación entre verticales y aplicaciones. Esta potencial interconexión entre en dispositivo y el modo en el que interactúa con el entorno conduce a un contexto más amplio que las comunicaciones M2M, lo que es llamado IoT. Las comunicaciones M2M se pueden considerar un predecesor de IoT ya que sientan las bases de la conectividad en la que IoT se desarrolla.

El amplio rango de casos de uso para IoT conlleva que los Indicadores Clave de Desempeño (KPIs) en IoT difieran en gran medida con respecto al costo, el tiempo de vida de la batería del dispositivo, cobertura, rendimiento, capacidad de la red, seguridad y la confiabilidad.

En este contexto hay que considerar que las características de tráfico MTC

difieren en gran medida del tráfico generado por los humanos, las diferencias principales se listan a continuación:

- Tráfico dominante en el enlace ascendente.
- Baja o no movilidad. Muchos dispositivos son estacionarios o tiene una movilidad menor que los dispositivos usados para las comunicaciones entre humanos (H2H).
- Nuevos perfiles de tráfico: Los dispositivos MTC generalmente envían paquetes periódicos o en ráfagas. Además, mientras que el tráfico MTC es generado uniformemente a lo largo del día, el tráfico H2H se concentra en las horas de luz y la tarde.
- Calidad de servicio (QoS): los requisitos de MTC y las comunicaciones H2H pueden divergir significativamente en términos de latencia, confiabilidad, consumo energético, o seguridad.

A.1.2 Conectividad en el IoT y desafíos

Dentro de la infraestructura de IoT, los sistemas de comunicación inalámbricos son esenciales ya que suponen un menor costo que la infraestructura de comunicaciones cableada. Para conseguir el desarrollo de la visión de IoT, una gran variedad de tecnologías de comunicación inalámbricas ha emergido gradualmente. El gran panorama de conectividad IoT refleja los diversos requisitos de comunicación para satisfacer los KPIs de IoT. Algunas de las opciones disponibles se listan a continuación:

Identificación por Radiofrecuencia (RFID) es una tecnología de identificación. Esta tecnología se suele usar en logística de mercancías. Sin embargo, investigaciones recientes están explorando el uso del corto alcance de RFID para detección de proximidad y localización dentro de IoT.

ZigbBee es un estándar inalámbrico abierto usado para crear pequeñas redes inalámbricas de área personal de bajo consumo y con baja tasa de transmisión de datos.

Bluetooth Low Energy (BLE) es una pila de protocolos completa para comunicaciones de corto alcance. Bluetooth fue inicialmente orientado a sustituir los cables en los dispositivos móviles, y ahora ha evolucionado para ser usado en muchas más aplicaciones. BLE es una versión de bajo consumo de energía y costo comparado con Bluetooth.

WiFi es una tecnología de interconexión para redes de área local inalámbrica. WiFi soporta altas tasas de transmisión de datos y es ampliamente usado en las conexiones móviles.

Redes de Baja Potencia y Largo Alcance (LPWAN) sin licencia se han convertido recientemente en un habilitador clave para IoT. Estas redes son capaces de proporcionar comunicaciones para una tasa de transmisión de datos baja, de baja potencia a largas distancias. Ejemplos de este tipo de tecnologías son SigFox, Ingenu, y LoRaWAN.

Redes celulares han sido históricamente una mala opción para los casos de uso de IoT debido al alto consumo energético que el dispositivo tiene mientras se comunica con la estación base y el coste por unidad. Sin embargo, en muchos despliegues se ha usado 2G para hacer pequeñas transmisiones de datos. Recientemente, el diseño de 5G se ha orientado a cubrir casos de uso no contemplados anteriormente en las redes móviles como son las comunicaciones críticas con muy baja latencia y *massive MTC* (mMTC).

LPWA en redes celulares especialmente diseñados para proporcionar los requisitos de conectividad de MTC masivo (gran cobertura, reducción de la complejidad del dispositivo, y gran tiempo de vida de la batería). Para poder conseguir estos objetivos, el 3GPP ha introducido tres tecnologías en sus estándares: *Long Term Evolution* Cat-M1 (LTE-M), *Narrowband IoT* (NB-IoT), and *Extended Coverage GSM IoT* (EC-GSM-IoT).

Dependiendo del despliegue IoT, una o varias de estas tecnologías podría proporcionar el set de requisitos necesarios. Centrándonos en las redes celulares, el despliegue de los casos de IoT es desafiante ya que las redes no fueron diseñadas para MTC. En general, los principales desafíos se pueden resumir en los siguientes:

- Conexiones masivas: dependiendo de la aplicación de IoT, un gran número de dispositivos MTC podrían estar conectados en una pequeña área. Esto podría conllevar a un gran número de solicitudes de acceso por parte de estos dispositivos, debido a la escasez de recursos radio, esta situación podría congestionar la interfaz radio de la red.

- Consumo de energía extremadamente bajo: esta característica es esencial para los dispositivos MTC alimentados con una batería o con acceso limitado a fuentes de alimentación que tienen interacciones infrecuentes con la red. Para soportar esta funcionalidad, las redes celulares necesitan actualizar varias funcionalidades como los procedimientos de control o los modos reposo y conectado.

- Diversidad: para abordar el mercado IoT, las redes celulares tienen que soportar eficientemente diversos requisitos de varios casos de uso de IoT. Esto significa ser capaz de soportar desde dispositivos estacionarios a dispositivos con gran movilidad, de comunicaciones tolerantes al retardo a extremadamente exigentes en la latencia, o de tráfico infrecuente a un continuo flujo de datos. Para soportar esta diversidad, la red debe optimizar la parte radio y el núcleo de red para incluir más funcionalidades como funcionalidades de ahorro de energía y gestión de la movilidad, congestión, control de la sobrecarga, etc.

A.1.3 Nuevos estándares celulares para IoT

Entre otras tareas, parte del trabajo realizado por el 3GPP es responsable de la estandarización de diferentes tecnologías de acceso radio desarrolladas durante la evolución de las generaciones celulares, como por ejemplo LTE en 4G y *New Radio* (NR) en 5G.

Cuando se estaba desarrollando LTE, el foco de atención estaba en los servicios móviles de banda ancha con requisitos exigentes de altas tasas de datos, baja latencia, y alta capacidad. Desde su introducción, LTE ha seguido avanzando considerablemente a lo largo de los años. Esta evolución ha conseguido mejorar las características de LTE desde su despliegue inicial con una velocidad de tasa de datos máxima de 300 Mbps en un ancho de banda contiguo de 20 MHz, al soporte de velocidades de tasa de datos máximas de varios Gbps conseguidos a partir de

mejoras en la tecnología de las antenas, agregación de portadoras radio, etc. Además, la evolución de LTE ha incluido también el soporte de nuevos casos de uso más allá de las comunicaciones de banda ancha iniciales, como por ejemplo, el soporte de MTC, comunicaciones a grupos de dispositivos, comunicaciones críticas, comunicaciones directas entre dispositivos.

A pesar de que LTE sigue evolucionando, el camino hacia la próxima generación de red celular, denominada 5G, se está convirtiendo rápidamente en el centro de atención. El objetivo general de 5G es proporcionar un ecosistema tecnológico de redes inalámbricas para ofrecer un medio de comunicación sin fisuras para cualquier tipo de dispositivo. 5G proporcionará unas características de red más avanzadas y mejoradas comparadas con LTE (4G). Para ello, 5G se basa en diferentes pilares: evolución de las tecnologías de acceso radio actuales, *Self Organizing Network* (SON), ondas milimétricas, virtualización, compartición de espectro, etc.

Para poder optimizar el soporte de IoT por las redes celulares, el 3GPP ha incluido 3 tecnologías en el segmento LPWA:

- EC-GSM-IoT es una evolución de GSM optimizada para IoT.
- LTE-M es una evolución de LTE optimizada para IoT.
- NB-IoT es una nueva tecnología de acceso radio basada en LTE y optimizada para dispositivos de muy bajo costo y consumo energético.

Teniendo en cuenta MTC está dentro de los principales casos de uso considerados en el diseño de 5G, se prevé que LTE-M y NB-IoT continuarán evolucionando como parte de la familia de tecnologías de 5G.

A.2 Objetivos

Dado el reciente avance de las redes celulares para abordar el mercado de IoT, y los desafíos que presenta la inclusión de MTC en estas redes, el objetivo principal de la presente tesis es el estudio de la inclusión de mMTC en las redes celulares. Específicamente, el uso de NB-IoT para dar servicio a mMTC dentro de las redes celulares. Para ello, la realización de esta tesis aborda las siguientes cuestiones:

1. Centrándonos en las redes LTE actualmente desplegadas pero no optimizadas para el tráfico IoT, ¿cuáles son las principales consecuencias las redes LTE experimentarán intentando dar soporte a IoT? ¿Cómo se podría gestionar el tráfico IoT previsto en el núcleo de la red para satisfacer los diversos requisitos de los diferentes tipos de dispositivos?

   Para contestar a estas cuestiones, vamos a modelar y evaluar el rendimiento de diferentes perfiles de tráfico en la red LTE. Dentro de este objetivo desarrollaremos y evaluaremos modelos analíticos y de simulación para estimar la carga de señalización del plano de control de LTE. Para ello, este objetivo está dividido en los siguientes subobjetivos:

   1.1 Revisión y análisis de las características de MTC y de los procedimientos de señalización de LTE.

   1.2 Diseño y desarrollo de modelos analíticos y de simulación para el plano de control de LTE. El modelo analítico estará basado en teoría de colas.

2. Considerendo la nueva estandarizada tecnología de acceso radio NB-IoT, ¿son suficiente los nuevos mecanismos de NB-IoT para conseguir la mejora de cobertura substancial deseada? ¿cuáles son los beneficios e inconvenientes de cada técnica usada en NB-IoT?

   Para contestar a estas preguntas, analizaremos el rendimiento de NB-IoT cuando se extiende la cobertura considerando diferentes escenarios y suposiciones. Para ello, este objetivo está dividido en los siguientes subobjetivos:

   2.1 Diseño y desarrollo de expresiones analíticas para estimar el rendimiento de las técnicas de extensión de cobertura en NB-IoT. Estas expresiones se basan en la descripción de la transmisión en NB-IoT en términos de la relación señal ruido (SNR), la utilización del ancho de banda, y la energía per bit transmitido. Las expresiones se basarán en la teoría de Shannon.

   2.2 Revisión de la estimación realista del canal y los factores no ideales a tener en cuenta en el rango de bajas SNR que NB-IoT puede llegar a operar.

   2.3 Desarrollo de simuladores para obtener la relación entre la SNR y el error de estimación de canal.

2.4 Derivación de expresiones analíticas para modelar la estimación de canal realista en NB-IoT.

3. Además de la nueva tecnología de acceso NB-IoT, dos procedimientos de señalización fueron introducidos en el estándar para optimizar la transmisión de datos, ¿son estas optimizaciones eficientes para IoT y bajo qué circunstancias? ¿cuál es el rendimiento del dispositivo esperado usando NB-IoT y esos procedimientos?

   Para responder a estas preguntas, estudiaremos el rendimiento de un dispositivo NB-IoT en términos de tiempo de vida de la batería y latencia. Para ello, este objetivo está dividido en los siguientes subobjetivos:

   3.1 Diseño y desarrollo de un modelo analítico para estimar el consumo energético de un dispositivo NB-IoT y los recursos radio consumidos. El modelo analítico se basará en una cadena de Markov.

   3.2 Evaluación empírica de dispositivos NB-IoT en un banco de pruebas controlado. Se configurará un banco de pruebas con dispositivos NB-IoT comerciales conectados a un emulador de estación base para medir el consumo energético del dispositivo.

   3.3 Validación del modelo analítico para NB-IoT. Para validar el modelo, se configurará el banco de pruebas y el modelo con los mismos parámetros y se compararán los resultados en términos de tiempo de vida de la batería y latencia.

A.3 Conclusiones

Las conclusiones más relevantes extraídas del trabajo realizado en esta tesis se resumen en las siguientes:

Ch2 La Virtualización de las Funciones de Red (NFV) permite que las funciones de red se ejecuten como componentes de *software* en *hardware* general. Esto consigue varias ventajas como el uso de *hardware* comercial más barato que *hardware* especializado, o mayor flexibilidad en la creación, escalado y despliegue de las funcionales de red cuando sean necesarias permite que se pueda usar.

El paradigma que ofrece NFV se presenta como un habilitador para proporcionar nuevas arquitecturas de redes celulares capaces de hacer frente al incremento de tráfico previsto. A pesar de que el tráfico de *enhanced Mobile Broadband* (eMBB), se considera una extensión directa del servicio de banda ancha de 4G, la inclusión de los caso de uso URLLC y mMTC requiere dar soporte a una gran heterogeneidad de servicios. Dentro de este contexto, NFV proporciona la escalabilidad y flexibilidad necesarias para estos nuevos casos de uso.

Dentro de una red LTE, el *Mobility Management Entity* (MME) debe gestionar varios paquetes de señalización por cada subscriptor vinculado con él. Este rol central del MME junto con la inclusión de nuevos tipos de tráficos a gestionar puede incurrir en tormentas de señalización y congestión de la red si el nuevo tráfico no se gestiona eficientemente. Este capítulo se centra en el MME al ser una de las entidades de LTE más afectadas debido al creciente desarrollo de IoT dentro de las redes celulares.

Para superar este desafío, en este capítulo se ha propuesto la virtualización del plano de control de LTE, específicamente del MME. Para ello, el nuevo MME virtualizado (vMME) se descompone en tres niveles: *Front End* (FE), *Service Logic* (SL), and *State Database* (SDB). El FE es el responsable de la comunicación con las otras entidades de la red y balancea la carga entre los SLs. Los SLs se encargan del procesamiento de los paquetes entrantes. Y la SDB guarda el estado de la sesión del usuario.

La evaluación realizada estudia cuatro diseños diferentes del vMME y el rendimiento resultante de cada diseño en términos de recursos necesarios, coste basándonos en el modelo de Amazon EC2, y el tiempo de respuesta para cada clase de tráfico considerado. Los resultados muestran que el diseño del nivel SL tiene un gran impacto en el tráfico que se puede gestionar con respecto a los recursos disponibles, afectando significativamente al rendimiento de los otros tráficos. Si el tráfico mMTC se gestiona compartiendo recursos con otros tráficos, los recursos necesarios para satisfacer el retardo objetivo de URLLC incrementan significativamente. Para poder evitar esto, es necesario aislar los recursos de procesamiento para clases de tráfico con requisitos conflictivos. Considerando los posibles cuellos de

botella en los diseños evaluados, el nivel de la SDB es crítico debido a sus limitaciones a la hora de poder escalar por la arquitectura de todo compartido escogida en el análisis.

Ch3 El caso de uso mMTC describe escenarios con accesos masivos a la red por parte de un gran número de dispositivos. Un dispositivo mMTC suele tener baja complejidad (y por lo tanto es de bajo coste) y está alimentado por una batería, y permanece la mayor parte del tiempo durmiendo y se activa intermitente para enviar pequeños paquetes de datos ocasionales. Dadas estas características, muchos dispositivos mMTC dependerán de LPWAN para su conectividad con Internet.

La introducción de NB-IoT en los estándares del 3GPP está pensada como una solución dentro del segmento LPWA en las redes celulares. NB-IoT está específicamente diseñado para MTC y tiene cuatro objetivos de diseño clave: máxima latencia de 10 segundos para una transmisión en el enlace ascendente, extensión de cobertura con un máximo de pérdidas de propagación para el enlace (MCL) de 164dB, duración de la batería de más de 10 años, y soporte para conexiones masivas.

Para analizar los beneficios y limitaciones de NB-IoT para mMTC, en este capítulo se ha propuesto un modelo analítico para estudiar el consumo energético de un dispositivo NB-IoT. El modelo analítico se basa en las probabilidades de los estados estacionarios de una cadena de Markov. Este modelo no solo relaciona las métricas de rendimiento de la duración de batería y latencia con los parámetros de entrada, sino que además proporciona una metodología de análisis que puede ser fácilmente extendida para incluir escenarios más complejos o más procedimientos de control. Además, se han comparado la ganancia en capacidad cuando el dispositivo usa los nuevos procedimientos optimizados para la transmisión de pequeños paquetes y el procedimiento convencional de transmisión de datos de LTE denominado *Service Request* (SR). Los dos procedimientos de transmisión de datos optimizados consiguen reducir la señalización requerida para la transmisión de datos entre el dispositivo y la estación base. El primero de ellos, denominado *User Plane optimization* (UP) guarda el contexto de la interfaz radio del dispositivo y a partir de resumir y suspender la conexión

radio en vez de borrarla por completo al terminar la transmisión consigue reducir la señalización. La segunda opción, denominada *Control Plane optimization* (CP) usa el plano de control para transmitir el paquete de datos, de esta forma no se necesita establecer el plano de datos y se requiere menos señalización. La evaluación realizada también considera cuatro escenarios de transferencia de datos, dependiendo de si el dispositivo realiza una transferencia de datos en el enlace ascendente o descendente, y si se necesita una confirmación de la transferencia.

Desde el punto de vista de la duración de la batería, los resultados muestran que cuando el dispositivo tiene una condición radio pobre y cambia a un nivel de cobertura peor en NB-IoT, lo que se denomina *Coverage Enhancement Level* (ECL), hay una reducción en la duración de la batería muy significativa, lo que hace aún más desafiante alcanzar el objetivo de 10 años. Además, el intervalo entre los paquetes, denominado *Inter-Arrival Time* (IAT), tiene un impacto considerable en la duración de la batería. Conforme el IAT incrementa, el dispositivo puede pasar periodos más largos en modos de ahorro energético, como el *Power Saving Mode* (PSM), y la duración de la batería puede prolongarse.

Comparando los procedimientos de transmisión de datos (SR, UP, y CP), CP consigue los mejores resultados debido a la señalización del *Release Assistance Indication* (RAI) que permite a la red saber cuándo se puede liberar la conexión radio para liberarla cuanto antes. Por ejemplo, para el ECL 0, esta notificación ayuda a mejorar la duración de la batería usando CP desde un 78% cuando el IAT es de 30 minutos a un 4% cuando el IAT es de 24 horas, comparado con UP. Hay muchos factores que hacen más desafiante satisfacer los 10 años de duración de la batería. En nuestros resultados, este objetivo se cumple para la gran parte del rango de IATs evaluados mientras el dispositivo pertenece al ECL 0. Sin embargo, para los ECLs 1 y 2, este objetivo se cumple cuando el dispositivo tiene un IAT mayor que 6 horas y 13 horas, respectivamente.

En términos de ganancia en la capacidad, los procedimientos de transmisión de datos optimizados (UP y CP) consiguen una gran reducción en el número de recursos consumidos, lo que se traduce en una ganancia en capacidad

considerable, en comparación al procedimiento SR. Esto se repite en todos los escenarios de transferencia de datos considerados, siendo especialmente notoria la ganancia en capacidad conseguida en el escenario donde un dispositivo manda un paquete de datos en el enlace ascendente sin necesitar una posterior confirmación. Por ejemplo, en este caso, CP y UP consiguen ganancias de un 162% y un 120% en el ECL 0, respectivamente. A pesar de los beneficios de CP, este procedimiento no es recomendable para largas transmisiones de datos, ya que la red podría forzar al dispositivo a establecer el plano de usuario para continuar la transferencia de datos si se excede un máximo número de mensajes transferidos. Por lo tanto, para transmisiones largas UP sería más recomendable.

A partir de esta evaluación también se ha analizado qué canal radio es el que limita la capacidad del sistema, resultando que depende del tráfico considerado y el nivel de cobertura del dispositivo. Cuando el dispositivo pertenece a un buen ECL, la limitación de la capacidad la genera los canales radio del enlace ascendente, sin embargo, para peores ECLs, la limitación viene de los canales radio del enlace descendente. Comparando los resultados de capacidad en el enlace ascendente y el enlace descendente, las ganancias del enlace descendente alcanzan algo menos de dos veces las ganancias del enlace ascendente. Esto es debido a la señalización adicional generada por la ejecución del procedimiento *Tracking Area Update* (TAU) en los escenarios de transferencias en enlace descendente para gestionar el tráfico de bajada.

Ch4 El objetivo principal cuando se propuso NB-IoT era conseguir una extensión de la cobertura de 20 dB comparado a *General Packet Radio Service* (GPRS). Esto es debido a que para los dispositivos mMTC la eficiencia energética y buena cobertura son los KPI más importantes. Sin embargo, ambos KPIs están relacionados de manera conflictiva, cuando se usa una configuración más robusta para extender la cobertura, más recursos radio se usan y el consumo energético se incrementa. Además, la extensión de cobertura en NB-IoT cubre un rango de operación SNR muy bajo. En este rango de operación, las prestaciones de los mecanismos de estimación del canal pueden no ser suficientes para asegurar una estimación del canal precisa. Por lo tanto, el rendimiento de los canales físicos de NB-IoT se puede

ver limitado.

En este capítulo se ha estudiado el rendimiento de NB-IoT cuando se extiende la cobertura y se consideran factores no ideales como una estimación del canal realista. El rendimiento se analiza en tres escenarios diferentes para la estimación del canal (CE): CE ideal, CE realista usando la técnica de *cross-subframe*, y CE realista sin usar *cross-subframe*. La técnica de *cross-subframe* usa en la estimación varias subtramas consecutivas en la estimación del canal para reducir el efecto del ruido. Para este estudio se ha propuesto un sistema completo para la evaluación analítica de NB-IoT que conjuntamente obtiene los resultados de rendimiento de la extensión de la cobertura y la resultante duración de batería y latencia del dispositivo. Este sistema depende de tres elementos: la estimación de la SNR a partir del teorema de Shannon, un algoritmo de adaptación del enlace ascendente propuesto en este capítulo también, y el modelo analítico de estimación de consumo energético propuesto en el anterior capítulo de esta tesis.

Los resultados muestran que NB-IoT presenta una gran brecha entre su rendimiento y el límite de Shannon. Esto es debido a cómo NB-IoT está implementado y al esquema de repetición utilizado. Centrándonos en la adaptación del enlace ascendente, para un dispositivo en buena cobertura, la modificación del *Modulation and Coding Scheme* (MCS) y el número de recursos o la reducción del ancho de banda consiguen mantener la utilización del ancho de banda y pueden cubrir un rango limitado de la extensión de la cobertura. Para dispositivos con mala cobertura, la reducción del ancho de banda y las repeticiones se vuelven esenciales para alcanzar una mayor extensión de la cobertura. Sin embargo, en esta situación la ganancia conseguida debido a las repeticiones puede ser limitada.

La ganancia en SNR cuando se duplican las repeticiones es limitada en escenarios de CE realista comparada con los 3 dB ideales. Por ejemplo, para un número de repeticiones mayor que 16, la ganancia cuando se duplican las repeticiones usadas en el escenario de CE realista es menor de 1.5 dB. Este pobre rendimiento cuando se considera CE realista empeora cuando la MCL incrementa ya que aumenta el error presente en la estimación del canal. Para reducir este error, el uso de la técnica de *cross-subframe* en

la estimación del canal mejora el rendimiento del estimador del canal ya que se promedian varias subtramas. A partir de las simulaciones, para el mayor tamaño de ventana considerado cuando se usa *cross-subframe* (16 ms), el error decrece 6 dB comparado con la estimación a partir de una única subtrama.

Como resultado de una imprecisa estimación del canal para SNR bajas, los resultados de rendimiento del dispositivo en términos de duración de la batería y latencia empeoran significativamente en CE realista comparado con CE ideal. En términos de duración de la batería del dispositivo, para un MCL de 154 dB, CE realista con *cross-subframe* alcanza una reducción de la duración de la batería de un 50% y un 18% comparado con CE ideal, para los IATs de 2 horas y 24 horas, respectivamente. Para MCLs mayores, como por ejemplo 164 dB, que dependen de las repeticiones para extender cobertura, esta degradación del rendimiento del dispositivo empeora alcanzando un 90% de reducción en duración de la batería en CE realista con *cross-subframe* comparado con CE ideal.

Ch5 Las redes incluyendo NB-IoT eventualmente estarán preparadas para implementarse en despliegues reales. Sin embargo, todavía hay desconocimiento de si NB-IoT será capaz de hacer frente a los KPIs de IoT, ya que esta tecnología tiene el potencial de abarcar un amplio rango de aplicaciones IoT como agricultura, salud, seguridad, gestión de flotas, ciudades inteligentes, utilidades, etc. Cada aplicación IoT puede tener sus propias prioridades en sus KPIs, lo que resalta la necesidad de tener una clara visión del equilibrio entre los diferentes KPIs de NB-IoT para ayudar al mayor desarrollo y despliegue de NB-IoT.

En este capítulo se ha estudiado el rendimiento de NB-IoT analíticamente y experimentalmente. En esta parte se ha observado la configuración de una red viva de NB-IoT en Aalborg (Dinamarca), y se ha evaluado el rendimiento de un dispositivo NB-IoT bajo diferentes escenarios de cobertura y tráfico en un banco de pruebas con un entorno controlado. El banco de pruebas está compuesto por dispositivos NB-IoT comerciales conectados a un emulador de estación base NB-IoT y un analizador de potencia. La evaluación experimental considera tres dispositivos NB-IoT comerciales y

cuatro test. Cada test se centra en analizar el impacto de diferentes factores en el rendimiento del dispositivo en términos de consumo energético y latencia. Específicamente, cada test se centra en:

- Tamaño del paquete de datos en la transmisión en el enlace ascendente.
- Incremento del número de recursos asignados en una transmisión, es decir, el estudio de diferentes tasas de codificación.
- Incremento del número de repeticiones aplicadas en el enlace ascendente.
- Modificación en la asignación de ancho de banda en la transmisión.

Además, se ha validado el modelo analítico de NB-IoT anteriormente propuesto usando medidas experimentales con el mismo banco de pruebas. En este caso se han definido otros tres test y se han comparado los resultados obtenidos con las medidas experimentales y el modelo analítico fijando la misma configuración en ambos casos. Específicamente, cada test se centra en evaluar diferentes aspectos del modelo analítico:

- Las esperas debido al proceso de planificación de los recursos definido en NB-IoT.
- La extensión de las transferencias en la interfaz radio debido al uso de repeticiones en todos los canales radio.
- El impacto en la matriz de recursos y el cálculo de potencia de transmisión debido al cambio de subespaciado de portadora para las transmisiones en el enlace ascendente.

Los resultados muestran que hay diferentes causas que se apoderarán de una gran porción del consumo energético dependiendo del perfil de tráfico del dispositivo o el ECL al que pertenece. Por ejemplo, para dispositivos en buena cobertura, la configuración de la liberación de la conexión radio es importante. En este caso, soluciones como el *Release Assistance Indication* (RAI) que permite a la estación base conocer directamente cuando liberar la conexión radio, o la monitorización relajada de los canales de control pueden ahorrar una cantidad considerable de energía al dispositivo. Por

el contrario, para dispositivos en mala cobertura, el consumo energético está dominado por la transferencia de paquetes en la interfaz radio. En este caso, la reducción de la señalización y optimizaciones como *Early Data Transmission* (EDT) que permiten el envío de los paquetes de datos necesitando menos intercambios de paquetes entre el dispositivo y la estación base son esenciales. Como es evidente a partir de los resultados empíricos y lo que ya se vió en el Capítulo 3, la configuración de los tres ECLs tiene un gran impacto en el rendimiento del dispositivo.

Considerando los objetivos de duración de la batería y latencia de NB-IoT, los resultados indican que pueden satisfacerse mientras el dispositivo no esté en condiciones extremas de cobertura (por ejemplo, pertenecer al ECL 2) o no haya necesidad de un uso extensivo de las repeticiones de los paquetes (más de 16 repeticiones). Por ejemplo, para un dispositivo enviando paquetes UDP de 50 bytes, el objetivo de 10 años de duración de la batería se satisface si los paquetes tienen un IAT mayor de 4 horas, 11 horas, y 68 horas para los ECLs 0, 1, y 2, respectivamente.

Además, la validación del modelo analítico muestra que el modelo analítico tiene un buen rendimiento comparándolo con las medidas experimentales. Por ejemplo, el máximo error relativo comparando la estimación de la duración de la batería con el modelo con las medidas empíricas da unos valores de 21% y 11% para los IATs de 6 minutos y 24 horas, respectivamente. El error relativo disminuye conforme la energía consumida mientras se completa el procedimiento CP incrementa (por ejemplo, debido a repeticiones). Esto es debido a que el procedimiento se vuelve más importante que otras simplificaciones asumidas en el modelo como las sincronizaciones.

Bibliography

[1] F. Ghavimi and H. Chen, "M2M Communications in 3GPP LTE/LTE-A Networks: Architectures, Service Requirements, Challenges, and Applications," *IEEE Communications Surveys Tutorials*, vol. 17, no. 2, pp. 525–549, Secondquarter 2015.

[2] B. Research, "M2M/IoT Sector Map," 2017.

[3] G. Americas, "Cellular Technologies Enabling the IoT," White paper, Nov. 2015.

[4] A. El-Nashar, M. El-saidny, and M. Sherif, *Design, Deployment and Performance of 4G-LTE Networks: A Practical Approach*, 1st ed. Wiley Publishing, 2014.

[5] 3GPP TR 21.914, "Release description; Release 14 (Rel-14)," v14.0.0, 2018.

[6] 3GPP TR 21.915, "Release description; Release 15 (Rel-15)," v0.6.0, 2019.

[7] M. M. Do, "Timeline of 5G Standardization in ITU-R and 3GPP." [Online]. Available: https://www.netmanias.com/en/post/oneshot/11147/5g/timeline-of-5g-standardization-in-itu-r-and-3gpp

[8] *Detailed specifications of the terrestrial radio interfaces of International Mobile Telecommunications Advanced (IMT-Advanced)*, Recommendation ITU-R M.2083, ITU-R Std., 2015.

[9] 3GPP TS 24.301, "Non-Access-Stratum (NAS) protocol for Evolved Packet System (EPS) (Rel-15)," v15.2.0, 2018.

[10] Netmanias, "LTE EMM and ECM States." [Online]. Available: https://www.netmanias.com/en/post/techdocs/5909/ecm-emm-lte-mobility/lte-emm-and-ecm-states

[11] R. Ratasuk, N. Mangalvedhe, Y. Zhang, M. Robert, and J. Koskinen, "Overview of narrowband IoT in LTE Rel-13," in *2016 IEEE Conference on Standards for Communications and Networking (CSCN)*, Oct 2016, pp. 1–7.

[12] Y. . E. Wang, X. Lin, A. Adhikary, A. Grovlen, Y. Sui, Y. Blankenship, J. Bergman, and H. S. Razaghi, "A Primer on 3GPP Narrowband Internet of Things," *IEEE Communications Magazine*, vol. 55, no. 3, pp. 117–123, March 2017.

[13] O. Liberg, M. Sundberg, E. Wang, J. Bergman, and J. Sachs, *Cellular Internet of Things: Technologies, Standards, and Performance*, Elsevier, Ed. Elsevier, 2018.

[14] Y. D. Beyene, R. Jantti, K. Ruttik, and S. Iraji, "On the Performance of Narrow-Band Internet of Things (NB-IoT)," in *2017 IEEE Wireless Communications and Networking Conference (WCNC)*, March 2017, pp. 1–6.

[15] M. R. Palattella, M. Dohler, A. Grieco, G. Rizzo, J. Torsner, T. Engel, and L. Ladid, "Internet of Things in the 5G Era: Enablers, Architecture, and Business Models," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 3, pp. 510–527, March 2016.

[16] J. Mocnej, A. Pekar, W. K. Seah, and I. Zolotova, "Network Traffic Characteristics of the IoT Application Use Cases." [Online]. Available: https://ecs.victoria.ac.nz/foswiki/pub/Main/TechnicalReportSeries/IoT_network_technologies_embfonts.pdf

[17] G. A. Akpakwu, B. J. Silva, G. P. Hancke, and A. M. Abu-Mahfouz, "A Survey on 5G Networks for the Internet of Things: Communication Technologies and Challenges," *IEEE Access*, vol. 6, pp. 3619–3647, 2018.

[18] G. Wu, S. Talwar, K. Johnsson, N. Himayat, and K. D. Johnson, "M2M: From mobile to embedded internet," *IEEE Communications Magazine*, vol. 49, no. 4, pp. 36–43, April 2011.

[19] Netmanias, "LTE Identification I: UE and ME Identifiers." [Online]. Available: https://www.netmanias.com/en/post/techdocs/5905/lte-lte-identification/lte-identification-i-ue-and-me-identifiers

[20] 3GPP TS 33.401, "System Architecture Evolution (SAE); Security architecture (Rel-14," v14.2.0, 2017.

[21] 3GPP TS 36.213, "Evolved Universal Terrestrial Radio Access; Physical layer procedures (Rel-13)," V13.11.0, 2018.

[22] 3GPP TS 36.321, "Evolved Universal Terrestrial Radio Access; Medium Access Control protocol specification (Rel-13)," v13.9.0, 2018.

[23] 3GPP RP-151393, "NB LTE – Battery Lifetime Evaluation," 2015.

[24] 3GPP TS 36.101, "Evolved Universal Terrestrial Radio Access (E-UTRA); User Equipment (UE) Radio Transmission and Reception (Rel-14)," v14.3.0, 2017.

[25] 3GPP TS 36.104, "Evolved Universal Terrestrial Radio Access (E-UTRA); Base Station (BS) Radio Transmission and Reception (Rel-14)," v14.3.0, 2017.

[26] 3GPP R1-1703113, "On mMTC, NB-IoT and eMTC latency evaluation," 2017.

[27] 3GPP R1-1703112, "On mMTC, NB-IoT and eMTC battery life evaluation," 2017.

[28] 3GPP TS 36.521-3, "Evolved Universal Terrestrial Radio Access (E-UTRA); User Equipment (UE) conformance specification; Radio transmission and reception; Part 3: Radio Resource Management (RRM) conformance testing (Rel-14)," V14.4.0, 2018.

[29] 3GPP TS 36.211, "Evolved Universal Terrestrial Radio Access; Physical channels and modulation (Rel-13)," V13.10.0, 2018.

[30] E. Dahlman, S. Parkvall, and J. Sköld, *4G LTE-Advanced Pro and The Road to 5G*, third edition ed. Academic Press, 2016.

[31] Cisco, "Cisco Visual Networking Index: Forecast and Trends, 2017–2022," White paper, 2018.

[32] Ericsson, "Ericsson Mobility Report," 2018.

[33] 3GPP TS 22.368, "Service requirements for Machine-Type Communications (MTC); Stage 1 (Rel-14)," v14.0.1, 2017.

[34] M. Z. Shafiq, L. Ji, A. X. Liu, J. Pang, and J. Wang, "Large-Scale Measurement and Characterization of Cellular Machine-to-Machine Traffic," *IEEE/ACM Transactions on Networking*, vol. 21, no. 6, pp. 1960–1973, Dec 2013.

[35] N. Pineda, "M2M VS IoT: Know the Difference." [Online]. Available: https://www.peerbits.com/blog/difference-between-m2m-and-iot.html

[36] Ericsson, "Cellular networks for massive IoT," White paper, Jan. 2016.

[37] S. Amendola, R. Lodato, S. Manzari, C. Occhiuzzi, and G. Marrocco, "RFID Technology for IoT-Based Personal Healthcare in Smart Spaces," *IEEE Internet of Things Journal*, vol. 1, no. 2, pp. 144–152, April 2014.

[38] M. Bolic, M. Rostamian, and P. M. Djuric, "Proximity Detection with RFID: A Step Toward the Internet of Things," *IEEE Pervasive Computing*, vol. 14, no. 2, pp. 70–76, Apr 2015.

[39] Z. Alliance, "Zigbee 3.0." [Online]. Available: https://www.zigbee.org/zigbee-for-developers/zigbee-3-0/

[40] B. SIG, "Bluetooth market update," 2018.

[41] H. Wang and A. O. Fapojuwo, "A Survey of Enabling Technologies of Low Power and Long Range Machine-to-Machine Communications," *IEEE Communications Surveys Tutorials*, vol. 19, no. 4, pp. 2621–2639, Fourthquarter 2017.

[42] SigFox. [Online]. Available: https://www.sigfox.com/en

[43] LoRa Alliance, "LoRaWAN 1.1 Regional Parameters," 2017.

[44] LoRaWAN. [Online]. Available: https://lora-alliance.org/

[45] J. Navarro-Ortiz, S. Sendra, P. Ameigeiras, and J. M. Lopez-Soler, "Integration of LoRaWAN and 4G/5G for the Industrial Internet of Things," *IEEE Communications Magazine*, vol. 56, no. 2, pp. 60–67, Feb 2018.

[46] *Detailed specifications of the radio interfaces of international mobile telecommunications-2000 (IMT-2000)*, Recommendation ITU-R M.1457-11, ITU-R Std., 2013.

[47] *IMT Vision – Framework and overall objectives of the future development of IMT for 2020 and beyond*, Recommendation ITU-R M.2012-3, ITU-R Std., 2018.

[48] J. Rodriguez, *Fundamentals of 5G Mobile Networks*, 1st ed. Wiley Publishing, 2015.

[49] 3GPP TR 38.913, "Study on scenarios and requirements for next generation access technologies (Rel-15)," V15.0.0, 2018.

[50] 3GPP TS 22.261, "Service requirements for the 5G system; Stage 1 (Rel-15)," v15.0.0, 2017.

[51] GSMA, "Road to 5G: Introduction and Migration," 2018.

[52] ——, "3GPP Low Power Wide Area Technologies," White paper, 2016.

[53] R. Ratasuk, A. Prasad, Z. Li, A. Ghosh, and M. A. Uusitalo, "Recent advancements in M2M communications in 4G networks and evolution towards 5G," in *Intelligence in Next Generation Networks (ICIN), 2015 18th International Conference on*, 2015, pp. 52–57.

[54] N. S. Networks, "Signaling Is Growing 50[Online]. Available: http://docplayer.net/6278117-Signaling-is-growing-50-faster-than-data-traffic.html

[55] Alcatel-Lucent, "Managing the Signaling Traffic in Packet Core," Application note, 2012.

[56] C. Cox, *An Introduction to LTE: LTE, LTE-Advanced, SAE and 4G Mobile Communications*. Wiley, 2012.

[57] S. Ahmadi, *LTE-Advanced. A Practical Systems Approach to Understanding 3GPP LTE Releases 10 and 11 Radio Access Technologies*. Elsevier, 2014.

[58] 3GPP TS 23.401, "General Packet Radio Service enhancements for Evolved Universal Terrestrial Radio Access Network access (Rel-15)," v15.2.0, 2017.

[59] 3GPP TS 36.322, "LTE; Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Link Control (RLC) protocol specification (Rel 15)," v15.1.0, 2018.

[60] V. Nguyen, A. Brunstrom, K. Grinnemo, and J. Taheri, "SDN/NFV-Based Mobile Packet Core Network Architectures: A Survey," *IEEE Communications Surveys Tutorials*, vol. 19, no. 3, pp. 1567–1602, thirdquarter 2017.

[61] N. Omnes, M. Bouillon, G. Fromentoux, and O. L. Grand, "A programmable and virtualized network IT infrastructure for the internet of things: How can NFV SDN help for facing the upcoming challenges," in *Intelligence in Next Generation Networks (ICIN), 2015 18th International Conference on*, Feb 2015, pp. 64–69.

[62] ETSI NFV ISG, "Network Functions Virtualisation. An Introduction, Benefits, Enablers, Challenges & Call for Action." Introductory White Paper, 2012. [Online]. Available: https://portal.etsi.org/NFV/NFV_White_Paper.pdf

[63] U. N. Bhat, *An Introduction to Queueing Theory: Modeling and Analysis in Applications*. Birkhäuser, Springer, New York, 2015.

[64] H. Chen and D. D. Yao, *Fundamentals of Queueing Networks: Performance, Asymptotics, and Optimizatio*, 7th ed. New York: Springer, 2001.

[65] J. R. Jackson, "Networks of Waiting Lines," *Oper. Res.*, vol. 5, no. 4, pp. 518–521, Aug. 1957. [Online]. Available: http://dx.doi.org/10.1287/opre.5.4.518

[66] T. Taleb *et al.*, "EASE: EPC as a service to ease mobile core network deployment over cloud," *Network, IEEE*, vol. 29, no. 2, pp. 78–88, 2015.

[67] Y. Takano, A. Khan, M. Tamura, S. Iwashina, and T. Shimizu, "Virtualization-Based Scaling Methods for Stateful Cellular Network Nodes Using Elastic Core Architecture," in *2014 IEEE 6th International Conference on Cloud Computing Technology and Science*, Dec 2014, pp. 204–209.

[68] G. Premsankar, K. Ahokas, and S. Luukkainen, "Design and Implementation of a Distributed Mobility Management Entity on OpenStack," in *2015 IEEE 7th International Conference on Cloud Computing Technology and Science (CloudCom)*, Nov 2015, pp. 487–490.

[69] J. Prados-Garzon, J. J. Ramos-Munoz, P. Ameigeiras, P. Andres-Maldonado, and J. M. Lopez-Soler, "Latency evaluation of a virtualized MME," in *2016 Wireless Days (WD)*, March 2016, pp. 1–3.

[70] ——, "Modeling and Dimensioning of a Virtualized MME for 5G Mobile Networks," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 5, pp. 4383–4395, May 2017.

[71] J. Prados-Garzon, P. Ameigeiras, J. J. Ramos-Munoz, P. Andres-Maldonado, and J. M. Lopez-Soler, "Analytical modeling for Virtualized Network Functions," in *2017 IEEE International Conference on Communications Workshops (ICC Workshops)*, May 2017, pp. 979–985.

[72] P. Andres-Maldonado, P. Ameigeiras, J. Prados-Garzon, J. J. Ramos-Munoz, and J. M. Lopez-Soler, "MME support for M2M communications using network function virtualization," pp. 106–111, 2016.

[73] B. Urgaonkar, P. Shenoy, A. Chandra, P. Goyal, and T. Wood, "Agile Dynamic Provisioning of Multi-tier Internet Applications," *ACM Trans. Auton. Adapt. Syst.*, vol. 3, no. 1, pp. 1–39, Mar. 2008.

[74] J. Wan, J. Yi, X. Wang, and Z. Li, "Shared storage architecture for parallel database," *Journal of Computational Information Systems*, vol. 4, no. 1, pp. 375–382, 2008.

[75] P. Agyapong, V. Braun, M. Fallgren, A. Gouraud, M. Hessler, S. Jeux, A. Klein, L. Ji, D. Martin-Sacristan, and M. Maternia, "Simulation guidelines (Deliverable D6.1)," METIS, Tech. Rep., 10 2013.

[76] B. Hirschman, P. Mehta, K. B. Ramia, A. S. Rajan, E. Dylag, A. Singh, and M. Mcdonald, "High-performance evolved packet core signaling and bearer processing on general-purpose processors," *IEEE Network*, vol. 29, no. 3, pp. 6–14, May 2015.

[77] Amazon Web Service, "Aurora Pricing." [Online]. Available: https://aws.amazon.com/es/rds/aurora/pricing/

[78] J. Vilaplana, F. Solsona, I. Teixidó, J. Mateo, F. Abella, and J. Rius, "A queuing theory model for cloud computing," *The Journal of Supercomputing*, vol. 69, no. 1, pp. 492–507, 2014.

[79] A. Iosup, S. Ostermann, M. N. Yigitbasi, R. Prodan, T. Fahringer, and D. Epema, "Performance Analysis of Cloud Computing Services for Many-Tasks Scientific Computing," *IEEE Transactions on Parallel and Distributed Systems*, vol. 22, no. 6, pp. 931–945, June 2011.

[80] Amazon Web Services, "Amazon Aurora Performance Assessment," Technical Report. [Online]. Available: https://d0.awsstatic.com/product-marketing/Aurora/RDS_Aurora_Performance_Assessment_Benchmarking_v1-2.pdf

[81] "Ubuntu Documentation," available online: https://help.ubuntu.com/community/Installation/SystemRequirements (accessed on 03 January 2019).

[82] U. Raza, P. Kulkarni, and M. Sooriyabandara, "Low power wide area networks: An overview," *IEEE Communications Surveys Tutorials*, vol. 19, no. 2, pp. 855–873, Secondquarter 2017.

[83] 3GPP TR 45.820, "Cellular system support for ultra-low complexity and low throughput Internet of Things (CIoT) (Rel-13)," V13.1.0, 2015.

[84] J. Xu, J. Yao, L. Wang, Z. Ming, K. Wu, and L. Chen, "Narrowband Internet of Things: Evolutions, Technologies, and Open Issues," *IEEE Internet of Things Journal*, vol. 5, no. 3, pp. 1449–1462, June 2018.

[85] G. C. Madueño, J. J. Nielsen, D. M. Kim, N. K. Pratas, Č Stefanović, and P. Popovski, "Assessment of LTE Wireless Access for Monitoring of Energy Distribution in the Smart Grid," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 3, pp. 675–688, 2016.

[86] X. Wang, M. J. Sheng, Y. Y. Lou, Y. Y. Shih, and M. Chiang, "Internet of Things Session Management Over LTE -Balancing Signal Load, Power, and Delay," *IEEE Internet of Things Journal*, vol. 3, no. 3, pp. 339–353, June 2016.

[87] P. Andres-Maldonado, P. Ameigeiras, J. Prados-Garzon, J. J. Ramos-Munoz, and J. M. Lopez-Soler, "Optimized LTE data transmission procedures for IoT: Device side energy consumption analysis," in *2017 IEEE International Conference on Communications Workshops (ICC Workshops)*, May 2017, pp. 540–545.

[88] L. Feltrin, G. Tsoukaneri, M. Condoluci, C. Buratti, T. Mahmoodi, M. Dohler, and R. Verdone, "Narrowband IoT: A Survey on Downlink and Uplink Perspectives," *IEEE Wireless Communications*, vol. 26, no. 1, pp. 78–86, February 2019.

[89] P. Jörke, R. Falkenberg, C. Wietfeld, P. Joerke, and R. Falkenberg, "Power Consumption Analysis of NB-IoT and eMTC in Challenging Smart City Environments," in *2018 IEEE Global Communications Conference (Globecom 2018)*, 2018.

[90] R. Ratasuk, B. Vejlgaard, N. Mangalvedhe, and A. Ghosh, "NB-IoT system for M2M communication," in *2016 IEEE Wireless Communications and Networking Conference Workshops (WCNCW)*, April 2016, pp. 428–432.

[91] 3GPP TS 36.300, "LTE; Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall description; Stage 2 (Rel-14)," v14.6.0, 2018.

[92] D. R. J. Schlienz, "Narrowband Internet of Things Whitepaper," Rohde & Schwarz Whitepaper.

[93] A. Hoglund, D. P. Van, T. Tirronen, O. Liberg, Y. Sui, and E. A. Yavuz, "3GPP Release 15 Early Data Transmission," *IEEE Communications Standards Magazine*, vol. 2, no. 2, pp. 90–96, JUNE 2018.

[94] R. Ratasuk, N. Mangalvedhe, Z. Xiong, M. Robert, and D. Bhatoolaul, "Enhancements of narrowband IoT in 3GPP Rel-14 and Rel-15," in *2017 IEEE Conference on Standards for Communications and Networking (CSCN)*, Sept 2017, pp. 60–65.

[95] A. Hoglund, X. Lin, O. Liberg, A. Behravan, E. A. Yavuz, M. V. D. Zee, Y. Sui, T. Tirronen, A. Ratilainen, and D. Eriksson, "Overview of 3GPP Release 14 Enhanced NB-IoT," *IEEE Network*, vol. 31, no. 6, pp. 16–22, November 2017.

[96] 3GPP RP-171428, "Revised WID on Further NB-IoT enhancements," 2017.

[97] 3GPP R1-1811675, "Chairman's notes of AI 6.1.5 Maintenance of Release 15 Further enhancements of NB-IoT," 2018.

[98] C. M. Grinstead and J. L. Snell, *Introduction to Probability.* American Mathematical Society, 2006.

[99] 3GPP TR 37.868, "Evolved Universal Terrestrial Radio Access; Physical channels and modulation (Rel-11)," V11.0.0, 2011.

[100] S. Gachhadar, "Clock Error Impact on NB-IoT Radio Link Performance," Master's thesis, Aalto University, School of electrical engineering, 2018.

[101] 3GPP R1-151216, "PUSCH channel estimation aspects for MTC," 2016.

[102] 3GPP R1-150021, "PUSCH link performance for MTC," 2015.

[103] Sharetechnote, "LTE-NB : NRS (NB Reference Signal)." [Online]. Available: http://www.sharetechnote.com/html/Handbook_LTE_NB_NRS.html

[104] L. Hanzo, M. Münster, B. Choi, and T. Keller, *OFDM and MC-CDMA for Broadband Multi-User Communications, WLANs and Broadcasting.* Wiley Publishing, 2012.

[105] Y. S. Cho, J. Kim, W. Y. Yang, and C. G. Kang, *MIMO-OFDM Wireless Communications with MATLAB.* Wiley Publishing, 2010.

[106] S. Colieri, M. Ergen, A. Puri, and B. A, "A study of channel estimation in OFDM systems," in *Proceedings IEEE 56th Vehicular Technology Conference*, vol. 2, Sep. 2002, pp. 894–898 vol.2.

[107] J. . van de Beek, O. Edfors, M. Sandell, S. K. Wilson, and P. O. Borjesson, "On channel estimation in OFDM systems," in *1995 IEEE 45th Vehicular Technology Conference. Countdown to the Wireless Twenty-First Century*, vol. 2, July 1995, pp. 815–819 vol.2.

[108] O. Edfors, M. Sandell, J. . van de Beek, S. K. Wilson, and P. O. Borjesson, "OFDM channel estimation by singular value decomposition," *IEEE Transactions on Communications*, vol. 46, no. 7, pp. 931–939, July 1998.

[109] J. Lee, H.-L. Lou, D. Toumpakaris, and J. M. Cioffi, "Effect of carrier frequency offset on OFDM systems for multipath fading channels," in *IEEE Global Telecommunications Conference, 2004. GLOBECOM '04.*, vol. 6, Nov 2004, pp. 3721–3725 Vol.6.

[110] N. M. Balasubramanya, L. Lampe, G. Vos, and S. Bennett, "Low SNR Uplink CFO Estimation for Energy Efficient IoT Using LTE," *IEEE Access*, vol. 4, pp. 3936–3950, 2016.

[111] C. E. Shannon, "Communication in the Presence of Noise," *Proceedings of the IRE*, vol. 37, no. 1, pp. 10–21, Jan 1949.

[112] E. Dahlman, S. Parkvall, J. Skold, and P. Beming, *3G Evolution, Second Edition: HSPA and LTE for Mobile Broadband*, 2nd ed. Orlando, FL, USA: Academic Press, Inc., 2008.

[113] M. Lauridsen, I. Z. Kovacs, P. Mogensen, M. Sorensen, and S. Holst, "Coverage and Capacity Analysis of LTE-M and NB-IoT in a Rural Area," in

2016 IEEE 84th Vehicular Technology Conference (VTC-Fall), Sept 2016, pp. 1–5.

[114] B. Vejlgaard, M. Lauridsen, H. Nguyen, I. Z. Kovacs, P. Mogensen, and M. Sorensen, "Coverage and Capacity Analysis of Sigfox, LoRa, GPRS, and NB-IoT," in *2017 IEEE 85th Vehicular Technology Conference (VTC Spring)*, June 2017, pp. 1–5.

[115] A. Adhikary, X. Lin, and Y. P. E. Wang, "Performance Evaluation of NB-IoT Coverage," in *2016 IEEE 84th Vehicular Technology Conference (VTC-Fall)*, Sept 2016, pp. 1–5.

[116] M. S. Ali, Y. Li, M. K. H. Jewel, O. J. Famoriji, and F. Lin, "Channel Estimation and Peak-to-Average Power Ratio Analysis of Narrowband Internet of Things Uplink Systems," *Wireless Communications and Mobile Computing*, vol. 2018, p. 15, 2018, article ID 2570165.

[117] F. Rusek and S. Hu, "Sequential channel estimation in the presence of random phase noise in NB-IoT systems," in *2017 IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, Oct 2017, pp. 1–5.

[118] S. Liu, F. Yang, J. Song, and Z. Han, "Block Sparse Bayesian Learning-Based NB-IoT Interference Elimination in LTE-Advanced Systems," *IEEE Transactions on Communications*, vol. 65, no. 10, pp. 4559–4571, Oct 2017.

[119] A. Ali and W. Hamouda, "On the Cell Search and Initial Synchronization for NB-IoT LTE Systems," *IEEE Communications Letters*, vol. 21, no. 8, pp. 1843–1846, Aug 2017.

[120] R. Ratasuk, N. Mangalvedhe, J. Kaikkonen, and M. Robert, "Data Channel Design and Performance for LTE Narrowband IoT," in *2016 IEEE 84th Vehicular Technology Conference (VTC-Fall)*, Sep. 2016, pp. 1–5.

[121] N. Mangalvedhe, R. Ratasuk, and A. Ghosh, "NB-IoT deployment study for low power wide area cellular IoT," in *2016 IEEE 27th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, Sep. 2016, pp. 1–6.

[122] C. Yu, L. Yu, Y. Wu, Y. He, and Q. Lu, "Uplink Scheduling and Link Adaptation for Narrowband Internet of Things Systems," *IEEE Access*, vol. 5, pp. 1724–1734, 2017.

[123] J. C. Ikuno, S. Pendl, M. Šimko, and M. Rupp, "Accurate SINR estimation model for system level simulation of LTE networks," in *2012 IEEE International Conference on Communications (ICC)*, June 2012, pp. 1471–1475.

[124] Y. D. Beyene, R. Jantti, O. Tirkkonen, K. Ruttik, S. Iraji, A. Larmo, T. Tirronen, and a. J. Torsner, "NB-IoT Technology Overview and Experience from Cloud-RAN Implementation," *IEEE Wireless Communications*, vol. 24, no. 3, pp. 26–32, June 2017.

[125] 3GPP TR 25.943, "Universal Mobile Telecommunications System (UMTS); Deployment aspects (Rel-14)," V14.0.0, 2017.

[126] 3GPP R1-150672, "Discussion and Performance evaluation for uplink PSD boosting," 2015.

[127] P. Mogensen, W. Na, I. Z. Kovacs, F. Frederiksen, A. Pokhariyal, K. I. Pedersen, T. Kolding, K. Hugl, and M. Kuusela, "LTE Capacity Compared to the Shannon Bound," in *2007 IEEE 65th Vehicular Technology Conference - VTC2007-Spring*, April 2007, pp. 1234–1238.

[128] M. Simko, D. Wu, C. Mehlfuehrer, J. Eilert, and D. Liu, "Implementation Aspects of Channel Estimation for 3GPP LTE Terminals," in *17th European Wireless 2011 - Sustainable Wireless Technologies*, April 2011, pp. 1–5.

[129] Z. Tian, Q. Zhou, M. Zhou, and S. Luo, "Channel estimation for LTE downlink using RLS-based threshold judgement," in *2013 22nd Wireless and Optical Communication Conference*, May 2013, pp. 272–277.

[130] 3GPP R1-160480, "Consideration on uplink data transmission for NB-IoT," 2016.

[131] 3GPP R1-161860, "Further considerations on NB-PDSCH design for NB-IoT," 2016.

[132] H. Zarrinkoub, *Understanding LTE with MATLAB: From Mathematical Modeling to Simulation and Prototyping*, 1st ed. Wiley Publishing, 2014.

[133] R. Ratasuk, J. Tan, N. Mangalvedhe, M. H. Ng, and A. Ghosh, "Analysis of NB-IoT Deployment in LTE Guard-Band," in *2017 IEEE 85th Vehicular Technology Conference (VTC Spring)*, June 2017, pp. 1–5.

[134] 3GPP R1-161920, "NB-IoT PUSCH performance with 15 kHz and 3.75 kHz subcarrier spacing," 2016.

[135] 3GPP R1-160061, "NB-IoT PUSCH link level evaluation," 2016.

[136] L. Feltrin, M. Condoluci, T. Mahmoodi, M. Dohler, and R. Verdone, "NB-IoT: Performance Estimation and Optimal Configuration," in *European Wireless 2018; 24th European Wireless Conference*, May 2018, pp. 1–6.

[137] A. Azari, G. Miao, C. Stefanovic, and P. Popovski, "Latency-Energy Tradeoff Based on Channel Scheduling and Repetitions in NB-IoT Systems," in *2018 IEEE Global Communications Conference (GLOBECOM)*, Dec 2018, pp. 1–7.

[138] L. Feltrin, A. Marri, M. Paffetti, and R. Verdone, "Preliminary evaluation of NB-IOT technology and its capacity," unpublished. [Online]. Available: https://www.tugraz.at/fileadmin/user_upload/Projekte/Dependablethings/IRACON_WS_17/9_Feltrin.PDF

[139] J. Lee and J. Lee, "Prediction-Based Energy Saving Mechanism in 3GPP NB-IoT Networks," *Sensors*, vol. 17, no. 9, p. 2008, Sep. 2017.

[140] J. Liu, Q. Mu, L. Liu, and L. Chen, "Investigation about the paging resource allocation in NB-IoT," in *2017 20th International Symposium on Wireless Personal Multimedia Communications (WPMC)*, Dec 2017, pp. 320–324.

[141] H. Malik, H. Pervaiz, M. M. Alam, Y. L. Moullec, A. Kuusik, and M. A. Imran, "Radio Resource Management Scheme in NB-IoT Systems," *IEEE Access*, vol. 6, pp. 15 051–15 064, 2018.

[142] A. Puschmann, P. Sutton, and I. Gomez, "Implementing NB-IoT in Software - Experiences Using the srsLTE Library," Appears in the

proceedings of the Wireless Innovation Forum Europe 2017. [Online]. Available: https://arxiv.org/abs/1705.03529

[143] S. R. Systems, "srsLTE: Open Source LTE," Available at https://github.com/srsLTE.

[144] D. Troha and P. Krasowski, "Wireless system design: NB-IoT downlink simulator," Essay, 2017, [Online]. Available:https://uu.diva-portal.org/smash/get/diva2:1083434/FULLTEXT01.pdf.

[145] S. Martiradonna, A. Grassi, G. Piro, L. A. Grieco, and G. Boggia, "An Open Source Platform for Exploring NB-IoT System Performance," in *European Wireless 2018; 24th European Wireless Conference*, May 2018, pp. 1–6.

[146] G. Piro, L. A. Grieco, G. Boggia, F. Capozzi, and P. Camarda, "Simulating LTE Cellular Systems: An Open-Source Framework," *IEEE Transactions on Vehicular Technology*, vol. 60, no. 2, pp. 498–513, Feb 2011.

[147] B. Martinez, F. Adelantado, A. Bartoli, and X. Vilajosana, "Exploring the Performance Boundaries of NB-IoT," Unpublished. [Online]. Available: https://arxiv.org/abs/1810.00847

[148] C. Y. Yeoh, A. bin Man, Q. M. Ashraf, and A. K. Samingan, "Experimental assessment of battery lifetime for commercial off-the-shelf NB-IoT module," in *2018 20th International Conference on Advanced Communication Technology (ICACT)*, Feb 2018, pp. 223–228.

[149] M. Lauridsen, R. Krigslund, M. Rohr, and G. Madueno, "An Empirical NB-IoT Power Consumption Model for Battery Lifetime Estimation," in *2018 IEEE 87th Vehicular Technology Conference (VTC Spring)*, June 2018, pp. 1–5.

[150] 3GPP R1-162048, "Summary of evaluation results for NB-PSS," 2016.

[151] P. Andres-Maldonado, P. Ameigeiras, J. Prados-Garzon, J. J. Ramos-Munoz, and J. M. Lopez-Soler, "Reduced M2M signaling communications in 3GPP LTE and future 5G cellular networks," in *2016 Wireless Days (WD)*, March 2016, pp. 1–3.

[152] P. Andres-Maldonado, P. Ameigeiras, J. Prados-Garzon, J. Ramos-Munoz, and J. Lopez-Soler, "Virtualized MME Design for IoT Support in 5G Systems," *Sensors*, vol. 18, no. 8, p. 1338, Aug. 2016.

[153] P. Andres-Maldonado, P. Ameigeiras, J. Prados-Garzon, J. Navarro-Ortiz, and J. M. Lopez-Soler, "Narrowband IoT Data Transmission Procedures for Massive Machine-Type Communications," *IEEE Network*, vol. 31, no. 6, pp. 8–15, November 2017.

[154] P. Andres-Maldonado, P. Ameigeiras, J. Prados-Garzon, J. J. Ramos-Munoz, J. Navarro-Ortiz, and J. M. Lopez-Soler, "Analytic Analysis of Narrowband IoT Coverage Enhancement Approaches," in *2018 Global Internet of Things Summit (GIoTS)*, June 2018, pp. 1–6.

[155] P. Andres-Maldonado, P. Ameigeiras, J. Prados-Garzon, J. Navarro-Ortiz, and J. M. Lopez-Soler, "An Analytical Performance Evaluation Framework for NB-IoT," 2019, Accepted for publication in the IEEE Internet of Things Journal.

[156] P. Andres-Maldonado, M. Lauridsen, P. Ameigeiras, and J. M. Lopez-Soler, "Analytical Modeling and Experimental Validation of NB-IoT Device Energy Consumption," 2019, Accepted for publication in the IEEE Internet of Things Journal.

[157] ——, "Experimental analysis of NB-IoT performance trade-offs," 2019, To be submitted.

This thesis has been supported by the following projects and grants:

- National Research Project **TIN2013-46223-P** "Arquitectura para Redes Móviles 5G basada en Software Defined Networks" ("Architecture for 5G Mobile Networks based on Software Defined Networks") funded by the Spanish Ministry of Economy and Competitiveness and the European Regional Development Fund.

- National Research Project **TEC2016-76795-C6-4-R** "5G-City: Gestión Flexible de Servicios 5G Orientada a Soportar Situaciones Críticas Urbanas" ("5G-City: Adaptive Management of 5G Services to Support Critical Events in Cities") funded by the Spanish Ministry of Economy and Competitiveness and the European Regional Development Fund.